The hallmark of an evidence-based practitioner is one who reflects on their clinical decision making and uses research evidence to reduce clinical uncertainty and guide their practice. Understanding how the results of empirical research can be appropriately integrated into clinical practice requires a basic understanding of study design and statistical analysis. This article provides an overview of some of the key concepts related to study design and statistical inference that are important to accurately measure clinical outcomes and to appropriately interpret the results of studies within the context of evidence-based decision making. (KEYWORDS: bias, biostatistics, confounding, sampling error, randomized controlled trial)

Nurses want to do the best for their patients. But how does one know what is best? Evidence-based healthcare offers strategies to help healthcare professionals, including nurses, ensure that their practice is evidence-based for the whole of their careers. At its most fundamental, this means the integration of best research evidence with clinical expertise and patient values. Though not the only factor to take into account, this approach requires that research evidence be given due weight in the decision-making process.

An evidence-based practice requires that nursing practitioners search for the best research evidence to answer clinical questions and to appraise that research evidence for its validity (is it true?), impact (how big is the effect?), and applicability (usefulness in clinical practice). To do this effectively requires an understanding of the fundamental principles of study design and biostatistics, because biostatistics is the language in which the results of studies often are expressed. The risks of developing a disease and its outcomes and the effects of exposure to interventions are at the heart of clinical evaluative studies, and these findings usually are expressed in mathematical terms. To understand the results and the degree to which they should inform practice, one needs some appreciation of statistical language. There are few certainties in research or clinical decision-making, and statistical approaches can help us understand and interpret the uncertainty associated with the results of individual or group of studies. In addition, consideration of study design will help the reader to assess the reliability of reported results.

This article identifies some of the key biostatistics that should be examined when critically appraising a research study and considering how it can be used in practice. Understanding and critically evaluating the

[From Department of Health Studies, University of York, Heslington, York, United Kingdom.]

Reprint requests to Trevor A. Sheldon, MSc, MSc, DSc, Head of Department of Health Studies, University of York, Heslington, York YO10 5DD, UK (e-mail: tas5@york.ac.uk)
statistical approach and reported findings are part of the process that the evidence-based practitioner must use to determine both the quality and usefulness of the research. The focus of this article is not to provide a comprehensive account of statistical theory or to summarize the range of available statistical tests and analytical methods. These topics are reviewed in widely available health and biostatistics texts. The primary goal of this article is to introduce the core ideas of biostatistics and study design as they apply to the evidence-based approach to clinical practice, and to enhance the reader’s abilities in critically reviewing reports published in evidence-based journals. Methods for measurement and expression of health outcomes are considered, followed by a discussion of how to interpret the results of evaluations. Issues of bias and the concept of chance are highlighted as well as possible reasons why an estimate of effectiveness might not be reliable. The importance of study design for attributing cause and effect and differentiating between statistical significance and clinical significance is reviewed.

**Measures of Effect**

When evaluating a treatment effect, the outcomes need to be expressed in terms of some measure of health or welfare. These outcomes are divided for purposes of presentation and analysis into two types: continuous or discrete. When the outcomes of a study are continuous (e.g., temperature, blood pressure, arterial oxygen pressure, or urine output), the researchers usually are interested in the extent to which these values change after exposure to an intervention.

When data are normally distributed (bell shaped symmetrical curve, Figure 1), the mean value is used to summarize the data (i.e., adding all the values together and dividing by the number of observations). When the data are not normally distributed but instead are skewed, (Figure 2) it is more informative to use the median rather than the mean value to summarize the data. The median is the value of the middle observation when all the observations are put in order; 50% of the observations lie above the median and 50% lie below.

Because not everyone in a group responds to an intervention in the same way or to the same degree, it is important to know something about the extent to which the values are spread out (i.e., how much each individual’s values in each group differ from the mean). This can be done by describing the range of values in a group (the minimum and maximum values recorded). Other methods of recording this dispersion include: the interquartile range, which focuses on the range between the 25th and 75th quartiles in which the middle 50% of all the observations lie; or the standard deviation, which is a measure of the average amount each individual value differs from the mean in that group (the lower the standard deviation, the smaller the spread of values).

In contrast, health outcomes, especially in acute and critical care, are often discrete.
They describe whether or not a health event has occurred (e.g., alive or dead, conscious or not conscious, discharged or not). For a group of people, discrete outcomes can be summarized as the percentage or proportion of people who experience an event during the follow-up period. For example, in a study of care for acute stroke, 61 out of 110 stroke patients (55%) on general medical wards were dead within 7 years. The percentage expresses the probability or risk that a person in the group of interest died within 7 years. This summary measure can be extended to take into account not only whether people experienced the event, but also the rate at which they did so. For example, if 20% of the participants in a study are dead after 2 years, the risk of dying by 2 years would be 20% regardless of whether they all died by 1 year or 2 years. The rate or incidence rate, however, measures the pace at which new events occur in a population per unit of time. If all deaths occurred by 1 year, the rate would be 20 per 100 person years (20 deaths for each 100 people followed up for 1 year), whereas if they all died by 2 years, the incidence would be halved or 10 per 100 person years).

Continuous variables often are expressed as discrete variables in evaluative studies, especially if there is a threshold above or below that is clinically meaningful. For example, the Glasgow coma scale is continuous using 15 points. However, cut-off levels can be used to signify differing levels of impairment: 3 to 8 implies severe impairment, 9 to 12 is moderate and 13+ is mild and may be used to determine clinical action such as use of intracranial pressure monitoring. The first approach measures the degree of coma before and after treatment, whereas the second uses a threshold measure to classify whether more or less people were severely comatose at the end of treatment compared to before the treatment.

These measures of disease or health can be used to determine whether an intervention has any effect and the size and direction of that effect. This often involves comparing the outcomes of interest in a treatment group receiving the intervention of interest, and a control group which does not. The approach depends on whether the outcome is a discrete or continuous variable.

**Continuous Measures**

Studies that use continuous outcome measures often compare the mean (or median) values of the outcome or the average change in outcome for the intervention and control groups (e.g., length of stay, length or the probabilities of survival). If the average values (or changes) differ, then it suggests that there are differences in the effect of the treatment and control interventions. For example, a recent study examined patient sedation during mechanical ventilation and compared protocol-directed sedation administered by nurses with traditional nonprotocol sedation in critically ill patients with acute respiratory failure. They reported that the mean duration of mechanical ventilation was reduced from 124 hours to 89 hours and the mean intensive care unit length of stay fell from 7.5 days to 5.7 days, with no change in the mortality rate. In the same field, another trial showed that critically ill patients receiving daily interruption of sedative infusions while intubated and receiving mechanical ventilation spent a median of 2.4 fewer days on mechanical ventilation (4.9 versus 7.3) and 3.5 fewer days in the intensive care unit (6.4 versus 9.9). Studies showing such a potential savings in resource use with no apparent reduction in the quality of care would be important to consider for local implementation if the study setting was seen as sufficiently similar to be relevant.

**Discrete Measures**

The measure of effect when using discrete outcomes compares the risk ratio, or the rate of events in the intervention and control groups. This is illustrated in Table 1. The risk of an event in the intervention group (R̂) is

<table>
<thead>
<tr>
<th><strong>TABLE 1</strong></th>
<th>Events and Nonevents in Intervention/Control Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td><strong>Intervention (Exposure)</strong></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>Yes (i)</strong></td>
</tr>
<tr>
<td>Yes</td>
<td>a</td>
</tr>
<tr>
<td>No</td>
<td>c</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
</tr>
</tbody>
</table>
simply the proportion of people in the intervention group experiencing the event, \( R_i = \frac{a}{a+c} \) and the risk in the control group (\( R_c \)) is the proportion of controls experiencing the event \( R_c = \frac{b}{b+d} \).

The relative risk (RR) or risk ratio is the proportion of patients in the intervention group who experience the outcome divided by the proportion of patients in the control group who experience the outcome.

\[
\text{Risk ratio (RR)} = \frac{R_i}{R_c} = \frac{a/(a+c)}{b/(b+d)}
\]

If the intervention and control groups both experience the same effect, then the risk of the event will be the same in both groups, and the RR will be 1.0. If the risk of death is reduced in the intervention group compared to the control group, then the RR will be less than 1.0. If, however, the intervention is harmful, then the RR will be greater than 1.0. The further away the RR is from 1.0, the greater the strength of the association between the intervention and the outcome.

For a variety of statistical reasons, some studies express the outcome as the odds of the event (\( a/c \)) rather than the risk of the event (\( a/(a+c) \)). The odds ratio (OR) is the odds of the event in the intervention group (\( a/c \)) divided by the odds of the event in the control group (\( b/d \)):

\[
(\text{OR}) = \frac{a/c}{b/d} = \frac{ad}{bc}
\]

As with the RR, an OR of 1.0 means there is no difference between the groups and an OR < 1.0 means that the event is less likely in the intervention group than the control group. When the event being measured is quite rare, the OR and RR are numerically similar.

These concepts are illustrated with reference to the acute stroke unit study referred to earlier,\(^1\) which compared mortality rates in 220 patients, half of whom had treatment in a general medical ward and half in an organized stroke unit. The number of deaths is 41/110 in the stroke unit group and 61/110 in the general medical ward after 7 years of follow-up (Table 2).

The risk of death in the treatment group = 41/110 = 0.37 (37%).

The risk of death in the control group = 61/110 = 0.55 (55%)

### Table 2 - Results of Trondheim Stroke Unit Trial\(^1\)

<table>
<thead>
<tr>
<th>Outcome After 7 Years</th>
<th>Stroke Unit (Treatment)</th>
<th>General Ward (Control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dead</td>
<td>41</td>
<td>61</td>
</tr>
<tr>
<td>Alive</td>
<td>69</td>
<td>49</td>
</tr>
<tr>
<td>Total</td>
<td>110</td>
<td>110</td>
</tr>
</tbody>
</table>

It can be seen that the risk of dying over 7 years is lower in the intervention group compared to the control group. This can be expressed as a risk ratio:

\[
\text{risk of death in treatment group} = \frac{41}{110} = 0.37 = \frac{0.67}{0.55}
\]

Thus, the risk of death for those in the stroke unit was 0.67 or 67% of risk for those on a general medical ward (a 33% reduction in RR). In other words, for every 100 people dying in the control group, only 67 would die in the intervention group.

The OR is given by:

\[
\text{odds of death in treatment group} = \frac{41}{69} = 0.59 = \frac{0.48}{0.55}
\]

Thus, the odds of dying in the treatment group is 48% of the odds in the control group. An understanding of this statistic would prompt nurses managing the care of patients with acute stroke on general medical unit to question patient placement. The fact that the mortality rate of patients treated in general settings could be reduced by a third would prompt an evidence-based professional to join with colleagues to review the organization of stroke care and patient placement within their facilities. Before implementing a change in practice based on a single study, it would be important to consider the results of all trials evaluating stroke units to ensure that the body of reliable research evidence was broadly consistent, and to clearly appreciate those important components of a stroke unit.

**Importance or Potential Impact**

The OR and RR are measures of the strength of association between an intervention and an outcome; an OR of 0.6 shows a stronger effect than an OR of 0.95. However, these
The absolute risk reduction is calculated by subtracting the incidence or risk of the outcome in the intervention group from the incidence in the control group. For example, in the stroke unit study the risk reduction is 55% (risk in control group) – 37% (risk in the treatment group) = 18%. So, 18 deaths in every 100 patients will be avoided. However, if it had turned out that the patients were at a lower risk of death, where 1% die (ie, \( R^c = 0.01 \)) and a treatment reduces the rate of death to 0.67% (ie, \( R^t = 0.0067 \)), then the RR is:

\[
\frac{R^c}{R^t} = \frac{0.0067}{0.01} = 0.67 \text{ (the same as before)}
\]

A RR of 0.67 looks like a large effect, but the absolute difference in risk is:

\[
\frac{R^c}{R^t} = 0.01 - 0.0067 = 0.0033 \text{ or 0.33%}
\]

This is a risk reduction of only 3 events per 1000 people. This sounds a lot less impressive than a statement that it reduces the risk by one third. So, when reading a report of an intervention study, one needs to interpret the RR or OR within the context of how frequently the outcome occurs in the population. In the case of the stroke unit research discussed earlier, the evidence-based practitioner would want the data on the usual incidence of death after stroke to assess whether halving this rate would be clinically meaningful.

Another approach to expressing the effect of an intervention is the number needed to treat (NNT), which conveniently expresses the absolute effect of the intervention. This is simply 1 divided by the absolute risk reduction:

\[
\frac{1}{(R^c-R^t)} = \frac{1}{0.18} = 5.5
\]

The NNT represents the number of patients (in this case around 6) who, on average, need to be treated to prevent one additional event. It is a useful way of expressing clinical effectiveness—the more effective an intervention, the lower the NNT. If however, the death rate were only 1% (risk difference of 0.0033), then the NNT would be 330 and more than 300 would need to be treated to avoid a single death.

**Attributing Cause and Effect—The Importance of Study Design**

The strength of the association between an intervention and a health outcome can be measured. However, an association does not necessarily indicate causation. It may be that the patients in the comparison groups were different to begin with, or were managed differently in other respects, or that the outcomes were assessed differently. It also could be that differences occurred by chance; one of the purposes of statistical analysis is to determine whether this is likely to be the case (see discussion that follows).

The simplest way to evaluate an intervention is to measure some baseline indicator in a group of patients, intervene, and then repeat the measures to see if there has been a change (Figure 3).

The problem with a before-after design is that there are other explanations for changes that might be observed. For example:

- the condition may have improved anyway due to the natural history of the condition or in that group of patients selected because of random changes (regression to the mean);
- participants could have been taking some other treatments (co-intervention);
- the act of caring for the patients may have resulted in an improvement in their perceived or actual health (the placebo effect);
- those caring for the patients may have improved the quality of their care because they knew that the results were being measured (Hawthorne effect).

For these and other reasons, it generally is not possible to attribute change after an intervention to the intervention. The fundamental problem is that it is not known what would have happened with these patients if they had not received the intervention.

The best way to assess the changes in a group of patients receiving a treatment over and above what would have happened any-
way without treatment, is to introduce a comparison group of patients who do not have the treatment. These are sometimes referred to as controls or a control group. There are several ways of doing this. One approach is to compare the outcomes of patients receiving the intervention now with patients in the past who did not have the intervention because it had not been developed, so-called historical controls. The problem with historical comparisons is that one cannot be sure that the comparison is a fair one. The data on past patients may be less comprehensive and accurate than the data on current ones. The patients may have received different forms of care, other healthcare environmental factors are likely to have changed, and the patients themselves may have been different from the current intervention group—for example, in severity of illness.

Ideally, a concurrent control is preferred, that is, a control group that is contemporary with those patients receiving the intervention to be evaluated. There are two main ways of finding a concurrent comparator group: examining records of or following groups of patients who happen to either receive or not receive the treatment being treated at the same time (observational study); or conducting an experiment with patients who are recruited and assigned to either receive or not receive the intervention (experimental study). In both situations the outcomes, or the change in the outcomes, at the end of the follow-up period are compared between the groups and it is this difference that is then taken to be due to the intervention.

Observational Studies

In observational studies, the investigator does not control the intervention, but rather observes the effects of what is occurring under routine conditions. For example, a study might look at the death rates of patients receiving another surgical operation. Two main types of observational studies are used to evaluate treatments: the cohort study and the case-control study.

In a cohort study, subjects are divided on the basis of what treatments they received. The design can either be prospective (following up patients in real time from the time they were identified for the cohort) or retrospective (identifying a cohort of people now who had been treated or exposed in the past and then collecting data about them since they were treated) (Figure 4). A prospective study design focuses on the presence of a risk factor or intervention (exposures) that may increase or decrease the risk of an event in the future. In a case-control study, the disease status (cases and controls) is known already, but we examine the histories of individuals to determine if they had specific risk factors or interventions (exposures) that may account for their disease status.

Observational studies are “natural” in that the researchers do not interfere with what happens, they simply collect and analyze relevant data that can reduce the costs and improve the generalizability of studies. Despite these and other advantages, there are some serious problems with observational studies for evaluating treatments. Most of these stem from the fact that the study is observational:

- because the researcher cannot influence activity, comparisons can only be made where there is “natural” variation. If everyone with a condition receives the same treatment, then there will not be controls with whom to compare outcomes
- patients receiving different treatments may also differ with respect to other sorts of care they receive that may not be well documented
- patients receiving different treatments might not be comparable in terms of the severity of their illness and so the result may be biased
patients may receive different treatments because of factors that are linked to outcomes, and thus confound the results.

Confounding is the spurious association between two variables caused by another variable (the confounder), which is correlated with the other two. For example, a strong association between peoples' reading ability and the size of their feet is found. This finding however, is due to the effect of the confounding variable of age. Older children have bigger feet and also read better than younger children.

A good example of the effects of confounding variables is provided by a study using data derived from the UK Intensive Care Society to examine the association between the volume of patients admitted to intensive care units and subsequent mortality. The death rates of 26 units treating approximately 10,000 adult patients were compared. Results indicated that smaller units had higher death rates, which suggested that smaller units provided poorer quality care. However, it is likely that a small unit only admits the most critical cases, while larger units will admit patients who are slightly less critical and at lower risk of death. Thus, severity of illness is acting as a potential confounder. To improve the validity of the comparison, the authors adjusted the results to take into account severity of illness or case mix using the Acute Physiology and Chronic Health Evaluation scores (APACHE II). After adjustment, there was no difference in the risk-adjusted mortality rates. In many studies, however, there is no information provided on known confounding variables. This is true especially in administrative databases where one depends on data collected routinely, often without information about potential confounding variables.

The evidence-based practitioner would, in the light of this knowledge, be very cautious in interpreting and acting on data suggesting that their unit or facility was performing better or worse than another. This is because the patients they treat may be different. For example, some health systems measure the quality of nursing care by the incidence of pressure sores (decubitus). The problem with this indicator is that some units will have more patients at higher risk of developing a decubitus (eg, older, thinner, less mobile, poorer nutrition, etc.) than others. Although there are a number of risk prediction instruments (eg, the Braden Scale), none of these tools are sufficiently accurate to adjust for variations in risk.

Experimental Studies
Unlike observational studies, the researcher exerts control over the intervention in exper-
Experimental studies. Most importantly, experimental designs allow the investigator to construct two or more comparable groups of patients. A powerful research design that is commonly used in healthcare research is the randomized controlled trial (RCT). In RCT studies, patients are assigned randomly either to receive an intervention or be placed in a control (comparison) group (Figure 5). This can be done by random-number tables or more crudely with toss of the coin. There are more sophisticated ways to randomly assign individuals to groups that can be performed in advance to reduce bias of group assignment.

Advantages of Randomization

When random allocation or assignment is used and there are enough patients in the trial, one can be more confident that the intervention and control groups are comparable—both with respect to factors that are known to affect outcome as well as potentially confounding variables that are unknown. The effect of random allocation can be imagined by considering the effect of tossing a coin several times. Although for the first few tosses there might, by chance, be more heads than tails, after 50 or 100 tosses they will even up, resulting in an equal number of heads and tails. Similarly, random allocation of a large sample of subjects ensures that each participant has the same probability as other participants of receiving the intervention or the control, and will improve the chance that the groups are balanced.

Thus, the main advantage of an RCT research design is that the groups are likely to be comparable in all important respects and therefore any potential confounding variables are balanced. For this reason alone, the RCT is regarded by many to be the gold standard for evaluating the effectiveness of interventions.

However, it is still possible for bias to occur in a large randomized study, from other sources. For example, if patients know that they have been assigned to the intervention rather than the control group, they might feel better simply from knowing this. There may be a differential placebo effect. To avoid this, one tries to ensure patients are not aware of the group to which they have been assigned (blinded allocation). This design is called a single-blind RCT. When the trial is comparing a treatment with no treatment this can be achieved by using a placebo intervention that is not active on that condition but is difficult to distinguish from the active one.

It may also be important to blind the health professionals to which intervention patients receive. For example, the experimental drug and the placebo are supplied in bottles marked A or B. This is done to ensure that the care of patients is the same in all respects other than the treatment allocations (a double-blind RCT). Finally, it is important that the people measuring the patient outcomes are not aware of the group to which patients have been assigned. Ideally then, patients, health professionals, and outcome assessors should all be blinded to knowledge of subject allocation to intervention or control groups. A large triple blind RCT is likely to achieve balance between the groups being compared in terms of patient characteristics, placebo effect, overall healthcare, and measurement of outcome, and therefore produce more accurate estimates of effect.

Validity

Internal validity is the degree to which study results actually reflect the true effect of the intervention or risk factor in the sample of patients studied. Randomized controlled trials often have greater internal validity than

---

Figure 5. Randomized controlled trial.
other designs, because their results are less likely to be biased and less subject to the effects of confounding than are observational studies.

Internal validity can be threatened even in a blinded RCT due to other aspects of design, data collection, or analysis. For example, when comparing two drugs, one can be made to appear more effective if the comparison drug is given at lower than recommended (subtherapeutic) doses, thereby reducing its effectiveness. Alternatively, if the study aims to compare the adverse side effects of drugs, then one can create a disadvantage for the comparison drug by administering it in higher doses to exaggerate its side effects. Another source of bias is in the handling of patients who do not comply with treatment or drop out of the study. These patients generally are not representative of all the people in their treatment group (because they were not randomly selected to be noncompliant or to drop out). It is important, therefore, to analyze the results according to who was originally included in the group (intention to treat) and not what they subsequently did; otherwise, the groups may no longer be balanced and the purpose of randomization defeated. Thus, data for those who drop out should be analyzed in the group to which they were originally allocated.

When examining reports of trials, it is important for the reader to critically appraise the quality of the design, the data collection methods, and the analysis to be able to assess its reliability. For example, one would look for data to demonstrate that the intervention and control groups were similar at baseline, and that patients, healthcare providers, and assessors were blind to the allocation. A variety of research critique checklists are available for appraising RCTs and other study designs.

A weakness of the RCT design is volunteerism, which threatens the external validity of a study’s findings. Patients who volunteer and meet the entry criteria of the trial may not be representative of all the patients for whom one would like to use the intervention. The issue of external validity or generalizability is critical to the clinical significance of the research. The findings of clinical trials may be weakened when the subjects enrolled in the study do not reflect the population for whom results are intended. For example, a study that included only young healthy men in trials of treatments for heart disease would be of limited clinical significance for the actual population afflicted with cardiac disease, who are usually older, less healthy men and women. This has led to the use of pragmatic trials, which are more representative of the sorts of patients found in routine practice and are carried out in a manner that better reflects the real world.

**Sampling Error**

Even if a study has been carried out in a methodologically sound (unbiased) way, a study result such as “5% less deaths occur in people treated with drug A compared to those not receiving the drug” does not necessarily mean that this is a true treatment effect. It could be due to the play of chance. To illustrate, imagine that you are playing a game with dice. On average, each of the six numbers should come up an equal number of times in unbiased dice. However, when your friend throws 2 or even 3 sixes in a row, you would not automatically presume that the dice were loaded (biased) or that your friend is cheating. Instead, you would probably conclude that this was just luck—the play of chance.

This example demonstrates that even if there is no true effect (ie, the dice are not loaded), events that look like there is an effect can be observed, simply because of chance (sampling error). This is particularly the case when there are small numbers of observations. For example, if the number six came up in 2 out of 4 throws (ie, 50% of the time), one would assume it was because of chance. However, if it occurred in 100 out of 200 throws, then one would tend to reject the idea that this was just because of chance and instead conclude that the dice were loaded.

Exactly the same logic can be applied to the results of evaluations of clinical interventions. It is possible that a study result showing benefit or harm for an intervention is due to chance, particularly if the study has a small sample size. Therefore, when the results of a study are analyzed, it is important to assess
the extent to which they are likely to have occurred by chance. If the results are highly unlikely to have occurred by chance, we accept that the findings reflect a real treatment effect.

Statistical theory tells us that if we could repeat an experiment several hundred times on different samples with the same number of patients, the resultant measure of effect (e.g., the mean difference, difference in proportion, or RR) would not always be the same. If the results of the different experiments are plotted on a graph, the shape of the curve (i.e., the distribution of the results) would be approximately normal, or bell shaped (Figure 1). On average, the results of the studies would give us a correct estimate of the true treatment effect. However, any one study result could vary from this “true” effect by chance. The degree of variation from the “true” effect is given by the measure of spread or standard deviation of this distribution, which, because it indicates the amount of random error that is likely, is called the standard error (SE). The bigger the SE, the more individual study results will vary away from the true effect.

Confidence Intervals

Because of sampling error, one cannot be certain which is the true value of the treatment effect (since the observed value may vary from the true value by chance). Statistical theory tells us that 95% of all possible studies produce an estimate that lies 1.96 SE on either side of the true value. This means that in 95% of the repeats of this experiment, the true value of the treatment effect would lie 1.96 SE on either side of the estimate of effect found in a single study (Figure 6). There is a 95% probability that the range—1.96 SE on either side of the study effect size—calculated from the value found in a single study includes the true value. This is called a 95% confidence interval (CI) and is a plausible range within which one can be 95% confident the true value of the population being studied will lie. If one wants to be more confident that the interval includes the true value, it can be made wider. For example, a 99% CI lies 2.5 SE on either side of the estimate from our study. In this case there is only a 1 in 100 chance that the true value falls outside of this range.

The wider the CI, the less precise is our estimate of the treatment effect. This precision depends on the size of the SE. This is a measure of the spread of the sampling distribution, which in turn depends on the sample size: the fewer the number of patients in a trial or number of events observed (e.g., deaths), the greater will be the sampling error. The greater the sampling error, the more likely it is that any one experiment will differ by chance from the true or average value and so the wider will be the 95% CI. On the other hand, if the size of the study was increased so that the distribution becomes less spread out, individual study results will fall much closer to the true or average result and the SE and so the width of the CI is reduced (Table 3). This is the same as saying that one can be more confident in the results of throwing the dice lots of times than when they are thrown only a few times.

Presenting the uncertainty about the size of a treatment effect using CIs is now the method preferred by credible healthcare journals. Confidence intervals also can be used to answer the question of whether or not a treatment has any effect. If the CI of the difference in mean blood pressure reduction or the difference in the proportion of people who die includes the value zero (i.e., no difference in average or proportions), then we are not confident that the treatment is better than the control. Similarly, if the CI of an OR or RR includes the value 1 (same odds or risk in treated and untreated groups), we cannot be confident that there is a difference between the effects of treatment and control.

The importance of CIs can be illustrated with an example from two clinical evaluations:

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Larger Sample Sizes Give More Precise Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size (Each Group)</td>
<td>Estimated Reduction in Blood Pressure</td>
</tr>
<tr>
<td>50</td>
<td>6 mm Hg</td>
</tr>
<tr>
<td>100</td>
<td>6 mm Hg</td>
</tr>
<tr>
<td>200</td>
<td>6 mm Hg</td>
</tr>
<tr>
<td>1000</td>
<td>6 mm Hg</td>
</tr>
</tbody>
</table>
tions of pressure-relieving overlays on the operating tables used in surgery to decrease the incidence of subsequent pressure sores among surgical patients. One trial with 416 patients compared operating tables covered with a viscoelastic polymer pad to a standard table and found a marked decrease in the incidence of postoperative pressure sores associated with using the polymer pad for patients undergoing elective major general, gynecological, or vascular surgery (OR = 0.53). However, the CI raised uncertainty about this finding because the 95% CI was 0.28 to 0.82. Thus, the true value for the OR of pressure sore development among all surgical patients who might be placed on the intervention vs the control operating table could not be precisely known; we are 95% confident that it might be as low as 0.28 or as high as 0.82.

Another trial evaluated an alternating pressure overlay intra- and postoperatively (the Micropulse system) compared to a gel overlay in 198 patients. In this study, 7 of 100 developed a pressure sore in the gel overlay compared with 2 of 98 in the Micropulse system (OR = 0.32). Because this is a smaller trial with fewer pressure sores, the 95% CI is wider: 0.08–1.22. This interval is so wide that it includes the value 1.0 (and above up to 1.22). One cannot be 95% confident from the evidence of this study alone that the Micropulse system is more effective than the gel overlay. Upon critical review of the results, the findings do not justify a change in clinical practice. The evidence-based practitioner would need to look further for additional evaluations of this pressure-reduction technology; or if none can be found, consider conducting their own research to answer this clinical question, or at least continue to monitor the literature until additional evidence can be found.

Type I Error—The Risk of a False-Positive Result

By now it should be clear that, because of the play of chance, it cannot be known with complete certainty what the true treatment effect is, even when there is no study bias. One can however, make probabilistic statements that indicate how likely it is that the true value lies within a range. This means that there is always the possibility of being wrong. One could conclude that a treatment effect exists (because the CI around the OR does not include 1) when in fact there really is no treatment effect (type I error).

Another, related approach, commonly used to assess whether the treatment effect is really different from zero by a certain amount, or is likely to be due to chance, is called hypothesis testing. Here, instead of asking “What is the plausible range within which one can be confident that the true value lies?” the question is re-framed as

**Figure 6.** 95% of all possible results of studies fall 1.96 standard errors either side of the true result. Left: smaller standard error. Right: larger standard error.
“What is the probability that the estimated study treatment effect is different from zero by a certain amount because of chance?” Thus if the probability of the result being due to chance is small (e.g., less than 5% or less than 1%), then conventionally it is accepted that the treatment effect is not due to chance and we say the result is statistically significant at the 5% or 1% level ($P < .05$ or $P < .01$). If, however, the probability that the treatment effect is different from zero by a certain amount is larger, then one cannot confidently exclude chance as a reason for the finding, and it is said to be “not statistically significant.” The $P$ value that is quoted after hypothesis tests is simply the risk of type I error that we accept—the probability that the result is actually due to chance. The smaller this is, the more confident we are that the result is not due to chance and is a true effect. Similarly, the greater the confidence level we use for the interval (e.g., 99% instead of 95%) the wider will be the interval and the smaller the risk of a type I error.

Consider a recent study to assess which position is better for intubated patients who are being mechanically ventilated: supine or semi-recumbent position. Fewer of the 86 intensive care unit patients in the semi-recumbent position (5%) developed pneumonia than patients in a supine position (23%) ($RR = 0.22$; 95% CI 0.6–0.81). This shows that we are 95% confident that the true risk ratio is between 0.6 and 0.81. The hypothesis test is statistically significant ($P = .018$), in other words, there is only a 1.8% chance of having found this result or greater by chance if there was really no true effect. The NNT was 6 (95% CI 4–29). Pneumonia is quite common in mechanically ventilated patients and because the effect of position was so large, even in this relatively small study, the results have sufficient power to show the effect on pneumonia incidence as statistically significant. If instead we compare the mortality rate in these patients, we find that the rate in the semi-recumbent position was 18% compared with 28% in those who were supine, however, this difference was not statistically significant ($P = .3$) and it is unlikely that the study had the power to detect a mortality difference (and in fact the study did not have this as a primary outcome). With respect to these findings of patient position and the incidence of pneumonia in mechanically ventilated patients, nurses should consider their influence on patient positioning and initiate patient care protocols that ensure nursing practice is evidence-based.

There has been considerable debate about the use of simple arbitrary cut-off rules of statistical significance and sensible guidance has recently been published. In addition, this whole approach to considering research results is being challenged by a different school of statistics-Bayesian statistics. This method suggests an approach to understanding how the results from research studies can be considered in the light of prior knowledge, including expert views (valued in professional practice) in order to inform clinical decision making.

**Type II Error—The Risk of a False-Negative Result**

The uncertainty of a false conclusion also works the other way. The CI could include 1 resulting in an inference of no effect, when in fact there is an effect (type II error). Similarly, one could say that the result is not statistically significant when in fact it is real. A type II error (false negative result) results from too much sampling error. The wider the spread of the sampling distribution (because it has a large SE), the wider the CI and so the more likely it is that the line of no effect will be included in the interval. Thus, the potential for a type II error is higher. Type II errors often result from studies with too small a sample size or in those that have too few events and so a large standard error (SE). In these cases, the CIs are so wide that they do not exclude from the range a value indicating no effect. The study is then said to have a low power and will often not be able to detect a treatment effect as statistically significant when there really is one. The more rare the outcome or the smaller the likely difference between the interventions being compared, the larger the study sample needs to be to have sufficient power. For example the trial of the Micropulse system for reducing post-operative pressure sores referred to earlier reported an OR of 0.32 but a very wide CI (0.08–1.22). It is possible that if the study had a larger sample size then it would have had sufficient power to detect a statistically signifi-
cant difference; but as it is, chance cannot be eliminated as a reasonable explanation.

When a study is undertaken, the number of patients should be sufficiently large to allow the CI to be narrow enough to exclude the number indicating no effect (eg, 1 if using ORs) if, indeed, a treatment effect of clinical importance really exists. Researchers should, therefore, carry out and report the results of a power or sample size calculation to ensure that the sample size is large enough to provide a reasonable chance of finding a statistically significant treatment effect if it is really there.

One of the problems in clinical research is the plethora of studies that are too small and have insufficient power. In these cases, one cannot interpret a statistically non-significant result to mean that no treatment effect exists. Because so many studies are underpowered, it is possible that important clinical effects are being missed. The evidence-based practitioner, therefore, should not assume that when a study reports no statistically significant effect of an intervention, or no statistically significant difference between two groups, that there really is no effect. First one should look at the sample size of the study and the width of the CI to see if the study had sufficient power to estimate this effect precisely enough. Several interventions that were later shown to be highly effective, were initially thought to be ineffective because the early studies evaluating them contained too small a sample size to demonstrate a statistically significant effect, and were misinterpreted as "evidence of no effect," rather than "no evidence of effect." One approach for dealing with this is to pool or combine the results of similar studies to get an overall and more precise estimate of treatment effect. This approach is called meta-analysis and is thoroughly discussed elsewhere (see Acton's article in this issue).

Two RCTs have been conducted of the Micropulse alternating pressure relief system for use in the operating suite.8 Considered individually, neither is sufficiently large to indicate whether there is an effect of this system on the risk of developing pressure sores. However, when these are pooled together (in a meta-analysis),13 the combined power is increased and the pooled result is RR = 0.21 (95% CI 0.06–0.7) This range excludes the value of 1 (which would indicate no effect), and so one can be more confident that this system is effective in preventing pressure sores.

Statistical Significance Is Not Clinical Significance

Researchers and readers of research often focus excessively on whether a result is statistically significant (ie, not likely to be due to the play of chance). It is important to bear in mind that just because a test shows a treatment effect to be statistically significant, this does not mean that the result is clinically important. For example, if a study is extremely large (and therefore, has a small standard error), it is easier to find small yet clinically unimportant or non–cost-effective treatment effects that are nonetheless statistically significant.

For example, consider a hypothetical international multi-center randomized controlled trial comparing death rates in 12,000 patients admitted to intensive care units with multi-organ failure receiving a new inotrope, with 12,000 patients receiving usual care. If there was a 1% reduction in mortality in the treatment group (49% deaths versus 50% in the usual care group) this would be statistically significant even with a 99.9999% CI, because of the large sample size. However, it is unlikely that clinical practice would be changed on the basis of such a small reduction in mortality. It becomes the responsibility of the clinically experienced practitioner, using the principles of evidence-based practice, to distinguish between clinically important and trivial effects.

Information Sources for Additional Biostatistics Resources

Readers wanting to pursue these topics in more detail are advised to consult epidemiological texts.13-15 In addition, readers are urged to consult the Web site of the journal Evidence-Based Nursing (http://www.evidencebasednursing.com), and other Web resources in the evidence-based healthcare field (see http://www.york.ac.uk/inst/crd/sites.htm#ebh, http://www.shef.ac.uk/scharr/ir/netting/, and http://www.ebmny.org).
Summary

Research findings are used by evidence-based practitioners to guide their practice and reduce their uncertainty in clinical decision making. However, to understand how to interpret research results, it is important to be able to understand basic statistical concepts and types of study design. This article has described how the presentation of the effect(s) of intervention(s) depends on whether the outcome being measured is a discrete or continuous variable. If discrete, risk ratios and ORs are used to measure the strength of the effect, and the absolute risk reduction or the NNT can be used to measure the impact, or clinical importance, of this effect. For continuous measures, the difference in mean or effect size is used.

Further, interpretation of the effects interventions depends on the research design of the studies from which they are derived. If data are derived from high quality randomized controlled trials where the treatment and control groups are likely to be comparable, or from a cohort study where there is evidence of adequate adjustment for confounding and bias, then the reported effects are likely to be more reliable. However, even unbiased results can be subject to uncertainty from sampling error. Confidence intervals are used to express the degree of this uncertainty. If studies have too small a sample size, the CI may be very wide, which makes the findings less useful for guiding clinical practice. Larger studies, or pooled estimates from a meta-analysis of multiple studies, provide more precise estimates of effect with tighter CIs and are thus more informative in guiding clinical decision making. The emerging paradigm of evidence-based practice presents exciting challenges for advanced practice nurses; understanding how to interpret current research evidence is an essential skill for meeting these challenges.

References