15 Clinical Research in Surgery

Ricardo Pietrobon and Dennis A. Turner

CONTENTS

15.1 Introduction
15.2 Why Evidence-Based Surgery?
15.3 Clinical Outcomes
   15.3.1 Pathology
   15.3.2 Impairment
   15.3.3 Disability
   15.3.4 Patient Satisfaction Theories
   15.3.5 Measurement of Patient Satisfaction
15.4 Selecting Most Appropriate Clinical Trial Designs
   15.4.1 Randomized Controlled Trials
   15.4.2 Outcome Studies
   15.4.3 Outcome Scales
   15.4.4 Secondary Data Analysis
15.5 Fallacy behind Levels of Evidence
   15.5.1 Traditional Concept of Evidence-Based Medicine and Its Roots
   15.5.2 Why Surgery Is Different
   15.5.3 Other Limitations of Surgical Trials
15.6 Conclusions
References

15.1 INTRODUCTION

Clinical trials are the key interfaces of basic science findings, product development, rationalization of existing therapies, and their translation into effective clinical therapy (see Chapter 1). Thus, all translational research depends upon clinical trial data. However, clinical trial design is a specialty field in its own right and few basic scientists, clinician investigators, and clinicians have formal training in this field, particularly in neurosurgery and orthopedics. The performance of clinical trials and
the outcomes of those trials form the most severe bottleneck of new therapy development, particularly in surgical fields and device development, where many alternative pathways for product introduction exist and no specific criteria are available for trial format. Thus, particularly in product development (unlike drug development), the manufacturer may omit clinical trials of efficacy determination altogether to focus only on safety issues, rather than risk a failed clinical trial that could doom FDA approval because alternative pathways to approval exist.

Various motivations should be considered when initiating a clinical trial. As noted in many chapters of this book, promising new therapies emerging from basic science studies can appear ready to apply to initial human feasibility studies (in a true translational approach). An enthusiastic investigator may promote these initial clinical studies for a variety of reasons or a sponsor (such as a corporate entity banking on a marketable product) may initiate human product development.

The motivations in both cases are usually financial and the trial outcome is a key ingredient in financial success. Both parties are under considerable pressure to bias the trial outcome. Clearly clinical trials in such situations are best performed by independent investigators not connected to the enthusiastic investigator or corporate entity in an effort to decrease this bias. In many cases, FDA approval is initially based only on biased clinical data. Only later (after market introduction) are clinical trials initiated by skeptical groups of independent investigators (often with negative outcomes).

Other motivations include improvement in knowledge and data on existing therapy. This was the case with the multiple carotid endarterectomy trials, in which the basis for an existing therapy was challenged by both neurologists (who felt the therapy was dangerous at best and lacked efficacy compared to the natural history) and vascular surgeons (who felt surgical therapy was an unassailable necessary path and unethical to withhold). The uncertainty within the medical community regarding the relative benefits of treatments was sufficient that the level of equipoise favored randomized controlled trials. Since these trials involved an existing (insurance-funded) therapy, the cost of the trials mainly involved obtaining and analyzing data.

Another situation is development of a new surgical therapy, not necessarily based on a product, but an extension of an existing procedure or a new concept. Examples from the 1980s are embryonic allograft and adrenal autograft implantations in the basal ganglia for Parkinson’s disease (see Chapter 8). In such cases, no corporate sponsor exists and no insurance reimbursement may be possible because of the experimental nature (and unknown benefits and risks) of the procedure.

Eventually the NIH-funded clinical trials of embryonic allograft neural transplantation required a sham control (a partial burr hole) because of the enormous bias on the part of patients favoring improvement from such dramatic surgery, particularly if the patients funded most or all of the procedures. Even a sham control involves considerable ethical issues and the issue of the relative worth of partially blinding patients to treatments is still debated today.

Even after FDA approval or a large controlled trial, considerable issues remain in the extrapolation of the results to clinical treatments performed by a wide range of practitioners on a variety of patients, often far beyond the initial disease indications. These issues may often require further open label trials based on a community
setting for acquiring data on uses and risks in the population at large, but often the quality of data obtained from such open label trials is poor and difficult to truly analyze. The community of surgeons has considerable interest in such trials because of political concerns related to the focus of academic centers on randomized clinical trials. However, participation by a wide group of academic and private practice surgeons is difficult, and new trial formats clearly are needed to assess how treatments are really used in practice.

These primary motivations for considering and initiating clinical trials highlight the needs for well-designed clinical trial formats, for innovations at the FDA approval level, and in translational research as a whole. This chapter will review the key concepts involved in the design of a surgical study and the multiple choices faced by a designer. It starts by reviewing the importance of the current concept of “surgical practice based on evidence” in modern surgery. This is followed by an evaluation of some of the multiple clinical outcomes that may be measured in a surgical investigation and the most common epidemiological designs: randomized controlled trials, outcome studies, and population-based studies.

Although they are important for surgical research, designs such as decision analysis, surveys, health economic, and qualitative studies will not be covered. A special attempt will be made to demystify some of the concepts inherited from classical “evidence-based medicine” without further judgment. In particular, the role of randomized controlled trials for evaluating treatment efficacy will be scrutinized — emphasizing differences between surgical and nonsurgical research while stressing the importance of making a surgical study both feasible and generalizable to a “real-world” patient population.

15.2 WHY EVIDENCE-BASED SURGERY?

Evidence-based surgery is a current movement based on the application of scientific method to the whole body of surgical practice, including long-established surgical traditions that may never have been subjected to systematic scrutiny. In scope, it is similar to evidence-based medicine because it pursues “the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients.” The roots of evidence-based medicine and surgery can be found in the work of Professor Archie Cochrane, a British medical researcher whose book *Effectiveness and Efficiency: Random Reflections on Health Services* (1972) and subsequent advocacy increased acceptance of the evidence-based concept. Using scientific techniques from other fields such as meta-reviews of the existing literature, risk–benefit analysis, and randomized controlled trials, evidence-based surgery aims for the ideal that all surgeons should make “conscientious, explicit, and judicious use of current best evidence” in making decisions about patient care.

Practicing evidence-based surgery implies both clinical skill and expertise in retrieving, interpreting, and applying the results of scientific studies. It also involves communicating the risks and benefits of different courses of action to patients. Critics of evidence-based surgery claim that surgeons already follow this procedure; that good evidence is often deficient in many areas; that lack of evidence of benefit and lack of benefit are not the same; and that the more data are pooled and aggregated,
the more difficult it is to compare patients in studies with the patient in the office. Despite its problems, evidence-based surgery does not aim to exclude the individual clinical experiences of surgeons. The intent is simply to enhance their experience with information from scientifically sound studies.

The next sections discuss evidence-based surgery from the perspective of an active clinical researcher attempting to design and conduct a study — not from the perspective of a surgeon trying to understand the literature. After a discussion of trial formats and the clinical outcomes that can be measured, fallacies in this approach will be evaluated in detail.

15.3 CLINICAL OUTCOMES

This section describes the main clinical endpoints that can be measured in a surgical study, including disease (pathology, impairment, and disability) and patient satisfaction. The outcome of a surgical procedure varies considerably, depending on the viewpoint (see Figure 14.1). The same surgical procedure (for example, a lumbar discectomy) may be viewed from the perspective of the patient, the surgeon, or society.

The value of surgery and the outcome may be viewed completely differently by these three involved parties. For example, the surgeon may view the outcome as a complete success if the patient’s leg pain is resolved. The patient may view the same surgery as a complete failure if his or her back pain is not resolved (despite leg pain improvement), even though total resolution was not an expectation of the procedure. Society as a whole may view the surgery as not worth funding if patients do not commonly return to work, even though this expectation may be far different from expectations of the medical community. These differing viewpoints naturally lead on occasion to opposing philosophies as to the worth or overall benefit of a particular medical or surgical approach.

15.3.1 PATHOLOGY

Pathology has been defined as the interruption of normal cellular processes and the simultaneous homeostatic efforts of an organism to regain a normal state.6–8 Active pathology can result from infection, trauma, metabolic imbalance, degenerative disease process, neoplasia, vascular, or other etiologies reflecting the basic mechanisms of disease occurrence. Examples of such processes include cellular disturbances consistent with the onset of disease processes such as spinal osteoarthritis and cerebrovascular accidents. Although surgical research has focused on pathology since the 19th century, largely following the Virchow tradition,9,10 pathology is not linearly associated with the final clinical outcomes noticed by patients and surgeons. For example, the same degree of lumbar osteoarthritis noted on radiographic studies can affect patients’ symptoms in different proportions, from completely asymptomatic to highly symptomatic. It is therefore necessary to consider pathological findings with other outcome measures such as impairment.

Pathological measures are often primary methods (separate from the symptoms associated with the primary disease) for understanding whether a treatment is
working. For example, measurement of the size of a brain tumor on MRI scans can form a primary data source with which to compare various chemotherapy treatment regimens. However, these treatments may or may not improve other types of outcomes or patient survival. In spine studies, radiographic fusion is often used as a surrogate marker for success of a fusion procedure, even though this marker does not appear to correlate with patient outcomes in most other respects. Clearly, the use of pathological measures may be an important basis to decide on treatment efficacy at a basic level, but these measures likely require supplementation with other types of outcomes to decide whether a treatment on the whole is worthwhile at patient level.

15.3.2 IMPAIRMENT

Impairment is a medically evident loss of function or abnormality at the tissue, organ, or body system level. Active pathology may result in some type of impairment, but not all impairments are associated with active pathology (e.g., congenital loss or residual impairments resulting from trauma). Impairments can also occur at the primary site of the underlying pathology (e.g., muscle weakness around an osteoarthritic knee joint), although they may also occur in secondary regions (e.g., cardiopulmonary deconditioning secondary to inactivity). Impairments can usually be objectively specified by an observer such as a physician or surgeon, and are classified in a standard text, the *American Medical Association Guide to Impairment.*

Although impairment is a measure that is closer to the outcome as observed by a surgeon, a patient’s perception about his or her own outcome is not linearly associated with impairment. For example, a limitation in shoulder range of motion secondary to a cerebral vascular accident may greatly affect the life of an active patient and be of little importance to a sedentary elderly patient. It is therefore necessary to extend the concept of outcome to a classification that captures the real impact of a disease on a patient’s life — disability.

15.3.3 DISABILITY

Several schools of thought have defined disability and related concepts. We will focus our discussion on the disablement model developed by Saad Nagi, a sociologist, the *International Classification of Impairments, Disabilities and Handicaps (ICIDH-1),* and its current revision, the *International Classification of Functioning, Disability and Health (ICF).* The three concepts have in common the view that overall disablement represents a series of related concepts that describe the consequences or impact of a health condition such as lumbar arthritis on a patient’s body, his or her activities, and on the wider participation of the patient in society.

In this social perspective, disability may or may not be linked to medical impairment and the degree of disability may vary widely for the same impairment. A common example is a finger amputation, an easily observed medical impairment. It may not constitute a disability for some occupations (manual laborer) but would produce complete disability for others (concert pianist, surgeon).
According to the conceptual framework of disability developed by Nagi,6–8,10 disability is the expression of a physical or a mental limitation in a social context. Nagi specifically views the concept of disability as representing the gap between a person’s capabilities and the demands created by the social and physical environment — a product of the interaction of the individual with the environment. This is a fundamental distinction of critical importance to scholarly discussion and research related to disability phenomena.

Independent of Nagi’s work in the early 1970s, a group in Europe developed the first draft of what later became the World Health Organization’s International Classification of Impairments, Disabilities, and Handicaps. Similar to Nagi’s, this model differentiates a series of related concepts: health conditions, impairments, disabilities, and handicaps designated the ICIDH-1 concepts. We will not review the ICIDH-1 classification except to note that in principle this original system was designed as a model for coding and manipulating data on the consequences of health conditions. This classification system was revised, giving rise to the ICF (Table 15.1 and Table 15.2). The ICF has two sections, each with two complementary components. Part 1 covers functioning and disability including body functions, structures, activities, and participation. Part 2 covers contextual factors — environmental as

### Table 15.1
Function and Disability Sections of ICF

<table>
<thead>
<tr>
<th>Component Constructs</th>
<th>Body Functions and Structures</th>
<th>Activities and Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Changes in physiological function</td>
<td>Capacity: executing tasks in standard environment</td>
</tr>
<tr>
<td></td>
<td>Changes in anatomical function</td>
<td>Performance: executing tasks in current environment</td>
</tr>
<tr>
<td>Positive aspects</td>
<td>Functional and structural integrity</td>
<td>Activity participation</td>
</tr>
<tr>
<td>Negative aspects</td>
<td>Impairment</td>
<td>Activity limitation and participation restriction</td>
</tr>
</tbody>
</table>

### Table 15.2
Contextual Factors of ICF

<table>
<thead>
<tr>
<th>Component Domains</th>
<th>Environmental Factors</th>
<th>Personal Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constructs</td>
<td>Facilitating or hindering impacts of features of physical, social, attitudinal worlds</td>
<td>Internal influences on functioning and disability</td>
</tr>
<tr>
<td>Positive aspects</td>
<td>Facilitators</td>
<td>Impacts of personal attributes</td>
</tr>
<tr>
<td>Negative aspects</td>
<td>Barriers and hindrances</td>
<td>Not applicable</td>
</tr>
</tbody>
</table>

© 2005 by CRC Press LLC
well as personal. Each component consists of various domains and, within each domain, categories that are the units of classification.

Disablement models such as Nagi’s and the ICIDH-1 formulation present the disablement process as more or less a simple linear progression of response to illness and its consequences, thus leading to a static conceptualization of the whole process. This view fails to recognize that disablement is more often a dynamic process that can fluctuate in breadth and severity across the life course; it is anything but static or unidirectional.

More recent disablement formulations and elaborations of earlier models have explicitly acknowledged that the disablement process is far more dynamic. In these newer concepts, a given disablement process may lead to further downward spiraling consequences. Pope and Tarlov\textsuperscript{15} use secondary conditions to describe any type of secondary consequence of a primary disabling condition. Commonly reported secondary conditions include pressure sores, contractures, depression, and urinary tract infections, but it should be understood that they can be pathologies, impairments, functional limitations, or additional disabilities. Longitudinal analytic techniques now exist to incorporate secondary conditions into research models and are beginning to be used in disablement epidemiologic investigations.

15.3.4 Patient Satisfaction Theories

Patient satisfaction is an important outcome measurement since it can influence the delivery of medical care at both the societal (total consumption of health care resources) and individual (patient participation) levels. Because patient satisfaction is a multidimensional concept, it is important to start by understanding its multiple definitions. Patient satisfaction is a complex concept that may incorporate sociodemographic, cognitive, and affective components. Although many theories for patient satisfaction have been proposed, few have been extensively tested and validated in different health care settings. Moreover, few studies have been conducted to explain associations between patient satisfaction and patient characteristics or subsequent patient behaviors. Although theories of patient satisfaction are difficult to categorize in an organized and easily comprehensible fashion, one may group these theories into intrapatient comparisons (disconfirmation theory) and differences between individual patients and health care providers (attribution theory) or other patients (equity theory).

Intrapatient comparison theories explain the satisfaction phenomenon by a match between patient expectations and perceptions of medical care. Differences between what is expected and what is perceived to occur will contribute to patient satisfaction or dissatisfaction. This theory is the dominant model of nonmedical customer satisfaction in which consumers compare their perceptions of a product or service against prior expectations. The resultant size and direction of the disconfirmation results in satisfaction or dissatisfaction.

Equity theories are based on the premise that patient satisfaction relates to whether patients believe they have been fairly treated. Equity occurs when patients compare their balances of inputs (time and money) and outputs (medical care and its results) with those of other patients. Patient satisfaction occurs when people perceive they are treated fairly; it may increase when patients perceive their outcomes as more favorable than those of other patients with the same conditions.

© 2005 by CRC Press LLC
Attribution theories assume that any causes for failed expectations will be examined. Dissatisfaction may occur if a patient and provider assume different reasons for a failure. A related concept is gap analysis in which identification of differences between provider and patient perceptions of services occurs. Addressing potential gaps arising because providers focus primarily on delivery of medical care and patients focus on services used may increase patient satisfaction.

15.3.5 Measurement of Patient Satisfaction

Although patient satisfaction has become a common measurement in clinical settings, proper assessment of a patient’s cognitive evaluation of and affective response to medical care provided is an extremely complex task. The difficulty in measuring patient satisfaction lies in the fact that satisfaction is a multidimensional concept with inputs or determinants that are not yet clearly defined. One of the major criticisms of patient satisfaction research relates to methodologic issues including lack of psychometric standards, reliability, and validity of surveys. Many patient satisfaction survey instruments have not undergone rigorous psychometric construction, which is essential in the evaluation of all complex psychological phenomena.

Many patient satisfaction surveys lack discriminatory values to assess specific aspects of medical care. As an example of these biased measurements, studies have shown that the use of a single global measurement to evaluate patient satisfaction generally results in high (95%) satisfaction ratings. Unfortunately, these single-item questions cannot distinguish satisfaction with overall medical care from satisfaction with specific aspects of care. These rudimentary instruments often cannot accurately measure the multifaceted nature of patient satisfaction and may actually reflect satisfaction with other issues of medical care.

Implementation of a survey instrument may also be associated with potential bias. Possible problems include mode of administration (e.g., telephone, personal interview, mail, and structured versus unstructured interviews), timing of survey, nonresponders, and use of proxies. Although no consensus about an acceptable response rate seems to exist, different methods of administration may produce different response rates. The response rate is important in that missing data from nonresponders may affect the validity of the results. The differences between responders and nonresponders are important because patients who are less satisfied with their medical care may be less likely to respond to satisfaction surveys.

The timing of the survey and use of proxies may also introduce bias. There is a greater likelihood of recall bias with increasing lengths of time between hospital discharge and administration of the survey, which may affect patient responses. Proxies (e.g., family members and friends) may not accurately reflect the views of patients. Patient satisfaction assessments are typically considered as nonparametric data and appropriate statistical analysis should be conducted.
15.4 SELECTING MOST APPROPRIATE CLINICAL TRIAL DESIGNS

Selecting the most appropriate trial design is one of the most important steps in the design of a surgical study. In some situations, a cohort study may not appropriately control for baseline differences between two or more treatment groups, whereas a randomized controlled trial may be unfeasible or may over-select a study population that would no longer be truly representative of the group of patients seen on a daily basis. Therefore, the choice of a study design implies knowledge that is extraneous to a purist and a theoretical view of the field. Other issues include pragmatic factors determined by the experiences of previous researchers attempting to conduct similar research studies, available time and funding, and the general expectations of peers in the field. The goal of the study is also critical, for example, feasibility studies for initial human trials of a device or surgical procedure versus pivotal trials for FDA approval.

15.4.1 RANDOMIZED CONTROLLED TRIALS

**Definition** — A randomized controlled trial (RCT) is a design in which participants are allocated at random to receive one of several clinical interventions. Since treatments are randomly allocated, differences in baseline characteristics across groups tend to be balanced if the groups are large and, therefore, any differences in outcome can be attributed to the intervention.

**Classification** — RCTs can be broadly classified into the following categories according to exposure to the intervention:\(^{16}\)

1. In parallel RCTs, each group of participants is exposed to only one of the study interventions. Most surgical RCTs fall into this category.
2. Cross-over RCTs use a design in which each participant is given all the study interventions in successive periods; the order of the interventions is determined randomly. It is interesting to note that conclusions from parallel studies are produced by comparisons across groups; in cross-sectional RCTs, the comparisons are made from participants. Two issues are crucial for the success of a cross-over trial. First, the condition treated must be chronic and stable. Second, the interventions must be of short duration to avoid contamination of the groups that will follow.
3. Factorial design is performed when two or more interventions are used separately and also in combinations and compared against a placebo. For example, when comparing treatments A and B for a certain condition, four groups would be formed: one to receive only treatment A, one to receive only treatment B, one to receive treatments A and B, and a group that would receive no treatment.

**When to perform RCTs** — Unlike nonsurgical trials, logistical problems largely determine the ideal situation for conducting an RCT. These factors include:
1. A sufficient number of eligible patients should be available. The evaluation should be based on data rather than a surgeon’s opinion about the number of cases because surgeons tend to overestimate true numbers of patients.

2. The intervention should be stable, with the perspective that treatment will continue in a similar fashion until the trial is concluded. Since surgical RCTs tend to last longer than their nonsurgical counterparts, it is important to ensure that the therapy is not a simple fad, and that results will still be applicable when the study is completed. Otherwise, the results will be meaningless if the therapy is outmoded.

3. Participating surgeons should be truly involved in the study and fully believe in the presence of equipoise. Equipoise (state of genuine uncertainty) about the comparative efficacy of two different interventions is an ethical imperative for conducting a study.

**Important points** — If RCT is selected as the design of choice, some important concepts should always be kept in mind:

1. When a placebo is considered, it is important to ensure that no other intervention has been previously shown to be effective. If an efficacious therapy exists, it should be used in place of a placebo to avoid patient suffering.

2. It is important that the randomization be performed appropriately and include true randomization mechanisms such as random number generation. Mechanisms such as selection of treatments based on the first letter of a last name or day of the week are pseudorandom methods and should be avoided.

3. Surgical procedures and postoperative management should be identical across participating surgeons and institutions. Differences in surgical techniques will create clusters that can be impossible to control during comparative analysis across techniques.

4. Outcomes should be assessed by a researcher not involved with the previous stages of the study and, of great importance, they should never be assessed by the treating physician.

5. If an outcome can be interpreted by different individuals in different manners, such as the interpretation of radiographs, researchers should conduct previous inter- and intraobserver agreement studies to ensure that the measurement will not be biased by the opinion of a single evaluator.

### 15.4.2 Outcome Studies

**Definition** — *Outcome study* is a popular name commonly used in the surgical field to describe studies involving prospective cohorts of patients, particularly when standardized scales are used to compare functional outcomes across patients and
treatments. Although the intervention is not randomized, outcome studies are valuable tools for determining associations between treatments and disease outcomes. Of particular importance is that, because of the absence of randomization and its issues of selection bias, outcome studies are fundamental in determining how a certain intervention applies to patients in more natural settings.

Outcome studies usually vary in lengths of follow-up, many constituting collections of all patients attending a certain clinic and thus enrolling all patients until their last clinical visits. As could be expected, durations of follow-up will vary and results can only be appropriately analyzed by time-to-event methods to obtain the chance of a particular outcome for an individual with a particular treatment. One weakness of outcome studies is that they are usually inefficient for measuring rare outcomes where a secondary data analysis of a population-based study would be more appropriate.

**Classification** — Outcome studies can be generally classified as:

1. **Prospective outcome studies** — An investigator chooses or defines a sample of subjects to be studied. This selection is usually based on a specific diagnosis, surgical procedure, or interest. The same set of outcome scales is then applied to all participants before and after the surgical intervention. Finally, the comparison across treatment groups is usually performed by comparing postoperative outcomes as measured by the scales adjusted for baseline scores and other potential confounders of the association between treatment and outcome.

2. **Retrospective outcome studies** — The design of a retrospective outcome study is essentially the same as that of a prospective study. Retrospective studies, however, are usually based on longitudinal cohorts of patients from a given clinic; outcome scales are widely spread across all diagnoses and procedures. Although the outcome measurements are usually somewhat less specific than measurements from prospective designs, retrospective outcome studies have the advantage of providing a much larger population immediately. Those conducting retrospective outcome studies should pay particular attention to important factors such as number of missing observations within measurements (missing values for specific variables) and missing observations within the clinic (patients who may have attended but failed to complete outcome questionnaires).

3. **Multiple outcome studies and external controls** — This design compares several separate outcome groups. It is typically used when two or more groups are treated with different surgical techniques. The most important factor is to ensure that the baseline variables (functional levels, disease classifications, and sociodemographic factors) of the outcome groups are as homogeneous as possible. Having groups who overlap reasonably in the areas mentioned above will ensure that these differences can be controlled during the comparison across the different surgical interventions.
15.4.3 Outcome Scales

Because outcome scales are at the cores of outcome studies, it is important to understand what they are and how they are validated. Validated outcome scales are characterized by their reliability, validity, and responsiveness to clinical change. These properties ensure that the data are collected and interpreted in a systematic and reproducible way, allowing comparisons across different patient populations.

**Reliability** — This is the property that determines whether the instrument measures the outcome of interest in a consistent and reproducible way. Reliability is assessed by measuring internal consistency and temporal stability. Internal consistency requires the items constituting a scale to be highly intercorrelated and measure the same concept or construct. Scales whose items are all highly intercorrelated are considered to be one-dimensional because they measure only a single construct. If a scale measures more than one construct, its items are expected to correlate in clusters, and the scale is multidimensional. Internal consistency is usually expressed by Cronbach’s coefficient. Values above 0.7 are usually considered to express acceptable internal consistency. Score stability over time, on the other hand, refers to the consistency of scores obtained on different occasions by the same individuals.

One of the most common measures of temporal stability is test–retest reliability. For example, a scale demonstrates good test–retest reliability if patients with stable conditions tend to have similar scores over time. A common problem of test–retest reliability is that the assumption of a stable underlying condition often can be supported only if the time between the two evaluations is relatively short and if patients can be assumed to not be responding to items based on a recollection of their previous responses.

**Validity** — This is an indication that the scale primarily measures the construct it is intended to measure instead of another related construct. For example, a scale devised to measure neck pain or dysfunction should not capture dysfunction due to concomitant depression. The commonly reported types of validity are (1) face, (2) content, (3) criterion-related, and (4) construct validity. A scale is considered to have face validity if its content seems to measure what it is supposed to measure. This evaluation is usually performed by the scale designers rather than the target population without any quantitative evaluation, and therefore it can be biased. A scale demonstrates content validity when the items reflect all the significant aspects of the construct to be measured. Again, taking neck dysfunction as an example, work-related disability is only one of the dysfunctions caused by the underlying disease and a scale presenting items exclusively about work dysfunction would capture the entire scenario. Thus, while such a scale may have adequate content validity as a measure of work dysfunction, it would lack content validity as a measure of dysfunction conceptualized more broadly.

Criterion-related validity implies that a scale is able to predict some criterion variable, such as the course of the underlying disease. Criterion validity can be applied to situations where the criterion follows (postdictive validity), precedes (predictive validity), or coincides with (concurrent validity) the measurement in question. Finally, construct validity refers to a scale’s behavior in relation to
other related assessment tools. For example, one can reasonably hypothesize that neck pain would be associated with impaired quality of life. A neck pain scale would therefore be considered to have construct validity if a correlation between the neck pain scale and a valid quality-of-life questionnaire could be documented.

**Responsiveness** — Responsiveness is the ability of an instrument to detect small but important clinical changes such as minimal clinically important differences. This index is the minimal score difference able to detect a “clinically important change,” which is a subjective judgment made by a clinician or a patient independent of available treatment choices. Most experts would probably agree that it is important to define and assess the minimal clinically important differences for individual functional scales.

A crucial point in scale evaluation is that psychometric properties are not intrinsic to a specific instrument, but instead are highly susceptible to change as functions of the populations where they are used and how they are applied. As an example, athletes may perceive functional incapacities at levels of activity sedentary patients may never approach. In sum, it is important to consider how a scale may perform within a specific population of interest.

### 15.4.4 Secondary Data Analysis

**Definition** — Secondary data analysis is conducted by using data that has already been collected, usually with a broad purpose. Because it avoids the hurdles involved with primary data collection, the secondary data analysis approach provides a quick and efficient method of answering research questions.

In cases of research questions requiring very large populations, for example, national surveys, secondary data analysis is the only possibility because the cost of a prospective study would be prohibitive. Secondary data analyses can also serve as excellent resources for obtaining preliminary information on a research question that can later be further investigated through a prospective study with more specific clinical variables. Finally, large databases with variables that relate to latent variables such as disability or quality of life represent excellent resources for the formulation of outcome scales because their large numbers allow for the use of powerful statistical techniques such as item response theory.

**Formulation of research questions** — The formulation of research questions from secondary data can occur in one of two directions. First, researchers with preconceived research questions may look for large repositories of data in search of a database with the variables that may answer their questions. Second, researchers may navigate through a dictionary of available databases while trying to formulate research questions that are of interest to them, the clinical research community, and the public in general. In practice, research questions are usually formulated using a mix of the above-mentioned approaches.

Since the formulation of a research question depends on finding databases that will support the study question in both cases cited above, it is important to locate repositories containing information on multiple databases with easily retrievable information. Such a project was recently conducted by the Center for Excellence in
Surgical Outcomes of Duke University in Durham, NC. Known as QUESTFORM (QUESTion FORMulation), this web application aggregates detailed information on more than 50 different clinical databases. The first section of QUESTFORM contains extensive information about database characteristics, including primary purpose, validity of specific variables, details about data collection methods, year coverage, generalizability, total, and number of patient encounters.

QUESTFORM can be used in two modes. In the first mode, researchers can navigate through the data dictionaries of different databases while they are guided in the formulation of a well-formed epidemiological question. In other words, they are instructed on the selection of outcomes, primary predictor variables, confounders that can potentially distort the association between main effect and outcomes, and inclusion–exclusion criteria. A search tool for International Statistical Classification of Diseases and Related Health Problems (ICD) and Current Procedural Terminology (CPT) codes is provided to identify specific disease and procedure codes. Once the question is fully formulated, researchers can save the question in a graphical format known as a question diagram. Question diagrams can then be reviewed by other members of the research team for project feasibility (clinical epidemiologist), statistical approach (statistician), coding (statistical programmer), and literature review (participating students).

In the second mode, researchers can navigate through previously formulated question diagrams. These previous examples serve as templates that can be modified to generate new research projects. By navigating through previously formulated diagrams, researchers can learn from observing successful designs and also save time while creating new designs that bear structural similarities with previous question diagrams.

15.5 FALLACY BEHIND LEVELS OF EVIDENCE

Much has been said about RCTs for the evaluation of treatment efficacy because RCTs are considered the gold standard against which all other clinical research designs should be compared. In this section, we will argue that despite the several advantages of RCTs if one is to consider internal validity only, RCTs may lack external validity or feasibility in several surgical situations where other designs would be clearly more appropriate.

Because most RCTs are performed in academic medical centers, the number of surgeons involved in surgical trials is often limited and carefully monitored for quality. For example, the number of procedures performed prior to trial initiation and their outcomes and morbidities are often carefully audited to ensure that the surgical procedure is performed with the highest possible quality and consistency. Although this auditing ensures that the trial design covers a single surgical procedure (as far as can be specified), the same surgical procedure may actually be performed in many different ways and with far different outcomes once it is available to a wide range of nonacademic and academic centers after the trial conclusion. This can lead to severe problems with extrapolation or generalization of the trial results to the populations of both patients and surgeons.
15.5.1 **Traditional Concept of Evidence-Based Medicine and Its Roots**

Evidence-based medicine has traditionally claimed that RCTs provide better evidence when compared with other study designs. This is clearly demonstrated by Table 15.3. The clear preference for RCTs can also be distinguished by the designation of the Oxford Centre for Evidence-Based Medicine of nonrandomized studies as the “hurly-burly of real-world clinical care.” But is the choice of a research design something to be made with disregard for the logistics surrounding the study?

15.5.2 **Why Surgery Is Different**

Despite the now classical statements about RCT design superiority, we believe that the choice of research design is multifactorial and that, despite their internal validity, RCTs may in some situations achieve results that are of quality inferior to the quality of their nonrandomized counterparts. Several problems can threaten the validity of an RCT:

1. Restrictive enrollment criteria implemented to enhance internal validity in clinical efficacy trials can have the unintended consequence of excluding cases that would make the study sample truly representative of a real-world population (e.g., those with comorbidities).
2. Evidence obtained solely from RCTs may be misapplied by policy makers, payors and/or practitioners who misunderstand the approach and misinterpret it as prescribing a narrowly formulaic (“cookbook”) approach to healthcare.
3. Many patients are unwilling to be randomized to treatments, particularly when one assignment option involves, in their perspectives, inert or ineffective treatments.
4. When an RCT protocol involves a single sustained treatment, the design may fail to reflect usual practice in which shifts in treatment occur until a desired outcome is achieved and maintained.
5. Research documentation and reporting of critical phenomena, such as treatment delivery (fidelity), therapy process measures, and population reach are infrequent.
6. Relevant outcomes including functional status, quality of life, durability of change, potential negative or iatrogenic outcomes, cost of treatment, and client satisfaction may have been neglected in detriment of harder endpoints.
7. Overemphasis on treating or fixing presumably homogeneous disorders may detract from potentially more valuable efforts to understand what caused the problem originally, what contingencies now maintain it, how treatment influences biopsychosocial processes to produce desirable behavioral change, and what changes are needed to address more complex, comorbid problems.
<table>
<thead>
<tr>
<th>Level</th>
<th>Therapy/Prevention, Etiology/Harm</th>
<th>Prognosis</th>
<th>Diagnosis</th>
<th>Differential Diagnosis/Symptom Prevalence Study</th>
<th>Economic and Decision Analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>SR (with homogeneity) of RCTs</td>
<td>SR (with homogeneity) of inception cohort studies; CDR validated in different populations</td>
<td>SR (with homogeneity) of Level 1 diagnostic studies; CDR with 1b studies from different clinical centers</td>
<td>SR (with homogeneity) of prospective cohort studies</td>
<td>SR (with homogeneity) of Level 1 economic studies</td>
</tr>
<tr>
<td>1b</td>
<td>Individual RCT with narrow confidence interval</td>
<td>Individual inception cohort study with &gt;80% follow-up; CDR validated in single population</td>
<td>Validating cohort study with good reference standards; CDR tested within one clinical center</td>
<td>Prospective cohort study with good follow-up</td>
<td>Analysis based on clinically sensible costs or alternatives; systematic reviews of evidence; includes multiway sensitivity analyses</td>
</tr>
<tr>
<td>1c</td>
<td>All or none</td>
<td>All or none case-series</td>
<td>Absolute SpPins and SnNouts</td>
<td>All-or-none case series</td>
<td>Absolute better-value or worse-value analyses</td>
</tr>
<tr>
<td>2a</td>
<td>SR (with homogeneity) of cohort studies</td>
<td>SR (with homogeneity) of retrospective cohort studies or untreated control groups in RCTs</td>
<td>SR (with homogeneity) of Level &gt;2 diagnostic studies</td>
<td>SR (with homogeneity) of 2b and better studies</td>
<td>SR (with homogeneity) of Level &gt;2 economic studies</td>
</tr>
<tr>
<td>2b</td>
<td>Individual cohort study (including low quality RCT; &lt;80% follow-up)</td>
<td>Retrospective cohort study or follow-up of untreated control patients in RCT; derivation of CDR or validated on split sample only</td>
<td>Exploratory cohort study with good reference standards; CDR after derivation of validated only on split sample or databases</td>
<td>Retrospective cohort study or poor follow-up</td>
<td>Analysis based on clinically sensible costs or alternatives; limited reviews of evidence or single studies; includes multiway sensitivity analyses</td>
</tr>
<tr>
<td>Level</td>
<td>Type of Study</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2c</td>
<td>Outcomes research; ecological studies</td>
<td>Outcomes research</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3a</td>
<td>SR (with homogeneity) of case-control studies</td>
<td>SR (with homogeneity) of 3b and better studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3b</td>
<td>Individual case control study</td>
<td>Nonconsecutive study or lacking consistently applied reference standards</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3b</td>
<td>Individual case control study</td>
<td>Nonconsecutive cohort study or very limited population</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Case series (and poor quality cohort and case control studies)</td>
<td>Case series or superseded reference standards</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Expert opinion without explicit critical appraisal or based on physiology, bench research, or first principles</td>
<td>Expert opinion without explicit critical appraisal or based on physiology, bench research, or first principles</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Definitions:**
- RCTs = randomized clinical trials
- SR = systemic reviews
- CDR = clinical decision rule
- SpPins = high specificity so a positive result rules in diagnosis
- SnNouts = high sensitivity so a negative result rules out diagnosis

**Source:** [www.cebm.net/levels_of_evidence.asp](http://www.cebm.net/levels_of_evidence.asp)
8. Efforts to standardize treatments potentially support progression toward a restriction of treatment that will enable therapy to be delivered by paraprofessionals or by computers.

These points do not indicate RCTs have no place in surgical research, but mean that their advantages must be contrasted against their weaknesses. In other words, blindly applying the principles primarily developed for nonsurgical studies to surgical studies is, at least, a mistake.

15.5.3 Other Limitations of Surgical Trials

Particularly in small disciplines such as neurosurgery, the number of patients spread throughout the world may be insufficient for an adequate randomized trial of a treatment despite considerable interest in the outcome. Small patient populations can thus pose severe constraints particularly if they are widespread and hard to capture. The fragmented health care system in the United States likewise hampers patient access. One proposed solution is to limit surgical procedure reimbursement to patients enrolled in clinical trials. That would encourage both surgeons and patients to improve enrollment and enhance the number of trials. It is difficult to standardize surgical procedures to the point where the same technique is performed by many surgeons and studied at many locales, in contrast to studies of drug formulations and devices that can be standardized.

Because many device companies would rather avoid the expenses and the uncertain outcomes of randomized trials, alternate methods of FDA approval exist, but they severely limit the amount of data concerning device efficacy at the time of market introduction. Often fewer than 100 patients are studied adequately at one center. After market introduction, it is highly unlikely that a device company would sponsor an additional critical trial by skeptics, since negative trial results would lead to decreased revenues. As a result, companies have little incentive to perform trials of devices.

Since surgeons’ incomes depend on procedures, they have little incentive to compare surgical treatments to nonsurgical treatments and such comparisons are nearly forbidden in surgical circles. Rather, most surgical trials compare one form of device or procedure to another, rather than to alternative treatments. Development of a community consensus among surgeons about equipoise and when the time is correct to initiate a study is also very difficult. In many cases, surgeons almost have to be forced into studying surgical procedures by outside influences and circumstances.

The funding of studies is also problematic. The National Institutes of Health (NIH) fund only a small number of surgical studies and even pilot studies are difficult to initiate without initial funding. NIH’s determinations of what should be studied may also differ considerably from the topics surgeons would suggest, reflecting the discrepancy between society’s needs and those perceived by medical practitioners. All these factors combine to create a difficult environment in which many small, retrospective studies are performed, many without any lasting merit or contribution. However, taking the next step toward a prospective trial is almost prohibitive in terms of the enthusiasm needed to obtain patient and surgeon enrollment, funding, and consensus within the surgical community.
15.6 CONCLUSIONS

The motivation for initiation of clinical trials varies, depending on the goal of the study and who will benefit from it. Surgical clinical trials are very different from those usually designed for medical trials, including FDA approval studies and rationalization of existing therapies. Small populations, particularly in neurosurgery, fractionated health care systems, and lack of understanding of clinical trial formats all contribute to difficulty in initiating clinical trials of substantial, lasting benefit. While all neurosurgery practitioners desire valid information about the treatments they suggest to patients, the path to that information is highly convoluted and limited. Nevertheless, all translational therapy depends on clinical trial format, which in some cases (such as stroke trials), can be the limiting feature of new therapy introduction.

REFERENCES