Summer 2007 – Vol. 3/No. 3

"Evidence" Versus "Science" in Practice Guidelines

Elizabeth Genovese, MD

Early practice guidelines developed by medical specialty organizations, insurers, etc., often reflected only the consensus opinion of those "experts" who authored them. The movement in the 1980s and 90s,[1] toward making medical decisions more reliant on evidence, as demonstrated by the use of high-quality randomized controlled trials (RCTs) in the medical literature, rather than on doctors' opinions, led to the proliferation of a new type of guideline described as "evidence-based." Those involved in developing "scientifically based" guidelines and systematic reviews, as well as those with an understanding of the differences between the various types of studies in the literature and the ramifications of these differences, were generally loathe to classify guidelines as evidence-based unless the highest possible standards were used in classifying the strength of the scientific evidence. The clinical practice guideline, *Acute Low Back Problems in Adults*, published by the Agency for Healthcare Research and Quality in 1994,[2] and the first and second editions of ACOEM's *Occupational Medicine Practice Guidelines*,[3,4] adhered to these standards and permitted the strength of the evidence to be placed in one of only four possible categories:

Level A **–** *Strong research-based evidence:* Provided by generally consistent findings in multiple (more than one) high-quality RCTs.

Level B **–** *Moderate research-based evidence:* Provided by generally consistent findings in one high-quality RCT and one or more low-quality RCTs, or generally consistent findings in multiple low-quality RCTs.

Level C **–** *Limited research-based evidence:* Provided by one RCT (either high or low quality) or inconsistent or contradictory evidence findings in multiple RCTs.

Level D – *No research-based evidence:* No RCTs or trials that are conflicting.

The revisions to the second edition of the ACOEM *Guidelines* uses the same system. However, "Level D" has been renamed "Level I," to reflect its use to characterize situations in which the body of evidence is "insufficient" or conflicting.

To the extent that the literature has adequate high-quality RCTs on a given topic, it is possible to develop guidelines or conclusions regarding treatment and causation that are wholly based on scientific evidence. Many medical interventions have not been rigorously evaluated. Some have not been evaluated due to lack of funding for research (generally because there is no one interested in promoting these interventions) or are already widely accepted based on anecdotal evidence alone. Other interventions have either not been evaluated in RCTs or were evaluated in RCTs of such low quality (often sponsored by those with a vested interest in the intervention) as to be unacceptable even as Level C evidence. Inability to effectively "blind" subjects by using a credible placebo,[a] or failure to compare a given intervention to alternatives in addition to placebo (comparative efficacy), limits the number of studies of acceptable quality (and certainly applicability) even further. Consequently, while most scientific, evidence-based clinical practice guidelines cite the literature that evaluates a given topic, the final decision regarding the implications of the results (or lack thereof) is the consensus opinion of the authors/collaborators. This opinion reflects the

[a]Many placebos (sham forms of actual drugs or treatments) are easily identifiable as such by those who are given them. A "credible" placebo is one which is actually perceived as being the real treatment by subjects who receive it.

Continued on next page . . .

Coming next issue . . .

- **Literature Update:** Brief review of selected new articles relevant to the *Practice Guidelines* Recommendations
- **UR and the *Guidelines*:** What do you do with requests for additional PT?
- **How does one apply "no recommendation"?** (an option when the evidence is insufficient)
- **Subscriber Corner:** Questions from our readers

For Subscription Inquiries:
APG Insights/ACOEM Communications
25 Northwest Point Boulevard, Suite 700
Elk Grove Village, IL 60007-1030
Telephone: 847/818-1800
Fax: 847/818-9266
Web site: www.acoem.org

Subscription rate for 4 issues (1 year) is $105 for ACOEM members; $125 for non-members.

APG Insights is published 4 times a year by the American College of Occupational and Environmental Medicine (ACOEM). The College and its editorial staff disclaim responsibility for any injury to persons or property resulting from any ideas or products referred to in this newsletter.

ISSN#: 1553-9237
Printed in the USA.

application of the high-quality scientific evidence in the literature to current practice patterns, with the goal of using the data to reach conclusions regarding the extent to which these patterns are or are not supported.

There are few if any RCTs to support most of the interventions commonly used to manage occupational injuries. As a consequence, when faced with having to make recommendations regarding clinical practice, those participating in the development of the "scientifically based" ACOEM *Guidelines* (which only considers evidence as potentially high quality[b] if it is supported by RCTs) needed to base these recommendations on what are now referred to in the ACOEM Methodology as the "First Principles of Clinical Logic."

These "First Principles of Clinical Logic" are:

- Adherence to ACOEM's *Code of Ethical Conduct*.
- Accordance with evidence-based practice as described in the ACOEM Methodology, particularly with respect to prioritization of treatment modalities.
- Testing and treatment decisions should be the result of collaboration between the clinician and the patient with full disclosure of benefits and risks. The best treatment strategy should be recommended. In cases where the patient cedes that judgment to the clinician, the clinician's judgment as to the best treatment strategy should be implemented.
- Treatment should not create dependence or functional disability.
- Treatments should improve on the natural history of the disorder, which in many cases is recovery without treatment.
- Invasive treatment should, in almost all cases, be preceded by adequate conservative treatment and may be performed if conservative treatment does not improve the health problem and there is evidence of effectiveness.
- The more invasive and permanent, the more caution should be exercised in considering invasive tests or treatments and the stronger the evidence should be of efficacy.
- The more costly the test or intervention, the more caution should be generally exercised prior to ordering the test or treatment and the stronger the evidence of efficacy should be.
- When two treatment methods appear equivalent, the most cost-effective method is preferred.
- When there are options for testing or treatment available, the clinician should choose the option supported by clinical and statistical significance.
- Imaging or testing should generally be done to confirm a clinical impression prior to surgery or other major, invasive treatment.
- Tests should be performed when the results affect the course of treatment.
- Treatment should have specific, objective goals and should be monitored for achievement of those goals within a reasonable time.
- Failure to achieve a goal does not change the risk/benefit calculation for a subsequent treatment.
- Recommendations should be evidence-based with evidence of efficacy balanced with evidence of benefits and harms.
- Clinicians should disclose any conflicts of interest (including ownership or other financial arrangements) they may have with any of the testing or treatment methods.

[b]Use of the term "possibly" reflects the fact that many RCTs are so flawed as to be unreliable due to deficiencies in their design or statistical analysis.

Continued on next page . . .

The consensus panels for the first and second editions of the ACOEM *Guidelines* weighed all of these factors in reaching their final conclusions regarding whether an intervention was "recommended," "not recommended," or "optional." While the 9 categories of recommendations used in the *Guidelines* revisions (strongly recommended, moderately recommended, recommended, insufficient for, insufficient no recommendation, insufficient against, recommended against, moderately recommended against, strongly recommended against) are more explicit in demonstrating the degree to which they are based on both the strength of the literature and consensus, the principles behind ACOEM's recommendations have remained constant. The physicians involved in the second edition of the ACOEM *Guidelines* and its revision were (and are) well aware of those studies that are not RCTs, but are nonetheless cited by proponents of the interventions evaluated as "evidence" to support their position; but did not include in the formal assessment of the strength of the evidence as there was no scientific basis for doing so.

Other available guidelines include scientifically unsound studies and even opinion in rating the strength of evidence. For example, the American Association of Interventional Pain Physicians uses the following system:

> **Level I –** *Conclusive*: Research-based evidence with multiple relevant and high-quality scientific studies or consistent reviews of meta-analyses
>
> **Level II –** *Strong*: Research-based evidence from at least one properly designed randomized, controlled trial; or research-based evidence from multiple properly designed studies of smaller size; or multiple low quality trials
>
> **Level III –** *Moderate*: a) Evidence obtained from well-designed pseudo-randomized controlled trials (alternate allocation or some other method); b) evidence obtained from comparative studies with concurrent controls and allocation not randomized (cohort studies, case-controlled studies, or interrupted time series with a control group); c) evidence obtained from comparative studies with historical control, two or more single-arm studies, or interrupted time series without a parallel control group
>
> **Level IV –** *Limited*: Evidence from well-designed non-experimental studies from more than one center or research group; or conflicting evidence with inconsistent findings in multiple trials
>
> **Level V –** *Indeterminate*: Opinions of respected authorities, based on clinical evidence, descriptive studies, or reports of expert committees.[5]

While this may seem appealing since more studies are explicitly considered in the overall rating of the evidence, evidence other than RCTs is not scientific. Assimilating studies other than RCTs into ratings of the overall "strength of the evidence," and then determining that the evidence falls into Level III, IV, or V in those guidelines which formally do so, is making a distinction that is both arbitrary and invalid. If the literature is of unacceptable quality, and hence likely to be "wrong," it doesn't matter if it is wrong due to one critical flaw or as the result of several critical flaws. It also is irrelevant whether there was one poorly designed study of a given type (such as a case series) or several of the same type. This is particularly the case when a manufacturer wishes to have an intervention adopted; in such situations it may well be that a low-cost unproven clinical intervention that has been empirically used by most physicians for years has not been formally studied, while a new intervention that hasn't been shown to lead to better outcomes (and may carry an increased risk of harm) has been extensively evaluated in poor quality, industry-sponsored research. Does this mean that the latter is "better" and Level III? Of course not. However, guidelines which use levels to differentiate between information which shouldn't even be considered as evidence imply that this is the case, and consequently set up a scenario which lends a level of "credibility" to poor-quality studies (based on the number performed) that is completely unfounded.

This is not to say that ACOEM did not at times cite lower level studies in the second edition of the ACOEM *Guidelines*. In discussing diagnostic testing there were often situations in which there were no RCTs and the questions being answered (such as what is the correlation between a positive result on this test and a given clinical finding when it is given to a consecutive group of individuals presenting with a problem unrelated to the one under evaluation) could be in part addressed by the results of an observational study (generally high level prospective cohort or, in other situations, a case control study). ACOEM also did refer to lower level studies (in the context of the "first principles") when explaining the basis for recommendations regarding interventions that are widely used *despite* the absence of evidence, especially in those situations when we made "no recommendation" but still wanted to provide some degree of guidance for the practitioner. However, this was not routinely performed, and, as stated previously, articles other than RCTs (or "quasi" RCTs) were never considered in ratings of the overall strength of the evidence.

---

[c]Studies other than RCTs are sometimes cited in documents that look at causation, as it is not possible (and unethical) to set up situations in which people are exposed to putative toxins as part of an "experiment."

Continued on next page . . .

This apparent omission of "evidence" leads those who equate "science" with the number of articles cited to state that the ACOEM *Guidelines* is not scientifically based. These claims are specious and may reflect ignorance regarding how to differentiate between "generic" evidence-based guidelines and those evidence- based guidelines, such as ACOEM's, that are "scientifically based." This article will explain why only RCTs (or systematic reviews and meta-analyses that rely on RCTs) are considered as evidence in the ACOEM *Guidelines*, and will also go into some depth in explaining why even RCTs are often so flawed as to be classified as only low quality. There will also be further explanation of how the consensus panels involved in the revision of the ACOEM *Guidelines* then use this evidence (or work in the absence of evidence) to formulate recommendations, or reach a decision (when the literature is insufficient) to make "no recommendation."

TYPES OF STUDIES

Studies can be *experimental* (RCTs for example) or *observational*. Scientifically, an experimental model is the gold standard for evaluating the validity of a given hypothesis. In well-designed experiments, one is able to *control* all variables except the one being evaluated. This allows the investigator to assess the unique impact of this variable upon outcome(s). Experimental models using humans are obviously more difficult to set up than are those using inert elements, tissue samples, animals, etc., as there are no concerns in the latter group regarding the ethics of denying treatments that appear to be of benefit (based on earlier, non-controlled trials) to the patients chosen to be "controls." This is especially so when a treatment may potentially prolong or save lives. It is also far easier to match control and experimental groups with all relevant variables in non-human studies; follow-up is also not an issue under rigorous laboratory conditions. For these reasons, experimental studies are characterized as "RCTs" rather than as "experimental" per se.

Observational studies differ from experimental studies in that they draw conclusions with regards to the degree to which two events are associated by analyzing the response of existing populations (groups of people) to a given exposure or treatment. In observational studies, researchers do not randomly select subjects. Instead, they work with existing data, and are hence easily contaminated. The degree to which one can rely upon the results of observational studies is based both upon their design and the methodology applied to the analysis of data. Even when studies are impeccable in design, it is important to remember that observational studies can *never* be described as "proving" cause and effect, though they certainly provide information regarding the strength of a postulated association.

Observational studies can be *analytical* or *descriptive*. *Case reports* and *case series* are examples of descriptive studies. In these, an intervention is provided to one or a group of patients, after which outcomes are described. Case series are frequently used to evaluate treatment. They are easily done as they tend to use patients who are already candidates for the treatment under investigation, and are inexpensive since they do not, by definition, include individuals who will not receive the intervention or require use (or design) of control or alternative treatments. No matter how impressive the outcomes, it is critical to recognize that while one can use the results of case series to generate hypotheses regarding a potential relationship(s) between the treatment and the observed outcome(s), (which can then be tested in future studies), they can never be used as grounds to assert the existence of a causal relationship between the two. The phenomenon described as "regression to the mean" is of particular relevance in case series. This alludes to the fact that symptoms, especially pain, from most medical conditions are not static, but instead wax and wane over time. Case series simply assess the response to a procedure when it is performed without comparing the patient to another patient with identical symptoms who does not have the procedure. If the procedure is performed during a "natural exacerbation"(which is likely since most patients would not agree to a procedure during the period of time when they are relatively asymptomatic), and results are judged during a natural remission, "improvement" is spontaneous, and *not* because of the procedure. In an experimental study both the procedure and the "sham" or alternative procedure used as a "control" are done at one point in time thus eliminating the possibility that a positive result was simply reflective of the natural history of the condition as opposed to truly caused by the intervention.

Analytic observational studies are designed to evaluate hypotheses and therefore always compare individuals receiving a treatment (or with a particular exposure or disease) to others who did not. There are three general types of analytic studies – *prevalence*, *case-control*, and *cohort* studies. In *prevalence* (or cross-sectional) studies, data are collected at a single point in time and then analyzed; those with the condition (or who received the treatment) of interest are compared to those without it (or who did not); the differences are documented and analyzed. Since data is not collected longitudinally, statements regarding cause and effect are not valid.

In *case-control* studies, individuals who have received a given treatment (as in a case series) are matched to control persons who share characteristics that might impact on the outcome, after which the two groups are comparatively analyzed. Cases and controls should be predetermined, and equal, in order for the study to be valid, with the two groups matched for all relevant factors except the one under evaluation (which assumes that one knows all the relevant factors and can achieve this degree of matching). As this generally does not occur, case-control studies, although preferable to case series, are not considered as acceptable proxies for RCTs in asserting the existence of a causal relationship between a treatment and a given outcome.

Continued on next page . . .

TABLE 1 – COMPARISON OF RELEVANT OBSERVATIONAL STUDIES

STUDY CHARACTERISTICS	Cross Sectional	Case Series	Case Control	Nested Case Control	Prospective Cohort
Data Collected at Single Point in Time	Yes	No	No	No; cases are incident	No
Data Collected Longitudinally	No	Yes	Yes	Yes	Yes
Subjects Selected	No	Y/N	Y/N	No	No
Control Group	No	No	Yes	Yes	No
Matching Attempted	No	No	Yes	Yes	No
Recall Bias	Yes	Y/N	Y/N	No	No
Prone to Artifactual Associations	Yes +++	Yes +++	Yes ++	Yes +	Yes +
Expense	Low	Low	Low	Medium	High
Strength of Evidence	Low	Low	Low	Varies	Medium

The optimal type of observational study is a cohort study, a longitudinal prospective (or retrospective, though prospective is best) study of a group of healthy people in a well-defined source population. Although these are usually used to evaluate causation, they can also be used to assess therapeutic interventions, especially diagnostic studies and procedures. Baseline information about the persons under evaluation are collected before the intervention (or even the disease for which it would be used) occurs. The population (cohort) is then observed over time (with the period of follow-up dependent upon the treatment under investigation) to evaluate their status and response to treatment. It is, however, prohibitively expensive to follow large populations simply to see what happens. While it is possible to "nest" case-control studies inside large cohort studies to evaluate those who did and did not receive a treatment of interest and compare outcomes, it is impossible, even with careful matching of cases and controls, to eliminate many, if not most, of the factors that could affect outcome other than the treatment intervention per se. Inability to both adequately standardize (control) the intervention provided and control for the relevant differences between those who receive a treatment and those who do not is what makes reliance on use of even the highest quality observational studies in assessing the value of treatment interventions problematic.

SOURCES OF ERROR IN STUDY DESIGN

The demonstration of an association between two events (or a treatment and an outcome) in a study does not necessarily mean that the two are causally related. Associations can be causal, but can also reflect flaws (either inadvertent or intentional) in a study's design or statistical analysis. If one is going to review experimental or observational studies in the literature, or analyze one's own data to assess the relationship between a given treatment and clinical outcomes, it is mandatory to understand the multiple potential sources of error and the shortcomings of statistical analysis, even when done properly.

Artifactual Associations/Bias

Artifactual associations occur when the studies or analyses upon which the determination of a causal relationship between an intervention and a given outcome are based fail to appropriately consider sources of *bias* or are flawed in their statistical analysis of the data. Bias is a systematic or measurement error in the design or analysis of a study, which leads to an overestimation or underestimation of the strength and significance of a given association. While bias is more likely to occur in observational studies (as their design, especially when case series, is inherently less rigorous) it is also rampant in experimental studies. There are many forms of bias, one or several of which may be present in a given study.

There are several different types of *selection bias*, and this type of bias is often further broken down into *membership bias, procedure-selection bias*, etc. Membership bias occurs when those who volunteer (or are chosen) to participate in a study differ from those who do not. Those who participate in studies of an intervention obviously differ from the general population by virtue of having a health condition that ostensibly warrants the performance of the intervention under evaluation. They also may be more likely to believe in the value of medical interventions (or the particular investigation under evaluation – referred to as "expectation bias") than might the general population, or may be available to participate in a study because of factors that differentiate them from other individuals with similar medical conditions. All of these can potentially affect results. Another form of membership bias occurs when the characteristics of patients assigned to different treatment groups differ in a fashion that affects their response (such as whether an injury was sustained at work versus recreationally, whether those who received the "control" treatment had been denied access to the "experimental" treatment by their insurance carriers, etc.).

Continued on next page . . .

Procedure bias occurs when subject characteristics are matched across groups, but the groups of subjects do not receive identical treatment – e.g., when members of one group get a treatment other than the one being investigated that may potentially impact on results. *Attention bias* is a form of procedure bias. It occurs when the there is more time spent with (attention paid to) those who receive either an intervention or the control (generally the former). Given the substantive body of literature indicating that the spending time with a patient is therapeutic in and of itself, it would appear reasonable to control for this effect in designing studies.

Measurement bias occurs when instruments used to measure the results of an exposure or treatment are not of sufficient validity (low predictive value) to produce reliable results or are not relevant to the outcome of interest. An example of the latter would be studies evaluating chronic pain treatments in which results are often "measured" by using subjective criteria (pain relief) without reliance on more objective measures of patient benefit such as a meaningful increase in the ability to perform relevant activities of daily living or a meaningful decrease in reliance on analgesic medications or other forms of palliation. When this lack of reliability occurs (by chance if nothing else) in one group rather than another, outcome analysis will be based upon flawed data and will be flawed as well. A variant of measurement bias occurs when study analysts are not blinded to participant group assignments, as this is likely to subtly bias the analysis of results, especially when the analysis is at least in part based upon the significance one ascribes to "soft" findings.

When members in one group are more likely to recall certain events or exposures than those in another, *recall bias* is said to occur. This type of bias is frequently found in prevalence, case-control, and retrospective studies in general as those who have a condition, or have undergone a particular treatment (such as back surgery) are generally more likely to recall an exposure or focus on potential side effects of the procedure than do those who did **not** have the condition or receive the treatment. *Detection bias* also can lead to an increased rate of association between an exposure or treatment and a health outcome, but rather than reflecting patient recall, it occurs when diagnostic testing is differentially applied to one group more than to another. This type of bias occurs most frequently when a new treatment is accompanied by potential health risks, leading to an increased rate of screening in the treatment group and, in turn, detection of the condition of interest at an earlier stage than it otherwise would have been in when found (or when it is still asymptomatic).

Non-response bias can either falsely elevate *or* falsely reduce the association between a given treatment or exposure and a health outcome. In the case of exposure analysis, if those who do not respond fail to do so because they have *not* developed the condition of interest (despite exposure) or recovered from a disease, analysis solely of those who respond will lead to a falsely elevated estimates of the risk of the adverse outcome based upon its prevalence in the remaining population studied. When response to a treatment intervention is being studied, failure of those who have done well to respond will lead to underestimation of benefit. On the other hand, if non-response reflects the development of health outcomes or side effects so severe as to preclude those who did poorly from responding, the converse will obviously be the case – i.e., adverse effects will be underestimated.

Compliance bias is another form of bias, and is of major importance in the characterization of treatment effects. Non-compliance with the treatment arm of a study due to side effects or because one treatment is clearly easier to perform or follow than another may be justifiable, but clearly lessens the degree to which results can be extrapolated, especially if subjects who fail to comply with the treatment due to adverse effects are also lost to follow-up. In order to adjust for the characteristics and outcomes of those who left the study must be analyzed (in the case of experimental studies described as "intention to treat analysis"). If those who withdrew were not comparable to those who remained, or if withdrawals were preferentially from one group (experimental or control) rather than another, analysis of the study must compensate for this. Since patients who fail to actually receive a treatment or complete the study may differ from those who did, an "intention to treat analysis," in which study groups are compared in terms of the treatment to which they were initially allocated, should be performed in order to minimize the risk of bias. While this invariably results in a conservative estimate of the treatment effect, this is preferable to the overestimation of benefit that might otherwise occur.

Placebos and "Blinding"

In RCTs, one group of subjects (experimental group) receives the treatment being evaluated while the other group (or groups), the "control(s)," receive(s) an alternative. Classic RCT study design generally mandates that a treatment be compared to a credible placebo. This is due to the "placebo effect," i.e. the tendency for people to respond to interventions, regardless of whether they are real or sham, based upon their expectation that they will be helpful (which, as noted previously, can lead to "expectation bias").[6,7,8] This in turn appears to be dependent upon the practitioner providing the intervention, which is why it is important that all aspects of treatment delivery *except* the treatment itself be as standardized and scripted as possible.

There is extensive literature indicating that the placebo effect can mimic the response that a patient would have to a real treatment. This has been shown to include objective changes in physiology and on imaging studies,[9,10,11,12] and often lasts for a significant period of time before extinguishing.[13,14] As a placebo control will only be effective if patients believe it to be a real treatment (credibility), they must be blinded to whether they are receiving the true treatment or a placebo, as should be the individuals analyzing the data and those performing the actual intervention (though this is often not possible). The extent of blinding in a RCT affects its quality, as reflected in the three out of a possible total 11 points assigned to RCTs using ACOEM's Methodology (see Table 3).[15]

Continued on next page . . .

Placebo controlled studies are relatively easy to perform when the treatment is a medication, but far more difficult to set up when the treatment is a procedure. Placebo versions of surgical procedures are the most difficult to design, and are accompanied by significant questions regarding their appropriateness when "sham surgery" is performed. More often than not, the difficulty of obtaining a true placebo control for most procedures is such to have led some authors to recommend that terms such as "control group" and "placebo treatment group" be replaced by "the notions of "placebo effect maximizing group" and "placebo effect minimizing group," as this clearly indicates that what such trials measure is relative effect, and an "underestimation of the absolute placebo effect."[16] In addition, since often we are interested in both whether an intervention works when compared to placebo **and** whether it works better than alternatives, it is generally optimal for RCTs to include an arm in which they evaluate the intervention being evaluated against adequate amounts of another treatment that has been shown to be effective[17] (it is quite interesting to look at the choice of comparators in studies that have been designed by those with a vested interest in proving that their treatment is best).[18,19]

TABLE 2 – EXPERIMENTAL VS. OBSERVATIONAL STUDIES

STUDY CHARACTERISTICS	EXPERIMENTAL	OBSERVATIONAL
Subject Blinded to Treatment	Usually *	No
Data Analysis Blinded	Usually	Rarely
Treatment Provider Blinded	Sometimes	No
Investigator Controls Treatment or Exposure	Yes	Sometimes
Stringent Inclusion and Exclusion Criteria	Sometimes	Rarely
Subjects Randomly Selected	Yes	No
Treatment Allocation Concealed	Sometimes	No
Co-interventions Avoidable	Yes	Sometimes
Possible Membership Bias	Sometimes	Yes
Possible Procedure (Attention) Bias	Sometimes	Yes
Possible Detection Bias	Sometimes	Yes
Possible Expectation Bias	Sometimes	Yes
Possible Recall Bias	No	Yes
Possible Non-response Bias	Sometimes	Depends
Possible Confounders (Non-causal)	Sometimes	Often
Suggests "Cause and Effect" if Positive	Yes	Rarely **
Only Suggests "Association" if Positive	No	Yes

*Though degree to which blinding is "real" often controversial.
May suggest cause and effect if prospective study evaluating causation. Will **not if study is evaluating an intervention as no control for bias.

While comparing the intervention of interest to alternative interventions as well as the placebo intervention is important, it is equally important to avoid using concurrent interventions that may affect outcome as an adjunct to the treatment under investigation. This is especially so if co-interventions are to be differentially used on one group of subjects and not the other, as they then may serve as "confounders" (be the real source of an effect that is then erroneously attributed to the intervention being evaluated) or lead to attention bias (described above) if the interactions between the subjects and those individuals providing the treatment are not equal in both the experimental and the control groups, since this will generally predispose toward better outcomes in the former.

Other Sources of Error

Co-morbidities (e.g., osteoarthritis) and factors such as age/activity level, substance abuse, psychological problems, etc., may affect the degree to which individuals will respond to an intervention. Consequently, high-quality studies list both inclusion and exclusion criteria in order to demonstrate that they are evaluating the efficacy of an intervention in individuals who are similar to those upon which it is to be used outside of the experimental setting. Observational studies do not divide subjects into experimental and control groups. In RCTs, following selection of the initial group of subjects to be evaluated, randomization is performed to adjust for any differences between subjects that were not captured in the initial criteria which may affect the ultimate outcome. The larger the study, the greater the chance randomization will successfully balance out any differences between the experimental and control groups.

Continued on next page . . .

While many studies only require that subjects be "blind" to the randomization procedure and group assignment, higher level studies use a computer model to generate the randomization sequence and insist that those working with the subjects are unaware of how these subjects are allocated to various treatment and control arms of the study. Randomization and concealment of treatment allocation are each worth up to a point in the overall RCT article quality score assigned by ACOEM. An additional point is given if analysis of the study population(s) after study completion reveals them to have baseline comparability.

Drop-out rate (non-response bias) and compliance have already been noted as potential causes of artifactual associations. Although not ordinarily described as a type of bias, the timing of assessments is also of relevance as it must be identical for both the experimental and control groups. Length of follow-up is also important, as even if an intervention leads to positive outcomes, they are often of little value if the duration of benefit is relatively small when compared to the treatment's cost (or the risk associated with its use).

Statistical Analysis

It is clear that study design can have a major impact upon the accuracy of conclusions. The type of statistical analysis used in reaching conclusions is also of significance as goals are to maximize significance and minimize error. This article will not elaborate in detail upon the ways in which flawed statistical analysis can affect whether a study's outcomes are reported as positive or negative. However, there are a few major principles that are important to understand.

- All statistical analysis is based on observations (and calculations) indicating that most measured values will fall into a "normal" bell-shaped dispersion of values. The average value is called the mean; the dispersion of values is characterized by the results of calculations aimed at determining "variance" and "standard deviation."
- Analysis of statistical significance is based upon the premise that 95% of a given sample of "normal" values will fall within 2 standard deviations of the mean (average) and 99% will be within 3 standard deviations. In a two-tailed test (a test where the researcher has no expectation that values will necessarily be either lower or higher than expected), it is generally held that the lowest or highest 2.5% of values are 95% likely to represent a real difference from what would be "normal," with a study leading to results of this nature described as having a p value of 0.05. This still means that there is a 5% probability that these levels are reflective of a normal distribution (and simply represent outliers). Studies that yield results that fall only within the top 0.5% of values on either side of the normal curve (or 1% of values on one side of the curve if a two-tailed analysis is not deemed appropriate) are considered to be 99% likely to be representative of true differences (and p is 0.01.) Although many studies consider a p value of 0.05 (likelihood that the values would have fallen into the 5% of the population norms as a result of the intervention by chance) as significant, many believe that true statistical significance is not reached until the p value is 0.01 or less.
- Another way to use the data is by calculating confidence intervals (CIs), ranges of values that represent 2 ($p < 0.05$) or 3 ($p < 0.01$) standard deviations from where one would have expected the mean to have fallen. Results are statistically significant in demonstrating a difference (rejecting the null hypothesis of no difference) if the observed results after exposure or treatment do not fall into this confidence limit. The new hypothesis is then that the intervention (event) under evaluation was a causal factor in shifting the CI away from the "original" mean. Again, this is reflective of an association between the exposure or treatment and the outcome rather than absolute proof of causality. Alternatively, one can calculate likelihood or odds ratios indicating the chances of observing the value seen in a given study in the normal population. If the likelihood ratio is 1.0, then the outcome was no more likely to occur (or not occur) as a result of the exposure or intervention than it was in its absence. Likelihood or odds ratio values are always listed in conjunction with a range representing, as a rule, the 95% CI level for significance. Clearly, if the value 1.0 is included within the confidence limits of the calculated ratio, the observed value is most likely **not** of statistical relevance.
- Some studies do not use p values or CIs to characterize results, but instead use terminology such as "absolute" or "relative" risk reduction or focus on relative decreases (or increases) in whatever factor(s) are deemed relevant to the intervention or risk factors being evaluated. Studies of this nature should be viewed with extreme caution as an increase in disease-free survival or any other benchmark that has been chosen as indicative of treatment efficacy of 200% simply requires that the number of individuals who can be so characterized increases by a factor of two. While this may be important in a small study, it is obviously irrelevant, even if statistically significant, if the study population is large and the absolute numbers of those with the outcome small.

Error and Power Analysis

A causal relationship is generally postulated to exist between a treatment (or event) and given health outcome if the use of the treatment or occurrence of the event leads to an average (mean) result greater than 2 to 3 standard deviations from the mean of the values expected or found in a control or baseline group. Type I error is equal to the p value and occurs when one concludes there is a causal relationship between an exposure and a disease/treatment and clinical improvement when there really is not. Type II error, the converse, occurs when a difference or causal relationship is **not** found, but exists. The degree to which Type II error occurs is also referred to as the *power* of the study. Power is the ability of a test to appropriately reject the hypothesis of "no difference" when it is false.

Continued on next page . . .

In other words, power describes the ability of a study to detect a given difference of a given size between two outcomes if the difference really exists. Power analysis consists of determining how large a sample is required to detect an actual difference of specified magnitude. The larger the sample or size of the difference one is expecting, the greater the power.

A more detailed discussion of these principles and of statistical analysis as a whole is beyond the scope of this article. Suffice it to state that before ascribing significance to the results of a study it is important to know the p values (along with exactly what outcomes were and were not analyzed, since some studies intentionally omit some from analysis), and results of the power analysis, and then determine their significance based upon the parameters that are being evaluated.

ASSESSING THE QUALITY OF A STUDY: SUMMARY

Regardless of their shortcomings, well-designed experimental (randomized controlled trials) studies allow for better control of most variables and forms of bias than do observational studies. In the absence of a control intervention, positive outcomes in case series are particularly likely to represent a placebo response and are often contaminated by expectation and attention bias and the use of inadequate (if any) inclusion and exclusion criteria. The rating system for RCTs used by ACOEM is demonstrated in Table 3, and assigns points for randomization; concealed treatment allocation; baseline comparability of groups; patient, provider, and assessor blinding; avoidance of co-interventions; compliance; dropout rate; timing of assessments; and analysis by intention to treat. Low-quality studies are those with a score of 3.5 or less, with those with scores from 4.0 to 7.5 considered of intermediate quality whereas those with scores ranging from 8.0 to 11.0 are high quality. As observational studies such as case series would, at most, get points only for assessor blinding and avoidance of co-interventions (in the absence of a control or comparative treatment group many of the quality assessment indicators become irrelevant), they are, virtually be definition, neither high quality nor scientific. One can get a good sense of study quality by asking the following questions – some of which have been adopted from other sources[20]:

1. ***What was the aim (hypothesis) of the study?***
   - Was it an exploratory study, surveying a population for interesting cases of a disease in order to generate hypotheses (case-report or case series) or to assess whether a potential association exists between two concurrent events (cross-sectional)?
   - Was it an anecdotal report of a change in the status of a patient or a group of patients, in association with a particular treatment?
   - Was a specific hypothesis being examined, and, if so, how?
2. ***How were study subjects selected?***
   - Were the main features of the study population well described?
   - Were inclusion and exclusion criteria reasonable and relevant to the issues being addressed and the population(s) to whom the treatment is to be applied?
   - Was randomization performed? If yes, was treatment allocation concealed?
   - Were both treatment and control groups comparable at entry and all relevant variables controlled for (can not assess initially)?
3. ***How were treatments or exposures defined and quantified?***
   - Was there a placebo group? If so, were placebos indistinguishable from treatment both initially and subsequently, so as to allow for blinding? Was the credibility of the placebo assessed?
   - If there was no placebo group, or if the treatment or exposure was one that could not be adequately concealed, how was this adjusted for or considered in the analysis of data – and did this adjustment or consideration have merit?
   - Was the treatment compared to an alternative intervention (along with or instead of placebo)? If so, was the intervention well described and the dose (or equivalent) and duration adequate to yield the desired outcome?
   - Were those who provided the treatment blinded? If not, could this have affected the results?
4. ***How were outcomes measured?***
   - Were they subjective or objective? If the former, what tools if any were used to measure subjective responses and of what reliability are they? What allowances were made for potential bias (this is of major import when outcomes are defined subjectively)?
   - If outcomes were measured objectively, how was this performed? If diagnostic tests were used to assess outcome, were they valid – i.e., of sufficient positive (or negative) predictive value in the population analyzed to accurately reflect the presence or absence of pathology? Do the tests or interpretations of the tests, exhibit high degrees of inter-rater and intra-rater reliability?
   - Over what period of time were outcomes measured? Was length of follow- up reasonable given the treatment or exposure under analysis? Was it identical in both treatment and experimental arms of the study?

Continued on next page . . .

5. ***Were there any obvious potential sources of error in initial outcome analysis?***
   - Were those who analyzed the study results blinded to subject treatment or exposure status both at the initiation of the study and subsequently?
   - Were the outcomes of participants who withdrew (intentionally or unintentionally) or were lost to follow-up described and included in the analysis?
   - Did the analysis discuss whether these individuals differed in any substantive fashion from those who continued in the study (intention to treat analysis)?

6. ***What type of statistical analysis was used, and how were the results conveyed?***
   - Specifically, were "p" values, confidence limits, likelihood estimates or odds ratios consistent with the existence of a statistically significant relationship between the exposure and the disease?
   - If so, how significant were they? (The greater the significance the less the likelihood that there was a Type I error.)
   - Was the sample size large enough to make meaningful conclusions?

7. ***Does the association make sense and correspond with what is already known or with what one might suspect to be true?***
   - If the study was evaluating the relationship between an exposure and a health effect, did the cause precede the effect in time? Does increased exposure increase the risk – i.e., is there a dose-response relationship? Does decreased exposure, or the elimination of exposure, reduce or eliminate the risk of disease?
   - If the study was evaluating a treatment option, is there a scientifically sound rationale linking the treatment to the observed outcome? Would one expect similar outcomes if a physiologically equivalent treatment were offered? Have there been other studies of this treatment leading to similar outcomes?

8. ***Regardless of the strength of the study, were the outcomes concrete enough, and of sufficient clinical relevance, to suggest a change in, or maintenance of, clinical practice?***
   - Was the study comparative in nature, or without controls and/or placebo? If so, was the treatment, test or exposure in the comparison group one which has already been investigated via high-quality studies? If not, how can one be certain that any difference, or lack of difference, between study groups means anything, or is any way clinically applicable?
   - Did outcomes represent a significant change in what was experienced or held to be true previously?
   - If so, was the change of such magnitude as to mandate an immediate reconsideration of current practice patterns?
   - If the documented change in outcome was small, is the health condition under treatment or evaluation sufficiently serious as to make even a small change in outcome based upon an intervention or exposure worthy of serious consideration?

9. ***How were potential sources of bias or error accounted for in the final study analysis?***
   - Did the investigators describe sources of bias (especially selection and recall bias), and state whether they felt that these may have led to artifactual associations?
   - Did the study account for possible confounding or co-existing factors (by considering all potentially relevant prior or current exposures or treatments that subjects may have experienced)?
   - Were all other possible sources of error described, with focus upon their potential impact upon study results?

9. ***Who were the authors?***
   - Were they invested financially in the outcome of the study?
   - If not, could their results be reflective predominantly of practice skills and other factors that are unique to them or their situation?

10. ***What are the costs (both direct and indirect) that would be associated with adoption of a new test, procedure, or means of reducing or mitigating exposure?***
   - Is the intervention more difficult to perform, or does it require more staff, than those currently used?
   - Is the intervention more harmful or potentially risky than other treatment options that are equally, or only slightly less, effective?
   - Will the intervention lead to additional testing or treatment that would not have been required otherwise? If so, what is the cost (both financial and emotional)?
   - Overall, is the incremental cost of adopting the intervention justified by the clinical benefit gained?

Continued on next page . . .

TABLE 3 – ACOEM CRITERIA FOR RATING RCT QUALITY

CRITERIA	RATING DESCRIPTION
Randomization	Assessment of the degree that randomization was both reported to have been performed and successfully achieved through analyses of comparisons of variables between the two groups.
Treatment Allocation Concealed	Concealment of the allocation scheme from all involved, not just the patient.
Baseline Comparability	Measurement of how well the baseline groups are comparable (e.g., age, gender, prior treatment).
Patient Blinded	Blinding of the patient/subject to the treatment administered.
Provider Blinded	Blinding of the provider to the treatment administered.
Assessor Blinded	Blinding of the assessor to the treatment administered.
Co-interventions Avoided	Assessment of the degree to which the study design avoided multiple treatments. This includes either a study design that includes combinations of interventions (e.g., a combination of exercise and anti-inflammatory medication) or patient self-administration of other treatments that may plausibly alter the results.
Compliance Acceptable	Measurement of the degree of non-compliance.
Dropout Rate	Measurement of the drop-out rate.
Timing of Assessments	Assessment of whether the timing of measurements of effects is the same between treatment groups.
Analyzed by Intention to Treat	Ascertainment of whether the study was analyzed with an intention to treat analysis.

The rating is converted into a quality grade – low (0-3.5), intermediate (4.0-7.5), or high quality (8.0-11.0).

UNDERSTANDING ACOEM'S RECOMMENDATIONS

While many "evidence-based" guidelines cite studies that are not RCTs as evidence, scientific evidence-based practice guidelines are mandated to only consider RCTs, or high level systematic reviews or meta-analyses of RCTs, as evidence that can support or refute an intervention. Lower level evidence can be used to get a sense of side effects, costs, and potential benefits, but can ***never*** be used as grounds for stating that the intervention is scientifically supportable or of documented clinical efficacy. ACOEM reflects this in the rating system for strength of the evidence in both the initial first and second edition of the *Guidelines* and in the recent revisions of the latter. The revised treatment guidelines have been designed to be more transparent with regards to the basis of recommendations than were previous editions. Thus the strength of the evidence is clearly linked to recommendations regarding interventions which are: "strongly recommended" (strength of evidence A), "moderately recommended" (strength of evidence B), or "recommended" (strength of evidence C). A similar progression is used to characterize situations when interventions are not recommended.

If the evidence is insufficient, a decision still needs to be made whether to recommend, not recommend, or make no recommendation regarding an intervention. At this point, the consensus panels (Evidence-based Practice Panels) become critical. The panels represent the wide range of physicians and other health care providers who will employ the interventions discussed, as reliance upon only one type of physician or only 1-2 individuals would introduce an unacceptable level of bias into the process. If a panel feels, based on the ACOEM First Principles, the clinical experience of its members, and occasionally, information from lower level trials not rigorous enough to be used as evidence, that a procedure is low in cost and risk, is widely used clinically, and may be of benefit in selected populations, the intervention is generally recommended. If the panel believes the cost or risk of harm is too high (in light of no proven benefit), the procedure is not recommended. Finally, when there is no consensus or information supporting benefit, and the intervention is used clinically, the panel generally decides that "no recommendation" can be made. This then leaves the determination whether to allow the procedure, and for how long to the jurisdiction or insurer using the guideline.

CONCLUSIONS

Although the number of practice guidelines described as evidence-based has proliferated during the past decade, the only ones that can be classified as scientific are those that exclusively consider RCTs as acceptable evidence. While this does not preclude reaching decisions about whether to recommend, not recommend, or make no recommendation regarding the use of interventions for which the evidence is insufficient, the reader of a high-level scientific evidence-based practice guideline should be easily able to identify the degree to which recommendations reflect the strength of the literature or the consensus of those writing the guideline. The rationale for choosing to recommend (or not) an intervention when the literature is insufficient should also be provided.

Continued on next page . . .

Reliance on a strict scientific interpretation of the evidence does not allow ACOEM *Guidelines* to state that there is support for interventions that are not supported by RCTs, although panel members can choose to recommend them based on other factors (which may include, as stated previously, the results of lower level studies) as long as doing so in consistent with our "First Principles." Those who do not understand the critical differences between types of evidence will invariably consider the absence of citations from studies other than RCTs when rating the strength of the evidence as reflecting an inadequate review of the literature. But because RCTs are themselves often flawed, to consider evidence from anything less as a scientific basis for treatment decisions would be unwise at best. ACOEM is the professional medical organization representing physicians and other providers involved in optimizing worker health. As such, it is unwilling to disseminate a guideline that is anything other than scientific, evidence-based and clear in describing the composition of our consensus panels and the principles upon which they base their recommendations.

REFERENCES

[1]Sackett DL, Rosenberg WM, Muir Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ.* 1996;312(7023):71-2.

[2]Bigos S, Bowyer O, Braen G, et al. *Acute Low Back Problems in Adults*. Clinical Practice Guideline No. 14. AHCPR Publication No. 95-0642. Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, U.S. Department of Health and Human Services; 1994.

[3]Harris J, ed. *Occupational Medicine Practice Guidelines: Evaluation and Management of Common Health Problems and Functional Recovery of Workers*. 1st ed. Elk Grove Village, Ill: American College of Occupational and Environmental Medicine; 1997.

[4]Glass LS, ed. *Occupational Medicine Practice Guidelines: Evaluation and Management of Common Health Problems and Functional Recovery of Workers*. 2nd ed. Elk Grove Village, Ill: American College of Occupational and Environmental Medicine; 2004.

[5]Boswell MV, Trescot AM, Datta S, et al. Interventional techniques: evidence-based practice guidelines in the management of chronic spinal pain. *Pain Physician.* 2007;10(1):7-111.

[6]Hunter P. A question of faith. Exploiting the placebo effect depends on both the susceptibility of the patient to suggestion and the ability of the doctor to instill trust. *EMBO Rep*. 2007;8(2):125-8.

[7]Kaptchuk TJ. The placebo effect in alternative medicine: can the performance of a healing ritual have clinical significance? *Ann Intern Med*. 2002;136(11):817-25.

[8]Linde K, C. M. Witt, Steng A, et al. The impact of patient expectations on outcomes in four randomized controlled trials of acupuncture in patients with chronic pain. *Pain*. 2007;128(3):264-71.

[9]McRae C, Cherin E, Yamazaki TG, et al. Effects of perceived treatment on quality of life and medical outcomes in a double-blind placebo surgery trial. *Arch Gen Psychiatry*. 2004;61(4):412-20.

[10]Mercado R, Constantoyannis C, Mandat T, et al. Expectation and the placebo effect in Parkinson's disease patients with subthalamic nucleus deep brain stimulation. *Mov Disord*. 2006;21(9):1457-61.

[11]Mayberg HS, Silva JA, Brannan SK, et al. The functional neuroanatomy of the placebo effect. *Am J Psychiatry*. 2002;159(5):728-37.

[12]Kuehn BM. Pain studies illuminate the placebo effect. *JAMA*. 2005;294(14):1750-1.

[13]Turner JA, Deyo RA, Loeser JD, Van Korff M, Fordyce WE. The importance of placebo effects in pain treatment and research. *JAMA*. 1994;271(20):1609-14.

[14]Gold M. Study design factors and patient demographics and their effect on the decline of placebo-treated subjects in randomized clinical trials in Alzheimer's disease. *J Clin Psychiatry*. 2007;68(3):430-8.

[15]ACOEM Methodology Document

[16]Hrobjartsson A. The uncontrollable placebo effect. *Eur J Clin Pharmacol*. 1996;50(5):345-8.

[17]Teutsch SM, Berger ML, Weinstein MC, et al. Comparative effectiveness: asking the right questions, choosing the right method. *Health Aff (Millwood)*. 2005;24(1):128-32.

[18]Pocock SJ, Hughes MD, Lee RJ, et al. Statistical problems in the reporting of clinical trials. A survey of three medical journals. *N Engl J Med*. 1987;317(7):426-32.

[19]Sackett DL, Oxman AD. HARLOT plc: an amalgamation of the world's two oldest professions. *BMJ*. 2003;327(7429):1442-5.

[20]Bombardier C, Kerr MS, Shannon HS, Frank JW. A guide to interpreting epidemiologic studies on the etiology of back pain. *Spine*. 1994;19(18 Suppl):2047S-2056S.