All men by nature desire to know.

Aristotle, *Metaphysics*

The practice of medicine has long been characterized as a combination of art and science; exploration into interactions of art and science in modern critical care medicine is a worthwhile endeavor because it informs how we acquire knowledge and apply it in practice. With respect to the so-called “art of medicine,” we believe that a better way to express this concept is with the term *expertise,* an essential quality manifest in individuals or small groups. Readers who have worked in an intensive care unit (ICU) will immediately recognize the importance of the individual as well as the collective expertise of teams of nurses, physicians, social workers, nutritionists, respiratory therapists, pharmacists, etc., in caring for acutely ill patients. While experts are often prepared with comprehensive formal education about their specialty, didactic instruction and reading alone are insufficient to acquire practical expertise; the key is extensive and involves ongoing, direct experience in performing the activity in question. Expert practitioners learn from personal—sometimes bitter—experience about what works and what fails in their setting. This helps explain the growing popularity of “hospitalists” and “intensivists” during the past decade (1,2).

Herein, we discuss fundamentals integral to understanding the science of statistics and its application to medical research. While many of the ideas we present are not readily explained in a single book chapter, we hope to stimulate the reader to both seek expert assistance initially and pursue further study. We have provided a number of our favorite resources as suggested references.

To make our explanations as concrete as possible, we cite several papers from core journals that deal with the problem of ventilator-associated pneumonia (VAP) since the problem of VAP is a common concern of all critical care practitioners. VAP serves as a useful example of a disease process with a specific definition, straightforward epidemiology, clearly articulated prevention strategies, and simple treatment options (i.e., various antibiotics). The study designs reported in these papers will include cross-sectional, cohort, and randomized clinical trials, and will illustrate various didactic points covered in the body of the chapter.

**FROM INFORMATION TO KNOWLEDGE**

How does science inform expertise in the ICU and how should practitioners use the “product” of science—journal articles—to improve care? First, it is helpful to define science as a collaborative process for acquiring, validating, and disseminating knowledge. The last part of our definition—knowledge—deserves some elaboration. A common definition of knowledge is *justified true belief* (JTB) (3). It should be self-evident that to know something (we will refer to it as “P”) means that one must believe P, and that P must be true. A harder concept to understand is that true belief must be justified in order to qualify as knowledge. To put it succinctly, true belief without justification is simply a “lucky guess.” This leads right back to defining the scientific method as an ongoing process for justification through empirical verification of shared belief. The goals of science are distinct from the method, and explicate the uses (applications) for the knowledge once it is acquired and corroborated. Most classic descriptions of scientific purpose include description, explanation, prediction, and control of the phenomena under consideration (3). Medicine is certainly an applied science, and the four goals are directly applicable when the phenomena of interest are human health and disease.

Journal articles and presentations at meetings are the main mechanisms for scientists in any field to advance, debunk, or corroborate various theories and hypotheses (i.e., advance and test knowledge claims). Critical care medicine is no exception, and this means that one ought to view individual papers that describe results and make claims from original research as part of a large work in progress, rather than any kind of established truth. Literature reviews, meta-analyses, and texts such as this one are written with the implicit acknowledgment of the contingent and dynamic nature of medical science and the knowledge it produces. An example of this dynamic is found in the controversy surrounding a clinical trial of ventilator settings for acute respiratory distress syndrome (ARDS) patients that was halted after an interim analysis, the results of which were released in advance of publication by the *New England Journal of Medicine* (4). Within weeks, rather intense and quite public criticism of the results and conduct of the trial was forthcoming from research subject advocacy groups as well as from within the academic community. These critics took issue with the trialist’s choice of treatment arms, claiming that they “excluded the middle” in comparing tidal volumes of 6 versus 12 mL/kg when most practitioners generally used an intermediate setting (8 to 10 mL/kg).

Effective reading of the medical literature requires an understanding of the role that journal articles play in scientific progress as described above. Equally important is a facility for critical thinking, tempered by a healthy dose of skepticism. By critical thinking, we mean being able to make and understand logical arguments consisting of premises and conclusions. In medical literature, these premises are often descriptions of empiric evidence that are made in quantitative terms (i.e., statistics). The key to critical reading—and effective writing—of medical literature is to not get lost in the numbers and to focus on assertions of evidence—methods and results—and
how they are used to support conclusions in the abstract and discussion sections. Skepticism may be restated as having an active bullshit (BS) detector. We use this term in all seriousness and with due deference to the philosophical work of Harry Frankfurt (5). He asserts that BS is increasingly common in modern society, proposing a simple and useful conceptual framework to handle it. Frankfurt articulates three distinct ways that people relate to the truth in what they say and write. These include telling the truth, lying, and BS. The difference between lying and BS is crucial and relates to the motives of the speaker/writer. Deliberately stating something that one believes to be untrue (i.e., a lie) implies an understanding and concern for what is actually true. In contrast, BS is produced with little or no regard for the truth status, coherence, or relevance of its content. To be successful, BS only has to be formulated and stated so as to sound good to the audience. By Frankfurt’s definition, under the pressure of publish or perish, medical literature has its fair share of BS—and readers would do well to keep this in mind.

We firmly believe that readers of this text are principled and ethical professionals who would never knowingly make or condone untrue (or BS) statements concerning any aspect of patient care. Unfortunately, a few members of the industries associated with the practice of medicine have deliberately told partial truths and sometimes even outright falsehoods. This may occur during overzealous marketing of drugs and medical devices to both physicians and patients. Perhaps more difficult for the average reader of peer-reviewed medical journals to guard against is the undue influence of a large and increasing amount of industry sponsorship of clinical research. Editors of biomedical journals and local research oversight committees both share a growing concern about this issue and have policies in place to ensure disclosure of potential conflicts of interest. A growing social phenomenon, closely related to BS, is that of deliberate ignorance about potentially difficult truths in some industries. Dr. Robert N. Proctor from Stanford has even coined a name for the study of organizational ignorance: *Agnatology* (6). US corporate culture has recently suffered from its tendency to ignore accounting and other structural problems until they threaten the existence of the company and land top executives in prison. Closer to home, we find high-profile cases of pharmaceutical companies brought to task for apparently ignoring or downplaying evidence of significant adverse events related to highly profitable drugs. By way of contrast, the physician culture seeks to expose difficult truths and learn from adverse events in the form of morbidity and mortality conferences. Another example of systematic local evidence is a database of case mix and clinical status. This database can be retrospectively mined for quality indicators, clinical trends, and outcomes. As electronic medical records systems and computerized critical care management applications gain traction, some of the clinical data collection and storage will be automatically performed. In the meantime, we encourage intensivists to create, maintain, and routinely consult a robust database of clinical information about all of their patients. Simply sharing information among a critical care team about such things as rates of VAP and skin breakdown stimulate both formal and informal efforts to standardize interventions and improve outcomes. Such local analyses can also help to focus reading of current literature as it emerges, as well as guide searches for published evidence, to answer specific clinical questions. Payers, regulators, and accreditation bodies are increasingly asking hospitals to report “quality indicators” that require exactly the sort of hospital-based process and outcomes analysis that we recommend above.

Below, we describe and illustrate some basic principles of study design and statistical analysis. One purpose of this explanation is to provide readers with tools to critically analyze the published literature pertaining to medical practice in the ICU setting. Some readers engaged in or planning a career in critical care or related research will undoubtedly find this material to be rather rudimentary. However, in keeping with our proposed “middle way,” where intensivists routinely collect data about their own patients, simple analytic methods can be quite useful. These include calculating rates, proportions, incidence, prevalence, and risk, as well as simple bivariate statistics from cohorts of critical care patients.
THE BAYESIAN–FREQUENTIST
DIALECTIC

Empirical observation about almost anything we encounter can be predicated in one of three ways: always, never, or sometimes. Our experience in medicine—and, indeed, most things—is generally of the “sometimes” variety. Thus, to make sense of complex and varying evidence over time, it was necessary to start recording, counting, and tabulating things and events. It can be argued that this is the main reason humans developed methods of counting, numbers to represent the results, and eventually mathematics. Quantifying observations, incorporating the resulting numbers into premises, and using normative methods for making and validating inferential conclusions are the raison d'être of probability theories and related statistical methods. Using these tools, enumerative inductions can be quantified with increasing sophistication. Over the past three centuries, western scientific methods have been spectacularly successful at describing, explaining, predicting, and sometimes controlling natural phenomena including human disease, disability, and death. During this time, two separate ways of conceptualizing about and calculating with quantitative empirical data have been articulated. These are, of course, the frequentist and Bayesian paradigms. What follows will draw on concepts from several fields including philosophy of science, biostatistics, and medical decision-making. Experts in these disciplines may feel that we are oversimplifying or distorting some of these concepts; for that, we apologize.

The frequentist paradigm is related to scientific realism, a view holding that the universe is deterministic. Theories are further from, or closer to, some absolute reality, and accepting a theory implies belief that it actually is true. Population parameters are defined as being real, fixed, and singular (i.e., they are constants rather than variables); causal and correlational relationships are likewise fixed and uniform. Any variance or error in our estimations of parameters and relationships between them come from only three sources: sampling, unobserved factors, and measurement. In this view, if one could “simply” measure the right things precisely and often enough, a complete understanding of a “clockwork” universe would be attained. This translates into a very particular way to view the results of experiments and statistical inference. Hypotheses are tested by calculating the probability of observing the results of an experiment or sample measurement given a specific answer or parameter value. A clinical trial of drug treatment for a particular disease is a classic example. The assumption is that there actually is a fixed and immutable answer to a question of the form: Does drug A work better than placebo in treating disease X? Analyzing the trial produces a single yes/no answer to that question, and the uncertainty is expressed as the chance that the answer would be wrong if we could repeat the experiment over and over (type I and type II errors).

Clinicians function in a largely frequentist world. They make dichotomous (yes/no) decisions about disease status and categoric (choice between types) decisions about treatment based on what they believe to be definitive diagnoses. The basic logic behind clinical decisions parallels that of frequentist hypothesis tests: “I am going to make a choice under uncertainty about the disease that my patient has and how to treat it. I know that I will be wrong some of the time and seek to minimize that frequency but will never eliminate it.” The relative frequency of making the wrong diagnosis or selecting the wrong treatment is analogous to frequentist type I and type II errors.

The Bayesian paradigm is related to empiricism. This philosophical view holds that all we can ever know about is what we have observed. There is absolutely no way of knowing whether the data-generating process is deterministic, stochastic, or chaotic. Further, for any given set of observations, there is an infinite set of theories that could explain it—this is called underdetermination. In selecting theories, the main criterion is that they are empirically adequate (agree with the data); belief in a theory in any absolute sense simply because it agrees with observations is never justified. Population parameters are viewed as not necessarily fixed or even ever knowable. Formal Bayesian theory holds that parameters have distributions of possible values. In contrast, the frequentist theory relies on fixed parameters and only allows samples to have distributions. Under Bayesian reasoning, hypothesis testing takes the form of calculating the probability that the parameter takes a certain value given the observed data and our past experience. Note that this is exactly opposite to the frequentist ordering where we take the population parameter as given and ask questions about the probability of observing the data.

The Bayesian world view is often represented as being incompatible with frequentist thinking. Modern Bayesians advocate a revolutionary paradigm shift in biostatistics and medical decision-making (14,15). This is modeled after Thomas Kuhn’s description of scientific revolutions exemplified by heliocentric replacing geocentric cosmology, and Einstein’s relativity versus Newtonian physics (16). In fact, Bayes introduced the concepts of prior and posterior probabilities about a century before Fisher and Pearson struggled over frequentist methods for analyzing agricultural experiments (17). One reason that frequentist methods for statistical testing initially became popular was that the required calculations were relatively easy to perform by hand. With the advent of powerful computers in the late 20th century, Bayesian statistics became possible to compute, and their conceptual advantages could be realized in practice. The ideas behind Bayesian decision-making seem to be more understandable to most physicians, as evidenced by the common misapprehension of frequentist core principles such as the p-value and confidence intervals (18,19).

Most physicians are familiar with a small subset of the Bayesian paradigm, though it is known by the deceptively inclusive name of “Bayes theorem.” The theorem quantifies how prior probability for disease X is modified by a Bayes factor (the likelihood ratio) derived from knowledge about test performance and the current patient’s test result (20). The product of this operation is the posterior probability of disease X in our patient, given the test result (positive or negative). The posterior probability is known as the positive or negative predictive value of the test. To illustrate the Bayes theorem, consider the following example taken from Armitage (21). From genetic theory, it is known that a woman with a hemophilia carrier has a probability of being a carrier of a hemophilia gene. A recombinant DNA diagnostic probe provides information contributing toward discriminating between carriers and noncarriers of a hemophilia gene. The prior probability of the woman being a carrier for the gene is 0.50. This probability will be modified after the DNA test result is known, resulting in the posterior probability that the woman is a carrier given either a negative or positive test result.
MOTIVATIONS FOR READING MEDICAL LITERATURE

There are many reasons for reading medical literature. Patients and family members often desire to educate themselves in order to become better advocates. Medical students and residents read articles for problem-based learning sessions and journal clubs. Physicians read articles in order to “keep up” with the medical literature and also to learn more about treating an individual patient. Physician researchers stay abreast of their research fields by regularly reading journals in their area of expertise. Physician researchers may also read articles in a peer-review process. Finally, policy-makers at various levels (e.g., government, payers, hospital, ICU directors) read and synthesize medical research in order to write guidelines and make informed policy decisions. Various reasons for reading medical literature will determine the extent of statistical expertise and rigor required for critical assessment. Herein, we will touch on basic statistical concepts, discuss common study designs, and provide guidelines for critical reading of medical literature. Interpreting the Medical Literature (22) is a useful resource for many consumers of medical journal articles. We also recommend The Handbook of Research Synthesis by Cooper and Hedges (23) for more detail on methodology for in-depth medical literature reviews including meta-analysis.

Basic Statistical Concepts

Statistics is the science of collecting, describing, and analyzing data, that provides the analytical framework for transforming information into knowledge. Of course, this knowledge is imperfect unless, perhaps, a biologic mechanism is identified that completely explains a phenomenon. Statistical methods quantify the imperfection of knowledge by providing results with an associated measure of error (e.g., level of significance, p-value, margin of error).

There are statistical concepts underpinning nearly every aspect of research, a process that includes the following six steps:

1. Pose a research question and formulate into statistical terms
2. Design a study
3. Collect data
4. Describe data
5. Analyze data using statistical inference
6. Answer the research question

Basic statistical concepts in the research process will be outlined below, but first, we give some vocabulary:

- **Population**—entire group of individuals of interest
- **Sample**—a subset of the population
- **Data element**—a measurement or observation on an individual
- **Population parameter**—a summarizing characteristic of all possible data values such as a population mean or proportion (the value of a population parameter is usually the object of a research question)
- **Statistic**—any quantity calculated from data
- **Inference**—process of extending or generalizing information known about a sample to the entire population
- **Probability distribution**—assignment of probability to the possible values of data that could be observed
- **Sampling distribution**—assignment of probability to the possible values of a statistic that could be observed

Table 7.1 gives the notation for common population parameters and the corresponding statistics that estimate them. Figure 7.1 shows the relationship between a probability distribution (represented by the dotted line) and a histogram. The probability distribution of a set of all possible values of a measure (e.g., all possible ages, all possible values of the PaO2/FiO2 ratio, all possible plateau airway pressures, etc.) is conceptual and not observable, whereas a histogram can be graphed from sample data. We have insight into the actual shape of the probability distribution from observing the outline of the sample histogram; the more data we have, the more refined our histogram is, and the true shape of the probability distribution of the data emerges more clearly.

In the “big picture” view, the science of statistics turns information into knowledge by connecting the object of a

---

**TABLE 7.1 Nomenclature and Symbols of Standard Parameters and Associated Statistics**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Population Parameter</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>µ</td>
<td>( \bar{X} )</td>
</tr>
<tr>
<td>Median</td>
<td>( \eta )</td>
<td>( \hat{X} )</td>
</tr>
<tr>
<td>Variance</td>
<td>( \sigma^2 )</td>
<td>( S^2 )</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>( \sigma )</td>
<td>( S )</td>
</tr>
<tr>
<td>Proportion</td>
<td>( \pi )</td>
<td>( p )</td>
</tr>
<tr>
<td>Correlation</td>
<td>( \rho )</td>
<td>( r )</td>
</tr>
<tr>
<td>Regression coefficient</td>
<td>( \beta )</td>
<td>( b )</td>
</tr>
</tbody>
</table>

1Population generally refers to the group of individuals of interest, but in the term population parameter, population refers to the “population” of all possible values of a measure (i.e., data element).
research question (an unobservable population parameter) into evidence provided by a research study (observable data) through statistical theory. Here we will assume the frequentist paradigm. For example, an investigator asks a research question such as the following: “What is the incidence of ventilator-associated pneumonia?” The research question is then “translated” statistically into a question about a population parameter. The “incidence of VAP” is a population proportion (i.e., \( \pi \)). A research study will be designed, and data from a sample will be collected. To analyze the data, an inference will be made. In other words, information from the sample will be generalized to the entire population. Figure 7.2 illustrates a hypothetic population and sample. Each circle represents an individual in a critical care setting. Dark circles are patients who have VAP. As Figure 7.2 indicates, a valid inference depends on obtaining data from a sample that fairly represents the population of interest—hence, the concept of random sample, a sample free from systematic bias in the way it is chosen or retained in a study.

Valid inference also depends on the selection of an appropriate statistical method to analyze data. Statistical methods are based on statistical and mathematical theory, assuming certain conditions (e.g., random sample, large sample size, normally distributed data, etc.) in order to provide valid results. Researchers (typically a team of physicians and biostatisticians) are responsible for choosing appropriate statistical methods and checking that no violations of assumptions have occurred. Statistical methods work like this: My research question is about a population parameter (e.g., incidence of VAP, a population proportion, \( \pi \)). Using my data, I can compute a statistic (e.g., incidence of VAP in my sample, a sample proportion, \( p \)). I will then apply the correct theory that will connect the statistic I have observed to the population parameter of interest. The theoretical connection occurs through the mathematical knowledge of the sampling distribution of a statistic. To better understand the idea of sampling distribution, refer to Table 7.2.

Suppose that 20 research teams are interested in estimating the incidence of VAP, and each team can observe 30 patients. Table 7.2 contains the raw data and the sample proportion (i.e., observed incidence of VAP) obtained in each study. In this conceptual framework, we can think of the statistic (e.g., sample proportion, \( p \)) as having a probability distribution (i.e., sampling distribution), which can be visualized by graphing a histogram (Fig. 7.3).
In the case of a sample proportion, the Central Limit Theorem (CLT) tells us that under certain conditions (i.e., random sample and large sample size), the shape of the sampling distribution will be normal, centered at the value \( \pi \). Note that the histogram in Figure 7.3 has an approximately normal shape. The CLT also tells us that the center of the normal curve is \( \mu \), the true incidence rate of VAP. Here is the power of the statistical method; even if we do not know the probability shape of the original data, under conditions that are not too difficult to achieve, we (approximately) know the probability shape of the statistic and its connection to the population parameter of interest. We can then use this connecting theory to conduct inference. This “central” idea (hence the namesake of the CLT) is depicted in Figure 7.4 for the case of numerical data—say, the ages of patients who develop VAP in a critical care unit. In this case, the CLT tells us that the shape of the sampling distribution of the sample mean \( \bar{x} \) will be normal, centered at the value \( \mu \)—the population mean—even though the probability distribution of the original data is not normal.

The theory underlying statistical methods generally requires the understanding of probability and calculus, and thus is something of a “black box” for many. For a more in-depth discussion about statistical inference, see Cox (24).

There are three basic goals of statistical inference:

1. Estimation
2. Test of hypothesis
3. Prediction

In the estimation type of inference, the research question is concerned with estimating some characteristic or feature of a population of measurements (e.g., what is the incidence of VAP?). In the test of hypothesis type of inference, the research question is concerned with testing a relationship (e.g., does protocol-driven weaning reduce the incidence of VAP?). Finally, the prediction inference type of research question estimates a characteristic or feature of a population of measurements that will be observed in the future (e.g., what will be the incidence of VAP in 10 years’ time?). The form of statistical results will depend on the inference goal. In the case of estimation, the result will be reported in terms of a confidence interval. A confidence interval (CI) gives a plausible range of values for a population parameter such as a mean, a proportion, or an odds ratio, along with a measure of method success (i.e., the confidence). Tejerina et al. (25) found that VAP was present in 439 out of 2,897 patients, 13.2%; a 95% exact CI for VAP is given by (12.0%, 14.4%). This means that in 95 out of 100 studies, the true value of VAP incidence would be captured in the constructed confidence interval.1 The margin of error for the estimate is half the length of the CI (or 1.2%).

In the case of a test of hypothesis, the result will be reported in terms of a \( p \)-value.2 A test of hypothesis for testing a relationship works by setting up two competing hypotheses—the null (relationship does not exist) and the alternative (relationship exists)—and using the observed value of the data to provide evidence for rejecting or failing to reject the null. There are two potential errors that can occur: Type I (rejecting the null hypothesis when it is true) and type II (failing to reject the null hypothesis when it is false). The power of a test is the probability of rejecting the null hypothesis when it is false (the reverse of the probability of type I error). The observed value of the data (i.e., the value of the test statistic) provides a \( p \)-value that is compared to a predetermined level of significance, also known as alpha (\( \alpha \)) or the probability of type I error. When the \( p \)-value is smaller than the set level of significance (typically set at 0.05), there is evidence for rejecting the null hypothesis and concluding that a relationship exists between two factors. When the \( p \)-value is larger than the set level of significance, then the test conclusion is to fail to reject null hypothesis. Evidence for concluding that there is no relationship between two factors depends on designing a study with adequate power to detect a “meaningful” relationship.

Tejerina et al. conducted a series of tests of hypotheses to test the relationship of factors thought to be associated with the development of VAP (e.g., gender, neuromuscular disease, sepsis, type of ventilation), finding a number of factors related to VAP at the 0.05 level of significance. For example, sepsis was significantly associated with development of VAP (\( p \)-value

\[ ^1 \text{A 95% CI does not mean that there is a 95% probability that the true value of the population parameter falls in the interval. In the frequentist paradigm, the population parameter is considered a fixed, absolute value. The probability that this value falls inside the CI is either 0% or 100% (i.e., it either falls inside or outside).} \]

\[ ^2 \text{A } p \text{-value is the probability of the observed data (or data showing a more extreme departure from the null hypothesis) when the null hypothesis is true.} \]
Distribution of original data (ages) | Sampling distribution of \( \bar{x} \)

**FIGURE 7.4** Data is sampled from a probability distribution of any shape. In this illustration the original data comes from a distribution that is skewed right. Suppose multiple studies are conducted and for each study the statistic \( \bar{X} \) is calculated from the sample data. The Central Limit Theorem posits that the distribution of the statistic is normal with mean \( \mu \), the population mean of the original data.

less than 0.001). The odds ratio of sepsis and VAP (estimate of the strength of association) was given as 19.9 with 95% CI as (15.7, 25.4). This means that the odds of developing VAP in patients with sepsis were 19.9 times higher than the odds of developing VAP in patients who did not have sepsis. The 95% CI gives a plausible range of values indicating that a feasible range for the odds ratio is as low as 15.7 and as high as 25.4. In 95 out of 100 studies, the true value of the odds ratio will be captured in the 95% CI.

In the case of inference where prediction is the goal, the result will be reported in terms of a prediction interval. A prediction interval is analogous to a CI but will be calculated in such a way to reflect the additional source of variation due to estimating future values. Estimating the incidence of VAP in 10 years’ time would be an example of inference with prediction as its goal.

After collecting the data, we use two main tools to describe it: graphs and summary statistics. A list of summary statistics is given in Table 7.1. In the analysis step, we will use a statistical method that can generally be classified as a univariate, bivariate, or multivariate method, according to the number of data elements involved in the research question. For example, estimating VAP involves one data element and would be considered a univariate analysis. Investigating the relationship of type of ventilation and VAP would be a bivariate analysis. Analyses involving more than two data elements are commonly referred to as “multivariate.” The choice of statistical method will depend on the inference goal, the number of data elements in the analysis (i.e., univariate, bivariate, and multivariate), and the data type(s).

There are five data types:
1. Categorical nominal
2. Categorical ordinal
3. Binary
4. Numerical discrete
5. Numerical continuous

Data types are determined by the possible values a data element can have. Categorical data have values that are names or categories and not meaningful numbers. The categorical ordinal data type has values that can be ordered. The categorical nominal data type has values that have no natural ordering. Binary data have two values. Numerical discrete data when graphed are isolated points on a number line, while numerical continuous data have values that can be conceptually viewed as an interval on the number line. Table 7.3 gives examples of data types from the critical care literature. Table 7.4 lists the most common statistical methods with a brief description of each.

In 95 out of 100 studies, the true value of the odds ratio will be captured in the 95% CI.

For more details about statistical methods, see Mozhalsky (26), Armitage et al. (21), and D’Agostino et al. (27). Understanding how statistical methods are applied is crucial to understanding the “evidence” that EBM generates. As Gauch wrote, “Method precedes results; method affects results. Method matters” (28).

The sixth and final step of the research process—answering the research question—is of paramount importance. Researchers present results in the form of technical reports, abstracts, posters, oral or platform presentations, manuscripts, and books. Writing and presenting results clearly is an art form in and of itself. The research is not complete unless important patterns are summarized, the specific aims of a study are

### TABLE 7.3 Variable Types

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Examples</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical nominal</td>
<td>Race</td>
<td>(African American, Caucasian, other)</td>
</tr>
<tr>
<td></td>
<td>Causes of acute respiratory failure</td>
<td>(ARDS, postoperative, aspiration, sepsis, trauma, congestive heart failure, cardiac arrest)</td>
</tr>
<tr>
<td>Categorical ordinal</td>
<td>Ventilation type</td>
<td>(none, manual, mechanical)</td>
</tr>
<tr>
<td></td>
<td>Type of surgery</td>
<td>(elective, urgent, emergent)</td>
</tr>
<tr>
<td>Binary</td>
<td>Gender</td>
<td>(male, female)</td>
</tr>
<tr>
<td></td>
<td>VAP</td>
<td>(absence of VAP, presence of VAP)</td>
</tr>
<tr>
<td>Numerical discrete</td>
<td>Number of prior surgeries</td>
<td>(0, 1, 2, 3, …)</td>
</tr>
<tr>
<td></td>
<td>Number of central venous lines</td>
<td>(0, 1, 2, 3, …)</td>
</tr>
<tr>
<td>Numerical continuous</td>
<td>Age</td>
<td>(0–100+ years of age)</td>
</tr>
<tr>
<td></td>
<td>PaO(_2)</td>
<td>(80–100 mmHg)</td>
</tr>
</tbody>
</table>

ARDS: acute respiratory distress syndrome; VAP: ventilator-associated pneumonia.

---

*In technically precise terms, a multivariate analysis involves the analysis of outcome data that is multidimensional (e.g., cluster analysis, factor analysis, principal components, etc.). A common usage of “multivariate analysis,” though, is an analysis that includes more than two variables (e.g., multiple regression, logistic regression, etc.).*
There are four basic types of study designs used in clinical research: controlled trials, cohort studies, case control studies, and cross-sectional studies. These terms address the context and methods used in the study design, and the “story” of how the study that gave rise to that data was conducted. For more on writing about the results of statistical analyses, see The Chicago Guide to Writing about Multivariate Analysis, by Jane Miller (29,30).

Types of Research Studies

The statistical methods used to analyze research data depend on how the study that gave rise to that data was conducted. There are four basic types of study designs used in clinical research:

1. Cross-sectional
2. Case control
3. Cohort
4. Experimental (clinical trials)

For the following discussion about these study designs, we need to clarify some terms that are commonly used to describe medical research and epidemiologic findings. These may have different meanings depending on the context they are used in. These terms include sample, outcome, factor, exposure, treatment, and control.

- The term sample refers to the subjects being studied in the research. This reminds us that we are looking at a subset of a population of interest and that the purpose of the research is to apply what we find out about the sample to the population.
- The outcome of any research is handled in statistical analysis as the independent variable. The outcome is what we are primarily interested in understanding, treating, or preventing. In medical research, the outcome often relates to some disease or condition, with the simplest results being present or absent. In the following critical care–related examples, VAP (present or absent) will serve as the outcome. Defining and determining outcome (disease) status in clinical research and epidemiology is a large subject in itself. We will stipulate that there is an unambiguous and agreed upon way to measure the outcome status in the following discussion and examples.
- A factor is measured along with the outcome to determine if there is a correlation between them. In statistical parlance, factors are referred to as independent variables while the outcome is the dependent variable. The structure of relationships between factors and outcome is often quite complex, which, under the philosophy of scientific realism, reflects some underlying causal structure. In our example of VAP, factors that have positive association with VAP are considered to be risk factors. On the other hand, a protective factor has a negative association with the outcome (VAP). Note that we avoid directly asserting that an association between factor and outcome implies that the factor causes or prevents the outcome in deference to the old—and still true today—saying that correlation does not prove causation.
- With respect to a factor, exposure simply refers to the status (or level) of the factor in a particular subject. In the case of VAP (outcome), we would say that a patient in a coma is exposed to the risk factor of obtundation. Perhaps the best known example to both professional and lay public is lung cancer (outcome) with a risk factor of tobacco use to which a person is exposed if he or she is a smoker.
- In any clinical research, factors and outcomes are things that we seek to observe, measure, and record, but not influence. In contrast, a treatment (or intervention) is something that is actively controlled by the investigator. Conveniently, statistical terminology uses the term treatment in the same spirit as implied in clinical research. That is to say, treatment is an experimentally manipulated factor whose influence on the outcome we are interested in knowing.

In experimental studies, the term control group refers to subjects that do not receive any treatment. In case control studies (described further below), control subjects are those who do not have the outcome in question.

Table 7.5 summarizes the features of the four types of clinical research designs in their simplest forms, and they are each described below in more detail. We also include an example from current core literature in critical care medicine to illustrate
the principles. It would be useful for readers to obtain copies of these papers and review them along with our explanations. Note that these studies often used more complex and involved methods of statistical analysis including secondary outcome measures; we will focus only on primary outcomes, factors, or treatments, and simple relationships between them. The order in which we present the design types reflects increasing cost, time, and potential risk or inconvenience for patients. Thus, observational and retrospective studies (cross-sectional and case control) are discussed first, followed by prospective cohort studies and clinical trials.

**Cross-Sectional**

A cross-sectional study does not involve the passage of time. A single sample is selected without regard to outcome or risk factor exposure status. Information on outcome and exposure status is determined with data collected at a single time point. The status of the outcome can be compared between exposed and unexposed groups. It is important to understand that even though we may define one attribute as the “outcome” and others as “factors,” there is no logical way to determine anything other than the current relationship between them. Without additional information, there is no valid way to infer even the temporal order of factor levels and outcome status, let alone any causal connection.

Our example, “Prevention of ventilator-associated pneumonia: current practice in Canadian intensive care units”, comes from the *Journal of Critical Care* [31]. In this study, Heyland et al. wanted to know the current status of VAP prevention strategies in Canadian ICUs prior to disseminating a new set of clinical guidelines. There is not really a defined “outcome” in this study; rather, a number of factors of interest were simultaneously measured. Such a study design is sometimes called a “snapshot.” In this case, dietitians recruited by the investigators directly observed VAP prevention strategies in ICUs throughout Canada on a single date (April 18, 2001). Since the 66 observers recorded data for every patient currently in their assigned ICU, the unit of analysis was the single patient (*N* = 702). The investigators followed up the initial observations by manually abstracting the medical chart entries from April 18, 2001, for each of the 702 patients. Finally, surveys were filled out by 66 ICU directors that asked questions about the regular practices relating to VAP prevention on their unit; thus, the unit of observation for this part of the study was an individual ICU (*N* = 66). These three methods are good examples of the various ways that cross-sectional data can be gathered and aggregated.

From the many results presented in this paper, we present a few examples of the sorts of statistics that are typically reported in cross-sectional studies. From the survey of ICU directors, we get results like university affiliation (29/66 = 44%) and number of beds (mean = 13.9). From the observations and chart abstractions, authors report patient gender (women = 299/702 for 43% and men = 403/702 for 57%), intubated and ventilated (403/702 for 57%), and patient age (mean = 63.5 years). As for VAP prevention strategies from direct observation, we have elevation of head of bed (mean = 30 degrees) and kinetic bed therapy (22/702 for 3%). From the survey of unit directors, 61 of 66 (92%) responded that they never used special endotracheal tubes that allowed subglottic secretion drainage, and none stated that they used prophylactic antibiotics. This example is typical of cross-sectional studies in that only descriptive statistics (as just described) are necessary and usually suffice. Additional methods that might be used in analyzing a cross-sectional study such as this one would be simple statistics to quantify relationships (correlation) between measured factors (none was reported in the paper). For example, at the ICU level, authors could have used survey results to look for a relationship between university affiliation and VAP strategies such as subglottic secretion drainage. This would be done by cross-tabulating the two variables into a 2 × 2 table in this case. The appropriate statistical method to test for a significant relationship between two categoric variables is a chi-square test (or the Fisher exact test in the case of sparse cell sizes).

**Case Control**

In a case control study, the investigator compares individuals who have a positive outcome status (the cases) and individuals who do not have the outcome (the controls). In the simplest implementation of the case control method, one or more control patients (outcome negative) are chosen for each case (outcome positive). When the outcome being studied is relatively rare, there are often more available candidates for controls than for cases in the sample available for study. A subsample of available controls is usually selected to match the cases on characteristics (factors) that might be related to the outcome but are not of primary interest to the research question. For example, it is common to match on gender and age by finding a single control subject with same gender and similar age for each case. This is called 1:1 (control:case) matching. When there are many more outcome-negative (control) candidates, investigators may use 2:1, 3:1, or greater (control:case)
matching ratios while still maintaining similarity between each case and its controls (e.g., gender and age).

Investigators look backward in time (i.e., retrospectively) to collect information about risk or protective factors for both cases and controls by examining past records, interviewing the subject, or in some other way. Unlike with a cross-sectional study, we can get an indication of whether the factor status predated the development of the outcome by asking the questions carefully. However, case control studies are subject to well-known bias in assessing presence and timing of risk/protective factors. It is very important to understand that in a case control study, the outcome frequency is fixed in the design, which means that we cannot directly estimate risk of the outcome. We can estimate relative risk of the outcome between different status levels of the factors by calculating the odds ratio between cases and controls for each factor. This estimate of relative risk will be biased, depending on the actual prevalence of the outcome in the population of interest and the relative numbers of cases and controls. For example, in a 1:1 case control study, the outcome frequency is, by definition, 50%. If the actual frequency of the outcome in the population is less than 10%, the odds ratios will severely overestimate the relative risk increase or reduction. The raw odds ratios may be corrected for this bias to form a better estimate of relative risk, though this is rarely done in practice.

An example of a case control study looking at factors associated with VAP, “Epidemiology and outcomes of ventilator-associated pneumonia in a large US database,” can be found in Chest (32). The investigators used information from a large (750,000/year) database of inpatient hospital (N = 100) admission abstracts (MediQaul-Profile Database) for 18 months beginning in January of 1998. They first identified 9,080 patients having at least 1 day of mechanical ventilation in the ICU during their hospitalization without an admission diagnosis of pneumonia. Of these, 842 (9.3%) developed pneumonia after initiation of ventilation, thus meeting criteria for VAP. From one to three controls (VAP negative, N = 2,243) were selected to match each case (VAP positive, N = 842) on duration of ventilation, severity of illness on admission, and age. It is important to note that after this step, none of the matched factors can be meaningfully evaluated because their distributions were forced to be in direct proportion to each other during the process of matching.

Investigators calculated odds ratios between cases and controls for several factors that might alter risk of developing VAP in practice; these included gender, race, obtundation, and type of ICU admission (trauma, medical, surgical) among others. As an example of how such results are evaluated, the numbers of males and females in VAP-positive cases were 540 (64%) and 302 (36%) and for VAP-negative controls were 4,262 (52%) and 3,976 (48%), respectively. The odds ratio male/female is calculated by (540)(3,976)/(302)(4,262) = 1.67. This can be interpreted by stating that the odds of VAP in males are 67% greater than for females, which can be used as an estimate of the relative risk of VAP in males compared with females. Note that in the original sample of patients, the frequency of VAP was 9.3%, whereas in the casecontrol sample used to estimate relative risk in males, it was 27%. Thus, the estimate of relative risk should be revised downward by methods that are beyond the scope of this chapter.

**Cohort**

A cohort is a group of individuals. The term comes from Roman military tradition where legions of the army were divided into 10 cohorts, and each in turn divided into centuries. These cohorts “march forward together” in time. In a typical prospective cohort study, investigators follow subjects after study inception to collect information about development of the outcome (disease). In a retrospective study, outcome status is determined from records produced prior to beginning the study. The cohorts are articulated based on information about risk factors precluding the outcome determination. In both types, initially outcome-negative (disease-free) subjects are divided into groups (cohorts) based on exposure status with respect to a risk factor. Cumulative incidence (the proportion that develops the outcome in a specified length of time) can be computed and compared for the exposed and unexposed cohorts. The main difference between prospective and retrospective cohort studies is whether the time period in question is before (retrospective) or after (prospective) the study is begun. In terms of bias and error in measuring outcome and risk factors, the retrospective cohort design is more problematic because we must rely entirely on historical records to know that subjects were initially outcome negative, what their risk factor status was, and the subsequent outcome status over time.

Despite logistical difficulty and expense, prospective cohort studies are very attractive to investigators because they allow direct estimation of absolute risk for the outcome of interest as well as differences in outcome based on various risk (or protective) factors (relative risk). In general, outcomes that develop quickly are easier to evaluate prospectively since the study will be finished sooner, with less chance for subjects to drop out or be lost to follow-up. Critical care medicine lends itself quite well to prospective cohort studies for this reason.

An example of a multinational and quite complex cohort study, “Incidence, risk factors, and outcome of ventilator-associated pneumonia,” comes from the Journal of Critical Care (25) and reports results of a study encompassing 361 ICUs in 20 countries. The cohort in question included 2,897 consecutive ICU patients who were mechanically ventilated for more than 2 days, with reason for admission not being pneumonia. These patients were a subset from a larger (N = 5,183) and already completed prospective study. Even though the authors analyzed existing data, they properly labeled their study as prospective because the patients were entered into the original study at time of admission to ICU, and all information was sequentially recorded in a database as the research progressed.

The outcome of VAP was strictly defined using Centers for Disease Control and Prevention (CDC) criteria prior to study inception and measured for each patient on a daily basis as yes/no. Though it may seem a trivial point, it is important to note that all patients had a VAP status of “no” on the first day of their ICU stay. Multiple baseline and clinical factors were measured. In the results section, the authors first considered the entire sample as a single cohort and reported the incidence of VAP as 439/2,897 (15%). There is not a meaningful hypothesis for the simple question of VAP incidence in the whole cohort, and we are instead making an estimate of VAP incidence in the population. Because of the large sample size, the authors were able to place 95% confidence intervals on this estimate of 14% to 16%.
In reading the rest of the results, it is helpful to think of each separate factor as dividing the entire study group (N = 2,897) into “subcohorts.” For example, it would give two cohorts (male = 1,809 and female = 1,088) for which VAP incidence was 293 (16%) for the males and 142 (13%) for the females. When considering the factors of problem type (medical = 1,911 and surgical = 986), the VAP incidence was 322 (17%) for medical and 117 (12%) for surgical. The null hypothesis in each of these “subcohort” studies is that VAP incidence is equal between the factor levels (male/female and medical/surgical). In these two examples, the null hypothesis was rejected for each of the factors (gender $p = 0.02$ and problem $p < 0.001$). Below, we give an example of a randomized trial with a sample size in each of three groups of about 130. The percentages of VAP are similar in magnitude as are the intergroup differences (roughly 18%, 10%, and 13%). However, because of the smaller sample size, the results do not reach statistical significance (at the 0.05 level).

There are two problems with doing multiple separate tests for simultaneously measured factors in a big study such as this one. First, the measured factors quite probably interact with each other in a complex way in their combined effect on VAP development. Thus, in any single (univariate) analysis such as with gender and VAP, the calculated relationship may be confounded with one or more other factors, and therefore biased away from the actual effect. Second, it is problematic to look at the same set of data repeatedly using different factor combinations because by sheer chance, 1 in 20 such tests will be “significant” at the 0.05 level. The optimal solution to both these problems is to perform multivariable regression analysis where the joint effect of all variables is tested simultaneously. The authors of this paper did such analyses, though the details are beyond the scope of this chapter. Suffice it to say, gender showed no significant effect on VAP after accounting for all other factors while problem still did. Finally, it is important to note that multivariable analysis cannot rescue a study from being confounded by unmeasured factors. The only certain way to avoid confounding is through the random assignment of factor levels prior to measuring the outcome (i.e., an experimental study).

**Experimental (Clinical Trial)**

In an experimental study (called a clinical trial in medical research), the investigator selects a sample of subjects with the same outcome status and randomly assigns each to a treatment (or intervention) condition. Subjects are followed in time, and the status of the outcome is measured and compared between the treatment groups; experiments and clinical trials are, by definition, prospective. As described above, the term control group is used for subjects that do not receive any treatment or intervention. Sometimes patients who are given “standard” treatment are said to be in the control group. A randomized experiment is the only way to definitively establish causality by empirical means. Similarly, for testing medical treatments and interventions, randomized trials are the only sure way to determine clinical efficacy. Both necessary and sufficient conditions to establish causality between a factor or treatment and outcome status are met in a randomized trial. These include unambiguous temporal association of cause before effect and elimination of any potential confounding factors (measured or unmeasured) that might affect the outcome. This helps to explain what may seem to be near-worship of the “prospective randomized clinical trial” in many discussions about medical research.

A good example of a randomized clinical trial comes from the *American Journal of Respiratory and Critical Care Medicine* (33), entitled, “Oral decontamination with chlorhexidine reduces the incidence of ventilator-associated pneumonia,” which nicely describes the research question and the result. Koeman et al. wanted to determine if two oral decontamination regimens would reduce incidence of VAP in intubated patients. Performing a classic double-blinded, placebo-controlled clinical trial in which 385 eligible patients were randomized to three treatment arms: chlorhexidine 2% (CHX), chlorhexidine 2% and colistin 2% (CHX/COL), and water (PLAC). During the ICU stay, all patients got mouth swabbing at identical intervals with the type of solution unknown to those caring for the patients. The primary outcome was carefully defined and evaluated using chart abstraction by a team of physicians who did not know the treatment assignment. These two design elements satisfy the definition of a double-blinded study because neither patients and providers nor those determining the outcome knew what the treatment assignments were. The first table in the results (their Table 1) shows baseline characteristics grouped by treatment assignment. Such a table is always included in any complete report of a randomized trial. The baseline characteristics are measured and presented because they might also influence the outcome (VAP). One is reassuring to see relative equality of the baseline characteristics because it shows us that “randomization works.” Note that the true power of randomized treatment assignment lies in the fact that we know that any other factors or characteristics that were not anticipated and/or measured will, by definition, also be equally distributed between the treatment arms. Some might rightfully observe that such baseline characteristic tables are superfluous to the core logic of a randomized trial. We carry on presenting them because they are reassuring and otherwise informative.

The outcome can be most simply expressed as an incidence of VAP during the ICU stay with individuals having a yes/no answer. When tabulated, the results (VAP total) were PLAC = 23/130 (17.7%), CHX = 13/127 (10.2%), and CHX/COL = 16/128 (12.5%). The trialists performed sophisticated statistical techniques using days to onset of VAP and survival analysis, as well as interim analyses to allow early termination of the trial. These are beyond the scope of this brief chapter. The null hypothesis in this study is that the incidence (hazard in survival analysis) of VAP was identical among all three treatment arms. The alternate hypothesis is that one or more of the treatment groups had significantly different incidence (hazard) of VAP. A chi-square test on the simple incidence of VAP between the treatment groups gives a $p$-value at 0.20, which indicates that the null hypothesis cannot be rejected. However, using survival analysis, authors found that both CHX ($p = 0.012$) and CHX/COL ($p = 0.030$) reduced the hazard of VAP compared with PLAC.

This is a good example of how the type of outcome variable analyzed affects the power of a study to detect differences. For simple counting of VAP incidence, the outcome variable is dichotomous (yes/no), while for survival analysis, the outcome variable is quantitative (days to development of VAP). In general, outcomes defined and measured numerically have
“more information,” and thus greater power to detect small differences, than categoric or dichotomous ones. In the paper cited above, the authors did not report the results of the simple chi-square test, which failed to reject the null hypothesis. We can speculate that if the chi-square statistic on simple VAP incidence had shown significant differences, the authors would have reported it. Finally, we recall the results of our cohort study example with similar VAP percentage differences but much larger sample size. In that study, the achieved p-values were much lower, reflecting the greater power of large samples to demonstrate small effects.

**CRITICAL READING OF THE MEDICAL LITERATURE**

Articles in the medical literature generally follow a prescribed structure consisting of the following components: (a) title, (b) author list, (c) keywords, (d) funding source, (e) abstract, (f) objective and hypothesis, (g) background, (h) methods (includes study design, measures, and data analysis), (i) description of sample, (j) presentation of findings and results, (k) discussion and conclusions, and (l) references.

Statistical aspects permeate a large number of articles in the medical literature. Miller (30) describes the similarity of writing about statistical analysis to the presentation of a legal argument. She writes:

In the opening statement, a lawyer raises the major questions to be addressed during the trial and gives a general introduction to the characters and events in question. To build a compelling case, he then presents specific facts collected and analyzed using standard methods of inquiry. If innovative or unusual methods were used, he introduces experts to describe and justify those techniques. He presents individual facts, then ties them to other evidence to demonstrate patterns or themes. He may submit exhibits such as diagrams or physical evidence to supplement or clarify the facts. He cites previous cases that have established precedents and standards in the field and discusses how they do or do not apply to the current case. Finally, in the closing argument he summarizes conclusions based on the complete body of evidence, restating the critical points but with far less detail than in the evidence portion of the trial.

Good scholarly writing should resemble a good legal argument. Good writing in the medical literature should also provide transparency of method in sufficient detail to allow for results to be replicated. Research quality is difficult to evaluate but integral to the process of “taking information to knowledge”—the ultimate goal for reading medical literature. Evaluation of literature and judging research quality is itself an academic discipline within the larger science of literature review. In broad strokes, we can group indicators of quality that are relevant to statistical considerations into the following categories: sampling and participation, measurement, data management, analytic framework (includes study design and statistical analysis), and reporting of results. Below we outline the structure of an article and point out quality indicators to look for in the “anatomy” of a research article.

**Title, Author List, Keywords, and Funding Source**

A good title can convey important information about the topic of a manuscript and can let readers know what is new or different about the work. For example, in a classic paper published in the journal *Intensive Care Medicine* (34), the title speaks eloquently of the content: “Prevention of nosocomial pneumonia in intubated patients: respective role of mechanical subglottic secretion drainage and stress ulcer prophylaxis.” Much has been written about the content and ordering of author lists for scientific journal articles and it should suffice to say that “honesty is the best policy.” Keywords are popular and certainly useful, though they represent the author’s subjective decisions about what is important. Ultimately, the National Library of Medicine staff does an excellent job of generating structured abstracts (e.g., MESH terms) from biomedical journal articles using the whole paper to do so. These PubMed database entries have become the dominant means by which the scientific community searches through the medical literature. Given the discussion above concerning the influence of corporate support on the conduct of research, it is vital that funding sources and author conflicts of interest be clearly and completely stated in any published paper. In the case of public or nonprofit research support, virtually all such agreements require authors to acknowledge the funding source in any related publications.

**Abstract**

Although should spend time analyzing the medical literature, it is clear that, given the pressures of everyday life and the journals that appear with seemingly increasing frequency on our desks monthly, we are often tempted to read only the title and the abstract. One final caveat: There may be important disparities among the results, discussion, and abstract. One memorable report compared two forms of fluid resuscitation. Three patients in one group had been given from two to three times the amount specified in the protocol. With exclusion of these patients properly in the data analysis, as noted in the results section, there were no differences between the two groups. With inclusion of patients with protocol violations, there was a “statistically significant” difference. The abstract cited the “statistically significant” analysis without any reference to the patients who should have been excluded. The authors’ conclusion of a statistically significant difference in treatment modalities was, in fact, denied by their own results. If you are in a hurry, do not just read the abstract and move on; come back and read the article properly when you have enough time.

**Objective and Hypothesis**

Obviously, the most pertinent starting point is an understanding of the investigator’s objective. The investigator has the obligation to state clearly and specifically the purpose of the study conducted, but this may be difficult to discern. In such cases, we may question whether the author had, indeed, a clear objective. “Fishing expeditions,” that is, extensive data collection projects with the intention of exploring and identifying important relationships, achieve success when the captain knows where the fish are. In other words, the so-called gold
mine of data does not guarantee that statistical search will lead to “pay dirt” and reveal important new relationships. The author, or we as researchers, must formulate specific objectives and a clear-cut hypothesis for testing. Lack of an understanding of objectives handicaps the reader and the author in any assessment or interpretation of the results.

A more specific and somewhat more subtle question in assessing objectives is classification of a study as descriptive and exploratory versus analytic. Using epidemiologic terminology, descriptive studies are those that “describe” diseases, characterize disease patterns, and explore relationships, particularly in regard to person, place, and time. Such studies mainly serve the purpose of “hypothesis generation.” The specific hypothesis can then be tested by an analytic study, one whose primary objective involves the test of a specific hypothesis.

To illustrate this distinction, a descriptive study reported the use of high-level positive end-expiratory pressure (PEEP) in acute respiratory insufficiency in patients who developed severe, progressive, acute respiratory insufficiency despite aggressive application of conventional respiratory therapy (35). Later, the term optimal PEEP, introduced in the first study, was updated in another descriptive study of 421 patients reported in 1978 (36). The second study entailed treatment of a large group with respiratory failure using titration of PEEP in conjunction with intermittent mandatory ventilation, but using cardiovascular interventions to support cardiac function until a preselected end point of 15% shunt could be achieved. The first study represented a description of the development of a treatment regimen; in the second study, refinements in this treatment regimen were applied to a broader population. Later, a hypothesis was constructed to test whether, in moderate arterial hypoxemia, there was any improvement in patient outcome or resource utilization using “optimal PEEP” compared with similar modalities of therapy, with an end point defined as achievement of nearly complete arterial oxygen saturation at nontoxic inspired oxygen fractions.

The hypothesis that PEEP titration to achieve an intrapulmonary shunt of less than 20% would have a better outcome or would achieve faster resolution of the disease process could not be substantiated in the analytic study (37). The two descriptive studies (35,36) identified a specific hypothesis that the third or analytic study tested.

**Background**

The background is generally an introduction with rationale on why the research that is being presented is important. This section should also contain a literature review and argument to show how the current research fills a gap in previous work. Sufficient detail on how the literature review was conducted should be reported so that it can be reproduced. The background should also reference seminal papers in the research area.

**Methods: Study Design**

The reader should consider carefully the definitions of the groups studied and the population to which the investigators intend to refer their findings. For instance, in the three PEEP studies quoted, the reader might assume that the failure to prove the hypothesis in the third study invalidated the findings of the two earlier descriptive studies. The third, an analytic study, however, involved only patients with early and moderate arterial hypoxemia. The original group of patients that was studied specifically excluded these patients and concentrated on developing therapy for those who had persistent hypoxemia despite aggressive application of conventional respiratory therapy. Thus, a technique that reversed hypoxemia in patients who were refractory to the then “conventional therapy” of acute respiratory insufficiency was found not to be useful in another population that had only moderate hypoxemia and did not have true adult respiratory distress syndrome. If the authors do not state clearly the populations with which they are dealing, the readers can easily lose this important distinction. This has even greater importance in review articles that may omit the important qualifiers or modifiers found in the original reports. The fact that a particular form of therapy useful in advanced disease has no particular advantage in patients with mild disease indicates that therapy should be restricted to patients who can benefit from treatment, rather than arrive at some alternative conclusion that titration of PEEP to preselected end points has no advantage.

The reader should examine carefully the methods section for a description of the study design, a definition of inclusion/exclusion criteria, and information about data management. A sample size justification should be given in the methods section that indicates the expected precision when the objective of a study is to estimate an unknown quantity or a power analysis when the objective of a study is a test of hypothesis, such as testing the relationship of groups (e.g., treatment group vs. control group) on a certain measure.

Epidemiologically, there are two major classifications of study design: experimental and observational. Loosely defined, an experimental study is one in which the investigator has control over or can manipulate the major factor under study. The epitle of the experimental study is the randomized controlled trial in which the investigator demonstrates “control” over the factor under study by randomizing patients to various regimens. Many prophylactic and therapeutic studies tend to be experimental in design. It cannot be assumed that just because a study was experimental and the investigator may have randomized patients that the study was well done and its conclusions are valid. Experimental studies are prone to various sources of bias and to poor execution. The label randomized is not equivalent to assurance of high quality, nor does it alone add validity to the study. Thus, randomized studies also need careful assessment of their design, methods, analyses, and conclusions. One other factor, blinding, is often viewed as an attribute of the highest-quality studies. If subjective elements are used to judge the effectiveness of treatment, there is a compelling rationale to blind the investigators. If there are subjective assessments of the patients’ response, there is a compelling rationale to blind the subjects. If all of the outcome variables are objective, blinding, strictly speaking, is unnecessary. Thus, in the assessment of a new medication to relieve pain, double blinding (both subjects and investigators) is necessary.

When the investigators cannot manipulate the major factor under study, they must rely on what has been observed; this study is an observational study. We should not view observational studies as being inferior to experimental studies. Clearly, a tight, well-designed, well-executed experimental study carries the greatest strength of evidence, but observational studies can also provide substantial, sound medical evidence. In fact, a well-planned and well-executed observational study can be
much more informative than a weakly designed and poorly executed randomized study. There are various approaches to the design of observational studies, such as cross-sectional, case control, prospective cohort, and retrospective cohort. The interested reader should consult basic epidemiology or statistics textbooks for further descriptions of these various design strategies and for the relative strengths and weaknesses of each design format (38,39).

With respect to observational studies, the reader should determine whether the data collection was prospective or retrospective. The principal advantage of prospective data collection is that the researchers, having clearly identified the objectives, can ensure collection of this relevant information in a manner that they can determine. Retrospective analysis of medical records depends on what happens to appear in the record, often with no indication of the manner in which the information was obtained. For example, gender, age, and hospital outcome (survival or death) are key data elements that may not appear for every patient in a retrospective chart review. Clearly, without a specified protocol, the researcher cannot anticipate that a daily blood gas, serum creatinine, or any other intermittent measurement dependent on a specific order will appear in the chart. Everyone should attempt a retrospective study (at least once) to learn the pitfalls and the impossibility of obtaining a complete database. This would enable each of the then-frustrated researchers to read other retrospective studies both with a great deal of deserved skepticism and with empathy for the difficulties with such research.

Selection of the study group is another important step. The researcher should look for possible sources of selection that would make the sample atypical or nonrepresentative. A sample selected by a random selection mechanism is generally more representative than a “convenience” sample; however, this is difficult to achieve. Allocation of treatments by a random mechanism is more achievable, but even such seemingly “random” allocation of cases such as alternate days may introduce an unappreciated bias. For instance, the Trauma Service at the University of Miami/Jackson Memorial Medical Center had two separate teams that alternated coverage every 24 hours; patients admitted on alternate days, therefore, are cared for by different teams of physicians. A study that entailed alternate-day assignment to treatment groups would entail, as well, the factor of differences in physician practice style, a factor that could not be disentangled in analysis of study results.

We must also consider the nature of the control group or standard of comparison. We frequently encounter the “historical control” group that usually has a “poorer” result than the contemporary group. The problem, of course, is that the basic assumption that the modality of treatment under investigation is the only cause for the difference in results is clearly erroneous. It has been tempting to ascribe the remarkable reduction in wartime mortality from World War II to Korea to Vietnam to the marked diminution in delay between injury and treatment. However, the entire surgical training experience changed during that time, an almost completely new pharmacopeia was available in Vietnam, and, most assuredly, many other variables are yet unaccounted for between the two eras. In fact, the principal reason for randomization in a study is to attempt to distribute the unknown and potentially important variables equally among groups to avoid selection bias. We may also see this effect if subjects accrue slowly and the study thus runs over many years. Other aspects of therapy may change and have a greater impact on outcome than the original variable selected for study.

Two aspects of clinical research that sometimes perplex beginning researchers and inexperienced readers are validity and generalizability. Validity deals with the ability of a study to give a scientifically sound answer to the question posed. Insofar as possible, this answer should be free from bias, uninfluenced by the effects of other related or confounding variables and with good statistical precision. Only then is there a basis for a valid study result. Generalizability deals with extrapolation of study findings to a larger population or to other groups. Assessment of generalizability depends on the degree to which the study subjects are representative of some larger target population and how well the selection of study subjects simulates the process of drawing a random sample from a population.

The ideal is for studies to be both valid and generalizable. In practice, this is rarely the case. In the design of clinical research, investigators face many situations in which they must choose between validity and generalizability. When faced with a choice, undoubtedly they should opt for validity. Without a valid study, an investigator has little or nothing of scientific merit. The investigator may have actually drawn a random sample from a larger population and have virtually ideal generalizability. But, if in the process validity was threatened or compromised, the findings are worthless. With findings of questionable or doubtful validity, there is nothing of value to generalize. Generalizability plays a subordinate role and, in fact, should not surface until validity has been firmly established. Often the reader must assume the onus of assessment of generalizability and of whether findings can be extrapolated to other populations.

**Methods: Measures**

In the reporting of research results, clarity in the definitions of the terms and measurements made has great importance. The more clearly the authors (or we as potential researchers) define the terms, including diagnostic criteria, measurements made, and the criteria of outcome, the more likely it is that we, the readers, can interpret the findings correctly and gain a proper perspective. For instance, in the field of invasive catheter-related infection, terms such as colonization, contamination, and infection of the catheter abound. Authors often use these terms differently, leading to great difficulty in interpretation and synthesis of results from different studies. Furthermore, a “positive culture” may represent different bacteriologic methodologies: Some authors use a semiquantitative culture of an intracutaneous catheter segment (40), whereas others use blood cultures aspirated through the catheter (41). Clearly, results from one methodology may not be comparable to another, and interpretations based on differing methodologies may lead to different conclusions.

We must also try to evaluate the methods of classification or of measurement. The essential question is to assess whether inconsistencies in observation or evaluation could have sufficient impact to influence materially the results of the study. We also must evaluate the reliability and reproducibility of the observations; this is more difficult to assess. Frequently, some clues inform the reader of the author’s concern with and awareness of reproducibility and reliability. When a subjective
element enters into an assessment, an author often refers to and sometimes provides data on the results of evaluations by independent observers and their degree of agreement. *Interrater reliability* refers to the ability of two or more independent rater to make the same observations. *Intrarater reliability* refers to an observation made by the same rater over two or more different times. With respect to abstracting information from charts, interrater and intrarater reliability is usually in the range of only 80% to 90%. An author who devotes some attention to issues concerning measurement or laboratory error seemingly would be cognizant of the importance of reproducibility and reliability. It is well to be suspicious of results from a study that seems entirely devoid of concern with these elements, especially if some subjective element is clearly involved in diagnosis, observation, or assessment of outcome.

**Methods: Data Analysis**

In reality, the first question we, as readers, should ask is, Are the data worthy of statistical analysis? We must then examine the methods of statistical analysis to determine whether they were appropriate to the source and nature of the data and whether the analysis was correctly performed and interpreted. These questions are difficult to answer. However, we recognize that this is an entire field to itself for which this chapter should stimulate the reader to pursue more vigorous study.

One of the first issues that should cross the reader’s mind is to ask whether the observed and reported finding could result simply from chance, the luck of the draw, or sampling variation. An arsenal of statistical methodology is available ranging from simple (e.g., t-test, chi-square test) to sophisticated (multiple logistic regression, Cox proportional hazards model) to examine the role of chance in the analysis of study results. Each medical reader may not have sufficient expertise to assess whether the investigators have chosen their methodology appropriately and have correctly performed the statistical analyses. Authors should provide rationale and references for innovative or unusual methods. A discussion of loss to follow-up, detection of outliers, item nonresponse, and possible imputation of missing data should be included in the data analysis section. Authors should clearly describe the analytic framework of a research study. Readers should beware when multiple analyses (i.e., “data fishing”) are conducted without appropriate adjustments. We hope that the journal’s peer review process has included some form of assessment of the statistical aspects of the report. Until we, the readers, learn enough, we must solicit expert biostatistical assistance. A biostatistician can evaluate more complex issues in addition to assessing the appropriateness of statistical methods. These include model diagnostics such as fit indices and the results of sensitivity analysis (if performed).

**Description of the Sample**

The CONSORT statement (http://www.consort-statement.org) is an important research tool that has been endorsed by prominent medical journals such as *The Lancet, Annals of Internal Medicine*, and the *Journal of the American Medical Association*. The CONSORT guidelines offer a standard way for researchers to report clinical trials that is appropriate for adaptation by other types of research studies. Authors should provide readers with a clear picture of the progress of all study participants, from the time they are assessed for enrollment until the end of their involvement. Information about reasons for loss to follow-up should be clearly stated. When authors describe the sample, sociodemographic and other descriptive information relevant to the study (e.g., medical history, disease severity and duration) should be clearly reported.

**Presentation of Findings or Results**

Authors must walk the fine line of clear and concise data presentation in the results section without editorializing or drawing conclusions from the data they presented. Remember, the facts should speak for themselves. The author must still detour into enough necessary detail for the reader to judge the importance of the data. Important findings require proper documentation. If a small number of subjects are presented, a table listing the important demographic characteristics is useful so that the reader has a clear understanding of the population studied.

It is surprising how often numerical inconsistencies are contained within reports published in even the most reputable medical journals. This may be partly caused by the many drafts and revisions compounded by textual proofreading, computational and tabular proofreading, and other processes. Because of the frequency of these errors, the reader may wish to use some quick checks: Columns and rows should add up to their indicated totals; percentages of mutually exclusive categories should add up to 100%; numbers in tables and figures should agree with those in the text; and totals in various tables describing the same population should agree. With the ubiquitous presence of hand-held calculators and personal computers, we can even run some of our own statistical tests, especially when the reported results appear incompatible with our quick mental assessment or even personal bias!

Clarity and precision are important criteria to judge the overall scientific validity of an article. Assessments, comparisons, and judgments belong in the discussion section. However, when these are enthusiastically included in the results section, they strongly suggest bias in the author’s approach. Strictly speaking, investigators should undertake an analytic study when they can wholeheartedly support affirmation or rejection of the hypothesis under test. Thus, inclusion of subjective opinions (e.g., “markedly improved outcome”) in the results section may be a subtle indication that the investigators performed the study to confirm their pre-existing personal views.

However, three points should be remembered. First, it is the author’s responsibility to provide the reader with information on the specific statistical analysis used in assessing the role of chance. Second, whatever the level of significance reported, no matter how small the p-value, we can never rule out chance with certainty. An exceedingly small p-value (1 instance in 1,000) denotes that chance is an unlikely explanation of the result, but the possibility remains, although unlikely, that this is indeed that 1 instance in 1,000. The third point is that a statistically significant result is not necessarily important or indicative of a real effect, only that an effect of chance has been ruled out with some reasonable certainty.

As clinicians, we know that measurements of pulmonary artery occlusion pressure (PAOP) differ among observers. For instance, estimation of PAOP from a visual inspection of the oscilloscope tracing may be 3 to 4 mm Hg different from the
example, in an observational study comparing the mortality and the "outcome" (dependent variable) under study. For epidemiologic definition, a confounding variable is one that is referred to effects of one or more related variables. In its strict definition, it may result in a biased result, or there may be some other related variable that is causing the difference. A result may be highly significant statistically but not practically important. Case e is neither statistically nor clinically significant.

**Figure 7.5** Suppose two groups are compared on a numerical measure and the CI for the mean difference between groups is calculated. The threshold that corresponds to a meaningful difference between the groups is indicated by the dashed line. Five results are possible. Confidence intervals in cases a, b, and d capture or fall above the "importance" threshold, and thus are candidates for practical importance or "clinical significance." Cases a and b are statistically significant since they do not contain zero (confidence intervals containing zero indicate no difference between groups). Case c is statistically significant but not practically important. Case e is neither statistically nor clinically significant. (Adapted from Berry G. Statistical significance and confidence intervals. Med J Aust. 1986;144:618–619; reprinted in J Clin Pract. 1986;42:466–468.)

Results calculated electronically and displayed in digital form on the monitor. In reviewing the effects of a drug, however, some investigators may interpret a change of the same magnitude (3–4 mm Hg) as an "effect" of the therapy. Thus, in addition to deciding whether a particular result is "statistically significant," that is, if it represents a real event (or results from chance), we must decide whether it has any real clinical, biologic meaning. Figure 7.5, adapted from Berry (42), illustrates five possible relationships between statistically significant and clinically significant results. Confidence intervals depicted in the figure are intervals that give plausible ranges of the difference between two groups. Confidence intervals in cases a, b, and d capture or fall above the "importance" threshold, and thus are candidates for practical importance or "clinical significance." Cases a and b are statistically significant since they do not contain zero (confidence intervals containing zero indicate no difference between groups). Case c is statistically significant but not practically important. Case e is neither statistically nor clinically significant.

Furthermore, in our interpretation of study results we must, with reasonable certainty, rule out the possibility of confounding and bias. A result may be highly significant statistically but the study design and conduct could lead to a substantially biased result, or there may be some other related variable that also explains statistically significant results. Confounding refers to effects of one or more related variables. In its strict epidemiologic definition, a confounding variable is one that is associated with both the "exposure" (independent variable) and the "outcome" (dependent variable) under study. For example, in an observational study comparing the mortality experience of two modalities of treatment for head injuries, an obvious "confounding" variable would be the severity of the injury. Clearly, the severity of the injury relates to the dependent variable under study: Mortality. The injury severity, however, may also have an association with the independent variable: The choice of the particular modality of treatment. Thus, any finding of a difference in mortality between modalities of treatment, even if statistically significant, might be explained by the confounding effects of the severity of injury. The important point is to judge whether the authors have considered all of the pertinent known confounding variables in their analyses and have taken proper steps to account for their effects. The reader, without substantive knowledge of the particular field of study, may be unable to delineate what pertinent potential confounding variables should have been considered. We (authors and readers) must cautiously proceed with forming conclusions.

**Bias** refers to a systematic departure from the truth. Bias may exist in many forms, and many statistical and epidemiologic adjectives can precede the word "bias" to denote some specific hazard or snag that can lead to a departure from the truth. Sackett (43) provides a useful compendium of the various biases that lurk to ensnare the unwary investigator, and the unwary reader, in the conduct of biomedical research. We shall use the three adjectives: selection, observation, and analysis.

**Selection bias** refers to how subjects were entered into the study. Is the manner of selection of persons for study such that the study will result in substantial distortion of the truth? As a simple example, consider a study comparing outcome of surgery in patients who agree and volunteer to undergo the operation with those who refuse. Those who choose surgery may be better operative risks (at least from their own perception), probably with less comorbid disease than that found in the nonsurgical group. Of course, other factors may have influenced the other group to refuse surgery. Still, the difference in the outcome of surgery might be more likely to result from the selective nature of the groups rather than from any real effect of the surgical procedure.

**Observation bias** refers to the methodology for handling and evaluating subjects during the course of the study. If a therapeutic intervention group receives more attention, more supportive therapy, and more intense scrutiny than a control group, an observed difference in outcome might more likely be explained by observation bias rather than by any real effects of the intervention. Retrospective studies are particularly prone to observation bias.

**Analysis bias** refers to fallacies that exist in the choice of statistical methods to analyze data. An example is the "average age at death" fallacy. Calculation of average age at death among decedents does not measure longevity; it reflects mainly the age composition of the total members of the groups, mostly those who are alive. For example, consider a newspaper report of a study that compared the average age at death of US professional football players with professional baseball players (44). The report stated that football players died, on average, 7 years earlier than baseball players. It would be erroneous to conclude that this differential reflects the more hazardous and traumatic aspects of professional football compared with professional baseball. In fact, professional football is a much newer sport (dating from the mid-1920s) than professional baseball (dating from the 1860s). Consequently, the total group of professional baseball players is considerably older.
than the total group of professional football players. As an extreme example of this average-age-at-death bias, consider the result anticipated in a comparison of the average age at death in a children's hospital with that in a retirement community hospital.

When, in the assessment of a study, we can rule out with reasonable certainty that the finding does not result from chance, bias, or confounding, we are well on the road to determining a real and meaningful effect. Finally, it is important to emphasize that the interpretation of statistical significance does not in and of itself connote medical or biologic importance.

Discussion and Conclusions

In this section, the author provides an interpretation of findings. Here the author can attach clinical relevance to the reported statistically significant findings. The findings may be compared with those of other studies and interpretations. Possible explanations for results can be postulated and differences from other reports in the literature explained. Hopefully, the author bases conclusions on the findings, although this is not always the case. When we discuss results, we should consider whether they have any meaning in the real world of bedside practice. A “significant” but relatively small difference in cardiac performance discovered only in carefully controlled circumstances has little resemblance to the constantly changing status of the critically ill patient in whom such a finding may not have any real import. We must ask ourselves whether the demonstrated result is important in influencing or directing bedside practice. We must retain our skepticism and use it to balance enthusiasm.

This section should contain a discussion of limitations. Authors who conclude that results would have been statistically significant if only a larger sample had been available display their lack of foresight and preparation; clearly, the time to discover the proper sample size is at the outset, the study-planning phase. Rather, it would be refreshing to encounter conclusions that forthrightly admitted that the hypothesis was incorrect, that the study showed that therapy did not lead to improvement, or that the investigator headed off on the wrong track. Negative reports of this sort will prevent other investigators from pursuing ideas that turn out to be flawed and can also direct investigators, including themselves, along more fruitful pathways.

The reporting of negative studies has been addressed from an editorial standpoint (45). Angell states, “...it is widely believed that reports of negative studies are less likely to be published than those of positive studies and some data have been put forth to support this belief. ... It is assumed that editors and reviewers are biased against negative studies, considering them less inherently interesting than positive studies. However, a bias against publishing negative studies would distort the scientific literature” (45). Although she believes that the New England Journal of Medicine publishes fewer negative reports than positive ones, it is not a matter of policy. She asks, “Does it deal with an important question? Is the information new and interesting? Was the study well done? We feel a particular obligation to publish a negative study when it contradicts an earlier study we have published and is of a similar or superior quality. When a good study addresses an important question, the answer is interesting and the work deserves publication whether the result is positive or negative” (45).

Finally, we should consider whether the conclusions are relevant to the questions posed by the investigators. Far too many reports begin with “unwarranted assumptions” in the introduction, end with “foregone conclusions” in the discussion, and contain in between a mass of barely relevant data. If we care to spend the time necessary to review published reports and, in particular, to do the preparation necessary before we embark on our own clinical investigations, such discouraging assessments will occur much less frequently.

LITERATURE REVIEW AND RESEARCH SYNTHESIS

Practicing physicians, researchers, patients, and family members are inundated with unmanageable amounts of information; the need exists for literature to “efficiently integrate valid information and provide a basis for rational decision making” (46). Systematic literature reviews (SLRs) assemble, critically appraise, and synthesize the results of primary investigations addressing a topic of concern (23). Systematic reviews contain a summary of all past research on an area of interest using a methodology incorporating explicit methods in limiting bias (systematic errors) and reducing chance effects, thus providing more reliable results upon which to draw conclusions and make decisions (47). The steps of the SLR guide the researcher through the process of systematically evaluating the existing literature in light of a predetermined research question (48). Meta-analysis is often conducted as part of an SLR.

The Cochrane Collaboration is an international not-for-profit and independent organization, dedicated to making up-to-date, accurate information about the effects of health care readily available worldwide through the methodology of SLRs. It produces and disseminates systematic reviews of health care interventions and promotes the search for evidence in the form of clinical trials and other studies of interventions. More information can be found on their Web site, http://www.cochrane.org. A search of the Cochrane library for VAP gave this title: “Prevention of ventilated associated pneumonia in critically ill patients treated for stress ulcers.”

SUMMARY

In De Anima, Aristotle declared that “it is necessary, while formulating the problems of which in our further advance we are to find the solutions, to call into council the views of those of our predecessors who have declared any opinion on this subject, in order that we may profit by whatever is sound in their suggestions and avoid their errors.” This wisdom still holds today. In order to “council the views” of medical researchers, physicians must develop a competency for reading the medical literature and understanding basic statistical concepts. We humbly accept that a book chapter can only provide an introduction of topics, and we encourage the reader to seek out additional educational opportunities. The National Institutes of Health (NIH) has made the paradigm of interdisciplinary research a cornerstone of the NIH Roadmap (http://nihroadmap.nih.gov). We advocate, whenever and wherever possible, collaboration among physicians, nurses, biostatisticians, data managers, and support professionals to form research groups
Bayesian statistical methods and concepts are elegant and are empirically justified more so than frequentist ones. Bayesian methods can now be employed more widely than in the past because of newly available computational power of computers. However, real-world clinical decisions are, of necessity, frequentist in nature in that they entail a definitive choice of disease process and treatment strategy.

In reading or designing the analytic plan for any study, perhaps the most important thing to be clear about is exactly what the observational unit is. Observational units always have a time dimension and a scope dimension. Examples include; a hospital day, a patient year, a condition episode, a procedural episode.

Within an “observation” there are one or more measurements (outcome/dependent variables) and a complementary set of factors (independent variables) and the relationship between these is what statistical analysis is meant to quantify.

Independent variables fall into two categories depending on the purpose of the study. There is always at least one independent variable “of interest” for any study and quantifying the relationship between it and the outcome (dependent) variable is the primary purpose for doing the study. In studies that are not randomized trials, other independent variables may be used as “controls” to help isolate the effect of the variable of interest on the outcome with “all else equal.”

There are two main statistical “products” of any study. These are parameter estimates and hypothesis tests. Parameter estimates include things like incidence, prevalence, natural history of disease, and treatment effects. Hypothesis tests usually involve the relationship between two variables (outcome and independent variable of interest). By using confidence intervals, almost any hypothesis test can be framed as an “effect size” or “odds ratio,” thus circling back to parameter estimation.

ACKNOWLEDGEMENTS
Joseph M. Civetta, Theodore Colton, Renee Parker-James, and Huang Teng-Yu.

References
4. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome.