The clinician’s guide to the literature: Interpreting results
Barbara L. Greenberg and Mel L. Kantor
J Am Dent Assoc 2009;140;48-54

The following resources related to this article are available online at jada.ada.org (this information is current as of January 27, 2009):

Updated information and services including high-resolution figures, can be found in the online version of this article at:
http://jada.ada.org/cgi/content/full/140/1/48

Information about obtaining reprints of this article or about permission to reproduce this article in whole or in part can be found at:
http://www.ada.org/prof/resources/pubs/jada/permissions.asp
The clinician’s guide to the literature
Interpreting results

Barbara L. Greenberg, MSc, PhD; Mel L. Kantor, DDS, MPH, PhD

Oral health care professionals rely on up-to-date information to guide clinical practice. The most reliable and current information can be found in the peer-reviewed literature. The clinical research literature relies on statistical analysis of data to make inferences and draw conclusions. Although there are many statistical tests available to the researcher, there are fundamental concepts that underlie statistical testing. Understanding these concepts will greatly facilitate critical review of the clinical research literature by the busy clinician. This article addresses these fundamental principles.

Clinicians and clinical researchers try to make inferences about the state of affairs using

ABSTRACT

Background. Oral health care professionals must have up-to-date information to guide clinical practice. The peer-reviewed literature contains the most reliable and current information. The clinical research literature relies on statistical analysis of data to make inferences and draw conclusions. In this article, the authors explore the fundamental principles that underlie statistical testing.

Conclusions and Practice Implications. Having the fundamental tools to critically interpret the results presented in the literature is one of the essential elements for appropriately translating clinical research into practice.

Key Words. Variability; hypothesis testing; statistical interpretation; confidence intervals; sample size.


Dr. Greenberg is interim associate dean for research, Office of Research, New Jersey Dental School, University of Medicine and Dentistry of New Jersey, 110 Bergen St., Newark, N.J. 07103, e-mail "greenbl@umdnj.edu". Address reprint requests to Dr. Greenberg.

Dr. Kantor is a professor, Department of Diagnostic Sciences, New Jersey Dental School, University of Medicine and Dentistry of New Jersey, Newark.
available information. Given that this information often is incomplete or uncertain, what tools can be used to ensure that the inferences made are valid? Clinicians use diagnostic tests and researchers use statistics. When clinicians use diagnostic tests, they would like test results that are always positive when the disease is truly present and results that are always negative when the disease is truly absent. However, diagnostic test results do not always reflect the patient’s true and often hidden status; there almost always is some uncertainty and, as a result, some misclassification (false-positive or false-negative results). Nonetheless, on the basis of the diagnostic test results (and even a physical examination can be thought of as a diagnostic test), clinicians draw conclusions and often must take action based on those conclusions. They use measures of diagnostic efficacy that quantify the uncertainty of the test, such as sensitivity and specificity, to interpret the test results to draw the “correct” conclusion.

Clinical researchers draw a study sample of subjects or patients from a target population of interest. Researchers would like it if the results observed in the sample were an accurate reflection of the results that would be observed if the entire population could be studied. Similar to diagnostic testing by clinicians, researchers conduct statistical tests on the observable data to make inferences about some underlying truth. The statistical test provides a measure of the likelihood of misclassification (false-positive or false-negative results) or uncertainty regarding the results obtained from the study sample. Researchers use this information to interpret the study results and to draw the “correct” conclusion.

The purpose of statistical testing is to use data collected in a study to make inferences about the relevant overall population. One of the strengths of statistical testing is the ability to quantify the uncertainty of the results when making inferences. Researchers use two approaches to quantify this uncertainty: hypothesis testing, which is associated with $P$ values, and estimation, which is associated with confidence intervals (CIs). Fundamental to quantifying uncertainty is estimating variability.

**VARIABILITY**

Variability is the source of uncertainty in clinical studies. Variability is a result of the limits of precision when a measurement is made, as well as the natural variation or randomness within a group. Variability commonly is quantified by two measures: the standard deviation (SD) and the standard error (SE).

The SD quantifies the variability of individual measurements in a study sample. The SD characterizes the distribution of all of the sample data points around the sample mean (that is, the arithmetic average). For data that are normally distributed (Figure 1), the area under the curve represents all of the observations in the distribution, and the mean $\pm 1.96$ SD defines the area under the curve where 95 percent of the individual observations lie. A sample data set with a mean of 15.5 and an SD of 3.0 indicates that approximately 95 percent of the sample data points fall between 9.62 ($15.5 - (1.96 \times 3.0)$) and 21.38 ($15.5 + (1.96 \times 3.0)$).

A small SD indicates little variability in the sample data, with the data points being tightly clustered around the sample mean. A large SD indicates a lot of variability in the sample data, with the data points being widely distributed around the mean. The SD is the most appropriate statistic for reporting measurement variability in a study for normally distributed data.

While the SD refers to individual observations in a study sample, the SE quantifies the

**ABBREVIATION KEY.** N: Noise. S: Signal.
variability of observed sample means if the study were repeated many times. Consider the following example: A sample is drawn from a population, the sample mean is calculated and the sample is returned to the population. Another sample of the same size is drawn from the population, its mean is calculated and this sample is returned to the population. Repeating this process many times and plotting the sample means would yield a normal curve (Figure 1). The SE quantifies the variability of sample means that would be obtained. In a theoretical normal distribution of all possible sample means for a given population, the mean ± 1.96 SE defines the area under the curve where 95 percent of the repeated sample means lie. The remaining 5 percent is distributed equally in the two tails.

Investigators should always report measurement variability in a research paper. The SD is the most appropriate statistic for reporting measurement variability of a sample statistic such as a mean. In some cases, authors incorrectly use the SE to indicate measurement variability in the study sample. Because the SE is derived by dividing the SD by the square root of the sample size, the SE is always substantially smaller than the SD. It is misleading to use the smaller SE to describe the measurement variability in a study sample, and researchers should avoid this practice. In summary, the SD is used when describing a sample and the SE is used when making inferences either through hypothesis testing or estimation. The table presents a theoretical data set with the mean, SD and SE.

Statistical tests involve the use of the SE to determine whether an observed difference between two groups or an observed association between two factors is greater than what would be expected by chance alone. Given that the SE can be calculated for many sample statistics in addition to the mean, it plays a major role in hypothesis testing and estimation for making statistical inferences.

**HYPOTHESIS TESTING AND THE P VALUE**

Hypothesis testing is one of the fundamental approaches for making inferences about a proposed relationship between an exposure or intervention and an outcome. There are two types of hypotheses: the null hypothesis and the alternative or research hypothesis. The null hypothesis is the assumption that there is no significant association between a factor of interest (an exposure or intervention) and the outcome of interest (a disease). The alternative or research hypothesis states that there is a “real” or “true” association between the exposure and the outcome. Box 1 presents an example of null and research hypotheses. The ultimate goal is to test the validity of the proposed hypothesis. The question is this: Are the observed data due to chance alone or is it more likely that the data reflect a true relationship?

Statistical testing is aimed at rejecting or not rejecting the null hypothesis rather than proving or disproving the alternative hypothesis. If the data suggest that the null hypothesis is false, then the null hypothesis is rejected, a de facto acceptance of the alternative hypothesis that the data indicate a true relationship. However, the alternative hypothesis is never tested directly.

Statistical inference commonly is done by means of hypothesis testing using P values. The P value answers this question: How likely is it that the results observed in the study are due to chance alone? If the likelihood that they are due

### Table

Hypothetical data for the number of decayed surfaces detected in 10 subjects.

<table>
<thead>
<tr>
<th>SUBJECT</th>
<th>NO. OF DECAYED SURFACES IN SUBJECTS WHO USED FLUORIDE RINSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

Mean*: 5.9

Standard Deviation (SD)†: 4.43

Standard Error (SE)‡: 1.40

---

* Mean \( \overline{X} = \frac{\sum X}{N} \)

† SD = \( \sqrt{\frac{\sum (X - \overline{X})^2}{N-1}} \)

‡ SE = \( \frac{SD}{\sqrt{N}} \)
to chance alone is “sufficiently low,” the null hypothesis is rejected. The investigator must establish a threshold for “sufficiently low.” The threshold is known as the α level and is most commonly set at 5 percent.

**Two-tailed versus one-tailed test.** In a two-tailed test, the 5 percent is distributed equally in the high and low ends (or tails) of the sampling distribution (Figure 1). The two-tailed test allows for a significant difference in either direction; for example, the new treatment could be better or worse than the standard treatment or placebo. Alternatively, in a one-tailed test, the 5 percent is in one of the two tails. The most conservative approach to hypothesis testing is to use the two-tailed test. The one-tailed test can be used only when it is known that the new treatment is better than the standard treatment or placebo, an assumption that rarely can be made.

If a P value is below the threshold (P < .05), the null hypothesis is rejected and the observed effect is considered a “true” relationship. A P value less than .05 indicates that an outcome as extreme as the observed effect is likely to be due to chance alone less than 5 percent of the time, suggesting that the effect is not a result of chance. The more certain an investigator wants to be, the lower the α level. To use our earlier analogy of diagnostic testing in clinical practice, the P value quantifies the likelihood of drawing a false-positive conclusion from a statistical test.

**Interpreting a P value.** Interpretation of the P value is basically a binary decision. When the P value is less than the α level, the result is statistically significant; when the P value is equal to or greater than the α level, the result is not statistically significant. It is incorrect to say a smaller P value is more significant than a larger P value. However, many investigators do not follow this interpretation and incorrectly refer to results as “very” or “extremely” significant when P values are small (P < .001). A small P value means the data are less likely to be due to chance alone, thus reducing the chance of misclassification or making a false-positive conclusion. A P value greater than the α level is not proof that the null hypothesis is true, but it does indicate that the evidence is not strong enough to reject it. Box 2 is an example of hypothesis testing.

Hypothesis testing using the P value also can be conceptualized as a signal detection task that can be represented as a signal-to-noise ratio. Statistical tests compare the strength of an associa-

---

**BOX 1**

**Example of null and research hypotheses.**

Researchers conducted a study to assess the efficacy of daily use of fluoride rinse in addition to performing regular oral hygiene to prevent the development of new carious lesions in 20 children aged 8 through 16 years across a one-year period. They included a control group that received a placebo rinse, which did not contain fluoride, in addition to performing regular oral hygiene.

**NULL HYPOTHESIS**
The null hypothesis assumes that there is no difference between the fluoride rinse and the placebo rinse in preventing dental caries.

**NONDIRECTIONAL RESEARCH HYPOTHESIS**
The nondirectional research hypothesis states that there is a significant difference between the group using the fluoride rinse and the group using the placebo rinse.

**DIRECTIONAL RESEARCH HYPOTHESIS**
The directional research hypothesis states that the fluoride rinse is significantly better than the placebo rinse in preventing the development of new carious lesions.

---

**BOX 2**

**Example of hypothesis testing.**

In the example of a fluoride rinse versus a placebo rinse (Box 1), consider a study of 10 patients using the fluoride rinse and 10 patients using the placebo rinse. Across the one-year study period, the mean number of new carious lesions among those using the fluoride rinse was 2.9, while the mean number among those using the placebo rinse was 5.9; the appropriate statistical test yielded a P value = .03.

Therefore, the null hypothesis of no difference between the fluoride and the placebo rinses is rejected, and the investigators concluded that there is a statistically significant difference between the fluoride and placebo rinses. The mean number of new carious lesions in the fluoride rinse group is statistically less than the mean number of new carious lesions in the placebo rinse group.

---

**ESTIMATION AND CONFIDENCE INTERVALS**

While statistical inference is done most commonly with hypothesis testing using a P value, an
alternative approach is to use the CI. The CI not only indicates significance, it also quantifies variability or uncertainty of the result used to make the statistical inference.

It is important to recognize the fundamental difference between statistical inference using hypothesis testing and CI. Hypothesis testing with the use of a P value is based on a single value or point estimate derived from the data, and it results in a binary “reject” or “not reject” decision. For example, when the means of two groups are compared, the means and the difference between the means are point estimates. Given that data from any single study are derived from a sample of the larger population of all eligible subjects, a different study sample could yield different point estimates. If the study were repeated using different samples, the resulting multiple point estimates could be tightly clustered or widely distributed. Therefore, a potentially more informative approach involves the use of an interval estimate that incorporates a range that quantifies the natural variability in the data; this is known as a CI. The strength of the CI is that the range of the interval reflects the variability that would be observed if the study were repeated many times.

The CI is used to estimate the upper and lower limits of the variability in the sample data. Given that the SE quantifies the variability of the study sample statistic (for example, sample mean) from the true population parameter (for example, population mean), it is used to calculate the CI. The sample statistic (for example, mean) + 1.96 SE defines the upper limit of the 95 percent CI, and the sample statistic −1.96 SE defines the lower limit of the 95 percent CI. The greater the variability in the sample data, the wider the CI—that is, the less precisely we have estimated the population mean.

The width of the interval also is a function of the degree of confidence one wants to have in the interval. Similar to the convention of setting α at .05, a 95 percent CI is used most commonly. Thus, the 95 percent CI includes the true population value 95 times of 100; the true population value will be outside this interval five times out of 100. This is the margin of error that the investigator is willing to accept. The margin of error that is widely reported for the results of opinion or political polls is nothing more than the 95 percent CI for the percentage of respondents who feel a certain way or will vote for a particular candidate.

Interpreting a CI. The CI is used to test significance and to estimate the magnitude of an observed association. In epidemiologic and clinical studies, estimates of the magnitude of an observed association are referred to as measures of disease association. To illustrate how to interpret a CI, we use the measure of disease association known as the relative risk (RR).

The RR provides information about the
strength of the association as well as the direction of the association. The RR is the risk of developing disease among those exposed divided by the risk of developing disease among those not exposed. Given that the RR is a ratio, if the risks are the same between the exposure groups, then RR equals 1. An RR of 1 is the same as the null hypothesis of “no association.” If the risk among exposed subjects is greater than the risk among nonexposed subjects, the RR is greater than 1 and indicates a positive association between the exposure and the outcome. If the risk among exposed subjects is less than the risk among nonexposed subjects, the RR is less than 1 and indicates a negative association between the exposure and the outcome; the exposure then is considered a putative protective factor.

Using the observed RR (our sample statistic of interest), one calculates the upper and lower limits of the 95 percent CI around the RR by using the formula (sample statistic ± 1.96 SE). A 95 percent CI around the RR that includes 1 indicates that the true value of the RR could be 1. Therefore, the RR is not considered to be statistically significant and the evidence does not support an association between exposure and disease. If the 95 percent CI does not include 1, the result is considered to be statistically significant and provides evidence of a significant association between exposure and disease. Box 3 shows an example of interpreting a CI.

**Further interpretation of the 95 percent CI.** The upper and lower limits of the CI denote the possible effect size and potential clinical significance. In some instances, the results may not be considered statistically significant (that is, the 95 percent CI around the RR includes 1), but the potential magnitude of the effect may be great. For example, a 95 percent CI of 1 to 35 is not statistically significant, but it does suggest that the magnitude of the effect could be large (upper limit of 35). In contrast, a 95 percent CI of 1.07 to 1.09 is considered statistically significant, but the magnitude of the effect is very modest (upper limit of 1.09), and it may be of relatively little clinical importance. A statistically significant result does not necessarily imply a clinically important result. Therefore, it is advisable to look at the potential effect size as well as the significance of the results. All investigators should indicate what is considered a clinically important outcome, and practitioners should use clinical judgment when interpreting results.

**Example of interpreting a confidence interval (CI).**

In the study example (Boxes 1 and 2), researchers determined the incidence of new carious lesions (disease) during the one-year study period for those who used the fluoride rinse (exposure) and for those who used the placebo rinse (not exposed).

**RELATIVE RISK**
The relative risk (RR) of developing new carious lesions for subjects in the fluoride rinse group compared with subjects in the placebo rinse group was 0.55, suggesting that the fluoride rinse is protective against the development of new carious lesions.

**CONFIDENCE INTERVAL**
The 95 percent CI for the RR is 0.33-0.87. Because the 95 percent CI does not include 1, the null hypothesis of no difference in caries prevention between the fluoride rinse and the placebo rinse is rejected.

The fluoride rinse resulted in significantly fewer new carious lesions compared with the placebo rinse.

**STATISTICAL INFERENCE AND SAMPLE SIZE**

For the statistical inference to be valid, an adequate sample size is necessary to ensure enough power to minimize misclassification and to correctly decide whether to reject the null hypothesis. Power is the probability that the study will be able to detect a difference or an association if, in fact, one exists (by analogy, a true positive diagnostic test result). The desired level of power is 80 percent or greater. Without sufficient power, a study result of no statistical significance could be incorrect owing to an inadequate sample size rather than being a valid reflection of the data.

Therefore, if the study reveals no significant outcome \( P \geq .05 \) or a 95 percent CI that includes the null value), two questions should be asked:

- Was there enough power? Was the sample size large enough? Sample size calculations should always be done at the outset of the study during the design phase. Establishing the clinically important difference or outcome is fundamental for sample size calculations, and the outcome should be generated on the basis of data in the established literature combined with clinical expertise.

In general, the larger the sample size, the smaller the difference or the weaker the association that can be detected as statistically significant. However, there also is danger from the other side of the spectrum, where very large samples can result in a statistically significant finding for even a very small effect that may be of
minimal or no real clinical significance. A statistically significant difference does not necessarily imply a clinically important difference.

**SUMMARY**

This article covered several of the basic concepts that underlie the majority of statistical analyses reported in the clinical research literature: variability, hypothesis testing, CIs and sample size. Having the fundamental tools to critically interpret the results presented in the literature is one of the essential elements for appropriately translating clinical research into practice.

**Disclosure.** Drs. Greenberg and Kantor did not report any disclosures.

The authors acknowledge the valuable input of Dr. Curt Bay, A.T. Still University, Mesa, Ariz., and the inspiration of friend and colleague Dr. Leo Lee Lewis.

**Suggested Readings**