Introduction to Statistics

Roger J. Lewis, MD, PhD

Director of Research
Department of Emergency Medicine
Harbor-UCLA Medical Center
Torrance, California

Presented at the 2006 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California.

Contact Information:

Roger J. Lewis, MD, PhD
Department of Emergency Medicine
Harbor-UCLA Medical Center, Box 21
1000 West Carson Street
Torrance, CA 90509

Telephone: (310) 222-6741
FAX: (310) 782-1763
email: roger@emedharbor.edu

Table of Contents

Introduction

The most powerful tool available for improving the effectiveness of emergency medical care is the information obtained from well-designed and properly analyzed clinical research. Because emergency medicine is a relatively young specialty, and because the clinical responsibilities of academic emergency physicians tend to be quite heavy, there are relatively few experienced mentors available to train new investigators in the statistical design and analysis of clinical studies. The purpose of this lecture is to teach concepts underlying the statistical design and analysis of a clinical trial.

Statistical Concepts

Classical Hypothesis Testing

Data from clinical trials are usually analyzed using *p* values and classical hypothesis testing.[1-3] In classical hypothesis testing, two hypotheses which might be supported by the data are considered. The first, called the null hypothesis, is the hypothesis that there is no difference between the groups being compared with respect to the measured quantity of interest.[2] For example, in a study examining the use of a new sympathomimetic agent for blood pressure support in patients with sepsis, the null hypothesis might be that there is no difference between the systolic blood pressure achieved with the test agent and with the control agent. The alternative hypothesis, on the other hand, is the hypothesis that the groups being compared are different.[2] The alternative hypothesis should also define the size of the difference, namely that the test agent results in a 10 mm Hg greater systolic blood pressure than the control agent. The difference between the two groups defined by the alternative hypothesis is called the treatment effect.

The magnitude of the difference defined by the alternative hypothesis must be set prior to data collection. In an interventional study, the difference between the groups defined by the alternative hypothesis is usually the minimum treatment effect thought to be clinically significant. Sometimes a larger treatment effect is defined, because designing a study to reliably detect the smallest clinically-significant treatment effect would require too large a sample size (see below).[4-7]

Once the null and alternative hypotheses are defined, the null hypothesis is "tested," to determine which hypothesis (null or alternative) will be accepted as true. This process of testing the null hypothesis consists of calculating the probability of obtaining the results observed, or results more inconsistent with the null hypothesis, assuming the null hypothesis is true. This probability is the *p* value.[2]

If the *p* value is less than some predefined value, denoted α, then the null hypothesis is rejected and the alternative hypothesis is accepted as true. In other words, if the null hypothesis being true would result in data like that obtained, or even more extreme results, less than α of the time (where α is usually 5% or 0.05), then the null hypothesis is rejected as false. The steps in classical hypothesis testing are shown in Table 1. Some definitions related to terms used in classical hypothesis testing are shown in Table 2.[8]

Type I Error

A type I error occurs when the investigator concludes that a difference has been demonstrated between two groups when, in fact, no such difference exists.[2,3,9] It is a type of “false positive.” When data are analyzed using *p* values, a type I error occurs when a statistically significant *p* value is obtained when, in fact, there is no underlying difference between the groups being compared. Since the *p* value is the probability of obtaining results equal to or more extreme than those actually observed, assuming there is actually no difference between the groups being compared, the risk of a type I error is equal to the maximum *p* value considered statistically significant, when there is actually no difference between the groups.[2,9]

Type II Error, Power, and Sample Size

A type II error occurs when there is a difference between the two groups, the difference is as great as that defined by the alternative hypothesis, and yet a non-significant *p* value is obtained.[2,4-7] A type II error is a type of “false negative.”

The power of a trial is the chance of detecting a treatment effect of a given size, if that treatment effect truly exists.[4-7] Studies are usually designed to have a power of 0.80 or greater. Since the power of the trial is the chance of finding a true treatment effect, the quantity (1 - power) is the chance of missing a true treatment effect, i.e., the risk of committing a type II error. The risk of committing a type II error is denoted β.[4-6] The value of α, the power, and the magnitude of the treatment effect the clinical trial is designed to detect (defined by the alternative hypothesis) determine the sample-size required for the study.[7]

The smaller the size of the treatment effect the study is designed to detect, the larger the required sample size will be, for a given value of α (the maximum significant *p* value), and power. For any size of treatment effect, a smaller maximum significant *p* value or a larger power will require a larger sample size as well. It is still surprisingly common for clinical studies to be

Step	Description
Define the null hypothesis	The null hypothesis is that there is no difference between the groups being compared. For example, in a clinical trial the null hypothesis might be that the response rate in the treatment group is equal to that in the control group.
Define the alternative hypothesis	The alternative hypothesis would be that the response rate in the treatment group is greater than that in the control group, by a given amount.
Calculate a *p* value	This calculation assumes that the null hypothesis is true. One determines the probability of obtaining the results found in the data, or other results even more inconsistent with the null hypothesis. This probability is the *p* value.
Accept or reject the null hypothesis	If the probability of observing the actual data, or more extreme results, under the null hypothesis is small ($p < \alpha$), then we should doubt that hypothesis. The idea is that if the probability under the null hypothesis of observing the actual results is very small, then there is a conflict between the null hypothesis and the observed data, and we should conclude that the null hypothesis is not true.
Accept the alternative hypothesis	If we reject the null hypothesis, we accept the alternative hypothesis by default.

Table 1. Steps in Classical Hypothesis Testing.

published, with negative results, which did not have adequate sample size to reliably detect a clinically-significant treatment effect.[4-7]

In many cases, the appropriate sample size for a study can be determined from published tables. Two useful sources of such tables are references 10 and 11. Several commercial software programs (e.g., PASS 2005 [see http://www.ncss.com/pass.html], Power and Precision [see http://www.power-analysis.com/]) can also be used.

Statistical Tests

Depending on the characteristics of the data being analyzed, different statistical tests are used to determine the *p* value. The most common statistical tests are described in Table 3. Student's t test and Wilcoxon's rank sum test are used to compare continuous variables (i.e. numbers like serum glucose, respiratory rate, etc.) between two groups of patients. If there are three or more groups of patients, then one-way analysis of variance (ANOVA) and the Kruskal-Wallis test are used to compare continuous variables between the groups. The chi-square test and Fisher's exact test are used to detect associations between treatment and outcome when both the treatment and the outcome are categorical variables (placebo versus active drug, lived versus died, admitted versus discharged, etc.).

Term	Definition
α	The maximum *p* value to be considered statistically significant; the risk of committing a type I error when there is no difference between the groups
α Error	A type I error
Alternative Hypothesis	The hypothesis that is considered as a alternative to the null hypothesis; usually the alternative hypothesis is that there is an effect of the studied treatment, of a given size, on the measured variable of interest; sometimes called the test hypothesis
β	The risk of committing a type II error
β Error	A type II error
Null Hypothesis	The hypothesis that there is no effect of the studied treatment on the measured variable of interest
Power	The probability of detecting a treatment effect the size of the treatment effect sought (i.e. obtaining a p value $< \alpha$), given α, and the sample size of the clinical trial; power = $1 - \beta$
p value	The probability of obtaining results similar to those actually obtained, or results more inconsistent with the null hypothesis, if the null hypothesis were true
Type I Error	Obtaining a statistically significant *p* value when, in fact, there is no effect of the studied treatment on the measured variable of interest; a false positive
Type II Error	Not obtaining a statistically significant *p* value when, in fact, there is an effect of the treatment on the measured variable of interest that is as large or larger than the effect the trial was designed to detect; a false negative

Table 2. Definitions of Terms Commonly Used in the Design of Clinical Trials.

Student's t test and one-way analysis of variance are examples of parametric statistical tests. Parametric statistical tests make assumptions about the underlying distribution of continuous variables. Both Student's t test and analysis of variance assume that the data are normally distributed, and that the different groups yield data with equal variance.

When the data to be analyzed are not normally distributed, then the *p* value should be obtained using a non-parametric test.[2,12,13] Non-parametric tests are distribution-free in that they do not rely on the data having any particular underlying distribution. The non-parametric alternative to a t test for unpaired samples is the Wilcoxon rank sum test or the Mann-Whitney U test. For a statistical comparison of paired measurements, one can use the Wilcoxon signed rank test (it is unfortunate that the two names are so similar). The non-parametric alternative to one-way analysis of variance is the Kruskal-Wallis test.

Statistical Test	**Description**
Student's t test	Used to test whether or not the means of measurements from two groups are equal, assuming that the data are normally distributed and that the data from both groups have equal variance.
Wilcoxon rank sum test (Mann-Whitney U test)	Used to test whether two sets of observations have the same distribution. These tests are similar in use to the t test, but do not assume the data are normally distributed.
Chi-square test	Used with categorical variables (two or more discrete treatments with two or more discrete outcomes) to test the null hypothesis that there is no effect of treatment on outcome. The chi-square test assumes at least 5 expected observations of each combination of treatment and outcome, under the null hypothesis.
Fisher's exact test	Used in an analogous manner to the chi-square test, Fisher's exact test may be used even when less than 5 observations are expected in one or more categories of treatment and outcome.
One-way ANOVA*	Used to test the null hypothesis that three or more sets of continuous data have equal means, assuming the data are normally distributed and that the data from all groups have identical variances. The one-way ANOVA may be thought of as a t test for three or more groups.
Kruskal-Wallis	This is a non-parametric test analogous to the one-way ANOVA. No assumption is made regarding normality of the data. The Kruskal-Wallis test may be thought of as a Wilcoxon rank sum test for three or more groups.

* Analysis of variance.

Table 3. Common Statistical Tests.

Confidence Intervals

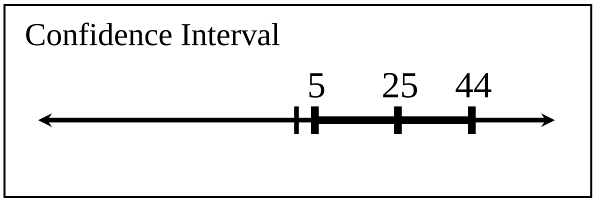
[This section is based on references 14 and 15]

Suppose we wish to test whether one vasopressor is better than another, based on the mean post-treatment systolic blood pressure (SBP) in hypotensive patients and, in our trial, we observe a mean post-treatment SBP of 70 mm Hg for patients given vasopressor A and 95 mm Hg for patients given vasopressor B. The observed treatment difference (mean SBP for patients on vasopressor B minus mean SBP for patients on vasopressor A) is 25 mm Hg. If the *p* value is less than 0.05, we reject the null hypothesis as false and we conclude that our study demonstrates a statistically significant difference in mean SBP between the groups. That the *p* value is less than 0.05 tells us only that the treatment difference that we observed is statistically significantly different from zero. It does not tell us the size of the treatment difference, which determines whether the difference is *clinically* important, nor how precisely our trial was able to estimate the true treatment difference. The true treatment difference is the difference that would be observed if all similar hypotensive patients could be included in the study.

Studies comparing two groups often yield a single number, such as the difference in mean SBP. This single number “estimates” the true difference between the groups, and is termed a “point estimate.” If, instead of using hypothesis testing and reporting a *p* value, we report the point estimate and the corresponding confidence interval surrounding it, we give readers the same information as the *p* value, plus information on the size of the treatment difference (and therefore its clinical importance), the precision of the estimated difference, and information that aids in the interpretation of a negative result.

The *p* value answers only the question, “Is there a statistically significant difference between the two treatments?” The point estimate and its confidence interval also answer the questions, “What is the size of that treatment difference (and is it clinically important)?” and “How precisely did this trial determine or estimate the true treatment difference?” As clinicians, we should change our practice only if we believe the study has definitively demonstrated a treatment difference, and that the treatment difference is large enough to be clinically important. Even if a trial does not show a statistically significant difference, the confidence interval enables us to distinguish whether there really is no difference between the treatments, or the trial simply did not have enough patients to reliably demonstrate a difference.

Returning to our example, a treatment difference of 0 is equivalent to the null hypothesis that there is no difference in mean SBP between patients on vasopressor A and patients on vasopressor B. In our trial, the confidence interval of 5 to 44 mm Hg does not include 0, and therefore, a true treatment difference of zero is not statistically consistent with our data. We conclude that the null hypothesis that there is no difference is not statistically consistent with our observed data and we reject the null hypothesis. If a 95% confidence interval does not include a zero treatment difference, this demonstrates that the results are statistically significant, equivalent to $p < 0.05$.



Our point estimate of 25 mm Hg gives an estimate for the size of the treatment difference. However, our results are also statistically consistent with any value within the range of the confidence interval of 5 to 44 mm Hg. In other words, the true treatment difference may be as little as 5 mm Hg, or as much as 44 mm Hg. If vasopressor B has many more severe side effects than vasopressor A, a reader may conclude that even an elevation of SBP as much as 44 mm Hg does not warrant use of vasopressor B, although the treatment difference is statistically significant. Another reader may feel that even an elevation in mean SBP of 5 mm Hg would be beneficial, despite the side effects. With *p* values, authors report a result as statistically significant or not, leaving us with little basis for drawing conclusions relevant to our clinical practice. With confidence intervals we may decide what treatment difference we consider to be clinically important, and reach conclusions appropriate for our practice.

We may also use confidence intervals to obtain important information from trials that did not achieve statistical significance (so called “negative” trials). Suppose we found the 95% confidence interval for the difference in mean SBP to be -5 mm Hg to 55 mm Hg, with the same point estimate of 25 mm Hg. Our results are consistent with vasopressor B raising mean SBP as much as 55 mm Hg *more* than vasopressor A, or as much as 5 mm Hg *less*. Because the confidence interval includes 0 (a zero treatment difference), equivalent to the null hypothesis that there is no treatment difference, the results are not statistically significant and $p > 0.05$. Since $p > 0.05$, we may be tempted to conclude that there is no advantage to using vasopressor A or B in our clinical practice. However, our data are also consistent with vasopressor B raising SBP as much as 55 mm Hg more than vasopressor A. Although $p > 0.05$, there remains the possibility that a clinically important difference exists in the two vasopressors' effects on mean SBP. Negative trials whose results are still consistent with a clinically important difference usually occur when there is too small a sample size, resulting in low power to detect an important treatment difference.

It is important to know how precisely the point estimate represents the true difference between the groups. The width of the confidence interval gives us information on the precision of the point estimate. The larger the sample size the more precise the point estimate will be, and the confidence interval will be narrower. As mentioned above, negative trials that use too small a sample size may often not show a statistically significant result yet still not be able to exclude a clinically important treatment difference. In this case, the confidence interval is wide and imprecise, and includes both zero, or no treatment difference, and clinically important treatment differences. Conversely, positive trials that use a very large sample size may show a statistically significant treatment difference that is not clinically important, for example an increase in mean SBP from 70 mm Hg to 72 mm Hg.

If a confidence interval includes zero, or clinically unimportant treatment differences, as well as clinically important treatment differences, we can not make any definitive conclusions regarding clinical practice. It is important to remember that the data are statistically consistent with the true value being anywhere within the entire range of the confidence interval from a trial.

Multiple Comparisons

Whenever two groups of patients are compared statistically, even if they are fundamentally identical, there is a chance that a statistically significant *p* value will be obtained.[2,9] If the maximum significant *p* value (α) is 0.05, then there is a 5% chance that a statistically significant *p* value will be obtained even if there is no true difference between the two patient populations. This risk of a false positive *p* value occurs each time a statistical test if performed. When *multiple* comparisons are performed, either the pairwise comparison of more than two groups of patients, or the comparison of many different characteristics between two groups of patients, the risk of at least one false-positive *p* value is increased, because the risk associated with each test is incurred multiple times.[16-22] The risk of obtaining at least one false-positive *p* value, when comparing two groups of fundamentally identical patients, is shown in Table 4 as a function of the number of independent comparisons made. For up to 5 to 10 comparisons, the overall risk of at least one type I error is roughly equal to the maximum significant *p* value used for each individual test,

Number of Comparisons	**Probability of at Least One Type I Error**
1	0.05
2	0.10
3	0.14
4	0.19
5	0.23
10	0.40
20	0.64
30	0.79

Table 4. Probability of at Least One Type I (False Positive) Error when Performing Multiple Independent Comparisons Between Identical Groups with $\alpha=0.05$.

multiplied by the total number of tests performed. This rough equality is the basis for the Bonferroni correction.[2,16]

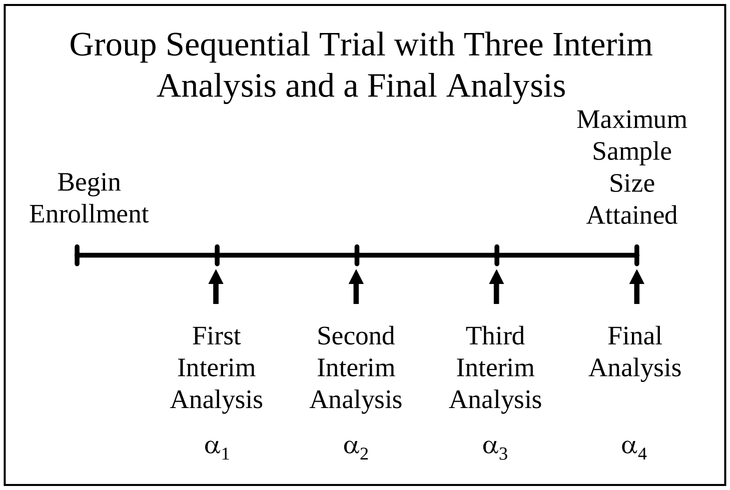
The Bonferroni correction is a method for reducing the overall type I error risk for the whole study by reducing the maximum p value used for each of the individual statistical tests (the testwise risk). The overall risk of a type I error that is desired (usually 0.05) is divided by the number of statistical tests to be performed, and this value is used as the maximum significant p value for each individual test. For example, if five comparisons are to be made, then a maximum significant P value of 0.01 should be used for each of the five statistical tests.

The Bonferroni correction controls the overall (studywise) risk of a type I error, at the expense of an increased risk of a type II error. Since each statistical test is conducted using a more stringent criteria for a statistically significant p value, there is an increased risk that each test might miss a clinically-important difference, and yield a p value that is nonsignificant using the new criteria of $p < 0.01$.

Specific statistical tests have been developed to compare three or more groups of patients. Examples include analysis of variance (ANOVA), the Kruskal-Wallis test, the chi-square test, and Fisher's exact test (Table 3). These tests, which do not use the Bonferroni correction, can be useful for detecting differences between three or more groups with relatively high power, while controlling the overall risk of a type I error. Their disadvantage is that, while they may detect a difference among three or more groups of patients, they do not define which pairwise differences are statistically significant.

Interim Data Analyses

During the conduct of a clinical trial, data accumulate sequentially, containing more and more information on the effectiveness of the treatments being compared. Often, however, these data are not analyzed until the trial has been completed and all patients enrolled. This type of fixed-sample-size design has the disadvantage that more patients then necessary to obtain a reliable result may be enrolled. It is often advantageous to plan one or more interim analyses of the data, which are conducted before the full sample size has been reached, to see if a reliable conclusion may be drawn from the data and the trial terminated early. Such interim analyses of the data must be planned in advance to avoid increasing the type I error rate, as this is a type of multiple comparison. Essentially, a more stringent criterion for statistical significance must be used at each interim analysis, to control the cumulative probability of a type I error occurring at any of the interim analyses.[8] Although the details of this type of trial are beyond the scope of this lecture, the interested reader may read one of several readily available resources.[8,22,23]



Subgroup Analysis

Any group of patients is heterogeneous. This is especially true for groups of patients defined by presenting signs and symptoms in an emergency department setting. Some of the patients within such a group may have a more severe form of the disease in question, and some patients may have a co-existing disease that modifies the original disease process.

Because of this heterogeneity, a treatment effect detected using the entire group in a clinical trial may or may not exist for a particular subgroup of the original population.[24-26] Because of this, the data from subgroups of patients are often analyzed separately. In some circumstances this is important to determine which therapies are most effective in clinically-important subgroups of patients.[24,25]

Unfortunately, several problems can occur when subgroups of patients are analyzed. First of all, analyzing subgroups of patients requires the use of multiple statistical comparisons, increasing the chance of a type I error. Secondly, since each subgroup of patients is smaller than the entire study population, the statistical tests used in analyzing these subgroups may have low statistical power, increasing the chance of a type II error. These problems arise whether or not the subgroups were defined prior to the acquisition of any of the clinical trial data.

Additional problems may occur if the subgroups of patients are not defined properly.[24,25] A "proper" subgroup of patients is defined by signs, symptoms, or laboratory results available at the initial presentation or which are not modified by the treatments being compared. An "improper" subgroup of patients is defined by signs, symptoms, or laboratory findings which can, in principle, be modified by the treatments being administered. For example, in a study comparing different volumes of fluid resuscitation for patients with septic shock, an improper subgroup of patients might be defined by a low systolic blood pressure after fluid administration (the so-called refractory shock group). A proper subgroup would, on the other hand, be defined by a low systolic blood pressure prior to any fluid administration.

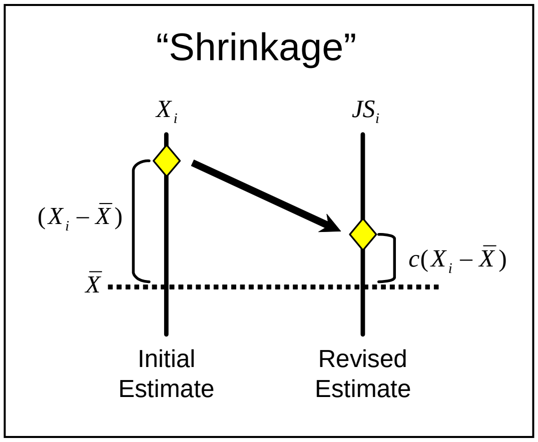
Because of the possible association between the subgroup assignment and the treatment being administered in an improperly defined subgroup, improper subgroups cannot be used to reliably assess the effectiveness of therapies. It is a common error in many retrospective studies to compare such improper subgroups of patients.

James-Stein Effect and Subgroups

In addition to the problem of increased risk of Type I and Type II errors when analyzing multiple subgroups in a clinical trial, there is a more subtle but pervasive issue which leads to overestimation of the variability in the treatment effect among the subgroups. This issue is sometimes called the James-Stein effect, because the mathematical tool used to correct the effect is called the James-Stein estimator. Here, I will provide a qualitative explanation of the effect, while avoiding most of the mathematical explanation.[27]

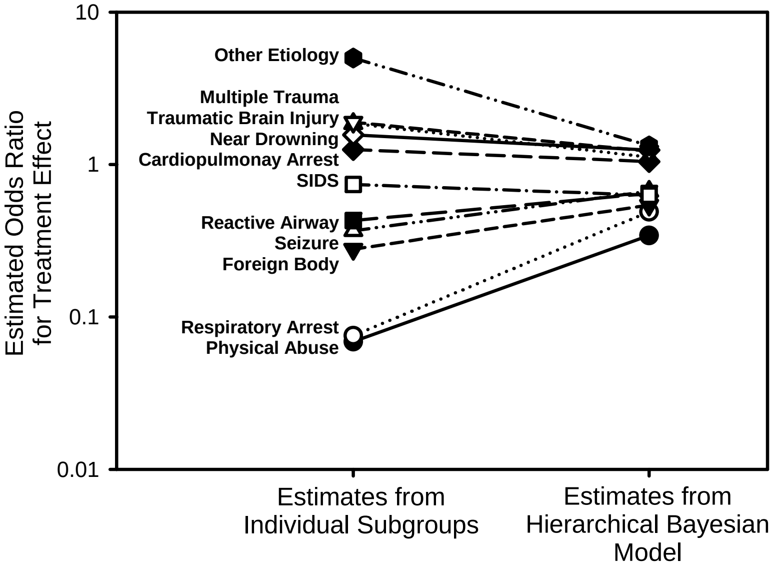
Consider a randomized clinical trial to determine the effect of a new treatment, compared to placebo, in which the true treatment effect is really the same in three or more distinct subgroups of patients (e.g., perhaps defined by the initial severity of disease, the clinical site at which they were entered or some other initial characteristic). When the trial is completed, the treatment effect for each of the subgroups is usually estimated by calculating each subgroup-specific difference in outcome, for example, by calculating separate odds ratios for each subgroup. Even though, in this example, the odds ratios should all be exactly the same, just by chance there will tend to be some variability. Thus, the variance or variability in the *observed* odds ratios is larger than the underlying variability in the treatment effects within the subgroups.

This implies that, when we report the odds ratios calculated for each of the subgroups as the observed treatment effects, we are reporting values that have too much “spread” or variance, and we tend to overestimate the treatment benefit in subgroups that appear to benefit from the new treatment and to overestimate the degree of harm in subgroups that appear to suffer harm. This is the James-Stein effect.[27]



One approach to correcting this problem is to use an expression for the treatment effect in each of the subgroups which “shrinks” each of the individual estimates towards the overall common treatment effect for the study population as a whole. To picture this shrinkage, the figure to the right represents the observed treatment effect in each subgroup (denoted by the subscript *i*) as X_i, so the difference between the observed treatment effect in each subgroup from the overall treatment effect is the difference X_i minus $\overline{X}$. After shrinkage, however, the deviation of each estimated treatment effect is reduced by a factor *c*, which is less than one, so all estimates are moved closer to the overall average. The value of *c* depends on the sample sizes of the subgroups, the variability in the subgroups, and other factors beyond the scope of this discussion. While there are a number of statistical modeling approaches which can be used to obtain treatment estimates with shrinkage, such as Bayesian hierarchical modeling, the details of these approaches are also beyond the scope of this discussion. The important point, however, is that one should be skeptical of observed treatment effects in subgroups that deviate significantly from the effects in the other subgroups or the overall treatment effect, unless there is a strong physiologic or mechanistic argument to explain the difference.

The next figure shows the results of applying this approach to the estimates of treatment effects in a study comparing endotracheal intubation to bag-valve-mask ventilation in the prehospital treatment of critically-ill infants and children.[28-30] The vertical axis is the odds ratio for survival to hospital discharge, so a value of 1 denotes no effect. While the raw observed treatment effects might suggest that there are subgroups of patients who experience benefit from endotracheal intubation and other subgroups who experience harm, application of the appropriate shrinkage factor reveals that there is no evidence of benefit or harm in any of the defined subgroups.



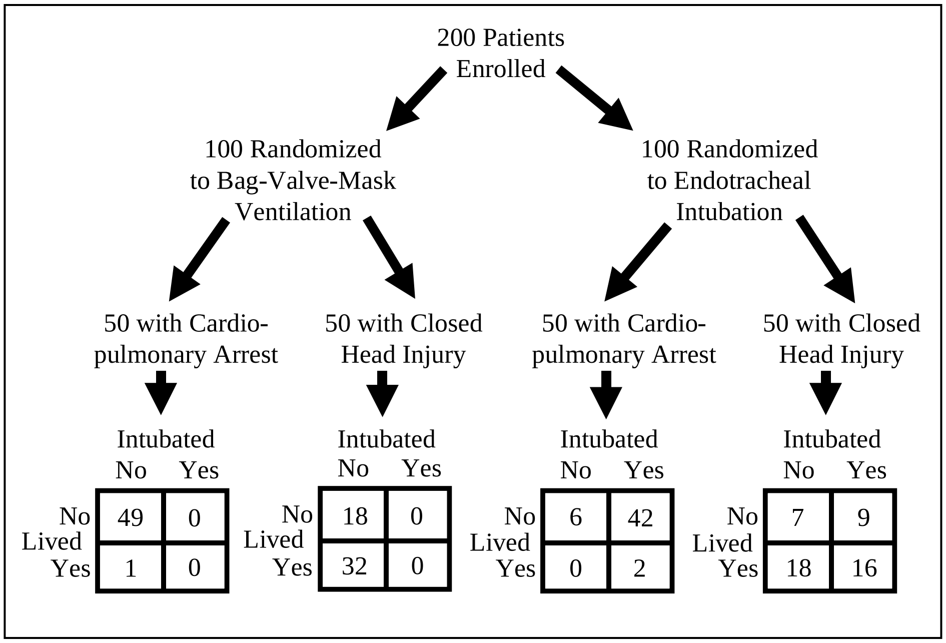
Intention-to-Treat Analysis

The effectiveness of any therapy is determined both by the therapy's inherent efficacy, and by one's ability to administer the therapy to the patient. For example, in the case of an oral medication the effectiveness will be decreased if patients find the side effects of the medication intolerable and are noncompliant. Similarly, an invasive procedure will be less effective if it can only be performed successfully in a minority of patients. To accurately estimate the effectiveness a therapy will have in clinical practice, one must properly account in a clinical study for those patients for whom a procedure is initiated but cannot be completed, or those for whom the medication is prescribed but not taken. This is the purpose of an “intent-to-treat” analysis.[31]

In an intent-to-treat analysis, patients are considered to be members of whichever treatment group to which they are originally assigned, regardless of whether or not they take the prescribed medication or whether or not the appropriate therapy is successfully administered to them.[31] For example, in a trial examining the use of an oral agent, patients would be considered to be part of the test group if they were originally prescribed the test medication, even if they freely admit to never having taken any of the tablets.

Another example is shown in the figure below, which shows a diagram of results obtained from a randomized trial of prehospital intubation in a pediatric population.[28,29] In this study, children were randomized to either receive bag-valve-mask ventilation alone, or bag-valve-mask ventilation followed by endotracheal intubation. The population of children was inherently heterogeneous, with some children suffering from cardiopulmonary arrest (e.g., sudden infant death syndrome), and some children suffering from closed head injury. The initial population was well randomized to the two interventions and the results shown in the Figure were obtained. It is important to note that children in cardiopulmonary arrest, because of their flaccid state, are substantially easier to intubate in the prehospital setting than children with closed head injury. Furthermore, it should be noted that the overall survival rate of children with closed head injury is much higher than the survival rate of children with cardiopulmonary arrest.

The goal of the analysis of the data from the figure is to determine the effect of out-of-hospital endotracheal intubation on survival. The correct way to group the data is to consider all patients initially assigned to the endotracheal intubation group as belonging to that group, whether or not they were successfully intubated. Thus there were 100 patients assigned to the endotracheal intubation group and 100 patients assigned to the bag-valve-mask ventilation group. Considering all patients in the endotracheal intubation group, whether or not


200 Patients Enrolled
100 Randomized to Bag-Valve-Mask Ventilation
100 Randomized to Endotracheal Intubation
50 with Cardio-pulmonary Arrest
50 with Closed Head Injury
50 with Cardio-pulmonary Arrest
50 with Closed Head Injury

Lived	Intubated No	Intubated Yes
No	49	0
Yes	1	0

Lived	Intubated No	Intubated Yes
No	18	0
Yes	32	0

Lived	Intubated No	Intubated Yes
No	6	42
Yes	0	2

Lived	Intubated No	Intubated Yes
No	7	9
Yes	18	16

they were successfully intubated, the overall survival rate is 36%. Similarly, considering all patients in the bag-valve-mask ventilation group, the overall survival rate was 33%. Thus the results of the study show a small trend towards improved survival with prehospital endotracheal intubation.

Alternatively (but incorrectly), one might group the data according to the treatment actually received.[31] In this case, children initially assigned to the endotracheal intubation group, but not successfully intubated, might be considered to be members of the bag-valve-mask ventilation group, as that was the treatment they actually received. If one counts all children who received only bag-valve-mask ventilation, then the overall survival rate in that group is 39% (51/131). Similarly, if one considers only children who were *successfully* intubated has the members of the endotracheal intubation group, then the overall survival rate in that group is only 26% (18/69). This type of analysis would then suggest that endotracheal intubation is quite harmful.

What is wrong with the analysis by treatment received? Because children with cardiopulmonary arrest are both easier to intubate (more compliant with the treatment administered) and have a much lower expected survival rate, analysis by treatment received suggests the therapy is harmful. In other words, the subgroup of children assigned to the endotracheal intubation group that were more difficult to intubated (closed head injury) and had a better expected survival were selectively reassigned to the bag-valve-mask ventilation group, because they could not be successfully intubated. Whenever there is an association between compliance, or the ability to administer an intended procedure, and the patient's baseline chance of a successful outcome, an analysis by treatment received does not accurately estimate the effect of the therapy in an actual patient population.[31]

Using Statistical Consultants

Many clinical investigators use statistical consultants to aid in the design of their clinical research, and in the analysis of the data obtained. With preparation, most investigators find such consultation extremely valuable. Often, potential problems with the proposed study can be anticipated by statistical consultants, allowing changes to be made in design or data collection methods before significant effort is wasted.

Following several rules will make such statistical consultation most valuable:

1. Try to define the single most-important question to be answered by the study, in terms of quantities that can be measured. Such quantities might include the percent survival in each of two patient groups, or the change in mean systolic blood pressure at a certain time. For a comparative study, define how big a difference you are looking for.

2. Get as much information as possible about what you *expect* to find in the control group. Estimates of the standard deviation of important continuous outcome variables, or estimates of the percent survival in the control group, are necessary for sample size calculations.

3. Decide on the α value, the power of the study, and what your maximum feasible sample size would be.

4. Consider important subgroups of the study population, and the variables that can be used to identify those subgroups.

5. Consider whether or not there are multiple important comparisons that will need to be performed at the end of the trial.

6. If possible, identify and bring for the statistical consultant examples of published studies that illustrate aspects of what you are trying to accomplish with your study.

7. Consider the feasibility of performing planned interim analyses of accumulating data, once you begin the study, so that the study may be stopped as soon as a reliable conclusion can be drawn.

Acknowledgements

Some of this material has been adapted from: Lewis RJ. The Medical Literature: A Reader's Guide. In: Emergency Medicine: Concepts and Clinical Practice. Fourth Edition. Rosen P, Barkin R, Danzl DF, Hockberger RS, Ling LJ, Markovchick V, Marx JA, Newton E, Walls RM, Editors. St. Louis: Mosby, 1998. Other material has been adapted from: Young KD, Lewis RJ. What is Confidence? Part 1: The Use and Interpretation of Confidence Intervals. Annals of Emergency Medicine 1997;30:307-310 and Young KD, Lewis RJ. What is Confidence? Part II: Detailed Definition and Determination of Confidence Intervals. Annals of Emergency Medicine 1997;30:311-318.

References

1. Menegazzi JJ, Yealy DM, Harris JS. Methods of data analysis in the emergency medicine literature. Am J Emerg Med 1991;9:225-227.
2. Lewis RJ, Bessen HA. Statistical concepts and methods for the reader of clinical studies in emergency medicine. J Emerg Med 1991;9:221-232.
3. Gaddis GM, Gaddis ML. Introduction to biostatistics: Part 4, statistical inference techniques in hypothesis testing. Ann Emerg Med 1990;19:820-825.
4. Brown CG, et al. The beta error and sample size determination in clinical trials in emergency medicine. Ann Emerg Med 1087;16:183-187.
5. Detsky AS, Sackett DL. When was a 'negative' clinical trial big enough? How many patients you needed depends on what you found. Arch Intern Med 1985;145:709-712.
6. Freiman JA, et al. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 "negative" trials. N Engl J Med 1978;299:690-694.
7. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122-124.
8. Lewis RJ. An introduction to the use of interim data analyses in clinical trials. Annals Emerg Med 1993;22:1463-1469.
9. Brown GW. Errors, types I and II. Am J Dis Child 1983;137:586-591.
10. Fleiss JL. Statistical Methods for Rates and Proportions. Second Edition. New York: John Wiley & Sons, 1981.

11. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Second Edition. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1988.
12. Gaddis GM, Gaddis ML. Introduction to biostatistics: Part 5, statistical inference techniques for hypothesis testing with nonparametric data. Ann Emerg Med 1990;19:1054-1059.
13. Lewis RJ. Parametric statistical tests: unnecessary assumptions, computers, and the search for the trustworthy p value. Academic Emergency Medicine 1998;5:1048-1050.
14. Young KD, Lewis RJ. What is Confidence? Part 1: The Use and Interpretation of Confidence Intervals. Annals of Emergency Medicine 1997;30:307-310.
15. Young KD, Lewis RJ. What is Confidence? Part II: Detailed Definition and Determination of Confidence Intervals. Annals of Emergency Medicine 1997;30:311-318.
16. Smith DG, et al. Impact of multiple comparisons in randomized clinical trials. Am J Med 1987;83:545-550.
17. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 1. Introduction. Mayo Clin Proc 1988;63:813-815.
18. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 2. Comparisons among several therapies. Mayo Clin Proc 1988;63:816-820.
19. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 3. Repeated measures over time. Mayo Clin Proc 1988;63:918-920.
20. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 4. Performing multiple statistical tests on the same data. Mayo Clin Proc 1988;63:1043-1045.
21. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 5. Comparing two therapies with respect to several endpoints. Mayo Clin Proc 1988;63:1140-1143.
22. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 6. Testing accumulating data repeatedly over time. Mayo Clin Proc 1988;63:1245-1250.
23. Lewis RJ, Bessen HA. Sequential clinical trials in emergency medicine. Ann Emerg Med 1990;19;1047-1053.
24. Yusuf S, et al. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991;266:93-98.
25. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992;116:78-84.
26. Lee KL, et al. Clinical judgement and statistics: lessons from a simulated randomized trial in coronary artery disease. Circulation 1980;61:508-515.
27. Efron B, Morris C. Stein’s Paradox in Statistics. Scientific American 1977;236:119-127.
28. Gausche-Hill M, Lewis RJ, Gunter CS, Henderson DP, Haynes BE, Stratton SJ. Design and Implementation of a Controlled Trial of Pediatric Endotracheal Intubation in the Out-of-Hospital Setting. Annals of Emergency Medicine 2000;36:356-365.
29. Gausche M, Lewis RJ, Stratton SJ, Haynes BE, Gunter CS, Goodrich SM, Poore PD, McCollough MD, Henderson DP, Pratt FD, Seidel JS. Effect of Out-of-Hospital Pediatric Endotracheal Intubation-Effect on Survival and Neurologic Outcome: A Controlled Clinical Trial. JAMA 2000;283:783-790.

30. Lewis RJ, Gausche M, Abdi M. Subgroup Analysis of Data from a Prospective Randomized Study of Prehospital Airway Management in Children Using Classical and Bayesian Techniques (Abs). Academic Emergency Medicine 1998;5:428.
31. Lee YJ, et al. Analysis of clinical trials by treatment actually received: is it really an option? Statistics in Medicine 1991;10:1595-1605.