Errors in the Archives of Dermatology and the Journal of the American Academy of Dermatology From January Through December 2003

Julie A. Neville, MD; Wei Lang, PhD; Alan B. Fleischer, Jr, MD

Objective: To assess the frequency of statistical errors in the dermatology literature.

Design: Original studies published in the Archives of Dermatology and the Journal of the American Academy of Dermatology from January through December 2003 were analyzed for correctness of statistical methods and reporting of the results.

Results: Of 364 studies published, 155 included statistical analysis. Of these, 59 (38.1%) contained errors in the methods or omissions in reporting of the statistical results. Fourteen percent of the articles with statistical analysis contained errors in the methods used (considered to be more significant errors), 26.5% contained errors in the presentation of the results, and 2.6% contained errors in both.

Conclusions: The misuse of statistical methods is prevalent in the dermatology literature, and the appropriate use of these methods is an integral component of all studies. Readers should critically analyze the methods and results of studies published in the dermatology literature.

Arch Dermatol. 2006;142:737-740

Statistics are frequently used when reporting the results of studies in the medical literature, yet errors commonly occur in the correct use and presentation of statistical findings. Previous reviews in other medical disciplines have demonstrated high error rates, ranging from 45% to 95%. A review of 100 articles in the dermatopathology literature found an error rate of 36% in 25 articles that contained statistical analysis. Inadequate power is also prevalent, with the results of 1 study indicating that most clinical trials with negative conclusions in dermatology did not have an adequate sample size to detect a difference between treatment groups. A more comprehensive review of the general dermatology literature to detect other statistical errors has not been conducted, to our knowledge.

For this reason, we performed a retrospective review of the statistical methods from all studies published in the Archives of Dermatology and the Journal of the American Academy of Dermatology in 2003. These 2 journals were chosen because they are well-respected peer-reviewed journals in the dermatology literature.

See also page 761

METHODS

Articles published in the Archives of Dermatology and the Journal of the American Academy of Dermatology from January through December 2003 that included statistical methods were reviewed for errors. The articles included in this study were those containing statistical analysis from the sections of these journals publishing scientific studies. In the Archives of Dermatology, these included the Studies, Observations, Correspondence, Evidence-Based Dermatology, and Reviews sections, and in the Journal of the American Academy of Dermatology, these included the Reports, Therapy, Laser Surgery, Dermatologic Surgery, Dermatopathology, and Brief Reports sections. Despite the inconsistent definition of what constitutes a statistical error, we chose to include those errors that were highlighted in previous reviews from other medical disciplines. The 2 groups of statistical errors considered were in the use of a statistical test and in the presentation of the results.
### Statistics

**Table. Statistical Tests Used in Articles With Statistical Analysis**

<table>
<thead>
<tr>
<th>Statistical Test</th>
<th>No. (%) of Articles (n = 155)</th>
<th>Proportion Incorrectly Applied*</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$ Test</td>
<td>46 (29.7)</td>
<td>3/46 (6.5)</td>
</tr>
<tr>
<td>Unpaired t test</td>
<td>29 (18.7)</td>
<td>11/29 (37.9)</td>
</tr>
<tr>
<td>Analysis of variance</td>
<td>26 (16.8)</td>
<td>3/26 (11.5)</td>
</tr>
<tr>
<td>Fisher exact test</td>
<td>23 (14.8)</td>
<td>0</td>
</tr>
<tr>
<td>Paired t test</td>
<td>16 (10.3)</td>
<td>5/16 (31.3)</td>
</tr>
<tr>
<td>Survival analysis</td>
<td>15 (9.7)</td>
<td>0</td>
</tr>
<tr>
<td>Wilcoxon rank sum test</td>
<td>13 (8.4)</td>
<td>0</td>
</tr>
<tr>
<td>Spearman rank correlation</td>
<td>12 (7.7)</td>
<td>0</td>
</tr>
<tr>
<td>Mann-Whitney test</td>
<td>11 (7.1)</td>
<td>0</td>
</tr>
<tr>
<td>Pearson product moment correlation</td>
<td>11 (7.1)</td>
<td>2/11 (18.2)</td>
</tr>
<tr>
<td>Wilcoxon signed rank test</td>
<td>11 (7.1)</td>
<td>0</td>
</tr>
<tr>
<td>Kruskal-Wallis test</td>
<td>4 (2.6)</td>
<td>0</td>
</tr>
<tr>
<td>McNemar test</td>
<td>3 (1.9)</td>
<td>0</td>
</tr>
</tbody>
</table>

*Errors occurred in the use of more than 1 test within a single article. Data are given as number/total number (percentage) of articles.

### Errors in the Use of a Statistical Test

Because most articles did not provide the raw data necessary to determine the distribution, we assumed that sample sizes smaller than 30 in each group would not have a normal distribution and that a nonparametric test should be used. While parametric statistical tests assume that the data collected have a normal, continuous, bell-shaped (Gaussian) distribution, nonparametric methods are free of this assumption and work well for small sample sizes and data with skewed distributions. The sample size of 30 was selected because it is used in Basic & Clinical Biostatistics by Dawson-Saunders and Trapp\(^1\) as an arbitrary cutoff for differentiating between data sets with a normal or nonnormal distribution. Exceptions to this were the rare occasions when the authors stated that they performed a visual or statistical test to ascertain the distribution of the data and then used the appropriate test based on these results. It is also possible that sample sizes larger than 30 may not have a normal distribution and that a nonparametric test should be used with these data. Given the lack of a clearly defined cutoff value and data necessary to determine if the correct test was chosen, we considered these to be errors in the use of a statistical test, although arguably, they may be questionable. In addition to using the appropriate test for the data distribution, we also evaluated the correct use of unpaired and paired t test, the use of analysis of variance (ANOVA) for multiple comparisons, the use of Fisher exact test for small sample sizes, and the pooling of variance.

### Errors in the Presentation of the Results

Minor errors in the presentation of the findings included failure to report the type of statistical test used in the article and whether it was 1-sided or 2-sided. Another error was presenting the results in the format of “a ± b” without reporting if b represents the standard deviation or the standard error of the mean. Although not considered an error in our analysis, we checked for the inclusion of details about the statistical analysis package and the power of the study.

During this 1-year period, 155 (42.6%) of 364 articles published in these 2 journals contained statistical analysis. Most articles that did not include statistical analysis were descriptive studies in which statistics would not have contributed any additional information to the article. Thirty-three (21.3%) of the articles used parametric methods only, 49 (31.6%) used nonparametric methods only, and 45 (29.0%) used a combination of these methods, and 28 (0.2%) used other methods. The most frequently used tests were $\chi^2$ test (29.7%), unpaired t test (18.7%), ANOVA (16.8%), Fisher exact test (14.8%), and paired t test (10.3%) (Table).

Of those studies that included statistical analysis, 59 (38.1%) of 155 contained errors or omissions in statistical methods or the presentation of the results. Twenty-two articles (14.2%) contained significant errors in the use of a statistical test that could potentially change the validity of the study results, 41 (26.5%) contained errors in the presentation of the results, and 4 (2.6%) contained errors in both.

Thirty-eight of the errors occurred in the Journal of the American Academy of Dermatology, with 14 errors in the use of a statistical test, 26 errors in the presentation of the results, and 2 errors in both. In the Archives of Dermatology, 21 errors occurred, constituting 8 errors in the use of a statistical test, 15 errors in the presentation of the results, and 2 errors in both.

Errors in the statistical test chosen included 3 articles (1.9%) not using Fisher exact test when analyzing a $2 \times 2$ contingency table, as should have been performed when the expected cell count for at least 1 of the cells was fewer than 5. Other errors included using an unpaired t test with paired data (3 articles [1.9%]), using t test or z test to compare multiple samples when a test such as ANOVA should have been used (2 articles [1.3%]), and comparing multiple studies without using the correct methods for pooling variance (1 article [0.6%]). The questionable error of using a parametric test (often t test) on sample sizes smaller than 30 without indicating the use of a test for normality occurred in 16 articles (10.3%).

In the presentation of statistical results, failure to state if a test was 1-sided or 2-sided was the most common omission, occurring in 32 articles (20.6%). Eight articles (5.2%) provided statistical results and P values without disclosing the statistical test used. Two articles (1.3%) did not state if they were reporting standard deviations or standard errors of the mean. Although not considered an error in our analysis, 92 articles (59.4%) did not report the statistical package used for the analysis, and only 16 (10.3%) of the articles included any information about the power of the study.

We evaluated industry sponsorship of studies to see if this affected the rates of errors. Of those studies with errors in the use of a statistical test, 4 (18.2%) of 22 were sponsored by industries, as were 10 (24.4%) of 41 studies with errors in the presentation of the results.
In this review, 59 (38.1%) of 155 studies using statistical tests contained errors in statistical methods or in the presentation of the results. Most of these errors were minor omissions in the presentation of the results, but 22 studies (14.2%) used an incorrect statistical test. Without the original data, it is impossible to know if these errors invalidate the results of studies, but such errors should prompt the reader to question the results.

Most articles that were published in these 2 journals used nonparametric statistics, often in conjunction with parametric methods. The error rate of 38.1% is consistent with error rates in published studies from other medical disciplines and in the study by Flotte et al in the dermatopathology literature.

Three studies used unpaired t tests with paired data. Typically, this results in a falsely elevated P value and can lead to failure in detecting a significant difference when one exists. Two studies did not use a test for multiple comparisons, which can result in finding a spurious difference between 2 groups.

Only 10.3% of the studies included information on the power of the study, usually to determine the sample size necessary to detect a statistical difference before study initiation. Studies with inadequate sample sizes have an increased risk of type II error (failing to find a difference when one actually exists). Although not necessary in all studies, power should be reported in studies reaching negative conclusions because the inadequate sample size may have resulted in the lack of significance. A less significant omission occurred in the failure to report the statistical package used for analysis, although we did not consider this an error because some authors only include the package details if relevant.

This study is limited by the inconsistent definition of what constitutes a statistical error. We chose to include those errors that were analyzed in reviews from other medical disciplines, but some of these errors can be considered questionable. One example of this was considering the use of a parametric test (usually t test) with small sample sizes to be an error unless the authors stated that they performed a test for normality. t Test is robust, and the results likely would not be changed by minor deviations from normality. Therefore, it is difficult to know without the raw data if the use of this test with small sample sizes affected the study results. In addition, some journals may only report normality testing if the data required transformation as a result. In studies with sample sizes smaller than 30 in each group, we recommend using a nonparametric test or reporting the performance of normality testing, through visual inspection of the plotted data or by means of a statistical test.

To correct these errors that occur within peer-reviewed journals, it has been suggested that a statistician review all articles before submission or that a statistician be included as a reviewer. Although this adds an additional burden of time and expense to the publication process, statistical reviewing has been shown to decrease the number of statistical errors in medical publications. The onus is worthwhile to ensure the validity of the study results.

Most statistical analyses are based on a limited battery of tests taught in an introductory statistics course, but many dermatologists may not be familiar with the correct statistical test that should be used with their data. As a result, unless authors are familiar with the statistical test that they are performing, they should consult a statistician before submission of their study results. In our analysis, we found a lower rate of errors in industry-sponsored studies, which typically include statisticians in the data analysis. In addition to these measures, a statistical checklist should be referenced before submission of any journal article that includes statistical analysis.

It may also be beneficial to incorporate training in statistics into dermatology residency programs or as a continuing medical education program. These programs would be offered to increase awareness about the importance of critically analyzing journal articles and recognizing common statistical errors when interpreting the results.

In summary, the appropriate use of statistical methods is an integral part of all studies performed and published. Errors in statistics frequently occur in the dermatology literature, as in many other disciplines of medicine, and readers should be critical of statistical methods and conclusions drawn from studies with incorrect or incomplete statistical analysis.

Accepted for Publication: July 25, 2005.
Correspondence: Alan B. Fleischer, Jr, MD, Department of Dermatology, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157-1071 (afleisch@wfubmc.edu).

Author Contributions: Study concept and design: Neville and Fleischer. Analysis and interpretation of data: Neville, Lang, and Fleischer. Drafting of the manuscript: Neville. Critical revision of the manuscript for important intellectual content: Lang and Fleischer. Statistical analysis: Neville, Lang, and Fleischer. Study supervision: Fleischer.

Financial Disclosure: None.

Previous Presentation: This study was presented as a poster at the 63rd Annual Meeting of the American Academy of Dermatology; February 18-22, 2005; New Orleans, La.

REFERENCES


ARCHIVES Web Quiz Winner

Congratulations to the winner of our March quiz, Mahmoud Saeed Khalil Abou-Hamidah, MD, dermatology and venereal specialist, Shaqura Hospital, Shaqura, Kingdom of Saudi Arabia. The correct answer to our March challenge was hydroa vacciniforme. For a complete discussion of this case, see the Off-Center Fold section in the April ARCHIVES (Ee HL, Wong SN, Khoo LSW, Tan SH. Bullae on the wrists of an 11-year-old boy. Arch Dermatol. 2006;142:515-520).

Be sure to visit the Archives of Dermatology Web site (http://www.archdermatol.com) to try your hand at the interactive quiz. We invite visitors to make a diagnosis based on selected information from a case report or other feature scheduled to be published in the following month’s print edition of the ARCHIVES. The first visitor to e-mail our Web editors with the correct answer will be recognized in the print journal and on our Web site and will also receive a free copy of The Art of JAMA II.