Management and Interpretation of Data Obtained from Clinical Trials in Pain Management

Brian R. Theodore, MS; Robert J. Gatchel, PhD, ABPP
Department of Psychology, College of Science, The University of Texas at Arlington, Texas, U.S.A.

■ **Abstract:** Conducting a clinical trial involves various stages of planning and implementation. The three major components involved in clinical trials are the management of data, the quality control to ensure data integrity, and the interpretation of the data at the conclusion of the trial. Although each process is distinct and involves different levels of effort and knowledge to implement, all processes are intimately linked. Data management techniques include the process of data entry and the implementation of an organized, comprehensive approach to quality control. Some guidelines for quality control screening are recommended to address various common issues related to clinical data, such as missing data, invalid cases, subject "outliers," and violation of distributional assumptions relevant to statistical analyses. In order to aid in interpreting the data, conditions that need to be met to make causal inferences are discussed. Taking into account baseline characteristics of the patient sample is also discussed as an extension to maintaining the internal validity of the study. Additionally, some common threats to statistical conclusion validity, including Type I error inflation and the problem of overpowered tests, are highlighted. Finally, the concept of the effect size as an important complement to statistical significance and how the various types of effect size measures can be interpreted within the context of a clinical trial are discussed. ■

Address correspondence and reprint requests to: Robert J. Gatchel, PhD, ABPP, Department of Psychology, College of Science, The University of Texas at Arlington, 313 Life Science Building, 501 S. Nedderman Drive, Arlington, TX 76019-0528, U.S.A. E-mail: gatchel@uta.edu.

DOI. 10.1111/j.1533-2500.2008.00240.x

Pain Practice, Volume 8, Issue 6, 2008 461–472

Key Words: data management, clinical trials, data interpretation

INTRODUCTION

The last two decades have been marked by an increasing emphasis on evidence-based medicine and employment of rigorous research methods in both clinical research and practice. As discussed by Lipman,[1] clinicians and researchers in the field of pain management have also been active in implementing evidence-based guidelines based on reviews of available evidence in the published literature. Presently, an international collaborative group is spearheading a comprehensive review of available evidence related to pain management as part of the Cochrane Collaboration's *Pain, Palliative, and Supportive Care Group*.[2] However, along with the increasing focus on the "end-product" of the evidence from clinical research, the importance of the "front-end" mechanisms and components of the research endeavor should also not be neglected. Therefore, this section will elaborate on several principles that are central to issues of *data management*, *quality control*, and *interpretation* of the data obtained from clinical trials.

Ban, Guy, and Wilson[3] noted that of the three stages of clinical trials corresponding to design, data collection, and data analysis, very little emphasis on the data collection stage is provided in the published literature. Unfortunately, not much has changed since 1983, and relatively recent peer-reviewed publications on

conducting clinical trials also focus on the design and analysis stages (eg, see Harden & Bruehl[4] and Mazumdar et al.[5]) However, this is understandable, given that data collection and management strategies are often dependent on organizational and logistic issues unique to an individual clinical practice or research laboratory. With this in mind, the subsequent sections will elaborate on some general principles related to data management and quality control that can be adapted to various settings that conduct clinical trials. Furthermore, it should be noted that these issues of management and quality control of data naturally lend themselves toward the ability to reliably interpret the data. In a similar vein as the issues related to data collection and management, the use of analytical techniques may also vary depending on the nature of the data collected, the design of the clinical trials, and the research questions and hypotheses being tested. Therefore, a general overview of issues related to interpretation of data from clinical trials is presented following the discussion on data management and quality control.

DATA MANAGEMENT

The first step in data management is establishing data collection protocols that are relevant to the clinical setting or research laboratory and the nature of the clinical trial being conducted. The nature of data collected often varies across several dimensions. This is an especially unavoidable fact when conducting research related to pain because pain is a complex biopsychosocial phenomenon with an interaction of multifactorial components consisting of physiological, psychological, social, environmental, and medico-legal factors.[6] Some data may be obtained from patients' medical history, either from medical charts, large registries, or communication with patients' primary physician. In addition, data are also often collected in the form of questionnaires that tap into measures of pain, psychological well-being, perceived disability, and health-related quality of life. In some trials, functional capacity or physiological data may be recorded from instruments that output these measures into a computer file. Finally, in those settings that do not require patients to be physically present at the clinic for follow-up evaluations, data may be obtained through telephone interviews or Internet-based, online data collection mechanisms.

Given this complex mix of data as described above, it is advisable to compile and organize the data into a single source that provides convenience of access and reduces the possibility of error when trying to tap into data from multiple sources for a given patient. Very often, the first step in compiling a database of clinical data is the process of data entry. Putting in place efficient and careful data entry protocols is the first line of defense against inaccurate interpretation at later stages of the clinical trial. When manually entering data, a general rule of thumb is to ensure all data from the various sources are double-entered by at least two different research or clinical personnel. This process ensures that transcribing errors can be detected and corrected when the two data sets are compared. However, in practice, large clinical trials may not have the personnel, resources, or time to implement double entry of data. In these cases, quality control mechanisms in the form of random checking of electronic data against the original hardcopy sources of data may be warranted. If the data have already been previously compiled into electronic files but are spread out across multiple individual files, then merging the data electronically using commercial or freeware database utility programs rather than attempting to manually enter all the data again into a single electronic file may be helpful.

The choice of the electronic database program used to manage and store data ultimately depends on the preference of the primary investigator and the research personnel, the size of the clinical trial and amount of data collected, budget constraints, and organizational policies on software licensing and use. Simple clinical trials may only require that data be entered into a standard spreadsheet application, but more complex and larger trials may require the functionalities of specific database utilities. Fortunately, there is a wide selection of both commercial and freeware database programs designed for the purpose of managing data obtained from clinical trials. A source that provides a comprehensive list of the available database programs for clinical data management is provided in Recommended Resources section at the end of this article. It should be kept in mind that both the implementation time and licensing costs in the case of commercial software, as initial investments for a good database utility program, will certainly pay off in the long run and ensure smoother, efficient access to data for later stages of the clinical trial dealing with quality control and analysis of data.

Finally, once the database has been put into place, periodic backup of the electronic database file should be implemented. The backup schedule may vary depending on the needs of the research setting and the available personnel resources, but having some form of backup is certainly much better than no backup at all. In our

experience, weekly backups of small data sets have been implemented in our research laboratory. For data on our larger clinical cohorts, the database file is hosted on a server that is backed up nightly in an automated process implemented by the in-house information technology personnel. With the data organized into a single source, the next step is to implement periodic quality control screening of the data.

QUALITY CONTROL

Quality control mechanisms are critical in ensuring accuracy of the data as well as the reliability and validity of subsequent interpretation of the data. Depending on the database software package implemented for use, there may be several types of quality control screening mechanisms built into the program. Therefore, it is a good idea for the primary investigator and the associated research personnel to read the accompanying product manuals and familiarize themselves with the features of the database program. However, there are some general principles of quality control screening that can be implemented by using simple methods via a spreadsheet application or data analysis software. In general, to ensure reliability of interpretation, quality control mechanisms involve the screening for the following problems: *missing data*, *invalid cases or outliers*, and *distributional assumptions*.[7]

Missing Data

Missing data can occur for a variety of reasons, including patient dropout during an intervention, noncompliance with filling out self-report measures, missing or misplaced questionnaire packets, gaps in a patient's medical history, or noncompliance during the follow-up period. With large-cohort clinical trials, there would be a logically greater likelihood of encountering these issues with missing data. However, as noted by Tabachnick and Fidell,[7] the amount of missing data is not so much the problem as the pattern of missing data. A general rule of thumb is that the threshold of concern for missing data is 5% for each individual variable within a data set, provided that the data are missing in a random pattern.

Determining the amount of missing data is a simple process that involves only dummy-coding a new binary variable to reflect all the cases that do and do not have missing data for each given variable in the data set. For smaller data sets, a simple "eyeballing" of the data set will suffice to determine the amount of missing data. To test the randomness of missing data for those variables that go beyond the 5% threshold, it is often useful to utilize prepackaged Missing Value Analysis utilities or computer macros that are available in most major statistical analysis software packages. These utilities for screening the randomness of missing data conduct exhaustive pairwise comparisons of variables and report significant *P* values for missing data patterns that are correlated between or among sets of variables.

In most cases where data are missing in small amounts and in random, the best option in dealing with these cases is usually the default option in major statistical analysis packages, that is, by excluding the case with the missing data from the analysis. However, as is often the case with randomized controlled trials (RCTs), a loss of even one or two cases in one group relative to another group results in an unbalanced design. Such instances are often handled by common statistical adjustments for unequal sample sizes.[7] These adjustments ultimately depend on the nature of the analysis and the test statistics used, and readers are provided with resources on research design issues in the Recommended Resources section of this article.

Another way of dealing with missing data is to impute the missing values based on some reasonable estimate. In clinical trials that experience patient dropout sometime during the treatment and, hence, unavailable for follow-up assessments, a common method of imputing missing values is known as the last observation carried forward (LOCF). The LOCF procedure simply uses the last known value for a given variable as the estimate of the missing data, under the assumption that the measure would remain stable given the dropout from the full treatment protocol.[8] However, studies on data simulation have revealed that the LOCF method often results in an underestimated treatment effect, and the impact on the power of tests and the probability of making statistical conclusion errors are greatly intensified when data are missing in a nonrandom pattern.[9,10]

More complex multivariate mechanisms for dealing with missing data involve creating missing data correlation matrices and regression-based estimation methods and are extensively reviewed by Tabachnick and Fidell,[7] Mallinckrodt et al.,[11] and Lane.[10] However, as noted by all these authors, none of these techniques hold up very reliably when data are missing in a nonrandom fashion. Such an issue is clearly related to the design, planning, and execution of the clinical trial, and no amount of statistical manipulation can provide a

quick fix for sloppy data collection or poorly planned and executed studies.

Invalid Cases and Outliers

Invalid cases can occur in one of two ways. As noted by Nyiendo et al.,[12] a common occurrence of invalid cases in clinical trials is when patients who are not qualified for treatment based on the study protocols are inadvertently recruited for the clinical trial. For example, if an intervention is targeting only a patient population that has not received any prior surgery, an inclusion of a small amount of surgical patients may end up biasing the results of the study. Prior to any analyses, a quick review of the characteristics of the patient cohort focusing on information relevant to qualifying protocols will provide an opportunity to flag these patients for removal from the analyses.

A second occurrence of invalid cases may occur at the level of the variable values. These can be generally screened for by looking for outliers in the data set. One instance of an outlier corresponds to values that are out of range. For example, if a measure of pain intensity ranges between values of 0 and 10, the value of 15 would be an outlying and invalid value on this variable. Such cases may stem from mistakes in data entry or from transcribing errors on the hard-copy source from which the electronic data was sourced and entered. Cross-referencing the hard-copy source may provide a quick correction of the value, if the hard-copy source is not in error in the first place. However, if the hard-copy source also reports a value that is clearly out of range and no other source of information is available to obtain the correct value, then the offending value within that variable for that particular case should be deleted in order to exclude it from subsequent analyses.

Finally, another case of outliers corresponds to values that are within the accurate range but may widely differ from the majority of cases simply because of random variation. For example, if at post-treatment all patients report a pain intensity rating that, on average, corresponds to the value of 2, an individual patient with a pain intensity rating of 9 may be a large outlier within the patient sample. If this outlying value is indeed accurate, based on cross-referencing the original hard-copy source, there are several available strategies in dealing with this as discussed by Tabachnick and Fidell.[7] For a small number of outliers, one strategy is to simply delete these outlying values, thus, excluding them from subsequent analysis. However, this strategy runs the risk of biasing the results in the opposite direction. Furthermore, the outliers themselves may be of interest for a variety of reasons relevant to some baseline characteristics of the outlying patients; therefore, it may often be a good idea to analyze these patients separately to look for any factors that may be predicting the outlying values. Another strategy to reduce the impact of outliers may be to apply numerical transformations to all values of the variable, that is, using square root or logarithm transformations. The disadvantage with transformations, however, is that they do not always result in a desirable distribution (sometimes making the scenario worse) and are often difficult to interpret meaningfully outside the properties of the statistical distribution itself.[7] If outliers do pose a significant problem in the data, a final option available may be to utilize complex robust estimation methods, such as Winsorized means or maximum likelihood estimation, on a case-by-case basis. Extensive discussions of robust estimation methods and cases where they may apply can be found in Hoaglin et al.[13]

Distributional Assumptions

Many of the commonly used statistical techniques rely on some basic assumptions about the distribution of the data. When distributional properties of the data depart widely from these basic assumptions, there may be an elevated risk of committing statistical conclusion errors, such as Type I and Type II errors. Although this issue is more closely related to interpretation of the data, it is presented and discussed within the section of quality control because some violations of these assumptions (as in the case of normality) may require additional recruitment of trial participants if the initial sample is relatively small. Similarly, violation of these assumptions may also be related to issues such as outliers, invalid values, or unbalanced sample sizes because of dropouts. The three major distributional assumptions relevant to conventional statistical techniques are *normality*, *linearity*, and *homoskedasticity*.[7]

The assumption of normality, as the name implies, requires that the distribution of data values for a given variable be approximately normally distributed, in the form of a bell-shaped, symmetrical curve. Common departures from normality often correspond to skewness in the distribution, where one tail of the distribution is stretched in a particular direction either to the left or right of the majority of data values. In clinical data, such skewness may be a product of floor or ceiling effects associated with self-report measures, or out-of-range values because of data entry errors. Another type

of departure from normality involves the concept of kurtosis, a distributional property in which the tails of the distribution are much thicker than that of a standard normal distribution. While there are built-in statistical tests for skewness and kurtosis available in most statistical analysis packages, these are often not recommended because of their overly sensitive nature to mild departures from normality.[7,14] Furthermore, almost all conventionally used test-statistics, such as the *t*- and *F*-statistics, are robust to substantial departures from normality and do not impact the interpretation of results. Additionally, an often neglected principle of statistical theory is the *central limit theorem*, which states that, given a large enough sample, the sampling distribution of means (upon which hypothesis tests are based) is normally distributed regardless of the distribution of the data values of the variable itself.[15] The size of the sample that can be considered sufficient, in terms of an approximately normal sampling distribution, has been demonstrated to be as little as 15 to 20 subjects,[15] and, in the case of analysis of variance (ANOVA) techniques, a degrees of freedom of 20 for the error term.[7] Therefore, there is little reason to fret over departures from normality. However, concerned researchers may want to try numerical transformations to align the distribution of their data closer to that of a normal curve, but the same problems with transformations as discussed with outliers apply here as well.

The assumption of linearity is concerned with a straight-line relationship between the values of two variables. An examination of scatterplots with each variable on a given axis of a 2-dimensional graph can indicate any substantial departures from linearity.[7] The problem associated with departures from linearity can result in the underestimation of correlation coefficients between two variables. Furthermore, if a nonlinear relationship exists between an independent/predictor variable and a dependent/outcome variable, the power of the statistical test may be substantially reduced. For substantial departures from linearity, a reliable method that can be applied in an analysis is to raise the power of the data values of the offending variable (usually the independent/predictor variable). For example, if the relationship on the scatterplot indicates a quadratic-type curved pattern, taking the square of the data values would ensure a more accurate fit between the two variables. This will also avoid the problem of the statistical test having reduced power.

The assumption of homoskedasticity is concerned with a similar spread, or variability, in one variable when compared with another variable. This is a consideration when looking at two continuous variables, with one being an independent/predictor variable while the other a dependent/outcome variable. When the independent/predictor variable is grouped, then homoskedasticity is referred to as the assumption of homogeneity of variance. Statistical tests for this assumption are built into all statistical analysis packages within each of the available parametric tests, and readers are encouraged to utilize these tests when conducting data analyses. Significant violations of this assumption show up at the conventional significant *P* value of <0.05. Violations of these assumptions often occur as a result of non-normality and severely skewed data. Furthermore, this violation may also occur as a result of unbalanced sample sizes across groups being investigated. In the case of two continuous variables, the violation of this assumption can often be ignored, as it only results in a slightly poorer fit in regression models.[7] However, recent simulation studies reveal that in the case of grouped data, violations of this assumption can result in either lower power or increased probability of Type I statistical conclusion error.[14] Applying data transformations or increasing the sample size (when possible) may rectify the problem, but not always. Moreover, in some cases, transformations or increases in sample size may result in worse outcomes with regard to statistical conclusion errors.[14] Furthermore, in the case of multivariate analyses, trying to rectify violations observed in pairwise comparison of variables may not necessarily translate into satisfying the more complex assumptions in a multivariate combination of variables.[7] Therefore, there are really no clear guidelines on a single bulletproof strategy when faced with a violation of homogeneity of variance. Given the complexity of the case when this assumption is violated, it is the reasoned judgment of the first author that, when faced with this situation in the data, analyses should be conducted as one normally would, but it should be noted in the results that violation of the assumption of homogeneity of variance was present in the data, and *P* values should therefore be interpreted with some level of caution.

An alternate approach to dealing with violations of distributional assumptions can be addressed at the analysis stage by using various nonparametric methods of statistical analyses. Extensive discussions of the types of analytical methods can be found in Gibbons and Chakraborti[16] and Wasserman.[17] However, the above discussion on the robustness of conventional parametric statistical methods with regard to violation of assump-

tions should be kept in mind. Furthermore, it should be noted that nonparametric-based methods do come at the price of reduced power for the statistical test.[18] Additionally, a recent simulation study evaluating nonparametric alternates to the analysis of covariance (ANCOVA) demonstrated that the ANCOVA was generally superior to nonparametric techniques even under conditions of violated assumptions.[19] Within this context, it should also be noted that the ANCOVA is among the parametric statistical techniques with the most restrictive assumptions.[7] Despite the temptation to automatically assume that nonparametric statistics may be more likely to yield significant *P* values relative to parametric methods under violation of assumptions, it should be kept in mind that research in clinical settings are less about chasing after a significant *P* value and more about demonstrating the efficacy and effectiveness of some type of intervention; and as will be discussed in the next section, statistically significant results can sometimes be clinically meaningless.

INTERPRETATION OF DATA

With the advances in computing technology and the evolution of statistical software packages, the process of data analysis has become infinitely easier over the past two decades. Modern data analysis packages offer a graphic user interface that allows data to be analyzed with the push of a button. However, even the most advanced software package will fail miserably at being able to apply reasoned judgments to the output of results, dissecting and separating statistically significant differences from clinically meaningful differences, and then draw informed conclusions that add sound evidence to the ever-increasing body of knowledge. The major issues related to interpretation of results from clinical trials can be classified as follows: *inferences of causality*, *controlling for baseline factors*, *statistical significance*, and *interpreting the effect size*. These issues will be discussed within the context of a clinical trial.

Inferences of Causality

Causality refers to a situation where some sort of systematic variation in treatment (ie, analgesic drug vs. placebo) can be inferred to have a direct causal relationship with the difference in outcomes observed between the treatment groups. In other words, if a statistically significant difference is observed between the treatment groups where the group receiving analgesic drugs reported more positive outcomes than the placebo group did, it can be concluded that the analgesic drug caused the positive outcome. However, it should be noted that it is not the statistical significance per se that establishes a causal link between the drug and the positive outcomes. The basis for inferring causality is pivoted upon the design of the study, through the use of a well-designed RCT that rules out other factors that may contribute to the difference in outcomes.

The ability to draw causal inferences from data breaks down when there are threats to the internal validity of the study. Internal validity is jeopardized when factors other than the systematically manipulated independent variable end up causing observed differences between treatment groups. As noted by Turk and Rudy,[18] examples of threats to internal validity include instrumentation or calibration problems, memory-based anchoring effects from repeated assessments, integrity of treatment implementation, patient attrition or noncompliance, crossover of patients between treatment groups, and experimenter or patient expectancy effects. Prior to conducting the trial, all these potential threats to internal validity should be addressed as part of the design and implementation of the research protocols. In the case of patient attrition and crossovers to another treatment group, an intention-to-treat (ITT) analysis should be incorporated as part of an RCT design. The ITT analysis treats all patients as part of the trial and as part of their assigned treatment group throughout the duration of the trial regardless of whether the patient has dropped out or crossed over.[20] Therefore, an ITT component as part of the design of the trial would require that data on all patients be collected throughout the duration of the trial. If there are certain underlying and unaddressed factors contributing to higher attrition or crossover rates in one treatment group relative to another, simply excluding the patients introduces a bias that increases the probability of committing Type I statistical conclusion errors.[20]

While the major advantage of a well-designed RCT provides the opportunity to draw causal inferences, it should be noted that the implementation of an RCT design per se does not guarantee the highest quality of evidence if internal validity is jeopardized or the wrong types of analyses are used.[21,22] Furthermore, the inability to utilize an RCT design does not absolutely disqualify the quality of evidence that may be obtained from certain well-designed nonrandomized trials.[23] In certain cases, randomly assigning patients to treatment groups may not be feasible or even ethically or legally permissible. When faced with such limitations, there are several other methodologies that may be profitably used

in designing the clinical trial. To aid with this purpose and to qualify any evidence obtained from nonrandomized designs, researchers should consult the criteria for levels of evidence published by the Oxford Center for Evidence-Based Medicine.[24] While RCT designs are rated as producing the highest level of evidence, these are followed, in descending order of quality, by: retrospective and prospective nonrandomized cohort studies, case-control designs, case series, and, finally, expert opinion. While the last two types of studies fail to provide reliable evidence, nonrandomized cohort studies are common in the published literature on clinical trials and, in some cases, have quality of evidence similar to that of RCT designs.[22] However, it should be noted that if the clinical trial was not an RCT design, no causal inferences should be made from the data. The only inferences that can be made are of associations between or among variables.

Controlling for Baseline Factors

Related to factors that can improve internal validity of a study is the effort to take into account baseline characteristics of the study sample. In almost all clinical trials, baseline measures on certain constructs are often collected. These may include initial pain scores, psychological well-being, or health-related quality of life measures. These same measures are subsequently collected at post-treatment. As is often the case, baseline measures are substantially correlated with subsequent post-treatment because both are the same measures tapping into individual differences on a given construct.[25] The individual differences present in the data contribute to "noise" or the overall error variance in data analysis. In order to reduce this error variance, thus, increasing the power of the test, the baseline scores can be profitably utilized as a covariate in an ANCOVA. In this case, the individual differences within the baseline measure (covariate) are assessed as a systematic effect within the ANCOVA, thus, removing it from overall experimental error. Therefore, in RCT designs, where it can be reasonably expected that the baseline measure in each group does not systematically vary with respect to each other, an ANCOVA should be utilized as a statistical test of group differences on post-treatment measures.

Another case where the ANCOVA may be useful is to adjust for pre-existing baseline differences that are significantly correlated with the outcome variable being studied. These pre-existing differences between groups often occur in nonrandomized studies, where naturally occurring groups are used in the clinical trial.[7,25] However, baseline differences may also be observed in well-designed RCTs simply by virtue of chance. In this second application of the ANCOVA, the idea is to assess group differences on post-treatment measures, assuming that all patients in both groups were equal on the baseline characteristic in question. Unfortunately, this second approach to the ANCOVA is not without its drawbacks. Caution should be applied in interpreting the results of an ANCOVA if there is a reason to believe that there is a causal link between the differences in the baseline measures and the presence in a given group.[25] In such cases, utilizing an ANCOVA may remove some of the systematic variance associated with the grouping variable, thus, reducing the power of the test.

Statistical Significance

The statistical significance of a hypothesis test is a major component of the decision making about whether a given treatment is associated with differences in outcomes relative to another type of treatment or control group. A brief description of the types of statistical conclusion errors and the power of statistical tests is warranted to provide some context to the discussion in the following sections.

Conventionally, the criterion for concluding a statistically significant relation is a probability value, *P*, of less than 0.05. This value represents the Type I error rate of statistical conclusions, that is, the probability of incorrectly concluding that there is a significant difference when no real differences exist (false positive). A counterpart to the Type I error is the Type II error, which corresponds to the probability of incorrectly concluding no significant differences when real differences exist (false negative). Related to the Type II error is the power of the test, which is defined as 1 minus the probability of a Type II error. Both the error rates are linked together in a trade-off relationship. Lower values of Type I error rates are associated with higher values of Type II error rates. In practical terms, this implies that when a test is more conservative (ie, a smaller Type I error), there will exist a greater probability of not being able to detect true differences (ie, a larger Type II error), thus, rendering the test less powerful. Conversely, if a test is more liberal(ie, a larger Type I error), there will be a greater probability of picking up trivial differences (ie, a lower Type II error) with a highly powered test.

The desired power of the test should be addressed earlier on during the design stage (and as a result, the Type II error as well), thus, providing at least the minimum required sample size for the trial. At the stage

of the analysis, all hypothesis tests are usually conducted at the conventional 0.05 level of the probability of a Type I error. Therefore, any difference between groups detected with a corresponding *P* value of less than 0.05 implies that there is less than a 5% chance of incorrectly rejecting the null hypothesis, given that no true differences exist. However, researchers should guard against potential inflation of the Type I error level during the analysis, which can make hypothesis tests more liberal than the conventional standard of a 0.05 probability level. The primary source for this problem is when multiple comparisons are conducted within the study. Generally, this case occurs in designs with more than two groups and where multiple exhaustive pairwise comparisons (or a subset thereof) are conducted on a given dependent variable or outcome. Any such comparisons must be controlled for Type I error inflation. The mechanisms for doing so are readily available in all statistical software packages. A common error-correction technique for multiple comparisons on a given outcome variable is the Tukey's Honestly Significant Difference test for pairwise comparisons. The Tukey Honestly Significant Difference test is a preferable method of error correction because of its greater power, compared with more conservative error correction techniques such as the Bonferroni test.[14]

If comparison with a control group is desired rather than multiple pairwise comparisons, then the Dunnet test may be more appropriate for multiple comparisons against a single control group. Apart from the straightforward case of multiple comparisons on a single outcome variable, there is still considerable debate on whether multiple tests on a range of different outcome variables should be corrected for Type I error inflation.[26] For example, if a study utilizes 5 outcome variables, should each test be corrected for the total number of 5 tests being conducted in the study? Such corrections often come at the expense of significantly reduced power because of the smaller Type I error threshold, and, as pointed out by Aicken and Gensler,[26] these types of stringent error-correction techniques may not warrant for preliminary or exploratory studies.

Another issue concerning the significance of statistical tests is closely related to power and sample size. While RCTs are generally moderate in their sample size because of the logistics involved in designing and executing the study, retrospective and prospective cohort studies can often have sample sizes that run into the thousands. Large sample sizes result in smaller standard error values in the sampling distribution of the parameter estimate of a statistical test.[15] Because power is inversely related to the standard error, large sample sizes can have a profound impact on the power of a test. As a result, large sample studies are often overpowered and have a high probability of picking up statistically significant differences that are trivial in magnitude. Therefore, it would not be very wise to rely solely on the significance value of $P < 0.05$ for a hypothesis test when running large sample studies. A statistically significant difference may be a trivial difference within clinical contexts, that is, a 3-point difference on a depression scale or a 1-point difference on a pain intensity scale. In such cases, it is wiser to shift the interpretation away from *P* values alone and start to look at standardized effect size statistics that describe the relative magnitude of the statistically significant effect.

Interpreting the Effect Size

As noted by Worzer et al.,[27] the effect size is a simple and well-established statistic for determining the practical importance of statistically significant differences on common measures used in clinical settings. Furthermore, it also facilitates a standardized comparison across different studies that assess similar or related constructs because the effect size is a standardized statistic that does not rely on raw score units. The type of effect size measure to be used ultimately depends on the nature of the data and the choice of analysis. In the most general sense, effect sizes vary depending on whether the analysis compares two continuous variables, a continuous variable between two groups, a continuous measure across multiple groups, and two categorical (or grouped) variables. The most common type of effect size for each of these scenarios is discussed.

***Pearson's* r.** The Pearson's product-moment correlation coefficient, *r*, is the most basic form of an effect size, and many of the other effect size measures can be derived from it. It is best used to describe the magnitude of association between two continuous variables. As noted by Cohen,[28] a general guide to interpreting the magnitude of an effect using *r* can be defined as follows: 0.1 for a small effect, 0.3 for a medium effect, and 0.5 for a large effect. In practical terms, correlation coefficients of 0.1 or less can often be interpreted as being a trivial association between two variables. However, one should keep in mind that in large sample studies, a correlation of around 0.1 may be statistically significant simply because of the nature of an overpowered test. Another common way of using the correlation coeffi-

cient as an effect size is to square it and report the coefficient of determination, r^2. The coefficient of determination simply describes the percent of variance in one variable accounted for by the other variable. For example, a large correlation coefficient of 0.5 corresponds to an r^2 of 0.25, or 25% of the variance that is shared between two variables. In clinical trials, a common use of r as an effect size measure is when two continuous variables are being evaluated in the study, that is, when analgesic drug dosage is assessed for its effect on post-treatment pain ratings. However, a major weakness of using r as an effect size is that there must be a fairly linear relationship between the two variables. Departures from linearity in the association between two variables results in an underestimate of the magnitude of effect because r captures the linear, straight-line relationship between two variables.[7]

***Cohen's* d.** This effect size measure is most commonly used for interpreting the analysis of two groups. It applies to both the independent-group's *t*-test as well as the paired-sample *t*-test. It is often provided as an option within the *t*-test analysis of most statistical analysis software packages but can also be computed easily by hand. Cohen's *d* is basically the difference between two group means divided by an estimate of the variance. The estimate of the variance will depend on the nature of the design or the homogeneity of variances across the two groups, with the baseline variance used as an estimate when dealing with pre-versus-post comparisons and a pooled variance estimate used when variances in the groups are not homogenous.[27] As noted by Cohen,[28] the *d* statistic is a standardized score and corresponds to the following effect magnitudes: 0.2 for a small effect, 0.5 for a medium effect, and 0.8 for a large effect. For an example in clinical trials, the Cohen's *d* statistic will be a useful complement to statistical tests assessing the effect of treatment with an analgesic vs. a placebo on post-treatment pain ratings. Caution should be exercised when interpreting any statistically significant differences when effect size magnitudes are close to 0.2 or lower, especially in large sample studies.

***Eta-Squared* (η^2).** This effect size statistic, often denoted in text as η^2, is a common effect size measure used in all the ANOVA-related analysis techniques and is a standard feature in all statistical software packages. The value of η^2 can range between 0 and 1 and corresponds to the percent of variance in the dependent variable accounted for by all the independent variables assessed. For assessing the effect of individual independent variables, the *partial* η^2 is reported for each independent variable in order to represent its unique association with the outcome.[7] Cohen's[28] guidelines for interpreting effect size magnitudes of η^2 are as follows: 0.01 for a small effect, 0.09 for a medium effect, and 0.25 for a large effect. Within the context of a clinical trial, consider an analysis based on a two-factor ANOVA that assesses two different treatment modalities. The second independent variable assessed, in order to determine if receiving compensation impacted response to the treatment modalities (by testing the interaction of treatment group by compensation status), was whether patients were on any type of disability compensation. Suppose that significant P values of <0.05 were reported for both the effects of treatment group (partial $\eta^2 = 0.15$) and the treatment group by compensation status interaction (partial $\eta^2 = 0.02$). On the basis of significant P values alone, one would conclude that, while there was an overall difference in the outcome measure based on the type of treatment modality, receiving disability compensation moderates a patient's response to the treatment. However, upon looking at the effect size statistics, one observes a very weak association between the interaction effect and the outcome. Therefore, a discussion of the results should note that much stronger evidence is needed on the role played by disability compensation before any conclusions can be drawn about its impact on the response to the treatment modalities.

The Odds Ratio and Phi. Both these effect size measures are appropriate for describing the association between two categorical variables when conducting significance tests using the χ^2-test statistic. The odds ratio is the more common effect size measure reported in medical journals when evaluating binary variables (ie, the variable can take on one of two values, such as presence or absence of disease). Consider the case of two randomly assigned treatment groups (interdisciplinary pain management vs. standard primary care) assessed for the effect of predicting surgery rates during a 2-year follow-up period. At the end of the follow-up period, it was determined that 25 out of 50 patients in the standard care group ended up receiving surgery (an odds of 1 patient receiving surgery for every 1 patient who did not receive surgery), while only 10 out of 50 patients in the interdisciplinary treatment group received surgery (an odds of 0.25 to 1, or 4 patients not receiving surgery for every 1 patient who did receive surgery). Therefore,

the odds ratio is simply 1 divided by 0.25. Given a significant P value of <0.05 for the χ^2-test, it can be interpreted that the standard care group had 4 times greater odds of receiving surgery during the 2-year follow-up period, relative to the interdisciplinary rehabilitation group. However, it should be noted that there are no standardized benchmark values for interpreting the magnitude of the odds ratio. Therefore, using the phi statistic in addition to the odds ratio can provide an alternate effect size measure that also corresponds to standardized effect size magnitudes. Another advantage of the phi statistic is that it can be generalized to more than just binary variables by using the Cramer's phi statistic, and can handle multiple categories on either variable.[15] The phi and Cramer's phi statistics are also provided in all statistical software packages within the χ^2 analysis. In order to determine standardized effect size magnitudes for these phi statistics, they are first converted into Cohen's[28] w-index, given by the equation:

$$w = phi * \sqrt{(k-1)}, \quad (1)$$

where k corresponds to whichever is the smaller in the number of rows or columns (ie, the number of categories in either variable). The w-index then corresponds to the following effect size magnitudes: 0.1 to 0.3 for small effects, 0.3 to 0.5 for medium effects, and >0.5 for large effects.

SUMMARY AND CONCLUSIONS

This article focused on three major aspects of conducting a clinical trial. Data management plays a major role in facilitating accurate and reliable access to clinical data. Therefore, great care should be placed at the outset, before the trial begins, in planning the data collection and management strategies. While the specifics will vary across different settings, a common rule of thumb is to have trained personnel handle data entry meticulously and to have proper organization and storage of data in a central database. Once the data management protocols are implemented and the trial begins, periodic quality control should be conducted in order to maintain the integrity of the database. Prior to conducting any analyses, problems related to missing data, invalid values, outliers, and violations of distributional assumptions should be identified and addressed.

Upon analyzing the data and reviewing the results, one should be careful not to infer causality if the study was not based on an RCT. If the study was not an RCT, then conclusions about causal relationships should be avoided. Furthermore, it is advisable to assess if any previously unidentified threats to internal validity may undermine the conclusions that can be drawn. These may include statistically controlling for any baseline differences that may exist between treatment groups. In addition, if an ITT protocol was included as part of the study, all patients should be included in the analysis, regardless of whether they dropped out or crossed over to a different treatment group within the trial. Threats to statistical conclusion validity should be kept in mind, especially if multiple comparisons are utilized or if the study consisted of a large sample resulting in overpowered tests. Finally, interpretation of the data should not only be limited to conclusions drawn from statistical significance but also needs to incorporate the magnitude of observed effect in order to avoid emphasizing potentially trivial results.

While it is impossible to address every little detail that goes into conducting a clinical trial and interpreting the data, the discussion in this article provides some general guidelines that clinical practitioners and researchers may adopt for their own research projects within the clinical setting. A Recommended Resources section is provided to guide readers to the more detail-oriented aspects of conducting a clinical trial.

ACKNOWLEDGEMENT

Support, including Spanish translation, was provided by an unrestricted educational grant by Allergan.

REFERENCES

1. Lipman AG. Evidence-based pain management and palliative care: the Cochrane Collaboration and its Pain, Palliative, and Supportive Care Collaborative Review Group. *Am Pain Soc Bull*. 2004;14:12–19.

2. Wiffen PJ, Fairman FS. The Cochrane Collaboration Pain, Palliative Care and Supportive Care Collaborative Review Group. *J Pain Palliat Care Pharmacother*. 2002;16:69–79.

3. Ban TA, Guy W, Wilson WH. Organizing and conducting clinical trials. *Neuropsychobiology*. 1983;10:137–140.

4. Harden RN, Bruehl S. Conducting clinical trials to establish drug efficacy in chronic pain. *Am J Phys Med Rehabil*. 2001;80:547–557.

5. Mazumdar S, Tang G, Houck PR, et al. Statistical analysis of longitudinal psychiatric data with dropouts. *J Psychiatr Res*. 2007;41:1032–1041.

6. Turk DC, Monarch ES. Biopsychosocial perspective on chronic pain. In: Turk DC, Gatchel RJ, eds. *Psychological Approaches to Pain Management: A Practitioner's Handbook*. 2nd ed. New York: Guilford; 2002:3–29.

7. Tabachnick BG, Fidell LS. *Using Multivariate Statistics*. 4th ed. Boston, MA: Allyn & Bacon; 2001.

8. Siddiqui O, Ali MW. A comparison of the random-effects pattern mixture model with last-observation-carried-forward (LOCF) analysis in longitudinal clinical trials with dropouts. *J Biopharm Stat*. 1998;8:545–563.

9. Cook RJ, Zeng L, Yi GY. Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics*. 2004;60:820–828.

10. Lane P. Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches. *Pharm Stat*. 2008;7:93–106.

11. Mallinckrodt CH, Raskin J, Wohlreich MM, Watkin JG, Detke MJ. The efficacy of duloxetine: a comprehensive summary of results from MMRM and LOCF_ANCOVA in eight clinical trials. *BMC Psychiatry*. 2004;4:26.

12. Nyiendo J, Attwood M, Lloyd C, Ganger B, Haas M. Data management in practice-based research. *J Manipulative Physiol Ther*. 2002;25:49–57.

13. Hoaglin DC, Mosteller F, Tukey JW, eds. *Understanding Robust and Exploratory Data Analysis*. Hoboken, NJ: John Wiley & Sons; 1983.

14. Keppel G, Wickens T. *Design and Analysis: A Researcher's Handbook*. 4th ed. Upper Saddle River, NJ: Prentice Hall; 2004.

15. Hays WL. *Statistics*. 5th ed. Fort Worth, TX: Harcourt Brace; 1994.

16. Gibbons JD, Chakraborti S. *Nonparametric Statistical Inference*. 4th ed. New York, NY: Marcel Dekker; 2003.

17. Wasserman L. *All of Nonparametric Statistics*. New York, NY: Springer; 2006.

18. Turk DC, Rudy TE. Methods for evaluating treatment outcomes: ways to overcome potential obstacles. *Spine*. 1994;19:1759–1763.

19. Vickers AJ. Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Med Res Methodol*. 2005;5:35.

20. Lachin JM. Statistical considerations in the intent-to-treat principle. *Control Clin Trials*. 2000;21:167–189.

21. Faulkner C, Fidler F, Cumming G. The value of RCT evidence depends on the quality of statistical analysis. *Behav Res Ther*. 2008;46:270–281.

22. Poolman RW, Struijs PA, Krips R, Sierevelt IN, Lutz KH, Bhandari M. Does a "Level I Evidence" rating imply high quality of reporting in orthopaedic randomised controlled trials? *BMC Med Res Methodol*. 2006;6:44.

23. Gatchel RJ, Maddrey AM. Experimental design issues in clinical research of musculoskeletal pain disabilities. *Crit Rev Phys Rehabil Med*. 2000;12:91–101.

24. Phillips B, Ball C, Sackett D, et al. Oxford centre for evidence-based medicine levels of evidence. 2001. Available at: http://www.cebm.net/index.aspx?o=1047 (accessed June 15, 2008).

25. Nunnally JC, Bernstein IH. *Psychometric Theory*. 3rd ed. New York, NY: McGraw-Hill; 1994.

26. Aicken M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs. Holm methods. *Am J Public Health*. 1996;86:726–728.

27. Worzer W, Theodore BR, Rogerson M, Gatchel RJ. Interpreting the clinical significance of pain questionnaires. *Pract Pain Manage*. 2008;8:16–29.

28. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

Appendix

Recommended Resources

1. Capterra, Inc.—Clinical Trial Management Software Directory. http://www.capterra.com/clinical-trial-management-software
2. Keppel G, Wickens T. *Design and Analysis: A Researcher's Handbook*. 4th ed. New Jersey: Prentice Hall; 2004.
3. Nyiendo J, Attwood M, Lloyd C, Ganger B, Haas M. (2002). Data management in practice-based research. *J Manipulative Physiol Ther.* 2002;*25*(1): 49–57.
4. Oxford Centre for Evidence-Based Medicine—Levels of Evidence. http://www.cebm.net/index.aspx?o=1025