**Chapter 50  
Data handling and statistics essentials**

**Craig D. Newgard and Roger J. Lewis**

**Introduction**

Performing well-designed and properly analyzed clinical and systems research is the essential first step in improving out-of-hospital medical care, and only through such research can we better understand the effectiveness of both current and new practices. Data handling procedures and appropriate use of statistics are critical aspects of high-quality out-of-hospital research. The purpose of this chapter is to present and explore concepts underlying data handling and statistical analysis for clinical research in the out-of-hospital setting.

**Classical hypothesis testing**

Historically, research data have been analyzed using *p* values and classical hypothesis testing [1–3], although there has been an increasing emphasis on the use of confidence intervals in the presentation of results. In classical hypothesis testing, two hypotheses or conclusions that might be supported by the data are considered. The first, called the null hypothesis, is the idea that there is no difference between two or more groups, with respect to the measured quantity of interest [2]. For example, in a study examining the effect of a new EMS dispatch system on response time (response interval), the null hypothesis might be that there is no difference between the median response times before and after implementation of the new system. The alternative hypothesis is the idea that the groups being compared are different with respect to the quantity of interest and, ideally, the alternative hypothesis defines the size of that difference [2]. For example, the alternative hypothesis might be that the new EMS dispatch system decreases the median response time by 1 minute or more compared to the old dispatch system. The difference between the two groups defined by the alternative hypothesis is called the “treatment effect” or “effect size.” The size of the difference implied by the alternative hypothesis, the treatment effect, must be defined prior to data collection [4–7].

Once the null and alternative hypotheses have been defined, the study is conducted and the data are obtained. During analysis of the results, the null hypothesis is “tested” to determine which hypothesis (null or alternative) will be accepted as true. 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]. For example, it may be determined that the median response time was 6 minutes for the new dispatch system and 7 minutes for the old system. Testing is done to determine the probability that the 1-minute decrease in response time is due solely to chance and is not a reflection of improvements in the dispatch system.

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 results like those obtained would occur less than α percent of the time (where α is usually 5% or 0.05) assuming the null hypothesis to be true, then the null hypothesis is rejected as false. Using our previous example, the researcher might state that he or she is willing to accept a 5% probability (α) of falsely concluding that there is a difference between dispatch systems when in reality there is no difference. If a 1-minute difference between the dispatch systems were observed, this difference could be due solely to chance, especially if the *p* value is greater than 0.05. However, using a *p* value of less than 0.05 as a cut-off (i.e. “statistically significant”), the probability of saying there is a difference between the systems when there really is not (called a type I error – see later in the chapter) is less than 1 in 20. Rather than proving the alternative hypothesis that a difference exists between the groups, we eliminate from consideration the null hypothesis that there is no difference.

Although the *p* value provides evidence for rejecting the null hypothesis, it does not provide information regarding the magnitude of the treatment effect or precision of the estimated treatment effect [8]. Furthermore, the investigator arbitrarily assigns α, the “level of significance.” A level of α equal to 0.05 is most often used and represents nothing more than medical convention. Probabilities above this level (e.g. 0.06) may still suggest an important difference in measured outcome between treatment groups, although not reaching a generally accepted level of proof. *P* values should be interpreted as only one piece of statistical information and are best interpreted in the context of the study design, the sample size, and the credibility of the hypotheses being tested [8]. Definitions of terms used in classical hypothesis testing are shown in [Table 50.1](https://jigsaw.vitalsource.com/books/9781118990827/epub/OPS/Vol2/c50.xhtml#c50-tbl-0001).

[**Table 50.1**](https://jigsaw.vitalsource.com/books/9781118990827/epub/OPS/Vol2/c50.xhtml#R_c50-tbl-0001) Definitions of terms commonly used in classical statistical testing

Source: adapted from Lewis RJ. The medical literature: a reader’s guide. In: Rosen P, Barkin R, Danzl DF, et al. (eds) *Emergency Medicine: Concepts and Clinical Practice*, 4th edn. St Louis: Mosby, 1998; and Lewis RJ. An introduction to the use of interim data analysis in clinical trials. Ann Emerg Med 1993;22:1463–9.

| **Term** | **Definition** |
| --- | --- |
| α | The maximum *p* value to be considered statistically significant; the risk of committing a type I error |
| α error | A type I error |
| Alternative hypothesis | The hypothesis that is considered as an alternative to the null hypothesis; the hypothesis that there is an effect of the studied treatment, of a given size, on the measured variable of interest; the hypothesis that there is a difference between two or more groups 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; the hypothesis that two or more groups are the same with respect to the measured variable of interest |
| Power | The probability of detecting the treatment effect defined by the alternative hypothesis (i.e. obtaining a *p* value <α), given α, and the sample size of the clinical trial; power = 1 − β |
| *p* value | The probability of obtaining results similar to those actually obtained, or results more inconsistent with the null hypothesis, assuming the null hypothesis is 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 or that the groups being compared are not different; 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, or that there is a difference between the groups that is as large or larger than the treatment effect tested; a false negative |

**Type I error**

A type I error occurs when one concludes that a difference has been demonstrated between two groups of patients when, in fact, no such difference really exists [2,3,9]. It is a type of “false positive.” Using *p* values, a type I error occurs if a statistically significant *p* value is obtained when there is no underlying difference between the groups being compared. The risk of a type I error is equal to the maximum *p* value considered statistically significant or alpha, typically set at a level of 0.05 [2,9].

**Type II error and power**

A type II error occurs when a difference does exist between the two groups that is as large as, or larger than that defined by the alternative hypothesis, yet a non-significant *p* value is obtained [2,4–7]. That is, the researcher states there is no difference between the groups because the *p* value is greater than α, when in reality there is a difference. A type II error is a type of “false negative.” A common cause of a type II error is an inadequate sample size.

The “power” of a trial is the chance of detecting a treatment effect of the size defined by the alternative hypothesis, if one truly exists [4–7]. Studies are usually designed to have a power of 0.80, 0.90, or 0.95. Because the power of a trial is the chance of finding a true treatment effect, the value β (1 − power) is the chance of missing a true treatment effect (i.e. the risk of committing a type II error) if a true difference equal to the effect size actually exists [4–6]. The value of α, the power, and the magnitude of the treatment effect sought (defined by the alternative hypothesis) are all used to determine the sample size required for a study [7].

**Power analysis and sample size determination**

Determining the required sample size for a clinical study is an essential step in the statistical design of a project. An adequate sample size helps ensure that the study will yield reliable information. If the final data suggest that no clinically important treatment effect exists, an adequate sample size is needed to reduce the chance that a type II error has occurred and a clinically important difference has been missed. If the final result is positive, an adequate sample size is needed to ensure the treatment effect is measured with appropriate precision. In addition to the basic study design and planned analysis method, four parameters influence the power of a study: the sample size, the effect size defined by the alternative hypothesis, the variability of the results from patient to patient, and α. For each method of statistical analysis and type of study design (unpaired samples, case–control studies, etc.) there is a different formula relating sample size to power.

Because the treatment effect size sought by the study is a major determinant of the sample size required, choosing an appropriate effect size is the first step in sample size determination. Optimally, one designs clinical studies to reliably detect the minimum clinically relevant treatment effect (i.e. the smallest treatment effect that would result in a change in clinical practice). Defining the minimum clinically significant treatment effect is a medical and scientific judgment, not a statistical decision. For example, what difference in response times is large enough to warrant changing a dispatch system (e.g. 4 minutes, 2 minutes, 1 minute, or 30 seconds)? The researcher has the responsibility of proposing an effect size that he or she believes would be clinically or operationally relevant. The decision about the effect size to be studied should be based on the researcher’s judgment, available resources (sample size feasibility), and relevant information from similar published studies.

The smaller the treatment effect sought by a study, the larger the sample size required. Frequently, available time and resources do not allow a clinical trial large enough to reliably detect the smallest clinically significant treatment effect. In these cases, one may choose to define a larger treatment effect size, with the realization that should the trial result be negative, it will not reliably exclude the possibility that a smaller but clinically important treatment difference exists.

One frequently faces the problem of interpreting data from a negative study in which no power calculation was initially performed. Although tempting, performing a *post hoc* (i.e. after the study is concluded) power analysis in which one calculates the effect size that could have been found with the actual sample size and a given power is invalid and should never be done. The correct approach to analyzing such data is to calculate the 95% confidence interval for the difference in the outcome of interest, based on the final data [10].

Because there is some risk in participating in clinical research, it is unethical to enroll patients in a trial that has an inadequate sample size and therefore is unlikely to yield useful information. Such a study would have little chance of generating meaningful information about the hypothesis and may waste time and resources, as well as expose subjects to risk with a low likelihood of benefit to society. This issue warrants special consideration during the design of phase I and so-called “pilot” studies. Studies involving human subjects (or laboratory animals) should be designed with a large enough sample size so it is highly likely they will yield useful information.

Historically, learning the techniques of sample size determination and power analysis has been difficult because of relatively complex mathematical considerations and numerous formulas. There has been tremendous improvement in the availability, ease of use, and capability of commercially available sample size determination software. These programs now allow the determination of sample size and the resulting power for a wide variety of research designs and analysis methods. Some of the available programs include nQuery Advisor by Statistical Solutions ([www.statistical-solutions-software.com/](http://www.statistical-solutions-software.com/)), PASS by NCSS ([www.ncss.com](http://www.ncss.com/)), and Power and Precision by Biostat ([www.power-analysis.com](http://www.power-analysis.com/)).

**Statistical tests**

Before reviewing some common statistical tests [11], it is useful to define the different types of characteristics, or variables, we may wish to compare [12]. The ability to differentiate the types of variables and measurements is critical to selecting appropriate analysis methods. There are two general scales of measurement: numerical and categorical. With a numerical (or quantitative) scale, the size of differences between numbers has meaning. Variables measured in a numerical scale may be continuous, having values measured along a continuum (e.g. age, height, weight, time), or discrete, taking on only specific values (e.g. number of paramedic calls per shift). The mean and median of variables measured using a numeric scale are often used to summarize results or to compare characteristics between groups.

Categorical variables are used for measuring qualitative characteristics. The simplest form of a categorical variable is termed a nominal scale, in which observations fit into discrete categories that have no inherent order (e.g. race, sex, hospital name). When there are only two categories (e.g. male or female), the variable is termed dichotomous or binary. If there are more than two categories, (e.g. blood types A, B, AB, and O), the variable is polychotomous (or polytomous). One common, but potentially ill-advised, practice is to categorize continuous variables (e.g. systolic blood pressure ≤90 mmHg or age by decades) for statistical analyses. Although this may simplify the interpretation of the variable, this practice substantially reduces the available information in the data, frequently requires an arbitrary selection of the appropriate cut-point(s), and may reduce study power as well as introduce residual confounding [13].

When a variable is ordinal, there is an inherent order among different categories. Examples include the Glasgow Coma Scale (GCS) score and the Apgar scores. Although an order exists among categories (each possible score is a category) in an ordinal scale, the relationship between categories may differ throughout the scale (i.e. the clinical implication of a difference between a GCS of 4 versus 3 is not the same as a difference of 15 versus 14).

Many statistical tests exist for both continuous and categorical data. Depending on the type of data being analyzed, different statistical tests are used to determine the *p* value. The most common statistical tests and their assumptions are described in [Table 50.2](https://jigsaw.vitalsource.com/books/9781118990827/epub/OPS/Vol2/c50.xhtml#c50-tbl-0002). Selecting the appropriate statistical test requires identifying the type of variable to be analyzed and making reasonable assumptions regarding the underlying distribution of the data and the standard deviation or variance of the data in each group [12]. Parametric tests require more assumptions about the data – typically that numeric data follow a normal distribution and that different groups yield data with equal with equal variance. Non-parametric tests do not require these assumptions [14]. Considering the low power of available tests used to detect deviations from the normal distribution, it is prudent to use non-parametric methods of analysis when there is any doubt as to the underlying distribution of the data [14].

[**Table 50.2**](https://jigsaw.vitalsource.com/books/9781118990827/epub/OPS/Vol2/c50.xhtml#R_c50-tbl-0002) Common statistical tests and their assumptions

| **Statistical test** | **Description** |
| --- | --- |
| **Parametric tests** |  |
| Student’s t test | Used to test whether the means of a continuous variable from two groups are equal, assuming that the data are normally distributed and that the data from both groups have equal standard deviation or variance. A less common form of the t test can be used to analyze data from matched pairs (e.g. before and after measurements on each patient) |
| One-way analysis of variance (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 equal standard deviations or variances. The one-way ANOVA may be thought of as a t test for three or more groups |
| **Non-parametric tests** |  |
| Wilcoxon rank sum test (Mann-Whitney U test) | Used to test whether two sets of continuous data have the same median. These tests are similar in use to the t test but do not assume the data are normally distributed |
| Wilcoxon signed rank test | Used to examine data from matched pairs, similar to the matched pairs t test, but when differences in each pair are not normally distributed |
| Kruskal-Wallis | This is a test analogous to the one-way ANOVA, but no assumption is required regarding normality of the data. The Kruskal-Wallis test may be thought of as a Wilcoxon rank sum test for three or more groups or as a one-way ANOVA for non-normally distributed data |
| Chi-square test | Used with categorical variables (e.g. two or more discrete treatments or groups with two or more discrete outcomes) to test the null hypothesis that there is no association between treatment and outcome. The chi-square test assumes at least five 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 five observations are expected in one or more categories of treatment and outcome |

**Parametric tests**

Parametric tests are used to analyze numerical data and require both that the data follow a normal distribution and that the variance of the data from each group is equal. For example, the Student’s t test (also called “t test”) is used to compare the means of numerical variables (e.g. serum glucose, respiratory rate) between two groups of patients. If there are three or more groups of patients, one-way analysis of variance (ANOVA) can be used to compare means between the groups [15]. A form of Student’s t test for paired data can also be used to compare mean differences between pairs of matched data points. Such data frequently result from studies in which baseline and posttreatment measurements are performed on each patient or experimental animal.

**Non-parametric tests**

When the data to be analyzed cannot be assumed to be normally distributed, then the *p* value should be obtained using a non-parametric test [2,14,16]. Non-parametric tests can be used for normally or non-normally distributed numerical data and provide a more robust estimate of the *p* value (i.e. an estimate less affected by the underlying distribution of the data than would be obtained with a parametric test). The trade-off for not requiring data to be normally distributed is a slight loss of power for detecting a true difference between groups (i.e. greater chance of type II error). This difference in power is usually of little practical significance but may require a slightly larger sample size to achieve the same desired power (e.g. an additional 10% more subjects).

The non-parametric alternative to a student’s t test for unpaired samples is the Wilcoxon rank sum test (also called the Mann-Whitney U test). For a statistical comparison of paired measurements, one can use the Wilcoxon signed rank test. The non-parametric alternative to one-way ANOVA is the Kruskal-Wallis test. These methods compare the medians between groups (as opposed to a comparison of means with similar parametric tests).

The chi-square test and Fisher’s exact test are non-parametric tests used to detect associations between treatment (or exposure) and outcome when both variables are categorical (e.g. placebo versus active drug, lived versus died, admitted versus discharged). For these analyses, the data are arranged by creating a table with individual cells representing all possible combinations of the variables (the most basic being a 2 × 2 table; [Figure 50.1](https://jigsaw.vitalsource.com/books/9781118990827/epub/OPS/Vol2/c50.xhtml#c50-fig-0001)). The chi-square test requires that each cell in the table have five or more expected observations. If your data do not fulfill this assumption, Fisher’s exact test may be used.

[**Figure 50.1**](https://jigsaw.vitalsource.com/books/9781118990827/epub/OPS/Vol2/c50.xhtml#R_c50-fig-0001) Example of a 2 × 2 table. This figure shows data from a hypothetical study comparing the outcomes of patients with medical cardiopulmonary arrest, separated by the presence or absence of an automatic external defibrillator at the site of the arrest. The total number of patients is 496. The other numbers written outside the table (43, 453, 127, 369) are termed “marginal” totals. The *p* value associated with these data is 0.068 when calculated using the chi-square test and 0.098 when calculated using a two-tailed Fisher’s exact test.

Many common statistical software packages (e.g. SAS, Stata, SPSS) may be used for database management and analysis. Although these software programs are efficient and generally provide a wide range of options for optimal analysis, a general statistical background and thorough understanding of the appropriate tests, the assumptions required, and associated limitations are necessary to accurately interpret the results and avoid drawing invalid conclusions. A given statistical program may not warn the user if an invalid approach to the analysis is being used.

**Confidence intervals [17,18]**

Suppose we wish to determine whether a new dispatch system decreases response times, as in our previous example. In our study, we observe a median response time of 6 minutes for patients transported under a new dispatch system and 7 minutes for patients transported under the previous dispatch system. The observed treatment difference is −1 minute. If the *p* value associated with the null hypothesis that the median response times are equal (a difference of zero) is less than 0.05, we reject the null hypothesis as false and we conclude that our study demonstrates a statistically significant difference in median response times. That the *p* value is less than 0.05 implies the treatment difference is statistically different from zero. The *p* value does not tell us the magnitude 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 an infinite number of patients could be included in the study.

The treatment difference identified in a study (e.g. −1 minute) estimates the “true” treatment difference between the groups and is called a “point estimate.” It is also possible to estimate a range of values between which the true treatment difference will lie with some degree of certainty (e.g. 95%). This “interval estimate” is called a confidence interval [11]. If, instead of reporting only a *p* value as the result of a study, the researcher reports the point estimate and the corresponding confidence interval around it, he or she will communicate whether the measured treatment difference was likely due solely to chance, plus information on the size of the treatment difference (and therefore its clinical importance) and the precision of the estimated difference.

The *p* value answers the question, “Is there a statistically significant difference between the two treatments?” The point estimate and its confidence interval answer the questions, “What is the size of the treatment difference?” and “How precisely did this study determine the true treatment difference?” The effect of a study on clinical practice should depend on whether the study has definitively demonstrated a treatment difference and whether that 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 the difference.

Returning to our example, a treatment difference (i.e. the calculated difference in response time between the two dispatch systems) of zero is equivalent to the null hypothesis that there is no difference in median response time between patients transported under either of the two dispatch systems. If a treatment difference of −1 minute were calculated with a 95% confidence interval of −0.5 to −2.5 minutes, an interval that excludes zero, it would imply that a true treatment difference of zero is not statistically consistent with our observed data. We would conclude that the null hypothesis is not consistent with our findings and would reject the null hypothesis. If a 95% confidence interval does not include a zero treatment difference, this is equivalent to a *p* value less than 0.05.

Our point estimate of a 1-minute decrease in response times for two dispatch systems gives an estimate of the size of treatment effect. However, the confidence interval (i.e. between −0.5 and −2.5 minutes) is statistically consistent with the true treatment difference being any value between −0.5 and −2.5. In other words, the true effect of the new dispatch system on response time may be a decrease of as little as 30 seconds or as much as 2.5 minutes. If the newer dispatch system requires significant purchase and recurring costs, an EMS medical director may conclude that even a 2.5-minute reduction in response time does not warrant the associated costs, although the difference is statistically significant. Another EMS medical director may feel that even a decrease of 30 seconds would be beneficial, despite the costs. When authors report *p* values alone, they often leave the reader with little basis for drawing conclusions relevant to their clinical practice [8]. With confidence intervals, the reader can decide what treatment difference is clinically important and can reach conclusions appropriate to his or her system.

We may also use confidence intervals to interpret information from studies that did not achieve statistical significance (so-called “negative” trials). Suppose we found the 95% confidence interval for the difference in median response time to extend from –3.5 minutes to +0.5 minutes, with the same point estimate for the difference in response times (−1 minute). This confidence interval indicates that the new dispatch system could decrease response times by as much as 3.5 minutes over the current dispatch system, or it could increase times by as much as 30 seconds. Because  *p* is greater than 0.05, it is tempting to conclude that there is no advantage to using the new EMS dispatch system. However, our data are also consistent with the new dispatch system decreasing response time by as much as 3.5 minutes. Although *p* is greater than 0.05, there remains the possibility that an important difference exists in the two dispatch systems.

Negative trials whose results are still consistent with a clinically important difference usually occur because the sample size is too small, resulting in inadequate power to detect an important treatment difference. A larger sample size will decrease the width of the confidence interval, allowing greater certainty of either no clinically important difference or determination of a difference that was not uncovered by an analysis of the smaller sample size.

It is important to know how precisely the study data define 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 and the narrower the confidence interval. As mentioned earlier, a trial that uses an inadequate sample size may not show a statistically significant result, yet may not be able to exclude a clinically important treatment difference. In this case, the confidence interval will be wide and imprecise, including both zero (no treatment difference) and clinically important treatment differences. Conversely, a positive trial that uses a very large sample size may result in a statistically significant treatment difference that is not clinically important (e.g. demonstrating a decrease in response time of 5 seconds).

If a confidence interval includes either zero or clinically unimportant treatment differences, as well as clinically important treatment differences, we cannot make any definitive conclusions. If this is the case, it will be necessary to repeat the study with a larger sample size, a more homogeneous population, or a more statistically efficient study design in order to narrow the width of the confidence interval and to provide more definitive conclusions.

**Multiple comparisons**

Whenever a characteristic of two groups of patients is compared statistically, even if the groups are fundamentally identical (i.e. they were randomly selected from the same larger population), 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 fundamentally no true difference between the two patient populations. This risk of a false-positive *p* value occurs each time a statistical test is performed. If multiple comparisons are performed, the risk of at least one false-positive *p* value is increased because the risk associated with each test is incurred multiple times [19–25]. This increased risk occurs whether the multiple comparisons are pairwise comparisons of more than two groups of patients or comparisons of many different characteristics between two groups of patients. The risk of obtaining at least one false-positive *p* value when comparing two groups of fundamentally identical patients and assuming each result is statistically independent is shown in [Table 50.3](https://jigsaw.vitalsource.com/books/9781118990827/epub/OPS/Vol2/c50.xhtml#c50-tbl-0003) as a function of the number of comparisons made. The overall or “studywise” risk of at least one type I error is roughly equal to the maximum significant *p* value used for each individual test multiplied by the total number of tests performed. This rough equality is the basis for the Bonferroni correction [2,19].

[**Table 50.3**](https://jigsaw.vitalsource.com/books/9781118990827/epub/OPS/Vol2/c50.xhtml#R_c50-tbl-0003) Probability of at least one type I (false-positive) error when performing multiple independent comparisons between identical groups with α = 0.05

| **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 |

The Bonferroni correction is one method for reducing the overall type I error risk for the whole study (the studywise risk) by reducing the maximum *p* value considered statistically significant (α) for each of the individual 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 [2]. For example, if five comparisons are to be made, then a maximum significant *p* value of 0.01 would be used as the cut-off 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. Because each statistical test is conducted using more stringent criteria for a statistically significant *p* value (i.e. a smaller α), there is an increased risk that each test will miss a difference as big as that defined by its associated alternative hypothesis by yielding a *p* value that is non-significant using the new criteria for *p*. Some investigators consider correction factors, such as the Bonferroni correction, for multiple comparison analyses to be controversial [26].

Certain statistical tests can be used to compare three or more groups of patients simultaneously, without the need to control for multiple comparisons. Examples include ANOVA, the Kruskal-Wallis test, the chi-square test, and Fisher’s exact test (see [Table 50.2](https://jigsaw.vitalsource.com/books/9781118990827/epub/OPS/Vol2/c50.xhtml#c50-tbl-0002)). These tests, which do not use the Bonferroni correction, have relatively high power, while controlling the overall risk of a type I error. Their disadvantage is that, although they may detect a difference among three or more groups of patients, they do not define which pairwise differences are statistically significant. Once an overall difference is demonstrated, additional statistical tests can be used to determine where the differences exist.

**Interim data analyses**

During the conduct of a clinical trial, data accumulate sequentially, gradually containing more and more information on the relative effectiveness of the treatments being compared. Often, however, the data are not analyzed until all patients have been enrolled. This type of fixed sample size design has the disadvantage that more patients than necessary to obtain a clinically important result may be enrolled, raising ethical concerns that “extra” participants were denied or subjected to a treatment that may have been proved advantageous or harmful, respectively. It is prudent to plan for one or more interim analyses of the data, which are conducted before the full sample size has been reached, to see if a final 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 because this is a type of multiple comparison. Although the details of this type of trial are beyond the scope of this chapter, the interested reader may study one of several reviews [25,27,28].

**Subgroup analysis**

Any group of patients is heterogeneous. This is especially true for patients treated by EMS systems. Some patients within a group defined by an out-of-hospital complaint or dispatch category may have a more severe form of the disease in question or no disease at all, and some patients may have a coexisting disease that modifies the primary disease process. Because of this heterogeneity, a treatment effect detected using the entire group in an EMS study may or may not exist for a particular subgroup of the original population [29–31]. To detect this heterogeneity in treatment effect, the data from subgroups of patients are often analyzed separately. In some circumstances, this is important to determine which interventions are most effective in important subgroups of patients [29,30]. This concept is exemplified in a clinical trial on prehospital pediatric intubation where the effect of intubation on clinical outcome was compared by diagnosis groups [32].

Unfortunately, several problems can occur when subgroups of patients are analyzed separately. First, analyzing subgroups of patients involves the use of multiple statistical comparisons, increasing the chance of a type I error. Second, because each subgroup is smaller than the entire study population, the statistical tests used in comparing 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 the clinical trial data, unless such planned analyses were integrated into the selection of α for each comparison and into the sample size calculations.

Additional problems may occur if the subgroups of patients are not defined properly [29,30]. A “proper” subgroup of patients is defined by signs, symptoms, or other characteristics available at the initial presentation (e.g. on the arrival of EMS personnel) that are not modified by the interventions being compared. Signs, symptoms, or other characteristics, which in principle can be modified by the interventions being evaluated, define an “improper” subgroup of patients. For example, in a study comparing different volumes of fluid resuscitation for patients with undifferentiated shock, an improper subgroup of patients might be defined by a low systolic blood pressure after fluid administration. A proper subgroup would, however, be defined by a low systolic blood pressure prior to any fluid administration. In this example, because of the possible influence of the fluid administered on the final subgroup assignment (if the postresuscitation blood pressure is used to define the subgroup), there is no way to accurately assess the effect of fluids on blood pressure in this subgroup. Unfortunately, many retrospective studies inappropriately compare such improper subgroups of patients.

The limitations of subgroup analyses should be recognized so that appropriate strategies for such analyses (e.g. *a priori* identification of meaningful and properly defined subgroups, limits on the number of planned subgroup analyses, and sample size calculations that account for such analyses) can be integrated into the study design before studies are begun and to assure that results from subgroup analyses are interpreted appropriately [33].

**Intention-to-treat analysis**

The effectiveness of a therapy in practice is determined both by the therapy’s inherent efficacy and by one’s ability to administer the therapy to the patient. For example, an intravenous medication will be completely ineffective if it cannot be given because no intravenous line can be established in the field. Similarly, an invasive procedure will be less effective on average if it can only be successfully performed in a minority of patients. To accurately estimate the effectiveness of an intervention in the out-of-hospital setting, one must properly account for those patients in a study for whom a procedure is initiated but cannot be completed or those for whom the medication is ordered but cannot be administered. This is the purpose of an “intention-to-treat” analysis [34].

In an intention-to-treat analysis, patients are considered to be members of the treatment group to which they were originally assigned, regardless of whether the appropriate therapy or intervention was successfully administered [34]. For example, in a study examining paramedic use of endotracheal intubation in children, a patient would be considered to be part of the intubation group if he or she was originally randomized to that group, even if the patient could not be intubated [32].

**Multivariable analyses**

When multiple patient characteristics or variables potentially influence a given outcome, or if there are differences between groups that must be “controlled for” to obtain a valid estimate of the treatment effect (e.g. crash severity in assessing airbag-related injuries among pediatric passengers), multivariable analyses can be very useful [35,36]. The goal of a multivariable analysis is to quantify the separate effect of each multiple predictor (or “independent”) variable on the outcome of interest (the “dependent” variable). The type of multivariable analytic technique selected depends on the types of variables to be analyzed (e.g. numerical versus categorical), the goals of the analysis, and other assumptions regarding the data. The two general types of multivariable analytic methods are mathematical modeling and stratified analysis [37]. Multivariable modeling provides a mechanism for integrating several variables into the same analysis to account for the variety of factors that may influence a given outcome and to increase comparability between patient groups. Stratification involves separating the sample into two (or more) groups, based on a given variable (frequently a confounder), and analyzing these groups in parallel. Because a detailed discussion of multivariable analysis is beyond the scope of this chapter, we will focus our attention on one of the most common methods found in the biomedical literature: multivariable logistic regression.

Multivariable logistic regression is a form of mathematical modeling used with a categorical outcome and multiple predictor or confounding variables. The outcome is generally dichotomous or binary (e.g. survival versus death). The predictor variables can be numerical or categorical. Additional assumptions must be fulfilled for this statistical technique to be valid. This type of modeling allows an investigator to assess how a single predictor affects the outcome of interest, while assuming one could hold all other variables in the analysis constant (i.e. controlling for the effect of these additional variables). This technique is very useful when certain confounding variables must be controlled for in order to generate a valid estimate of the association between one variable and the outcome of interest. Generally, a measure of association (odds ratio) is calculated for each predictor variable, along with the confidence interval and *p* value for the null hypothesis that the odds ratio is 1 (i.e. no independent association).

**Clustering**

“Clustering” refers to correlated data and represents the tendency of subjects who have some features in common (e.g. the EMS agency that responds to their 9-1-1 calls) to have other characteristics in common as well (e.g. hospital disposition or receiving certain types of hospital care). This effect leads to correlated observations that violate the assumption of statistical independence required by most common statistical tests [38]. When clustering is present, subjects within groups that show common features tend to be more alike than subjects drawn from different patient groups. Some prehospital examples of clustering include patients evaluated by the same EMS agency, ambulance crew, or hospital. These “clusters” may exhibit similarity in patient characteristics, EMS care (e.g. a certain EMS agency or crew may have more or less experience with certain procedures than other crews), or hospital care (e.g. treatments and outcomes at certain hospitals may be better or worse than other hospitals). The phenomenon of clustering may exist at multiple levels simultaneously, a situation that represents hierarchical clustering.

Although appropriately accounting for clustering in statistical analyses can be complicated, the important point is that failing to account for these correlations can result in inappropriately narrow confidence intervals and artificially low *p* values (i.e. increased type I error rates). Standard analyses do not properly account for the smaller variance between subjects in a cluster as well as the variance between clusters [38]. Ignoring clustering may also bias results of the analysis. Fortunately, some analytic methods appropriately account for clustered observations. For a more detailed discussion of cluster analysis, the reader is referred to an article by Wears [38].

**Missing data**

Even in rigorous, prospective out-of-hospital research, it is typically impossible to gather every intended data element from every patient. Thus, even in well-executed studies, some data values will be missing in the final dataset. Although sometimes regarded as a simple nuisance or minor distraction, the inappropriate handling of missing values in statistical analyses can introduce substantial bias in study results, decrease the precision of estimates, and reduce study power, potentially leading to invalid conclusions. There are valid methods for analyzing data with missing values (e.g. multiple imputation, maximum likelihood estimation, Bayesian estimation), provided certain assumptions are met. One important assumption is that the underlying mechanism of missingness is “ignorable” [39]. That is, the probability that values are missing must be either independent of both observed and unobserved values (a situation called “missing completely at random”) or must be entirely explained by the observed values (a situation called “missing at random”). One key strategy underlying many valid methods for handling missing data is to use observed values to predict plausible values for missing data, while appropriately accounting for the uncertainty (i.e. variance) inherent in this process. Ideally, an investigator would provide a plan for handling missing values (including appropriate adjustment for sample size) before initiating a research project. Because a detailed discussion of analysis methods for missing data is beyond the scope of this article, the reader is referred to a reference text [39] and two articles that describe considerations and options for handling missing data in emergency care research [40,41].

## Using statistical consultants

Statistical consultants can be extremely valuable in designing, implementing, and analyzing clinical research and should be involved early in the course of planning a research study. The consultant’s expertise may be helpful in anticipating potential problems with the proposed study, allowing changes in design or data collection methods before unnecessary effort and time are wasted.

If several guidelines are followed, statistical consultation will be more efficient and productive. First, the investigator should clearly define the focus and purpose of the study, as well as the single most important question to be answered, in quantitative terms. Such quantification might include the response time for each of two patient groups transported under different dispatch systems or the change in systolic blood pressure after a pharmacological intervention. For a comparative study, the proposed effect size should be clearly defined and judged to be clinically important. Next, you should retrieve 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 percentages of a specific outcome (e.g. survival) in the control group, are necessary for sample size calculations. These values may be found in previous studies, preferably those carried out in the same (or a similar) sample population, or in existing data from the same sampling frame (e.g. the county served by a given EMS agency).

The investigator should attempt to identify potentially important subgroups of the study population and identifiers that could be used to specify those subgroups during the course of the study. Specification of the need for multiple comparisons and the number of comparisons that will be performed in the analysis should be made at the beginning of the study to avoid confusion in the correction for multiple comparisons. Examples of previously published studies that illustrate aspects of what the researcher is trying to accomplish with the planned study can be very helpful, especially to further refine the purpose of the project for a statistical consultant. Anticipated rates of missing values for key variables (e.g. outcomes), plus appropriate methods for handling missing values, should be specified before initiating a study and should be accounted for when doing sample size calculations. Finally, for interventional studies of clinical conditions with significant morbidity or mortality, one should consider performing interim analyses of accumulating data, so the study may be stopped as soon as a reliable conclusion can be drawn (i.e. definite benefit or harm from the intervention).

## References

1. 1 Menegazzi JJ, Yealy DM, Harris JS. Methods of data analysis in the emergency medicine literature. *Am J Emerg Med* 1991;9:225–7.
2. 2 Lewis RJ, Bessen HA. Statistical concepts and methods for the reader of clinical studies in emergency medicine. *J Emerg Med* 1991;9:221–32.
3. 3 Gaddis GM, Gaddis ML. Introduction to biostatistics: Part 4, Statistical inference techniques in hypothesis testing. *Ann Emerg Med* 1990;19:820–5.
4. 4 Brown CG, Kelen GD, Ashton JJ, Werman HA. The beta error and sample size determination in clinical trials in emergency medicine. *Ann Emerg Med* 1987;16:183–7.
5. 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;145:709–12.
6. 6 Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. 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;299:690–4.
7. 7 Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994;272:122–4.
8. 8 Gallagher EJ. P < 0.05: threshold for decerebrate genuflection. *Acad Emerg Med* 1999;6:1084–7.
9. 9 Brown GW. Errors, types I and II. *Am J Dis Child* 1983;137:586–91.
10. 10 Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med* 1994;121:200–6.
11. 11 Dawson-Sanders B, Trapp RG. *Basic and Clinical Biostatistics*, 2nd edn. Norwalk, CT:Appleton & Lange, 1994.
12. 12 Gaddis ML, Gaddis GM. Introduction to biostatistics: Part 1, Basic concepts. *Ann Emerg Med* 1990;19:86–9.
13. 13 Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006;25:127–41.
14. 14 Lewis RJ. Parametric statistical tests: unnecessary assumptions, computers, and the search for the trustworthy p-value. *Acad Emerg Med* 1998;5:1048–50.
15. 15 Gaddis ML. Statistical methodology: IV. Analysis of variance, analysis of covariance, and multivariate analysis of variance. *Acad Emerg Med* 1998;5:258–65.
16. 16 Gaddis GM, Gaddis ML. Introduction to biostatistics: Part 5, Statistical inference techniques for hypothesis testing with nonparametric data. *Ann Emerg Med* 1990;19:1054–9.
17. 17 Young KD, Lewis RJ. What is confidence? Part 1: The use and interpretation of confidence intervals. *Ann Emerg Med* 1997;30:307–10.
18. 18 Young KD, Lewis RJ. What is confidence? Part II: Detailed definition and determination of confidence intervals. *Ann Emerg Med* 1997;30:311–18.
19. 19 Smith DG, Clemens J, Crede W, Harvey M, Gracely EJ. Impact of multiple comparisons in randomized clinical trials. *Am J Med* 1987;83:545–50.
20. 20 O’Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 1. Introduction. *Mayo Clin Proc* 1988;63:813–15.
21. 21 O’Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 2. Comparisons among several therapies. *Mayo Clin Proc* 1988;63:816–20.
22. 22 O’Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 3. Repeated measures over time. *Mayo Clin Proc* 1988;63:918–20.
23. 23 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;63:1043–5.
24. 24 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;63:1140–3.
25. 25 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;63:1245–50.
26. 26 Savitz DA, Olshan AF. Multiple comparisons and related issues in the interpretation of epidemiologic data. *Am J Epidemiol* 1995;142:904–8.
27. 27 Lewis RJ. An introduction to the use of interim data analyses in clinical trials. *Ann Emerg Med* 1993;22:1463–9.
28. 28 Lewis RJ, Bessen HA. Sequential clinical trials in emergency medicine. *Ann Emerg Med* 1990;19;1047–53.
29. 29 Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;266:93–8.
30. 30 Oxman AD, Guyatt GH. A consumer’s guide to subgroup analyses. *Ann Intern Med* 1992;116:78–84.
31. 31 Lee KL, McNeer JF, Starmer CF, Harris PJ, Rosati RA. Clinical judgment and statistics: lessons from a simulated randomized trial in coronary artery disease. *Circulation* 1980;61:508–15.
32. 32 Gausche M, Lewis RJ, Stratton SJ, et al. Effect of out-of-hospital pediatric endotracheal intubation on survival and neurologic outcome: a controlled clinical trial. *JAMA* 2000;283:783–90.
33. 33 Wears RL, Cooper RJ, Magid DJ. Subgroups, reanalyses, and other dangerous things. *Ann Emerg Med* 2005;46:253–5.
34. 34 Lee YJ, Ellenberg JH, Hirtz DG, Nelson KB. Analysis of clinical trials by treatment actually received: is it really an option? *Stat Med* 1991;10:1595–605.
35. 35 Afifi AA, Clark V. *Computer-Aided Multivariate Analysis*, 3rd edn. Boca Raton, FL: Chapman & Hall/CRC, 1997.
36. 36 Hosmer DW, Lemeshow S. *Applied Logistic Regression*, 2nd edn. New York: John Wiley & Sons, 2000.
37. 37 Rothman KJ, Greenland S. *Modern Epidemiology*, 2nd edn. Philadelphia: Lippincott Williams & Wilkins, 1998.
38. 38 Wears RL. Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data. *Acad Emerg Med* 2002;9:330–41.
39. 39 Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, 1987.
40. 40 Haukoos JS, Newgard CD. Advanced statistics: missing data in clinical research—Part 1: an introduction and conceptual framework. *Acad Emerg Med* 2007;14:662–8.
41. 41 Newgard CD, Haukoos JS. Advanced statistics: missing data in clinical research—Part 2: multiple imputation. *Acad Emerg Med* 2007;14:669–78.