JCO INTERVIEWS

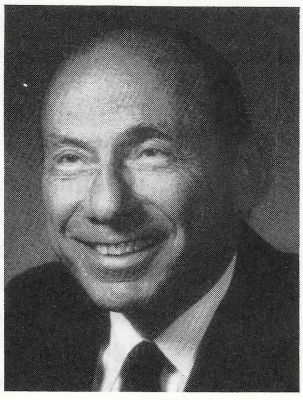
Robert G. Keim, DDS on Living with Statistics

DR. GOTTLIEB Let's face it: statistics is an arcane subject for most people. To the average journal reader, it may appear that the majority of "scientific" papers are not scientific and offer conclusions that are not validated by the statistical tests performed. Does that mean that the papers are without merit?

DR. KEIM Not really. To my mind, papers in which the statistical presentations are so complex and esoteric that the average practitioner cannot understand or interpret them are without merit outside of the world of academics. Incidentally, I do not agree with the statement that most scientific papers are not scientific. Statistics is not synonymous with the scientific method.

It is important that the reader avoid accepting the authors' conclusions as gospel truth simply because the writer has used statistical tests to prove significance. The reader must be critical. Statistically significant findings can be completely insignificant clinically. Vice versa, if a reader critically analyzes the data presented in a paper, he or she may well find information that would be very useful clinically, even if the authors failed to find statistical significance. Authors, reviewers, and editors are human and are quite capable of making mistakes. If readers are prepared to critically evaluate the scientific literature on their own, they can draw their own conclusions about published findings and may well find merit in even flawed papers.

Dr. Keim



Dr. Gottlieb

Dr. Keim is Book Editor of the *Journal of Clinical Orthodontics* and Associate Professor and Program Director, Department of Orthodontics, University of Tennessee College of Dentistry, 875 Union Ave., Memphis, TN 38163. He is in the private practice of orthodontics in Memphis. Dr. Gottlieb is Senior Editor of the *Journal of Clinical Orthodontics.*

DR. GOTTLIEB Is the first test of a journal paper a commonsense approach to the study design?

DR. KEIM Of course. Study design is an area of intellectual inquiry in and of itself, and there are numerous graduate-level courses, offered in many disciplines, devoted to the subject of research methodology and study design. Study design can be extraordinarily complicated, but in the final analysis, if a clinical study fails the test of common sense, it's essentially useless.

DR. GOTTLIEB Let's go over some of the things a reader should look for in articles that present data with statistical analyses. What is the implication of a large standard deviation?

DR. KEIM A large standard deviation implies that the individual subjects measured in the experiment vary widely from one another with respect to the variable that was measured. For example, let's say we measure the overbites of five different patients in each of two different practices. Let's say that in one, the scores for the

five overbites we measure turn out to be 1mm, 1mm, 1mm, 1mm, and 1mm. The average overbite here is 1mm. Now let's say that in the other practice, we get scores of 5mm, –10mm, –20mm, 10mm, and 20mm. Again, the average overbite is 1mm. The standard deviation in the first case is 0; the standard deviation in the second case is about 15.96. The overbites in the first case look a lot alike, those in the second don't look much alike at all, but in both cases, the average or mean overbite is the same. Those in the second practice display wide variability.

DR. GOTTLIEB What about the size of the sample?

DR. KEIM Sample size is important because it affects the power of a given statistical test to detect significant differences between the groups compared. Too small a sample size results in a low power, and it takes very large differences between the groups to show up as "statistically significant". On the other hand, a sample size that's too large results in finding statistically significant differences between groups when, in fact, no substantive differences exist. There are several ways to estimate valid sample size, including power analysis, tradition, and rules of thumb. For general use in clinical studies, a good rule-of-thumb sample size of 30 to 50 subjects per group is a safe bet.

DR. GOTTLIEB What is a true random sample?

DR. KEIM This is a sample wherein any and every subject has an equal chance of being assigned to any of the study groups. For example, let's look at a prospective study of Class II treatments, comparing effectiveness of cervical-pull headgear to bionator therapy. In order for this to be a true random sample, we would have to randomly assign each patient to one of the groups. We could do this by putting all of the subjects' names in a hat, then randomly drawing names to assign each patient to the headgear or to the bionator group, then treating them according to their group assignment. Assignment to either group could not be based on any diagnostic, behavioral, or treatment considerations other than the draw from the hat.

DR. GOTTLIEB How do you judge the validity of comparing one set of data to another?

DR. KEIM It depends on what you're comparing and why. Provided that the data sets were from similar populations to begin with and the factors that went into generating each set were similar, this can be a valid approach. To use the old cliche, you have to make sure that you are comparing apples to apples and oranges to oranges. You have to make sure that nothing other than the treatment could have affected one group differently from the other—in other words, you have to control extraneous variables. You have to make sure that the conditions under which the experiment was conducted were the same for both groups.

A classic example of the effects of extraneous variables is given in the area of research in education. A new approach to teaching reading was to be evaluated. The fourth graders in a school were to be the subjects. The students were randomly assigned to either the treatment group (the group that would be taught by the new method) or the control group (the group that would be taught by the conventional method). These students were from similar backgrounds and had similar demographics overall, and the experiment utilized random selection. The problem was that the new teaching technique was used in the morning, and the conventional technique was used after lunch. Any differences in effectiveness of the techniques could be due to a true difference in the efficacy of the techniques or due to something else, such as the possibility that kids might learn better in the morning. Which is it? The extraneous variable, time of day, could easily confound the results.

DR. GOTTLIEB How valid is it to compare results between different studies?

DR. KEIM The technique of meta-analysis

involves reviewing and combining the results of various previous studies. Provided the studies involved similar treatments and similar samples, and measured similar outcomes, this can be a useful approach.

DR. GOTTLIEB What do you think about treating one side of the mouth one way and the opposite side another to compare treatment effects?

DR. KEIM Scientifically, this is a good way to control for extraneous variables. Ethically, problems arise. What if the researcher started out with an experiment as you've described, then noticed shortly into the experiment that the treatment on one side was working much better than the other? Or even worse, what if one side was being damaged while the other was not? Any ethical operator would have to abort the experiment and render the better of the two treatments.

DR. GOTTLIEB Do twin studies have the same drawbacks?

DR. KEIM Again, they are a good way to control for extraneous variables, but the same ethical problems apply. What do you do when the treatment rendered to one twin is obviously working better than that being rendered to the other?

DR. GOTTLIEB Does clinical research by its very nature, with no controls, not lend itself to statistical analysis?

DR. KEIM First, it is important to point out that clinical research can indeed have controls. Provided that studies are conducted on a prospective basis, controlled clinical studies can be quite powerful. Uncontrolled clinical studies are of questionable validity, whether or not they are subjected to statistical analysis.

DR. GOTTLIEB Why do we need to prove the validity of clinical data?

DR. KEIM By definition, test validity is the extent to which inferences made on the basis of scores from an instrument are appropriate, meaningful, and useful. If you don't care whether the conclusions of the study you are reading are appropriate, meaningful, or useful, then there is no need to prove validity. On the other hand, if you want the material you read to be appropriate, meaningful, and useful, test validity is crucial.

DR. GOTTLIEB What kinds of clinical data can be supported by statistics?

DR. KEIM Statistics can be used to help the reader make a critical evaluation of virtually any quantitative data. What's important is whether the statistical techniques used are appropriate for the given experimental design.

DR. GOTTLIEB Most orthodontists might understand the mean, the median, and the standard deviation, but not necessarily when they should be applied. Please explain.

DR. KEIM The mean and the median are measures of central tendency. There is one more that is sometimes used—the mode. The mean is the arithmetic average of the scores in a distribution. The median is the point below which 50% of the scores fall. The mode is the most frequent score in a distribution. Each is appropriate under different circumstances.

For nominal data, only the mode is appropriate. Let's say we have 200 active patients in a practice, 150 female and 50 male. If we want to know whether the nominal average patient in our practice is male or female, we would look at the mode—there are more females. The median and the mean have no meaning here.

The median or the mean may be used for data that are measured rather than those used for classification, such as the gender example above. The median is the midpoint and is less influenced by skewed distributions. The mean is best used when the data set is symmetrical, as in a bell curve. If we were to weigh five patients in a practice and found values of 95, 100, 110, 115, and 685 pounds, the mean (or average) weight would be 221 pounds. The median would be 110 and, in

this case, would be more indicative of what most patients in the practice weigh. The mean was influenced by the extreme value of 685 pounds, so the data set was skewed.

In any set of data, each measurement differs somewhat from the average of those measurements. Some measurements differ more from the average measurement than others. How much these measurements differ from that average tells us something about how spread out our data are. The standard deviation is one way to express this dispersion.

DR. GOTTLIEB What is the null hypothesis, and how is it used?

DR. KEIM A hypothesis is an assertion subject to validation or proof. The null hypothesis is simply an assertion that no difference exists (the difference is null) between the groups in question. Hypothesis testing—that is, supporting or discounting differences between groups—is what inferential statistics is all about. The purpose of hypothesis testing is to help us draw conclusions about population parameters based on results observed in random samples. Marija J. Norusis described hypothesis testing as well as I've seen, in the 1990 user's guide for the Statistical Package for the Social Sciences.[1] I'll quote it verbatim:

"The procedure is about the same for tests of most hypotheses. It is as follows:

"1. A hypothesis of no difference (the null hypothesis) and its alternative are formulated.

"2. A test statistic is chosen to evaluate the null hypothesis.

"3. The test statistic is calculated for the sample.

"4. The probability (p value), if the null hypothesis is true, of obtaining a test value at least as extreme as the one observed is determined.

"5. If the significance level is judged small enough, the null hypothesis is rejected." Said differently, we assume that significant differences do exist between the groups.

DR. GOTTLIEB Please explain the use of "p" values.

DR. KEIM These represent the probability, if the null hypothesis is true, of obtaining a test value at least as extreme as the one observed. By convention in clinical research, when the "p" values for a given test statistic are less than .05—in other words, there are fewer than five chances out of 100 that the differences we observe are due to random chance alone—we assume that there are significant differences between the groups observed. This .05 value is used most often as the level of significance, but it is by no means a hard and fast rule.

DR. GOTTLIEB What is the standard error of the mean, and when is it useful?

DR. KEIM The standard error of the mean is found by dividing the standard deviation by the square root of the number of subjects in the sample. In reality, it is the standard deviation of the sampling distribution. A large standard error implies that we cannot be very confident that our sample statistics are really good estimates of population parameters. A small standard error allows us to feel more confident that our sample statistics are representative of population parameters. Authors frequently illustrate their findings by graphing the mean values they found with bars sticking out the top, like rooftop TV antennae, to represent one standard error. They hope to convey a sense of the variability in their data, but the standard error of the mean does not really indicate this. If they wanted to illustrate the variability in their data, it would be better to illustrate the mean plus or minus one standard deviation. Authors find the standard error of the mean more appealing because it is smaller than the standard deviation and thus makes their data look "tighter".

DR. GOTTLIEB Since clinicians treat individuals, is there any validity to treating them to population norms? Is the use of means to describe "normal" a valid concept?

DR. KEIM Trick question! The answer depends on how you define "normal". The American

Heritage Dictionary gives no fewer than nine definitions for normal, one of which is "an average". By this definition, yes, population means or "norms" are normal. Another definition is "functioning or occurring in a natural way". The validity of treating to population norms under this definition is questionable. Every case is, in a manner of speaking, an experiment with an "n" of 1. Each patient is his or her own average. My oldest son is 6-foot-2. The population norm for height of 17-year-old white males is 5-foot-10. Is he average? No. Is he normal? Yes. Should I have his legs shortened so that he fits the population norm? I don't think so. Six-foot-two is "normal" for him. Treating patients to some proposed population mean is not a good approach. Population means are best used as bases for comparison, not as treatment goals.

DR. GOTTLIEB Predictability relates to repeating the same experiment, but is it valid in another time, in another place, with another operator?

DR. KEIM You're confusing predictability and replicability. Predictability refers to the ability to accurately guess what's going to happen to one variable based on the effect of another variable or set of variables. Replicability refers to the ability to reach the same results in an experiment that were reached in a previous experiment of the same design. If the second experiment, conducted as you suggest in another place, at another time, by another operator, ends with different results from the first one, then what actually caused the observed results in both experiments comes into question. If both experiments were tightly controlled and the second researcher was very careful about replicating the experimental design of the first researcher, something other than the variables that were controlled caused the observed results. A situation wherein an operator can repeatedly get exactly the same results from repeated trials of an experiment is said to have high *internal validity.* Extraneous variables are tightly controlled. Examples of this would include laboratory experiments. Clinical studies, on the other hand, because they are conducted in the "real world", don't have as high internal validity. What is strived for here is *external validity*—the extent to which the results and conclusions can be generalized to other people and other settings. If the results of a trial cannot be replicated in a variety of settings, the usefulness of the experiment's conclusions to practicing clinicians is minimal.

DR. GOTTLIEB What are we to make of results that describe different ethnic groups—even ones that may be considered relatively close, such as British and American populations?

DR. KEIM This gets back to your question about common sense. If the results deal with variables that are common to British and American populations—for example, shear strength of adhesives—then what's good for Brits is probably good for Yanks. On the other hand, if you're dealing with variables that might be different between the two groups, such as caries rates, you have to be cautious about applying results of one group to the other.

DR. GOTTLIEB Even in a laboratory experiment, as with shear strength of adhesives, is there a reason to submit the data to statistical analysis?

DR. KEIM I've seen statistics defined as a collection of theory and methods applied for the purpose of understanding data. If the differences seen as a result of an experiment are obvious, statistical analyses really aren't needed to understand the data. Let's say we want to evaluate whether or not a new technique for molar distalization works for correcting Class IIs. We try several of them on overt Class II malocclusions, and all correct to Class I. The thing works, whether we subject the results to statistical analyses or not. But let's say we want to know a little more about the effects of the appliance on maxillary growth, or whether this particular appliance works faster than another. The differences here are not as obvious, and statistical analyses can help us figure out whether there are

indeed true differences, how big those differences are, and where those differences are. Regarding your example of shear strengths: If the differences between the adhesives were obvious—say it took only a few ounces of force to shear one type of adhesive and several pounds to shear another—statistical analyses would really not be necessary for us to understand the data. If, however, the differences were not quite so obvious—say within a few ounces of each other—statistical analyses could help us understand something that we might not see otherwise.

DR. GOTTLIEB I have seen the following in recent papers: t-test, unpaired Student t-test, Student t-test for matched samples, Student-Newman-Keuls multiple comparison test, Kruskal-Wallis test, analysis of variance, multiple analysis of variance, repeated measures analysis of variance, one-way analysis of variance, three-way analysis of variance, Bonferroni adjusted test for simple main effect, Tukey's procedure, Tukey multiple comparison test, Scheffe's test for multiple comparisons. I am going to guess that most orthodontists could not tell you what these tests mean. It is almost as if an author says, "I have run my data through a valid test, and trust me on the interpretation of it." What should a reader do who is not acquainted with the tests that are used?

DR. KEIM I agree with your guess about most orthodontists, and would add that most orthodontic researchers could not tell you what most of these tests mean. Several of the terms you mentioned are different names for the same test. It is very confusing, and your statement about the authors saying "trust me" is arguably correct. I say "arguably" because all reputable scientific journals, including orthodontic journals, put papers submitted for consideration through a process of rigorous peer review, wherein editors and reviewers who are qualified to judge the content of the papers evaluate the propriety of the content—including a review of the statistical analyses involved. The reputation of a journal hinges on the integrity of its reviewers. If authors are saying to the readers, "Trust me", editors are saying to the authors, "We'll see about that!" Increasingly, editors are seeking out statisticians to serve as reviewers, as with JCO. While some mistakes are made, editors are becoming much more sophisticated in terms of what they will accept for publication. While a reader does make a leap of faith when it comes to believing the conclusions of a paper, especially if the reader does not understand the statistical tests used, the process of peer review is meant to see to it that the journal in question is worthy of that faith.

DR. GOTTLIEB To a practicing orthodontist, even the idea of reading a statistics text numbs the mind. Is there a textbook you can recommend that covers basic statistics in a readable manner?

DR. KEIM There are several very good texts out. One of the most readable is *Applied Statistics for the Behavioral Sciences* by Dennis E. Hinkle, William Wiersma, and Stephen G. Jurs.[2] While the book is intended for students of the behavioral and social sciences, the fundamentals of statistics are the same for all disciplines. This book covers all basic topics in descriptive and inferential statistics and can take the reader from knowing next to nothing about statistics to being fairly sophisticated in critically evaluating scientific papers.

A much more detailed book specially written for the clinical sciences is *Biostatistics: A Methodology for the Health Sciences* by Lloyd Fisher and Gerald Van Belle.[3] This is the textbook used by our master's students at the University of Tennessee Department of Orthodontics. It is also used by the doctoral students in the Graduate School of Health Sciences here—students such as PhD-level biochemists, physiologists, anatomists, and the like. As I said, this text is extraordinarily detailed and takes a good deal of concentrated effort to read.

A recently published text that is intended specifically to help dental practitioners and students is *Critical Thinking: Understanding and Evaluating Dental Research* by Donald Maxwell Brunette.[4] While this is not a statistics text per se,

the book gives the reader the ability to recognize and (more important) criticize research papers. I recently reviewed this text for JCO [see p. 319], and I do recommend it.

DR. GOTTLIEB Any parting comments?

DR. KEIM Statistics should be regarded as a tool to help us understand the world we practice in. It should not be regarded as a ritualistic religion followed by cloistered, esoteric academics. It is the understanding that is important.

An excellent critique of the contemporary application of statistics to clinical science, by Dr. John M. Yancey, was reprinted in the May 1996 issue of the AJODO.[5] Yancey's 10 rules for evaluating scientific literature can be summarized as follows (his rules are in italics, and my interpretation follows each one):

1. Be skeptical. Make up your own mind about what the data presented really mean. Which brings up the second rule:

2. Look for the data. Yancey points out that in the past, authors published their data along with their analyses thereof. This is a cumbersome process when it comes to publication, but it allowed the readers to draw their own conclusions about the meanings of a given data set. Since most papers nowadays don't publish all the data, replication and further analysis of that data are usually impossible. This can make the authors' conclusions suspect.

3. Differentiate between descriptive and inferential statistics. Statistics that describe the data are referred to simply enough as "descriptive statistics". These would include averages (means), modes, medians, and standard deviations. They tell you something about the data. Inferential statistics, many applications of which generate "p" (probability) values, purport to allow the reader to infer *something* about the population from which the study sample was drawn. What that *something* is can be quite confusing, and interpreting that *something* is probably the most difficult task of the clinical researcher and reader alike. Be extraordinarily critical in evaluating what is inferred from the tests applied.

4. Question the validity of all descriptive statistics. Yancey quips about a man who has one foot frozen in ice and his other foot burning in flames. The man is quite comfortable—on the average. Obviously, this descriptive statistic (the average comfort level) means nothing relative to the true comfort level of the poor subject. Look for the real, physical meaning of the descriptive statistics presented in any paper.

5. Question the validity of all inferential statistics. Again, the reader has to be critical in evaluating what a given test infers about the data. If the conclusions drawn by an author don't jibe with the reader's own clinical experience and judgment, the reader should be very skeptical indeed. If the published findings, conclusions, and recommendations would make a difference in the way the reader practices clinically and the reader has some serious doubts about them, he or she should not hesitate to write to the author for a clarification. The patients are what matters.

6. Be wary of correlation and regression analyses. Correlation studies look for relationships between variables; regression studies are used to predict an outcome variable based on predictor variables. The problem with these studies occurs when the reader tries to make decisions or predictions about individual clinical circumstances, when these tests are meant to apply to populations instead of individuals. Be careful in applying inferences about populations to any patient's individual situation. Experience and clinical judgment are important here as well.

7. Identify the population sampled. Readers who are interested in what the findings of a particular paper mean in their own clinical practices must be sure that the population sampled is representative of the population that includes their own patients. For example, in a recent well-done paper, the authors plugged the noses of some young monkeys and measured the differences in the facial growth of those monkeys as compared to a control group without plugged noses. There were some significant differences between the two groups, supporting a current theory of how nasal obstruction influences facial

growth. However, the practicing orthodontist, while recognizing the value of animal studies, has to seriously question how similar the population of growing monkeys sampled is to the population of patients in his or her own practice.

8. Identify the type of study. Only truly randomized, tightly controlled studies—those in which the subjects were randomly assigned to treatment or control groups—can infer cause and effect.

9. Look for indices of probable magnitude-of-treatment effects. Make sure that the information provided by the authors gives you a good estimate of how big the differences they found between groups really are. The "p" values really don't tell you that.

10. Draw your own conclusions. You've got a good brain of your own, along with experience and clinical judgment. Make sure that you agree with the authors' conclusions, based on the evidence they presented, before you accept what they have to say.

REFERENCES

1. Norusis, M.J.: *Base System User's Guide,* SPSS, Inc., Chicago, 1990.
2. Hinkle, D.E.; Wiersma, W.; and Jurs, S.G.: *Applied Statistics for the Behavioral Sciences,* 3rd ed., Houghton Mifflin, Boston, 1994.
3. Fisher, L. and Van Belle, G.: *Biostatistics: A Methodology for the Health Sciences,* Wiley Interscience, New York, 1993.
4. Brunette, D.M.: *Critical Thinking: Understanding and Evaluating Dental Research,* Quintessence Publishing Co., Chicago, 1996.
5. Yancey, J.M.: Ten rules for reading clinical research reports, Am. J. Orthod. 109:558-564, 1996 (reprinted from Am. J. Surg. 159:533-539, 1990).