The Problem of Poor Statistical Reporting

"These reviews [of statistical errors] reveal a remarkable and depressing consistency, with typically around 50% of reviewed papers being found to contain clear statistical errors."

G.D. Murray, 1991

The Problem of Statistical Errors in the Literature

• Widespread
• Long-standing
• Potentially serious
• Largely unknown
• Concerns basic, not advanced, statistics

ERROR #1: Reporting measurements with unnecessary precision

Many measurements do not need to be reported with full precision:

• For all practical purposes, a patient weighs 60 kg, not 60.18 kg

• The mean age is 34.81 years, but how much is 0.81 of a year?

• The smallest P value that need be reported is P < 0.001.

ERROR #4: Using descriptive statistics incorrectly

• Use the mean and standard deviation ONLY to report normally distributed data:
  "Mean (SD) height was 72 cm (4.3 cm)."

• Use the median and interquartile range to report non normally distributed data:
  "Median (IQR) length was 9 cm (6 to 25 cm)."

ERROR #4: Using descriptive statistics incorrectly

The shape of the distribution (normal or skewed) may determine the class of statistical test used to analyze the data ("parametric" or "nonparametric," respectively).

Most biological data are not normally distributed; the median and IQR should be more common than the mean and SD.
ERROR #6: Reporting only P values for results

"Congratulations, Ms. Jones. Your drug has a P value of less than 0.01!"

P values: Have no clinical interpretation

- Have an either-or interpretation based on an arbitrary cut point (often 0.05)
- P values of 0.049 and 0.051 should be interpreted similarly, even though one is statistically significant and the other is not.

Confidence intervals:

- Are clinically interpretable
- Are sensitive to sample size
- Can indicate statistical significance in some circumstances if they exclude certain values

The importance of confidence intervals

"The effect of the drug was statistically significant."

Would you use this drug?

"The effect of the drug on lowering blood pressure was significant (P < 0.05)."

Now would you use this drug?

"Mean blood pressure in the treatment group dropped from 100 to 92 mm Hg (P = 0.02)."

Now?
The importance of confidence intervals

“The drug lowered systolic blood pressure by a mean of 8 mm Hg (95% CI = 2 mm Hg to 14 mm Hg; \( P = 0.02 \)).”

Now?

The moral?

“Never \( P \) alone.”

ERROR #9: Not accounting for all data or all patients

A schematic summary of the study can:

- Summarize the study design
- Show the number of patients at each stage
- Indicate denominators for proportions, percentages, and rates
- Present the main results of the study

A Schematic Summary

Patients Approached \( n = X \)

Patients Excluded \( n = X \)

 Patients Assigned \( n = 84 \)

Low-Air Loss Bed \( n = 43 \)

Foam Mattress \( n = 41 \)

Complete healers \( n = 21 \)

Complete healers \( n = 5 \)

ERROR #13: Ignoring uncertain results when calculating diagnostic test characteristics

- Intermediate results fall between a negative result and a positive result
- Indeterminate results indicate neither a positive nor a negative finding.
- Uninterpretable results occur when a test is not conducted according to standards.

Calculating Diagnostic Test Characteristics

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Disease Present</th>
<th>Disease Absent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>( A )</td>
<td>( B )</td>
<td>( A + B )</td>
</tr>
<tr>
<td>Negative</td>
<td>( C )</td>
<td>( D )</td>
<td>( C + D )</td>
</tr>
<tr>
<td>Total</td>
<td>( A + C )</td>
<td>( B + D )</td>
<td>( A + B + C + D )</td>
</tr>
</tbody>
</table>

Sensitivity = \( \frac{A}{A + C} \); Specificity = \( \frac{D}{B + D} \)
ERROR #15: Using a graph in which the visual message does not support the message of the data

Readers remember the visual impression of the figure better than the actual data

- The “lost zero” problem visually distorts the relationships between columns

ERROR #16: Confusing the “units of observation”

- In a study of 50 eyes, the number of patients could range between 25 and 50.
- What does a 50% success rate mean? Half the eyes improved, or half the patients?

ERROR #17: Interpreting underpowered studies with non-significant results as “negative”

Statistical power: the ability to detect a given difference if it really exists:

"The increase in infection rate using the new methods was not statistically significant . . . (and there was not 1 chance in 10 that we would have detected a 30% increase in rate)"

Frederick Mosteller

ERROR #17: Interpreting underpowered studies with non-significant results as “negative”

- Studies with nonsignificant results and low power are inconclusive, not negative.
- In studies with insufficient power, groups that are not statistically different cannot be said to be equivalent. “The absence of proof is not proof of absence.”

ERROR #18: Not reporting results in clinically useful units

Results expressed in absolute terms (the absolute or attributable risk reduction):

In the Helsinki study of hypercholesterolemic men, after 5 years, 84 of 2030 patients on placebo (4.1%) had heart attacks, whereas only 56 of 2051 men on gemfibrozil (2.7%) had heart attacks (P < 0.02), for an absolute risk reduction of 1.4% (4.1% - 2.7% = 1.4%).
ERROR #19: Not reporting results in clinically useful units

Results expressed in relative terms (the relative risk reduction):

In the Helsinki study of hypercholesterolemic men, after 5 years, 4.1% of the men on placebo had heart attacks, whereas only 2.7% on gemfibrozil had heart attacks. The difference, 1.4%, is a 34% relative risk reduction in the incidence of heart attack (1.4% ÷ 4.1% = 34%).

ERROR #19: Not reporting results in clinically useful units

Results expressed in an effort-to-yield measure, the number needed to treat:

The results of the Helsinki study of 4081 hypercholesterolemic men indicate that 71 men need to be treated for 5 years to prevent a single heart attack.

ERROR #19: Not reporting results in clinically useful units

Results expressed in another effort-to-yield measure:

The Helsinki study found that, after 5 years, about 200,000 doses of gemfibrozil were ingested for each heart attack prevented.

ERROR #20: Confusing statistical significance with clinical importance

"It has been said that a fellow with one leg frozen in ice and the other leg in boiling water is comfortable—on average."  J.M. Yancy

• In large samples, clinically irrelevant differences can be statistically significant.

• In small samples, large and important differences can go undetected as a result of low statistical power.

The Secret to Good Medical Writing:

Have something to say.

Say it.

Stop!

Contact Information

Tom Lang, MA
Tom Lang Communications and Training
Finely crafted medical writing—
Because publication is the final stage of research
tomlangcom@aol.com • 530-758-8716
www.TomLangCommunications.com
(cheap plug for my book)

*How To Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers, 2nd edition*

Thomas A. Lang, MA  
Michelle Secic, MS  
Foreword by Ed Huth, MD, MACP

*(American College of Physicians, 2006)*