Overview of the Role of Statistic Analysis in the Design of Spine-related Studies

Donna D. Ohnmeiss

Int J Spine Surg 2009, 3 (1) 26-29
doi: https://doi.org/10.1016/SASJ-2009-Comment1
https://www.ijssurgery.com/content/3/1/26

The International Journal of Spine Surgery
2397 Waterbury Circle, Suite 1,
Aurora, IL 60504, Phone: +1-630-375-1432

Overview of the Role of Statistic Analysis in the Design of Spine-related Studies

Donna D. Ohnmeiss, Dr.Med[a]

INTRODUCTION

We have entered a very exciting time in spine research. There are a myriad of new implants, minimally invasive techniques are continually being developed and refined, and disc tissue regeneration is looming on the horizon. Patients now have greater access to healthcare information and are exposed to much more marketing than of medicine than in the past which shapes their expectations and desires for particular avenues of care.

SAS Journal. March 2009;3:25–28. DOI: SASJ-2009-Comment1

[a]Texas Back Institute Research Foundation, Plano, Texas, and Special Deputy to the Editor-in-Chief of the *SAS Journal*

Address correspondence to Dr. Donna D. Ohnmeiss, Dr.Med., Texas Back Institute Research Foundation, 6020 W. Parker Rd. #200, Plano, TX 75093 (email: dohnmeiss@texasback.com)

Dr. Ohnmeiss is employed by a foundation that receives fellowship program support from Zimmer Spine, DePuy Spine, Synthes Spine, and Medtronic Neurological; and research project support from TranS1, Inc.

This sudden growth in treatment options has brought with it concerns about efficacy and costs. Guyer described this paradox in medicine.[1] Somewhere in all of this, one of the primary means to sort things out is what is viewed by many as the mundane and old-fashioned basics of solid methodology in the collection and analysis of data. The tools for conducting clinical research are becoming much more expansive as well. There are now automated data collection systems, electronic medical records, and extremely powerful software programs for analyzing data. With these resources, there has never been greater potential to meaningfully perform outcome studies and refining indication and contra-indications for the various emerging interventions. The tools of data collection instruments and statistical software are only as good as the manner in which they are applied. This may be viewed as analogous to the existence of the ideally designed spine arthroplasty device that fails to produce optimal results due to poor patient selection or poor execution of the surgical procedure. Any tool can only be as good as the manner in which it is used. As stated by Petrie, "The marked improvement, accessibility and ease of use of statistical software in recent years has led to a proliferation of errors which can be attributed to a lack of awareness of the consequences of inappropriate design and analysis rather than to mistakes in the techniques used."[2]

Medical school residents and clinicians seem to agree that understanding statistics is important to study design and reading the literature.[3, 4] Residents indicated that they understood the statistics in about 25% of the literature they read. In a study involving students and faculty at the Mayo Clinic, 23% of participants indicated they could determine whether appropriate statistical methods had been used in a study, 28% thought they could design their own study, and only 15% were confident that they could conduct their own statistical analyses.[3]

The purpose of this paper is to present some current issues impacting spine-related studies that are rooted in study design, including statistical analysis and reporting of results. A brief over-view of some tests and when they should be used is included. However, a discussion of all of the relevant statistical testing options and their use is well beyond the scope of this writing, as entire textbooks have been written on many of the individual analyses.

Statistics is a part of the study design

Perhaps one of the most important concepts is that statistics is not just a painful step toward the end of a study when a data file is provided to someone given the instructions to analyze it. It is very important for someone who is familiar with study design and data analysis to be included as a part of the study design from the inception. This goes well beyond just performing a power analysis for sample size. Ideally, someone familiar with studies in the particular discipline being investigated should be involved. They can provide valuable input into what types of data to collect, which versions of questionnaires to use, data file structure, discourage trying to collect too much data for which there are no plans on how to use, and may be able to provide insight as to problems encountered with previous studies in the same area.

Creating data entry files

There is likely no one better acquainted with the importance of data file structure that those who have to analyze the data.

SASJournal

Many small items such as recording height in inches, or having a feet field and an inches field that can be computed into one field later, rather that in mixed units of measure such as 5'10" can be addressed early if someone who is familiar with the format needed for analysis is involved with creating the files from the start. Each such entry has to be re-entered as a single numerical value. It is doubtful that unless someone has been through the exercise of having to re-enter multiple variables for a large number of patients can the frustration of this be appreciated. Sometimes it is helpful to design the database file with drop down menus for entry, particularly for string (text) data. This tends to create consistent entry and avoids time wasted going through the file to correct typographical errors, variation in terminology (such as using herniated disc, HNP, disc herniation all for the same diagnosis but will be recognized by software as three distinct diagnostic groups) and being consistent with capitalization. Many software programs will identify "male" and "Male" as 2 distinct values for a gender variable. For large multicenter trials, especially FDA IDE trials, the file structure is often determined by a database manager. It is important for this person as well as the statistician to have a role in the protocol development and design of case report forms and other data collection instruments, be they on paper or electronic format.

Electronic data capture has many potential advantages over traditional paper collection. First, there are no data entry errors when entering from other sources, patients or clinicians completing the forms cannot select 2 responses for the same item or write in qualifiers to explain their answer, and data is collected and available for review in real-time. However, despite the enthusiasm and benefits of electronic data capture, it should be made clear that it cannot overcome problems with using non-validated questionnaires, poor study design, and in some instances it actually introduces non-validated collection methods.

What test should be used?

Seeing all of the test options available in a statistical software package can be overwhelming. Not only are there many tests, but many have options to modify each by changing the default settings. The type of test(s) used to analyze data for a particular study should be derived based primarily on 2 key factors: 1) the characteristics of the data (continuous, categorical, parametric, etc.) and the question to be addressed (comparing groups, determining change over time or after intervention, looking for association between variables, etc.).

The characteristics of the data are essential in determining which test is most appropriate to use. Several terms are important in this arena. Continuous data are those that are measured on a continuous scale such as age or visual analog scale (VAS) value. Arithmetic properties apply; that is, one can add, subtract, calculate percentage change for variables with continuous values.

One frequently-overlooked item with continuous data is the distribution of the data. Generally, everyone is familiar with normally-distributed data or the "bell curve." However, not all continuous data fits into this category. The data may be distributed to be skewed to the left or right, meaning that a large number of patients tend to cluster at one end of the scale. The reason why this is important is that many commonly used analyses, such as a *t* test, are based on the assumption of normally distributed data. If this assumption is violated, the results from a *t* test may not be valid.

The other commonly encountered data type is categorical, which has several subgroups. Categorical data may be ordinal, which means that the values assigned imply a ranking. For example, a 4-point scale of poor, fair, good, or excellent for assessing outcome implies a ranking that provides information related to which patients did better, or worse, than others after treatment intervention. While numerical values may be assigned to the four categories, unlike continuous data, arithmetic properties are not applicable. One cannot assume that an "excellent" outcome is achieved by doing exactly twice as well as a patient with a "fair" outcome. Non-ordinal categorical data has no ranking and includes variables such as gender, insurance type, and diagnosis.

The other important factor in determining which test is most appropriate to use is what does one want to learn from the data. Such questions include: did patients improve after treatment, is there a difference in outcome between different treatment groups, and are 2 variables related.

Types of test to use for data analysis

One of the most commonly used tests is the very familiar *t* test. There are 2 basic versions available. The independent sample *t* test is used when comparing the mean values of a variable with normally distributed data from 2 groups, such as comparing VAS scores from investigational and control groups to determine if the groups are significantly different. If more than 2 groups are to be compared, ANOVA is preferred to multiple *t* tests between sets of 2 groups. A paired *t* test is used when comparing the values of the same variable at 2 different times such as VAS scores before and after treatment.

As described earlier, one of the often-overlooked underlying assumptions for applying the *t* test is that the data is normally distributed. If it is not, then testing involving median values, rather than mean values, is more appropriate. Although variables, such as VAS scores, may be normally distributed prior to an intervention, this does not mean that they will be normally distributed after intervention. This was well described

by Geisler when reviewing and re-analyzing data from the CHARITÉ FDA IDE trial.[5] The postoperative VAS and Oswestry scores were heavily skewed toward the lower end of the scale, indicating that the majority of patients had very low pain and disability scores after surgery. When reanalyzing the data adding the non-randomized training cases and using the non-parametric Wilcoxon rank sum test, which does not require the data to be normally distributed, it was found that the artificial disc group had VAS and Oswestry scores significantly less than those in the fusion group at all of the follow-up periods.

The Wilcoxon rank sum test, also called the Mann–Whitney U, is designed primarily to compare groups based on a variable with ordinal categorical or non-parametric continuous values. It is more appropriate to use than a *t* test to compare between groups outcome that is classified on an ordinal scale such as the four-point poor to excellent outcome scale described earlier.

One of the other commonly used tests is χ^2 (the chi-square test). This is most suitably applied to categorical data. It is designed to compare the proportion of subjects associated with a particular response in one group to that in another group. A Fisher's exact test may be used with the same type of data as a χ^2 but when one or more of the cells in the data table have an expected value of 5 or less.

Frequently-overlooked tests

One test that seems to be underutilized is the McNemar. McNemar may be loosely thought of as a non-parametric categorical version of the paired *t* test in that it is designed to test for change when subjects respond to the same question on more than one occasion. One application of this test would be in determining if there has been a significant change in the proportions of patients working and not working before and after an intervention.

One very powerful, yet seemingly rarely-used tool, is regression analysis. When working with multiple variables, this test can be used to identify which are significantly related to the variable of interest and which are not. For example, to identify factors related to the postoperative work status for patients enrolled in a particular study, the postoperative work status can be assigned as the dependent variable and all the variables of age, gender, educational level, smoking status, insurance type, length of time off work preoperatively, job demand classification, level(s) operated, preoperative work status, number of levels operated, surgery type, etc. as independent variables. Regression analysis is particularly helpful when some of the independent variables are possibly related such as insurance type, job demands, preoperative work status, and the length of time off work before surgery. The analysis will select the one of these variables that is most significantly related to postoperative work status and, then once the effects of that variable have been accounted for, determine if any of the remaining items are significantly related to the dependent variable of work status. This is preferable to multiple univariate analyses in which each variable is assessed separately to determine its potential relationship to the independent variable. As such several of the variables may be found to be related to the independent variable of interest, such as return to work. However the results may be somewhat misleading and too many dependent variables identified as significantly related to the independent variable if they are interrelated.

Logistic regression analysis (used to determine which variables if any are significantly related to a dichotomous variable – can only take on one of 2 values) was used by Guyer et al. in determining factors differentiating patients with the best and worst outcomes following lumbar total disc replacement.[6] In that study it was found that the length of time off work prior to surgery was the variable most strongly related to the best versus worst outcome classification. After the effect of that variable was accounted for, none of the other variables investigated had an impact on best versus worst outcome classification.

Lack of consistency in study design

There is little consistency in the design and reporting of results of spine studies. One good example of part of the problem are the recent total disc replacement (TDR) trials.[7, 8] When reviewed individually, the overall design of each of these studies was based on the traditional format for FDA device trials and a blending of real-life factors of available control groups, costs, patient recruitment, and likely input from key investigators and the clinical research organizations contracted to conduct the studies. However, due to the study designs used in the trials, a very great opportunity was likely missed. The studies used different versions of questionnaires, different methods of assessing satisfaction and work status, and different definitions of success. While there was certainly nothing wrong with the way the data were reported in either paper, it is frustrating for those who later want to do meta-analyses that the data cannot be combined from 2 studies that intuitively should be very similar.

The problem with inconsistency should improve through the use of electronic data capture systems. However based on the author's experience during the past several years, it is not safe to presume that all such systems are using validate instruments in a validated method. One should carefully review the questionnaires and how they are administered before committing to using any data collection system. One of the easiest items to check is which version of the Oswestry is used.[9] If this is not a current or validated version, use of the program should probably be avoided.

In spine, there seems to be broad acceptance for *t* tests and χ^2, with familiarity to a lesser extent for correlation

coefficients and ANOVA analyses. Unfortunately, it often seems that anything beyond this rather limited tool set is thought of as statistical voodoo or data manipulation to gets the desired results. While no doubt the "supermarket" approach to data analysis (that is, conducting various tests on the data until the one that produces the highly sought after $P < 0.05$ is identified) has been used abusively, there is nothing wrong with using techniques beyond most readers' comfort zones. Readers and reviewers need to become more comfortable with tests beyond those that are currently the most commonly used. There are occasions, although likely to be frowned upon by statistical purist, to present and analyze the data in more than one form. For example, if in the total disc replacement trials, to use the Wilcoxon analysis which is more appropriate than the *t* test based on the data distribution, but also report the mean values for the VAS scores since these have been used in so many other spine studies and may be very helpful to readers to interpret the data in the context of previous work. However, in such scenarios, the authors must make it very clear that the mean values were not used for analyses and are provided purely as reference data.

DISCUSSION

The demands for evidence-based medicine, the ability to compare interventions, identifying the most cost-effective treatments, identifying which patients are the best candidates for which procedures has never been greater for spine care providers. In evaluating treatment options, patients often have a simple question: "What is the success rate?" Unfortunately, in spine the definition of what constitutes a successful spine surgery remains elusive, with little in any agreement in how outcomes are classified, if the term "success" is mentioned at all.[10]

Part of the problem arises out of tradition and reporting results within the comfort zone of reviewer and readers. It is sometimes thought that not staying with the traditional *t* test and χ^2 is just a way to produce the all important "$P <$ .05" when these tried and true methods failed to do so or as it is sometimes phrased "just statistical mumbo-jumbo to make the data say what they want it to." Unfortunately, this view of data analysis only makes the situation worse and likely leads to missing much of the information that could be provided from a large data set simply because of lack of understanding on the part of the readers. The issues related to lack of consistency in using standardized, validated instruments and reporting results is not likely to change any time soon. Study design is often affected by issues such as what the FDA expects to see in device trials, the realities of the cost of doing research, patient safety and privacy concerns, and unfamiliarity of some of the people involved with study design with the finer points of data instrument selection, data entry process, and data analyses. Only by involving a team in the study design, execution, and interpretation of the data, with insight to the full breadth of a study will spine care providers move forward in addressing many of the current issues related to providing optimal spine care in an arena with so many new options and demands.

REFERENCES

1. Guyer RD. The paradox in medicine today-exciting technology and economic challenges. *Spine J* 2008;8:279-85.
2. Petrie A. Statistics in orthopaedic papers. *J Bone Joint Surg Br* 2006;88:1121-36.
3. West CP, Ficalora RD. Clinician attitudes toward biostatistics. Mayo Clinic proceedings 2007;82:939-43.
4. Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. *JAMA* 2007;298:1010-22.
5. Geisler F. Surgical Treatment for Discogenic Low-Back Pain: Lumbar Arthroplasty Results in Superior Pain Reduction and Disability Level Improvement Compared with Lumbar Fusion. *SAS Journal* 2007; 1:12-9.
6. Guyer RD, Siddiqui S, Zigler JE, Ohnmeiss DD, Blumenthal SL, Sachs BL, et al. Lumbar spinal arthroplasty: analysis of one center's twenty best and twenty worst clinical outcomes. *Spine* 2008;33:2566-9.
7. Blumenthal S, McAfee PC, Guyer RD, Hochschuler SH, Geisler FH, Holt RT, et al. A prospective, randomized, multicenter Food and Drug Administration investigational device exemptions study of lumbar total disc replacement with the CHARITE artificial disc versus lumbar fusion: part I: evaluation of clinical outcomes. *Spine* 2005;30:1565-75.
8. Zigler J, Delamarter R, Spivak JM, Linovitz RJ, Danielson GO, 3rd, Haider TT, et al. Results of the prospective, randomized, multicenter Food and Drug Administration investigational device exemption study of the ProDisc-L total disc replacement versus circumferential fusion for the treatment of 1-level degenerative disc disease. *Spine* 2007;32:1155-62.
9. Fairbank JC, Pynsent PB. The Oswestry Disability Index. *Spine* 2000;25:2940-52.
10. Ohnmeiss DD, Guyer RD, Rashbaum RF. What constitutes success in lumbar spine surgery? Spine Arthroplasty Society. Berlin, Germany; 2007.