Brief Notes on

Biomedical Literature Evaluation

Hisham S. Abou-Auda, Ph.D.

College of Pharmacy, King Saud University

Last Revision: September 2005
Primary Literature

- Original published or unpublished works that introduce new knowledge or enhance existing knowledge
- Experimental studies
  - Clinical trials
  - Pharmaceutical research
  - Educational assessments
- Observational studies
  - Cross-sectional
  - Case control
  - Cohort
- Descriptive Reports
  - Case reports
  - Case series
  - Pharmaceutical practice
  - Editorials
  - Letter-to-the-Editor

NOTE:

A review article published in Pharmacotherapy is not considered primary literature

Importance of Critical Literature Evaluation Skills

- Critical literature evaluation is the process of reading and evaluating primary literature journal articles in order to arrive at an interpretation that you can call "your own"
  - Are the results believable?
  - Are the results applicable to your practice?
  - Does the paper really support its claims?
- The ultimate interpretation and decision about the value of an article rests with the reader.
• Many studies are open to differing interpretation

• The majority of clinical decisions are based on primary literature reports.

• Literature evaluation skills are needed when therapeutic dilemmas arise, when information about a particular subject is conflicting, or when making decisions about drug policy (i.e., adding or deleting a drug from an institution's drug formulary).

• Patients often ask pharmacists to help them interpret information that they have obtained in the lay press.

• It has been estimated that 40 to 50% of published articles in the medical literature have serious problems with study design, statistical analysis, and conclusions.

Reasons for Publications of Poor Quality Research

<table>
<thead>
<tr>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The &quot;Publish or Perish&quot; dilemma</td>
</tr>
<tr>
<td>2. Investigator’s lack of knowledge in study design or statistical analysis</td>
</tr>
<tr>
<td>3. Peer-reviewer’s lack of knowledge in study design or statistical analysis</td>
</tr>
<tr>
<td>4. Preliminary results published to stimulate further research in area</td>
</tr>
<tr>
<td>5. Scientific fraud</td>
</tr>
</tbody>
</table>

From 1975 to 1983, the FDA conducted routine audits of investigators who were conducting clinical trials funded by the agency. Fraudulent data were found to have been submitted by 59% (24/41) of investigators disciplined for scientific misconduct.

Validity

• Internal Validity: Within the confines of the study, the results appear to be accurate, the methods and analysis are appropriate and the interpretations of the investigators appear supported.

• External Validity: The conclusions of the study can be applied to the reader's practice. Also referred to as "generalizability”of results.

Bias

• Bias is a systematic variation in which treatment groups under study are treated or measured differently on a consistent basis

• Bias can mislead one to conclude an erroneous outcome

• Not all types of bias can be avoided
Stages Where Bias Can Occur Within a Study
(From *Journal of Chronic Diseases* 1979;32:51-63)

1. Reading background information or the introduction of the study
2. Defining and choosing the study sample
3. Applying the experimental interventions
4. Measuring the outcomes
5. Analyzing the data
6. Interpreting the analysis and results
7. Publishing the findings

**Common Biases Found in Research**

(Adapted from Millares. *Applied Drug Information*. Vancouver; Applied Therapeutics, Inc.; 1998.)

**Bias of Rhetoric:** Rhetorical techniques used to convince the reader without reference.

**One-Sided Reference Bias:** Authors may restrict references to only those works which support their conclusions.

**Positive Result Bias:** Authors are more likely to submit and publish positive results.

**Hot Stuff Bias:** When a topic is "hot", investigators and editors may not be able to resist publishing additional results, no matter how preliminary or shaky the results may appear.

**Suspicion Bias:** Knowledge of a subject's prior exposure may influence subject selection or the outcome of an intervention.

**Sample Size Bias:** Samples which are too small can prove nothing; samples which are too large can prove anything.

**Admission Rate or "Berkson" Bias:** When hospitalization rates differ for different exposures/diseases, the relation between exposure and diseases can become distorted.

**Procedure Selection Bias:** Certain clinical procedures may be preferentially offered to those at low risk or those expected to have a favorable outcome.

**Missing Clinical Data Bias:** Certain clinical data may be missing because they were normal, negative, never measured, or never recorded.
Volunteer Bias: Volunteers from a specified sample may exhibit characteristics which may differ from non-volunteers.

Contamination Bias: This may occur when members of the control group inadvertently receive the experimental intervention.

Withdrawal Bias: Patients who withdraw from a study may differ systematically from those who remain.

Therapeutic Personality Bias: When an investigator knows what treatment the subjects are receiving the outcomes or measurements may be influenced.

Intensive Measure Bias: When outcome measures are incapable of detecting clinically significant differences.

Apprehension Bias: Certain measurements may be altered from usual values when the patient is apprehensive.

Obsequiousness Bias: Subjects may alter their responses based on what they perceive to be desired by the investigator.

Attention Bias: Subjects may alter their behavior if they know they are being observed (also known as the Hawthorne Effect).

Instrument Bias: Defects in the calibration or maintenance of measurement instruments may lead to systematic deviations in results.

Post-hoc Significance Bias: Alpha or beta error selected after the data have been analyzed

Data Dredging Bias: Data are reviewed for all possible associations without a prior hypothesis.

Tidying-up Bias: Exclusion of outliers or other "untidy" results that is not justified by statistical grounds.

Magnitude Bias: When interpreting a finding, the selection of a scale of measurement may markedly affect the interpretation.

Significance Bias: The confusion of statistical significance with clinical significance.

Correlation Bias: Equating correlation with causation
Confounding

- A variable that affects the dependent or independent variables within a study and makes it difficult to determine what effect it has on the measured outcomes.

- Confounding variables may hide a true association.

- It may be impossible to eliminate all confounding variables.

- Statistical methods can be used to control for some of the effects confounding.

Components of an Evaluative Journal Article

- Title: A brief description of the article's subject

- Authors: Informs the reader who conducted the study; the primary author or investigator is listed first

- Abstract: Identifies the purpose, design, and methodology of the study and briefly reviews its results and conclusions

- Introduction: Provides a framework for the purpose of the study through an overview of the literature and references other published studies that support the purpose

- Methods: A detailed description of the techniques used to conduct the study; enough information should be provided so that the reader could replicate the study. (This is the most important component the article and must be evaluated with careful scrutiny!)

- Results: Organizes and simplifies the raw data, but does not interpret the results

- Discussion: Authors interpret the data and debate the significance of their findings

- Conclusion: Provides a one- to two-sentence summary of the study's purpose and results

- References: Lists all references cited throughout the study
Evaluating the Journal

Characteristics of Reputable Scientific Journals

- Editorial policy that specifies the requirements for the types and formats of submitted manuscripts
- Peer-review policy which requires that all manuscripts be reviewed by consultants or researchers in the field of study prior to publication
- An editorial board that is composed of well-known researchers and leaders in their respective discipline (editorial board is usually listed on or near the title page)
- Does not contain an overabundance of advertisements
- Reputable journals usually publish supplements. Keep in mind that information found in supplements may not be peer-reviewed.
- Clinical studies are rarely published in "throw-away" journals. Throw-away journals are characterized by being free to readers, having a high advertisement-to text ratio, not being owned by professional societies, and having variable peer-review processes.

Evaluating the Investigators

- The investigators should have appropriate training and expertise in the area.
- The investigators should have a good track record of prior research.
- A biostatistician should be involved with the evaluation of data.
- The validity of studies authored entirely by investigators working in pharmaceutical industry is sometimes questioned because the goal of these individuals is primarily promotion of the company's drug products.
- The source of funding and potential conflicts of interest should be disclosed by the authors. Complete funding by pharmaceutical companies may be an issue of concern. However, potential bias does not necessarily negate the study results.
- The research site should have appropriate resources and technology to effectively conduct the study.
Evaluating the Title

- The title should be brief and catch the attention of readers interested in the topic.
- The title should indicate what the article is about without drawing any conclusions. A title that sounds like a newspaper headline may indicate bias by the authors.
- The title should not give the impression of answering questions that the study is not designed to answer.

Evaluating the Abstract

- Abstracts should provide a very brief overview of the study, and provide enough information for the reader to determine whether the study is of interest.
- Data presented in the abstract may not be discussed in the body of the article and important data may be omitted from the abstract due to space limitations. Abstracts should not be used as substitutes for careful analysis of the study, and clinical decisions should not be made based on information only from the abstract. (i.e., you must read the entire article to accurately evaluate the study)
- Inflammatory language may indicate the potential for bias
- Certain journals may have formats for "structured" abstracts to improve quality. Structured abstracts include: objective, research design, clinical setting, participants, interventions, main outcome measurements, results, and conclusions.

Evaluating the Introduction

- The introduction should provide valid reasons supporting the need to conduct the study; all factual statements must be referenced, and the authors should appropriately interpret the available literature for applicability and relevance.
- References should be up-to-date and mostly from the primary literature—not the tertiary literature. Authors should avoid citing only their own past research.
- A good introduction should provide a complete synopsis of the literature published to date and include the most important studies.
- Refer to original articles to confirm suspicious data. If the data are misquoted, then the credibility of the study or report is questionable.
• Conduct your own literature search to confirm that the authors have completely reviewed the literature. Authors may chose to ignore recent work that disagrees with their own.

• The study objective usually appears at the end of the introduction, and should be stated in a clear and concise manner. If the authors fail to state and objective, it may be an indication that the study was not well planned.

• A research and null hypothesis must be formulated for each study and should be stated in the introduction of an article:

  Alternative Hypothesis ($H_a$): A difference exists between groups
  Null Hypothesis ($H_0$): A difference does not exist between groups

**Evaluating the Methods** (the most important section!)

*Important Definitions*

**Randomization:** A procedure equivalent to flipping a coin that helps ensure that treatment groups are similar. When studies are randomized, subjects have an equal and independent chance of receiving any of the treatment modalities. Appropriate randomization techniques include the use of random number tables, computer-generated random numbers, or lotteries.

**Stratification:** A randomization procedure that divides subjects into equal groups to control for differences in confounding variables. Stratification allows separate estimates to make for groups of individuals who have the same values for the confounding variable.

**Open-label:** A study in which there is no blinding. Both the investigator and subjects are aware of the assignment of the treatment groups.

**Single-blind:** Either subjects or investigators are unaware of assignment of subjects to active or control groups.

**Double-blind:** Both subjects and investigators are unaware of assignment of subjects to active or control groups.

**Triple-blind:** Both subjects and investigators are unaware of assignment of subjects to active or control groups; another group involved with interpretation of data is also unaware of subject assignment.

**Double-Dummy:** A study in which two placebo are needed to achieve proper blinding of treatments.
Parallel study: All subjects receive only one treatment. Parallel studies are most appropriate when therapies are definitive or when disease states are self-limited (e.g., antimicrobials for infectious diseases).

Cross-Over Study: All subjects receive both treatments being studied and outcomes are assessed for each therapy. Cross-Over studies are appropriate when diseases are highly variable (e.g., pain) Advantages include: smaller sample size required and decreased error caused by variability.

**Important Considerations for Cross-Over Studies**
- Appropriate wash-out period
- Diseases with exacerbations and remissions
- Subjects must be randomly assigned to treatment order
- Blinded to time of cross-over
- Subject dropouts and deaths should be minimized

Controls: A group of persons used for comparison with a study group. Ideally, the control group is identical to the study group except that it has not been exposed to the treatment under study.

Placebo-controlled: A study in which the "control" group receives placebo that is identical to the study drug in terms of appearance, taste, and smell, but does not contain active drug.

Active-controlled: A study in which the "control" group receives treatment with another pharmacologically active medication.

Historical control: A study in which data from previously conducted trials or groups of previously treated patients are used for comparison with the current study population.

Run-In Period: Preinvestigation observation of patients usually designed to ensure that they are appropriate candidates for entrance into a randomized clinical trial. (to ensure adherence to therapy)

- The protocol must be approved the institutional review board of the institution to ensure patient safety and sound clinical research
- All subjects must agree to participate by signing an informed consent that clearly states the risk and benefits of participating in the study
• The inclusion and exclusion criteria for the subjects should be appropriate for the topic of study and clearly defined. Subjects enrolled in the study should have characteristics that are representative of patients with disease, and severity criteria should be appropriate.

• The methods section should describe in detail how the patients were selected to participate in the study. Where is the patient population from?

• Patients should be randomly assigned to treatment groups to ensure that groups are homogenous. All demographic information should be presented. (Always check to make sure that baseline characteristics between groups are similar)

• The medication doses, routes and frequencies of administration, and duration of treatment should be appropriate for the condition studied.

• All concomitant therapies should be clearly described, and diet and lifestyle characteristics should be similar between groups

• Compliance with therapy and adverse events should be monitored and recorded appropriately.

• All measurements should be standardized and conducted at appropriate intervals. Data collection forms and instruments should be validated.

• Power calculations and level of statistical significance should be determined a priori

**Evaluating the Results**

• Data should be presented in a clear and understandable format. Data presented in the abstract, charts, tables, or graphs should be consistent with what is described in the text.
Review of Statistical Methods used in Biomedical Literature

**Statistical Analysis:** The organization and mathematical manipulation of data used to describe characteristics that have been studied and formulate conclusions from the data.

**Statistically Significant:** The difference that exists between study groups cannot be explained by the role of chance alone.

**Clinically Significant:** The difference that exists between study groups is substantial enough to be clinically useful.

**Type I Error:** An error that occurs when statistical tests show that a difference exists between study groups, but in fact there is no difference (α-error)

**Type II Error:** Statistical tests conclude that no difference exists between study groups, but in fact a difference does exist (β-error)

**Type I Error**
- α is the probability of making a Type I error
  - Acceptable values for a error are selected before the study is conducted (a priori) to minimize the risk of an incorrect conclusion
  - The generally accepted value used for α error in biomedical studies is ≤ 0.05
  - An α ≤ 0.05 means that type I error will occur less than 5% of the time, and that 95% of the time investigators can be sure that the results cannot be explained by chance alone
  - After the study is completed, a p-value is calculated from the data collected to determine the actual observed significance
  - If the p-value is less than the preset α, it is assumed that the data supports the hypothesis 95% of the time

**Type II Error**
- β is the probability of Type II error and is directly related to the study's "power" (Power = 1 - β)
  - β can be calculated using the number of subjects (n) and the difference that the investigators are trying to detect (power) must be preset before the study is conducted (usually "a priori")
  - The arbitrarily accepted level of β-error is ≤ 0.2, which translates to a power of 80%
  - A power of < 80% (or 0.80) suggests that there may not be enough study subjects to detect a difference between groups
Information Needed for a Sample Size Calculation

- Mean difference the investigator wishes to detect
- Anticipated standard deviation within the study population
- Desired power (\( \beta \))
- Preset \( \alpha \)
- One- or two-tailed analysis

An example of sample size calculation formula for comparison of sample means:

\[
 n = \frac{2 \cdot \text{PI} \cdot \left( \frac{\sigma}{\mu_2 - \mu_1} \right)^2}{}\]

where
- \( n \) = desired sample size for each group,
- \( \text{PI} \) = desired power index,
- \( \sigma \) = anticipated standard deviation,
- \( \mu_2 - \mu_1 \) = the mean difference wished to detect

*(NOTE: This is an example, the specific equation used will vary depending the type of data, number of groups etc)*

The smaller the difference one wishes to detect between groups, the larger the sample size will be. For example, to detect a 20% difference in cure rates between two antibiotics, 20 patients will be needed to have 80% power. However, to detect a difference of 10%, 200 patients will be needed to have 80% power.

"One-Tailed" vs. "Two-Tailed" Analysis

- Refers to the distribution of the test statistic

  - One-tailed analyses are designed to detect deviations from the null hypothesis in only one direction. The test will only detect if drug A is more effective than drug B. The test cannot provide a valid determination of whether is drug A is less effective than drug B. One-tailed tests are typically used when comparing an active drug to placebo. In this situation, one is reasonably sure that the active drug may be similar to the placebo, but it is highly unlikely that the active drug will be worse than placebo.

  - A two-tailed test is designed to determine deviations from the null hypothesis in both directions. This type of analysis will determine if drug A is better than drug B, if drug A is equal to drug B, or if drug A is worse than drug B.
Descriptive Statistics
Measure of central tendency:
• mean
• median
• mode
Measure of variability:
• range
• standard deviation
• coefficient of variation

Incidence
Incidence is the probability that a healthy person will develop a disease within a specified period. It is the number of new cases of disease in the population within a specific period.

\[
\text{Incidence rate} = \frac{\text{Number of new cases of disease}}{\text{Population at risk}}
\]

Prevalence
Prevalence measures the number of people in the population who have the disease at a given time. It is the probability of people having a disease within a specified time frame.

\[
\text{Prevalence} = \frac{\text{Number of existing cases of disease}}{\text{Total Population}}
\]

Relative Risk
Relative risk is a measure of disease frequency when a specific factor is present or absent. An actual risk can only be measured when a cohort type of study design is used. Prospective studies allow for defining populations at risk and also allow for calculation of excess risk caused by exposure to a particular factor.

\[
\text{RR} = \frac{\text{Incidence rate among those exposed to a particular factor}}{\text{Incidence rate among those not exposed to the same factor}}
\]

Odds Ratio
Odds ratio is also a measure of disease frequency. An odds ratio is an estimator of relative risk and is calculated when prospective studies evaluating exposure to certain factors are not practical. Odds ratios are calculated when retrospective case-control studies are used. When calculating an odds ratio, three assumptions must be made; the control group is representative of the general population, the cases are representative of the population with the disease, and the frequency of the disease in the population is small.

\[
\text{Odds Ratio} = \frac{\text{Cases with exposure} \times \text{Controls without exposure}}{\text{Cases without exposure} \times \text{Controls with exposure}}
\]
95% Confidence Intervals
Statistical terms that represent the interval of numerical values within which you can be 95% confident that the population value you are estimating lies within the interval. This is also known as an interval estimate.

Inferential Statistics
Used to make assumptions pertaining to the null hypothesis and determine if the difference that exists can be explained by chance alone.

Type of statistical test that is used is based on:
- Data type utilized in the study
- Number groups for comparison
- Type of study design

Types of Data
Continuous: Characterized by having an infinite number of evenly spaced potential values between any two points.
Examples: Age, weight, S. Cr., Etc.

Nominal: Characterized by the arbitrary assignment of numbers to different characteristics that have a finite number of possible values.
Examples: race or sex

Ordinal: Data that is ranked in a specific order with no consistent magnitude of difference between ranks.
Examples: stage of disease or opinion scores (ranked 1-5)

Dependent variable
The outcome variable of interest in a research study. The outcome that one intends to explain or estimate.

Independent variable
Variables that will effect corresponding measurement of the dependent variable in a research study. Independent variables define the conditions under which the dependent variable is to be examined.

Parametric
Data collected from a sample that can be described using a "normal distribution". When plotted, the data is symmetrical, continuous, and forms a bell-shaped curve with its mean value corresponding to the highest point on the curve. Used to refers to interval or ratio data.

Nonparametric
Data collected from a sample that is not normally distributed. Used to refer to ordinal or nominal data.

Common Statistics for Nominal Measurements
- Chi-Square: A statistical test used to compare nominal data from independent groups when only two possible choices exist. This test should not be used if the expected value in any cell is less than 5.
• **Fisher's Exact test**: Another statistical test used to compare nominal data from 2 independent groups. (used when expected frequencies are < 5)

• **McNemar's test**: A variant of the Chi Square test used to compare nominal data from two matched or paired groups (e.g., when data is collected from the same patient at different time periods).

• **Contingency Table Analysis**: A test used to compare nominal data when there are three or more groups or three or more possible outcomes. Also referred to as "R X C" or "Row-by-Column".

• **Cochran Mantel-Haenszel Test**: A test used to compare nominal data when more than one independent variable exists. It is often used when data from several Contingency Tables are combined and analyzed together.

• **Bonferroni Correction**: A modification commonly used in statistical tests to make adjustments in p-values in order to minimize type I error when comparing several groups. (the more comparisons that are made, the greater the risk for type I error)

**Common Statistics for Ordinal Measurements**

• **Wilcoxon rank sum test**: A statistical test used to determine differences in ordinal measurements taken from 2 independent groups. The test involves combining the data from both samples and ranking them from smallest to largest.

• **Mann-Whitney U test**: A statistical test used to determine differences in ordinal measurements taken from 2 independent groups. This test involves comparing each data value from one group to the values obtained in the other group.

  (The Wilcoxon rank sum test and the Mann-Whitney U test use different methods to calculate the exact same p-values)

• **Kruskal-Wallis rank sum**: A single statistical test used to compare differences in ordinal data taken from 3 or more independent groups. If a difference is detected, then additional statistical tests are required to determine exactly which groups are different since there can be many comparisons.

• **Wilcoxon signed-rank test**: A statistical test used to compare ordinal data that is collected using repeated-measures. (e.g., when data is collected from the same individual at different time points, also known as a crossover study design)

• **Friedman's test** A statistical test used to compare ordinal measurements from 3 or more independent groups when the data is collected using repeated measures.
Common Statistics for Interval and Ratio Measurements

- **Student's t test:** A test statistic used to compare continuous (interval or ratio) data collected from 2 independent groups with equal variances. If the data are not normally distributed, then the appropriate corresponding nonparametric test statistic is the Mann-Whitney U test.

- **Paired t test:** A test statistic used to compare continuous (interval or ratio) data collected from the same individual using repeated measures. If the data are not normally distributed, then the appropriate nonparametric test statistic is the Wilcoxon signed-rank test.

- **Analysis of variance (ANOVA):** A single test statistic used to compare continuous data collected from 3 or more groups assuming the variances of all groups are equal and normally distributed. The Kruskal-Wallis test is the appropriate nonparametric test for this instance.

  - If the p-value calculated from an ANOVA test reports statistical significance (p<0.05), one can conclude that a significant difference among groups exists. However, to determine which groups are significantly different additional statistical tests (post hoc methods) are needed. Post hoc methods used to adjust multiple comparisons include:
    - Tukey’s test
    - Duncan’s test
    - Dunnett's test
    - Student-Newman-Keul's test (SNK)
    - Scheffé’s test
    - Fisher’s Least significant difference test (LSD)
    - Bonferroni test

- **Analysis of Covariance (ANCOVA):** A statistical test that compares differences in a continuous variable with variations that may result from differences between treatment groups (i.e., age, weight etc) The control variables used are known as covariates.

Regression Models

- **Regression models:** Statistical techniques used to determine if there are associations between two variables. A correlation coefficient is calculated (denoted by the letter "r"). The correlation coefficient spans between +1 and -1. Perfect correlation would result in a correlation coefficient of 1.00 (r=1.00), and if the r value is -1, then a perfect negative correlation exists.

- **Spearman's correlation coefficient:** A correlation coefficient used to assess possible associations when data for the regression model are nonparametric.
• **Pearson's correlation coefficient**: A correlation coefficient used to assess possible associations when data for the regression model are parametric.

• **Multivariate regression analysis**: A statistical technique used to determine if there is a relationship between multiple independent variables and a single dependent variable. A single continuous variable is chosen, and other variables are evaluated to determine their effect on the dependent variable.

• **Logistic regression**: A statistical technique also used to determine if there is a relationship between multiple independent variables and a single dependent variable. This method of analysis is designed for a nominal variable.

### Survival Analysis

• **Kaplan-Meier Method**: A method commonly used for estimating the effect a variable on survival beyond a certain time point. The outcome measure does not have to be survival; it can be any dichotomous (yes or no) outcome that is measured over a period of time. Subjects are followed until they experience the outcome under study or follow-up ends. Kaplan-Meier curves represent the probability of not experiencing the event over time. (note: survival curves are not regular proportions because of censored data)

  - Statistical tests commonly used to compare survival curves include the Wilcoxon rank-sum test, Mann-Whitney U test, and the logrank test.

• **Cox proportional hazards models**: A multivariable regression method used to determine if there is a relationship between multiple independent variables and a single dependent variable when the dependent variable is time to an event with censored data.

### General Considerations

• **Standard error of the mean (SEM)**: This value represents the "estimated" variation that would occur between means from repeated samples of the same size. The SEM is sometimes reported in the medical literature because it is smaller than the standard deviation and it makes the data appear to have a very narrow range. However, SD should always be used to describe variation within a sample.

\[
SEM = \frac{SD}{\sqrt{n}}
\]

• Bonferroni’s adjustment should be used when multiple comparisons are used to minimize the potential for α error.
• When a significant difference is reported using an ANOVA, post hoc results should be given.

• Parametric tests are generally considered to be "more powerful" than nonparametric tests.