CME Clinical Research Methodology 1: Study Designs and Methodologic Sources of Error

Daniel I. Sessler, MD,* and Peter B. Imrey, PhD†

Clinical research can be categorized by the timing of data collection: retrospective or prospective. Clinical research also can be categorized by study design. In case-control studies, investigators compare previous exposures (including genetic and other personal factors, environmental influences, and medical treatments) among groups distinguished by later disease status (broadly defined to include the development of disease or response to treatment). In cohort studies, investigators compare subsequent incidences of disease among groups distinguished by one or more exposures. Comparative clinical trials are prospective cohort studies that compare treatments assigned to patients by the researchers. Most errors in clinical research findings arise from 5 largely distinguishable classes of methodologic problems: selection bias, confounding, measurement bias, reverse causation, and excessive chance variation. (Anesth Analg 2015;121:1034–42)

For most of medical history, practice has been largely ad hoc, based on suboptimal evidence from personal experience and anecdotes shared among physicians. Although individual cases and experiences often yield invaluable insights to perceptive physicians, multiple anecdotes do not replace systematically collected data and organized comparisons, and practice mostly based on personal experience often has subsequently been shown to be suboptimal. (The pleural of anecdote is not data!) Voltaire nicely summarized early medical practice in 1760, writing that physicians "poured drugs of which they knew little, for diseases of which they knew less, into humans of which they knew nothing." One hundred years after Voltaire's comment but considerably before the microbial origin of disease was accepted, the British anesthesiologist John Snow, informed by work of John Graunt and formal vital statistics systems established in mid-16th century London,[1] used systematic, dispassionate data collection to infer that a contaminated water supply was the proximal cause of mid-18th century London cholera outbreaks.[2] During the next century, frameworks for systematic collection and evaluation of data, such as Koch's postulates for assessing microbial causation[3] and Hill's guidelines for inferring causality more generally,[4] fostered greater scientific rigor in conjunction with clinical observation.

Clinical research methods have since matured immensely. The meticulous work of Feinstein,[5–7] Sackett,[8] and others on the intellectual foundations of clinical epidemiology; the fostering of randomized clinical trials (RCTs) particularly by Hill,[9,10] Chalmers,[11,12] and Cochrane[13]; and the emergence of evidence-based medicine, midwifed by Sackett et al.,[14] have contributed greatly to revolutionary advances in clinical understanding and practice. The application of formal meta-analytic techniques from education and social science to medicine and epidemiology, through evidence-based medicine centers and the Cochrane Collaboration,[15,16] has systematized aggregation of research evidence for the medical community.

We thus now know much about how humans work and how they respond to disease and drugs. And we know even more about how cells and rodents respond. Physicians are thus encouraged to practice evidence-based medicine, which means that clinical decisions should be based on good evidence, preferably from relevant, high-quality, and reproducible studies in humans rather than on the physician's personal clinical experience alone.

However, any physician who tries restricting practice to methods established on the basis of strong evidence in humans quickly discovers that there is distressingly little basis for current medical care, although cardiology and oncology, with histories of strong National Institutes of Health funding and methodologic innovation, seem ahead of most specialties in this regard.[17,18] The purpose of clinical research is to bridge the remaining wide gap from the understanding of basic and animal science to the care of patients in an effort to improve medical outcomes. Systematic clinical research is necessary because humans have proven to be a poor model for rodents!

Even the best clinical studies have potential limitations, and it is helpful to understand the strengths and weaknesses of various approaches. By this, we do not mean that clinical research is generally problematic. However, it helps to understand how study design can influence results, and which types of studies are most reliable and thus the best basis for clinical decisions. In this and 2 subsequent articles, we will discuss aspects of clinical research methodology as a guide to understanding and interpreting reported results.

Even a series of articles can only have limited scope. Our focus will be on design choices in clinical research and the advantages and disadvantages of various approaches

From the Departments of *Outcomes Research and †Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio.

Accepted for publication April 7, 2015.

Funding: Internal departmental funds.

The authors declare no conflicts of interest.

Address correspondence to Daniel I. Sessler, MD, Department of Outcomes Research, Cleveland Clinic, 9500 Euclid Ave./P77, Cleveland, OH 44195. Address e-mail to DS@OR.org.

DOI: 10.1213/ANE.0000000000000815

to address clinical research questions. We focus on design because, although error can occur in design, data collection, and/or analysis and reporting of a study, poor study design generally cannot be remedied by subsequent steps.

A consequence of our focus on design is that we will not discuss operational issues, including ways to maintain blinding, electronic data acquisition, how to minimize and deal with missing data, or strategies for the prevention of fraud.[19] Nor will we discuss result reporting, which also is subject to various types of error and potential influence of competing interests. Furthermore, the 3 articles in this series will include only the most basic statistical approaches.

RESEARCH APPROACHES

Clinical research studies can be broadly categorized as retrospective or prospective. Retrospective studies use existing data on current and past patients to answer questions. These studies are conducted by assembling and organizing contemporaneously recorded information on past events, analyzing previously stored biosamples, and/or by returning to patients and physicians for further information about the past. Prospective studies answer questions by collecting new data on current and future patients over a future period during which medically relevant events occur, generally using methods specific to the intended research.

Prospective studies can be either observational (noninterventional) or experimental (interventional), in the sense of manipulating study-related treatments. For example, an observational study might involve determining the concentration of a blood biomarker and evaluating relevant outcomes. Conversely, experimental interventions might include the use of a novel anesthesia regimen, intraoperative monitoring device, or temperature management protocol. Clinical studies also can be characterized by timing. In cross-sectional studies, for example, exposure and outcome are evaluated simultaneously. Cross-sectional studies thus essentially survey the state of affairs at a particular time without looking forward or backward. For example, is hypertension more common among current cigarette smokers than among nonsmokers? Or, is reflux less common among patients who take antacids?

Cross-sectional studies, although useful for certain questions such as evaluating the prevalence of disease, are poor at capturing changes over time and hence provide little useful information for distinguishing causal from other relationships. Furthermore, cross-sectional studies may exclude important groups, such as people who die quickly, and thus are no longer represented in the population when the study is conducted. Consequently, cross-sectional studies are used only rarely in anesthesia research or in studies of treatment more generally, and we will not consider them further in this series.

In case-control studies, investigators ask whether people with a particular disease, or whose disease has progressed, had different previous exposures than people who remained free of the disease or whose disease remained stable or improved. For example, are surgical patients who develop anaphylactic reactions more likely to have been exposed to latex, or are patients who experience intraoperative malignant hyperthermia crises more likely to have a history of heat stroke?

In cohort studies, investigators look forward in time from exposure to outcome, by comparing frequencies and severities of outcomes in groups defined by different exposures. For example, do people with low vitamin D serum concentrations more often develop serious postoperative complications? The distinction can be nonobvious. For example, consider a study in which investigators compared anesthetic requirement in patients with and without fibromyalgia. The design might appear to be case-control because patients with and without disease are being selected. But, in fact, fibromyalgia is the exposure in this study and anesthetic requirement is the outcome. It is thus a cohort design.

An advantage of prospective studies is that they allow researchers to plan and manage data collection, which usually improves data quality. Prospective studies also allow investigators to answer specific research questions more directly than is usually possible in retrospective research. The trade-off for these benefits is higher research costs and longer waiting times for answers.

Comparative studies on treatments can be observational, when the researcher simply observes and describes current clinical practice. However, experimental studies generally are less prone to spurious findings than observational studies, and prospective cohort studies become experiments when researchers manage, rather than simply observe, the choice and use of treatments. When treatments are allocated randomly (i.e., "randomized") to patients, the result is a special type of cohort study called a randomized clinical trial (RCT).

Randomization commonly is used, often in conjunction with concealment of treatment assignments from study participants and some investigators, a process known as "masking" or "blinding" allocation. RCTs that are blinded to the extent practical provide the clearest evidence of a therapy's effects but are not always practical, and there are other legitimate approaches to error mitigation, including alternating intervention designs.[20] Two subsequent articles will discuss observational and randomized blinded studies in more detail.

SOURCES OF ERROR

There is no perfect study. All are limited by practical and ethical considerations, and it is impossible to control all sources of error—even in fully randomized and meticulously blinded trials. Multiple studies are thus usually required to convincingly confirm (or disprove) a hypothesis. Most errors in clinical research arise from 5 major sources of methodologic problems: selection bias, measurement bias, reverse causation, excessive random (chance) variation, and confounding (Fig. 1). Within each of these general classes are many specific types of error such as recall bias, attrition bias, confounding by indication, and so on.[21] Imperfect execution also hinders clinical research, as it does all human activities, but does not constitute methodologic error, and is hence beyond our scope here.

Selection bias, measurement bias, and reverse causation are systematic sources of error. They stem from intrinsic aspects of a study's design and would be expected to occur similarly in multiple repetitions of similarly designed studies. In contrast, chance or random error describes the net effect of idiosyncratic influences, human variation, and

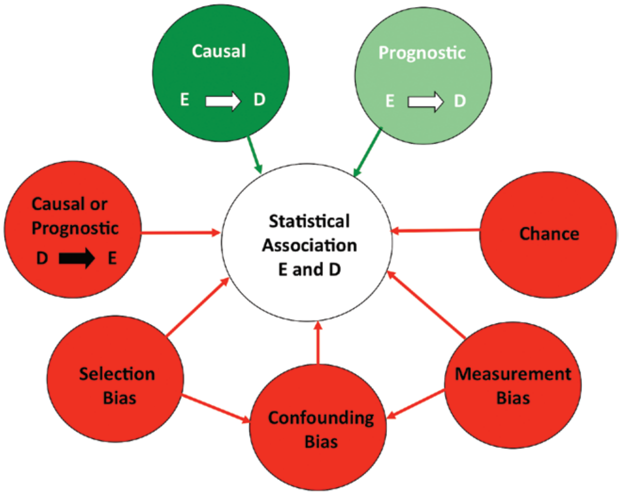

Figure 1. Sources of error. Investigators usually want to confirm a causal relationship between an exposure (such as obesity) and an outcome (such as postoperative respiratory complications); however, that relationship is only 1 of 7 possible causes of a statistically significant outcome; 5 of which are generally misleading—but often for subtle reasons. E = exposure; D = disease/outcome.

trend-free time-to-time fluctuations in measurements and measurement processes. Random errors because of such chance variation would not, therefore, be expected to similarly recur were the study to be repeated.

Confounding is a phenomenon—essentially, mistakenly attributing the influence of one exposure to another—that can be inherent in the medical situation being addressed or can arise as a consequence of selection bias, measurement bias, or chance. For example, pretzel consumption is associated with hepatic cirrhosis. However, pretzels do not cause cirrhosis; instead, it is beer consumed with pretzels that damages the liver. Statistical planning and analysis can be effective in controlling and assessing the effects of chance and in removing distortions because of anticipatable confounding. The other sources of error—selection bias, measurement bias, reverse causation, and possible confounding by unanticipated or even unknown factors—are most effectively addressed by strong study design.

Chance

Chance error refers to the fact that the results of any given study will differ somewhat from the true biological situation because of random variation. Such variability is known to occur at multiple levels in medical research: from one patient to another, from measurements of the same patient from one time to another, and from one measurement to another of the same patient at the same time. For example, results may differ when portions of the same serum sample are analyzed by multiple autoanalyzers in the same or different laboratories. Similarly, interpretations of images and biopsy samples by multiple radiologists and pathologists may differ, as might results from the same radiologists and pathologists at different times. Such errors are more common and of greater magnitude than generally assumed.[22–24] Undetected data recording error, from misplaced decimals, reversed digits, and transcriptions from incorrect fields on clinical report forms and spreadsheets are an omnipresent contributor to random variation in study results.

Another source of random error can be artifact from inaccurate or missing information because of random malfunctions or inappropriate settings of an automated data collection instrument: for example, when a blood pressure cuff is disconnected or is positioned incorrectly during surgery. Similarly, artifact can occur in nonautomated data collection because of random misunderstanding of definitions by database coders. A type of measurement error, artifacts are more common than generally appreciated and have considerable potential to degrade analyses. We will discuss them later in this review in connection with measurement bias.

In large, well-designed studies, the net effects of chance on estimated treatment effects and other primary study conclusions usually are limited. Sometimes, however, just by bad luck, and especially in studies with few patients and/or imprecise measurement methods, the influence of chance can be substantial and lead to an incorrect conclusion (i.e., no benefit or even harm, rather than benefit). The trouble is that without replication of a study or another form of confirmation, no one can know if its results faithfully represent the "true" biological situation or if chance (bad luck) led to a seriously mistaken inference.

Before starting a study, investigators develop a scientific hypothesis: a specific statement of the biological mechanism or clinical theory for which evidence will be collected. For example: "delirium is reduced by propofol vs. postoperative sevoflurane anesthesia." From this base, 2 "statistical hypotheses" are generated against which evidence collected by the study will be weighed.

The first, conventionally called the "null hypothesis H_0," is that only chance governs variation in patient responses and thus that some systematic relationship implied by the scientific hypothesis does not exist: for example, "H_0: Delirium is equally likely after propofol and sevoflurane anesthesia." The second is a statement about the sort of relationship anticipated if the scientific hypothesis were true, proffered as a logical alternative to chance variation alone as reflected in H_0, and hence called the "alternative hypothesis" H_A. Here, for instance, the alternative hypothesis would be "H_A: Delirium is less common after propofol than sevoflurane anesthesia." This is a predicted consequence of the initial scientific conjecture in large populations of similar patients receiving the 2 types of anesthesia for similar surgeries, and can be reasonably assessed by observing samples of patients.

A more general alternative hypothesis, appropriate to a less specific theory, would be H'_A: the delirium incidences after propofol and sevoflurane anesthesia differ. This is a "2-sided" alternative because it counts departures from the null hypothesis H_0 favoring either anesthetic as supportive, in contrast to the "1-sided" alternative specifying which best reduces delirium. (In general, investigators should a priori designate the most specific hypothesis consistent with their biological understanding, including expected directionality.) Researchers attempt to support alternative statistical hypotheses H_A, and consequently the scientific hypotheses generating them, by obtaining enough data consistent with H_A to falsify the corresponding null hypotheses

H_o. This is effectively done if the data are shown to be of a very unusual sort if only chance were involved.

Let us say that in the resulting study sample, delirium is actually less common after propofol than after sevoflurane anesthesia. The question then is whether the observed reduction is real (a true reflection of the biology) or whether it resulted from chance and/or from bias or confounding. The influence of bias and confounding is hard to evaluate and can be substantial; however, random error (chance) can be managed by the use of statistical tools, the 2 most common of which are *P* values and confidence intervals.

The *P* value is an index between 0 and 1 of how easily the data can be accounted for by pure chance variation. Specifically, *P* values reflect whether the observed data are compatible with what might be expected within the range of chance variation for a study of similar size when the null hypothesis is true. In the most common circumstances, when the null hypothesis represents lack of difference or equality of treatment effects, the *P* value is an index of compatibility of the data with biological inactivity or with identical average effects of competing treatments.

Thus, small *P* values are interpreted as representing data essentially incompatible with random chance, and hence "statistically significant" in supporting the hypothesized treatment effect by falsifying H_0. A *P* value <0.05 is conventionally considered statistically significant, with smaller values reflecting collections of data less and less compatible with chance, and hence for which chance is less and less plausible as a sole explanation. The conventional *P* value threshold of 0.05 is essentially arbitrary, and there are certainly situations (i.e., biologically implausible associations) in which it is reasonable to require smaller *P* values. Similarly, differences that are not statistically significant may well be clinically important.

We note though that a *P* value neither describes the actual magnitude of the clinical effect nor precludes the true effect differing considerably from that observed in a given study. The reason is that observed results are a combination of chance error superimposed on treatment benefit. Confidence intervals are thus used to describe the range of plausible treatment effects. For instance, a 95% confidence interval of 5% to 15% for the difference in the fractions of surgical patients who develop postoperative delirium after propofol or sevoflurane anesthesia is interpreted as meaning that the study data are compatible with a reduction in delirium risk anywhere between 5% and 15%.

Formally, a confidence interval is a range of estimates of an unknown numerical characteristic of a population from which the study sample is obtained, such as delirium incidences in patients given propofol or sevoflurane in the population of interest rather than the study sample. This range is determined from the study data using a method that provides a specified chance, called the "confidence coefficient," that the range will include the parameter's true value. Other things being equal, high confidence requires wide intervals; narrow intervals can be obtained by accepting a greater chance of missing the target and thus lower confidence. Assuming technical assumptions are correct, approximately 95% of 95% confidence intervals can be expected to include the corresponding true targeted values. Investigators who state, based on their data, that the true reduction in delirium risk with propofol is between 5% and 15% can thus be 95% confident that the statement is true, assuming chance is the only source of error. The use of 95% in defining confidence intervals is arbitrary but has become a widely accepted medical research convention, largely because of a useful connection between confidence intervals and hypothesis tests.

It is possible to test statistical effects, such as a difference in mean responses or a ratio of the fractions responding, by rejecting the hypothesized value if the confidence interval for the statistical effect excludes that value. For instance, we might hypothesize that the fraction of patients experiencing delirium after propofol anesthesia is half of the fraction experiencing it after sevoflurane anesthesia. That hypothesis would be rejected if the 95% confidence interval for the ratio of propofol to sevoflurane delirium risks generated by the data in our study excludes the value 0.5.

Hypothesis testing defined this way will have false-positive probability (i.e., type 1 error, symbolically α) equal to the amount by which the interval's confidence coefficient falls short of certainty, that is, short of 100%. Thus, a 95% confidence interval extends the result of a conventional $\alpha = 5\%$ level test of a statistical null hypothesis, by summarizing the results of similar tests of every possible other hypothesis: those with hypothesized values outside the interval are rejected, whereas those with hypothesized values in the interval are retained; this is the sense in which the latter are termed compatible with data (Fig. 2).

An additional, and sometimes serious, source of chance error results when investigators either informally or formally test various hypotheses, thereafter choosing one that is "significant" or consistent with their biases. This process, colloquially known as "data mining," is much more prone to false-positive error than is conveyed by the conventional $\alpha = 5\%$ associated with each individual test. This underlying "multiple testing" issue occurs in various guises. For example, the relationships of a disease to many possible risk factors (e.g., foods, occupational exposures, gene markers) may be simultaneously evaluated. Similarly, many different outcomes may be evaluated or a single outcome may be assessed at multiple time points. And finally, accumulating results of a study may be assessed periodically.

The problem is that if each of a number of tests has chance α of producing a false-positive result when no effect is present, then the chance of at least some false positives when no effect is present increases with the number of tests performed to the point where false positives become virtually certain. The trouble is that false positives cannot readily be distinguished from true positives. Several strategies are available to address this problem. Perhaps the most important is a priori designation of a single primary outcome or of multiple outcomes with appropriate statistical compensation to preserve total false-positive error at a specified level α. Similarly, the specific statistical approach should be designated a priori.

To assure a priori designation of these important design elements, most journals will only consider manuscripts describing clinical trials if the trials were publically registered before trial enrollment starts. Various study registries are available, but perhaps the most commonly used is Clinical Trials, which can be accessed at ClinicalTrials.gov. Despite its

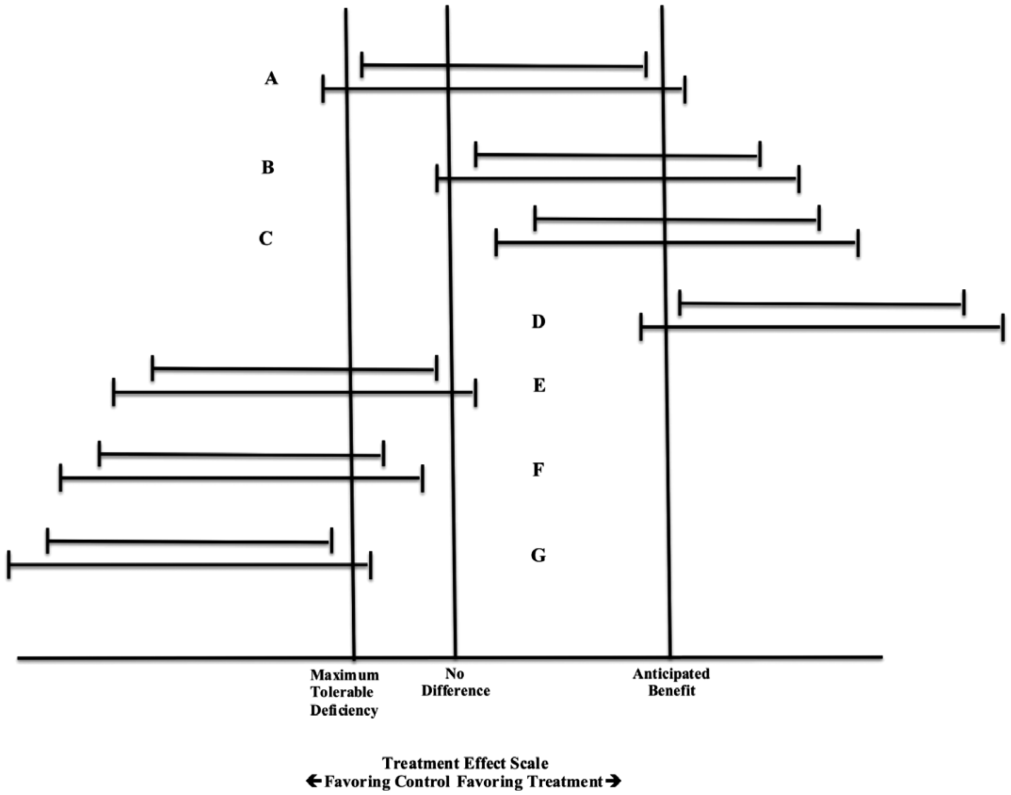
A
B
C
D
E
F
G
Maximum Tolerable Deficiency
No Difference
Anticipated Benefit
Treatment Effect Scale
←Favoring Control Favoring Treatment→

Figure 2. Confidence intervals for a difference or other comparative measure of outcomes between an experimental treatment and a control treatment. The horizontal axis is a scale of possible true values of this treatment effect, for which each confidence interval spans a range of plausible values compatible with the data from which the interval was generated. The vertical lines, from left to right, denote a relative deficiency of the experimental compared with control treatment that is the maximum acceptable in light of other advantages of the experimental treatment, such as reduced risk or costs, equality of treatments, and a benefit of the experimental treatment anticipated when designing the study. Each pair of lines portrays 95% (shorter) and 99% (longer) confidence intervals, summarizing, respectively, 5% and 1% level tests of significance of possible treatment effect values, with each line spanning the range of values which the data are insufficient to reject. Interpretations: (A) the observed treatment effect is not statistically significant, while at the 5% level of statistical significance the data are incompatible with either an intolerable deficiency or a benefit as large as anticipated, and at the 1% level cannot exclude such effects in either direction. B, The observed treatment effect is statistically significant at the 5% but not at the 1% level, compatible with the anticipated experimental treatment benefit, and excludes an intolerable deficiency. C, As for B and also statistically significant at the 1% level. D, The observed treatment effect is statistically significant, excludes an intolerable deficiency, and shows statistically significantly better than anticipated treatment benefit at the 5% but not the 1% level. E, The experimental treatment is statistically significantly worse than the control at the 5% but not the 1% level and compatible with a range of readily tolerable and clearly intolerable treatment deficiencies. F, As for E and also statistically significantly worse at the 1% level. G, The experimental treatment is statistically significantly worse than its maximum tolerable deficiency at the 5% level, and statistically significantly worse than control but still compatible with tolerable deficiency at the 1% level.

name, the registry is not restricted to trials and will accept cohort, case-control, and other types of clinical research. Trial registries contain important but limited information. Increasingly, investigators are separately publishing detailed methods of proposed studies, and some major journals are requiring submission and/or posting of study protocols, thus publically documenting a priori design and analysis choices.

Unfortunately, neither *P* values nor confidence intervals preclude or incorporate effects of systematic errors, which are often far more important than chance errors. In other words, the chance of meaningful random error can be very low, but the results still completely wrong because of bias and/or confounding.

Selection Bias

Selection bias occurs when otherwise-eligible participants are chosen nonrandomly for study inclusion and/or assigned to one treatment or another for nonrandom reasons, including reasons that might influence their response to treatment. This can result from subtle forms of the disease being missed or treatment being directed to patients thought most likely to benefit.

For example, patients with better education or stronger support systems may seek treatment earlier or be given more aggressive treatment. Similarly, patients may comply poorly with or even stop particular treatments, either for perceived lack of efficacy or because of side effects, essentially selecting themselves out of a study. For example, in a study of postoperative cognitive deficits, those with compromised executive function may simply be unable to organize a return visit for testing, with the consequence that attrition bias makes the tested cohort appear to have better cognitive function than the full study population. The extent to which any of these events occur nonrandomly usually is difficult to assess.

Directing a treatment to patients subjectively thought most likely to benefit is a perfectly natural and appropriate tendency in clinical care; however, the selection bias that results leads to what is known as "confounding by indication" in observational treatment comparisons, and the consequence is that one cannot be sure whether differences in treatment outcomes are because of differences in the effects of the treatments or initial constitutional differences in selected patients. In experimental studies, selection bias can largely be prevented by proper randomization followed by encouragement of patients, caregivers, and investigators to maintain the designated treatment allocation.

Table 1. Recall Bias, a Type of Measurement Bias

Reported parental arthritis history	Rheumatoid arthritis (%)	No rheumatoid arthritis (%)
Neither parent	27	50
One parent	58	42
Both parents	15	8

Otherwise-similar people with and without arthritis were asked whether their parents had arthritis. People with arthritis were far more likely to report that one or both parents also had arthritis. The difference was highly statistically significant, with a *P* value of 0.003. The subjects with and without arthritis, however, were siblings. They had exactly the same parents! (Modified from Schull WJ, Cobb S. The intrafamilial transmission of rheumatoid arthritis. 3. The lack of support for a genetic hypothesis. J Chronic Dis 1969;22:217–22.)

Measurement Bias

In studies that use retrospective data collection, the quality of records is often poor because most recording systems were designed for clinical or administrative purposes rather than research. The difficulty is that existing record quality may vary nonrandomly. In contrast, the quality of prospectively collected data can be excellent in well-conducted studies in which data collection methods can be tailored to the research rather than to administrative objectives external to it.

Measurement bias can result in any type of study, when any aspects of data quality, availability, or measurement method vary for reasons other than chance. For example, patients given new treatments may be watched more closely than those receiving conventional therapy, and enthusiastic clinicians may overestimate the benefits of new treatments or underestimate associated complications.

Data artifacts also can easily produce measurement bias when data are distorted or missing nonrandomly. However, vulnerability to bias depends on precisely what is meant by "nonrandom." Errors or missingness occur "completely at random" if there are no correlates of their occurrence and magnitude, "at random" if their occurrence and magnitude have correlates but are independent of the data missed or distorted, and "not at random" if occurrence or magnitude depends on the unobserved true values of the distorted or missing data.

Problems that occur completely at random increase chance variability but do not necessarily produce measurement bias, depending on the specific nature of the problem. In contrast, "at random" problems can easily generate bias, but such bias can be corrected by careful analysis if the correlates of the problem are known and have been measured, for example, if data are more frequently missing in the elderly but age is recorded. Measurement bias because of nonrandom artifact or missingness cannot be corrected in data analysis, and hence presents the greatest threat to research conclusions. The sensitivity of results to such problems, however, can be explored by "sensitivity analyses," that is, multiple analyses under different assumptions about the error-generating mechanism.

Many papers presenting registry analyses do not adequately describe how artifact, and data accuracy more generally, were evaluated and handled, and how much artifact was found. It is likely that detailed descriptions of artifact definition, quantity, and handling will soon be required in registry reports, along with sensitivity analyses when artifact has the potential to substantively influence conclusions.

Measurement bias in clinical data collection can be subtle and hard to detect. Consider, for example, a classic study by Schull and Cobb (Schull and Cobb 1969). The investigators asked an important question: Is arthritis hereditary? The experiment consisted of asking otherwise-similar people, with and without arthritis, whether their parents had arthritis. Their results are shown in Table 1. The results were clear: people with arthritis were far more likely to report that one or both parents also had arthritis. The difference was highly statistically significant, with a *P* value of 0.003.

There was just one problem. The subjects with arthritis and the subjects without arthritis were siblings; they had exactly the same parents!

So what happened here? Were some of the subjects lying? Unlikely. Most likely, people with rheumatoid arthritis thought much more about arthritis than those who did not. And they were far more likely to have discussed the issue with their parents and thus know (and remember) whether their parents had arthritis. This is "family information bias," a specific example of what is more generally called "recall bias." There are many other types of measurement bias, some of which are equally subtle.

An analog is that people with cancer are preoccupied with cancer and spend lots of time asking "why me?" Various environmental exposures, such as pesticides and workplace chemicals, are known to cause cancer in animal models and are thought or known to cause certain specific cancers in humans. Now let us say investigators are interested in whether environmental exposure contributes to the development of cancer.

The most obvious approach would be to find a group of patients with cancer and a similar group without cancer, and then ask about their environmental exposures. The results are predictable: those with cancer will come up with long lists of exposures because the question was already on their minds.

But when the investigators ask noncancer subjects if they have had major environmental exposures, the answers will usually be something like "Uh, no; not that I remember." The point is that such a difference in reported exposure would almost surely be statistically significant in a large study and almost certainly exaggerate the difference between cases and controls, if any, in real exposure. This is another example of recall bias.

Measurement bias also results from the placebo effect, which should never be underestimated. This is especially the case for subjective responses such as pain and quality-of-life, but placebo effects also have repeatedly been shown

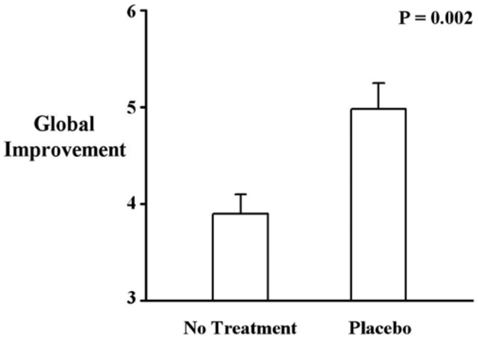

Figure 3. The placebo effect is well known and can substantially bias results. However, patients report benefit from placebo administration even when they know that their treatment is a placebo. Patients with irritable bowel syndrome were randomized to either open-label placebo pills presented as "placebo pills made of an inert substance, like sugar pills, that have been shown in clinical studies to produce significant improvement in IBS symptoms through mind-body self-healing processes" or no-treatment controls with the same quality of interaction with providers. Open-label placebo produced significantly higher mean global improvement scores (P = 0.02). Modified from Kaptchuk et al.[27] Reprinted under a Creative Commons Attribution License.

to influence supposedly objective outcomes[25–27] and, amazingly, patients report benefit from placebo administration even when they know that their "treatment" is a placebo (Fig. 3).[27] Taking a placebo—even one they know is a placebo—makes people feel better or at least believe and say they feel better!

Measurement bias in clinical trials can be largely controlled by blinding, sometimes called double blinding or triple blinding, to the extent that patients, clinicians who make treatment decisions and investigators who assess outcomes are unaware of which patients receive which treatments. Blinding does not remove placebo effects, which no one knows how to do, but distributes their benefits equally across patients in all treatment groups.

Confounding

Confounding is distortion of an apparent association between 2 factors that results from failure to account for a third factor associated with both. Confounding most often is contemplated and recognized when it produces a statistically significant relationship that is not clinically or biologically real; however, confounding can also hide a true association. Confounding is an important and sometimes subtle source of error because such alternative influences may not even be suspected. In retrospective studies, potential confounding factors may be well known but unavailable for analysis by virtue of not being included in clinical records.

To take a trivial example, the rate of mortality is much greater in Florida than Alaska. Is this because Florida is a more dangerous place to live? Should Florida retirees move to Alaska to enjoy a longer retirement? No. To do so would be to mistakenly attribute effects of biology to effects of geography. Mortality is greater in Florida because the median age of Floridians is 7 years greater than that in Alaska. In other words, the relationship of mortality to the state in which one resides is confounded by age, the factor that links exposure (the state of residence) with outcome (mortality). Of course, if investigators know that age is important, and know the ages of people in each state, then it is easy to compensate. For example, you might compare mortality of subgroups of people of similar ages in each state; you might also use statistical adjustments to compensate for differing ages.

The trouble is that potential confounding factors often are unknown. Alternatively, factors may be suspected confounders, but the data needed for adequate statistical compensation may not be available. To the extent that these factors influence the results, conclusions can be quite wrong, with the extent of the error being essentially unknown.

As an example of confounding by indication, blood transfusions are strongly associated with adverse outcomes, including mortality. However, blood transfusions are far more likely to be required by patients who are sickest in the first place. When one compares the mortality of groups differing in how many blood transfusions have been administered, one is implicitly comparing groups with very different levels of underlying illnesses. It is quite possible that underlying illness contributes more to subsequent mortality than the blood transfusions themselves.

Consider, for example, patients who have anemia, which often accompanies chronic diseases, including cancer, or patients having especially long and/or large operations. These patients are more likely than others to need a blood transfusion, and they are more likely to have bad outcomes. But the 2 are not necessarily directly linked; instead, they are indirectly linked (i.e., confounded) by the fact that sicker patients having larger operations are more likely to both need blood and have bad outcomes.

The important point is that a statistical association (i.e., $P < 0.05$) of transfusions with bad outcomes would not necessarily imply that restricting transfusions will improve outcomes because an entirely beneficial causal effect of transfusions may be concealed by the competing causal effect of the poorer initial conditions of patients who require them. Even an exceptionally statistically significant excess of adverse outcomes in transfused patients might be completely spurious, because of the separate associations of the need for transfusion and adverse outcomes with initially poor prognosis for unrelated reasons.

The extent to which these subtle and hard-to-quantify confounding factors contribute to research conclusions can be difficult to determine retrospectively. In other words, blood transfusions may actually worsen outcomes, but it is equally possible that outcomes are worse in patients given transfusions by virtue of factors that led to their being transfused. The distinction is critical because restricting transfusions will only improve outcomes if the first mechanism is accurate. Or, to make a stronger statement, basing transfusion policy on a spurious relationship would likely harm patients by denying them needed transfusions.

It is easy to confuse confounding, a statistical problem that can arise without a biologically or clinically meaningful basis, solely because of composition of a research sample, with "effect modification," which is a manifestation of nature's complexity that may be of critical clinical importance. Consider 2 examples.

First, suppose therapy A outperforms therapy B to a similar degree whether men or women are treated, but women generally do better than men. Suppose 10% of men but only 1% of women experience the least desirable outcome, regardless of which therapy they receive. If therapy A has an unpleasant side effect occurring mostly in men, so men receive B more than women, then the relative benefit of A will be exaggerated by comparing the more frequently female recipients of A with the more frequently male recipients of B. For instance, if 70% of those receiving therapy A are female, and 70% of those receiving therapy B are male, then the least desirable outcome will be experienced by only 0.3 × 10% + 0.7 × 1% = 3.7% of those on therapy A but by 0.7 × 10% + 0.3 × 1% = 7.3% of those on therapy B. To the naive who compare these proportions without regard to the different gender compositions of the patient groups, therapy A will appear to have halved the event rate relative to therapy B, although the performances of the 2 therapies are actually identical. This is confounding, (specifically, confounding by indication), a failure to compare appropriately similar groups, in a manner fully analogous to the earlier geographical example, with therapies A and B playing the respective roles of Alaska and Florida, and men and women playing the respective roles of seniors and younger folks.

Second, consider a situation in which, because of underlying biological mechanisms, therapy A is more effective and less toxic in men whereas therapy B is more effective and less toxic in women. Or, to take an extreme example, for a disease with case fatality of 50% in untreated patients, suppose for those receiving therapy A the case fatality decreases to 25% in men but increases to 75% in women, with these numbers reversed, case fatalities of 75% in men but 25% in women, among recipients of therapy B. This is effect modification, in which the relative effect of an exposure or treatment differs from group to group, in this case as a function of sex. In this example, the effect of therapy A relative to therapy B is to triple (75% vs 25%) the case fatality among women, but to reduce it by two-thirds (25% vs 75%) among men. Unlike confounding, which is essentially a research error, effect modification describes biological situations with distinct clinical implications and needs to be understood if treatments are to be optimally applied.

Aspects of the patient's condition or circumstances are candidates for effect modifiers. A well-known example is that the benefit of streptokinase and other thrombolytic agents for improving stroke outcomes depends on the time elapsed after stroke onset. Specifically, the substantial benefit achievable with early administration is modified (in this case largely lost) when a thrombolytic is given too late.

Reverse Causation

Reverse causation is a special and rare type of error in which the roles of cause and effect are misconstrued, so an effect is mistaken for a cause and a cause mistaken for an effect. For example, certain organisms are more commonly isolated from the esophagus in patients with esophageal cancer or precancerous conditions than from healthy controls. However, a conclusion that such organisms cause the cancer would be incorrect because the cancerous condition may produce tissue changes altering the microenvironment and consequently the esophageal microflora.

Reverse causation errors occur when investigators have difficulty identifying the timing, and particularly the ordering, of when a patient is exposed to a risk factor of interest and when the relevant disease outcome first occurred. Such errors are of most concern in studies with retrospective data collection. However, reverse causation error also can occur in prospective studies of long-term exposures and/or chronic disease, where data collected prospectively may reflect long-standing lifestyle and environmental factors and disease signs or symptoms that may have originated much earlier with sequence unknown.

For instance, prospective cohort studies of cancer etiology typically exclude cases diagnosed early during the follow-up period because they may have been latent at the study's start and been initiated before the study's assessment of the exposure. The problem this causes is evident in studies on smoking and lung cancer. In such studies, lung cancer is found more often among smokers who recently quit than among current smokers because some people with early symptoms of cancer stop smoking before accruing a formal diagnosis. Reverse causation thus makes it appear as if quitting smoking promotes cancer, which is not actually the case.

CONCLUSIONS

Clinical research can be categorized by the timing of data collection: retrospective or prospective. Clinical research also can be categorized by study design. In cross-sectional studies, exposure and outcome are evaluated simultaneously. In case-control studies, investigators compare previous exposures (including genetic and other personal factors, environmental influences, and medical treatments) between groups distinguished by later disease status (broadly defined to include the development of disease or response to treatment). In cohort studies, investigators compare the subsequent incidence of disease between groups distinguished by one or more exposures.

The major sources of error in clinical research are selection bias, confounding, measurement bias, reverse causation, and chance. Selection bias results from nonrandom allocation of patients to exposures in a way that influences outcomes. Confounding results when an apparent association between a particular exposure and disease actually results from their separate relationships with something else, termed a "confounder." Measurement bias results from nonrandom errors in assessing exposure and/or disease. Reverse causation errors occur when the timing of exposure and disease development are unclear, allowing a consequence of a disease process to be mistaken for a contributing cause. Chance error refers to the fact that the results of any given study will not perfectly reflect the true biological situation because of random variation from one patient to another, of the same patient from one time to another, and from one measurement to another.

Clinical research errors in general, and biases in particular, are best avoided by a strong study design coupled with thoughtful statistical analysis. Although the latter can compensate for confounding to the extent that factors are known and measured, formal statistical methods are most effective in coping with evaluating random variation. Hence, although design and analysis are both important, prevention by design is usually much preferable to correction by analysis. Our next article will discuss observational study designs and the nature of statistical corrections for confounding in observational studies in more detail. The third article will focus on experimentation, specifically on the design features of RCTs that attempt, and may reasonably be expected to succeed, in protecting against the 5 major sources of clinical research errors.

DISCLOSURES

Name: Daniel I. Sessler, MD.
Contribution: This author helped write the manuscript.
Attestation: Daniel I. Sessler approved the final manuscript.
Name: Peter B. Imrey, PhD.
Contribution: This author helped write the manuscript.
Attestation: Peter B. Imrey approved the final manuscript.
This manuscript was handled by: Steven L. Shafer, MD.

REFERENCES

1. Graunt J. Natural and political observations mentioned in a following index, and made upon the bills of morality. London: Thomas Roycroft, 1667. Available at: http://www.edstephan.org/Graunt/o.html. Accessed June 26, 2015
2. Snow J. On the mode of communication of cholera. London, England: John Churchill, 1855. Available at: http://www.ph.ucla.edu/epi/snow/snowbook.html. Accessed June 26, 2015
3. Koch R. Investigations into the Etiology of Traumatic Infective Diseases. London: The New Sydenham Society, 1880
4. Hill AB. The environment and disease: association or causation? J R Soc Med 1965;108:32–7
5. Feinstein AR. Clinical Judgement. Baltimore, MD: Williams & Wilkins, 1967
6. Feinstein AR. The Architecture of Clinical Research. Philadelphia, PA: W.B. Saunders Company, 1985
7. Feinstein AR. Clinimetrics. New Haven, CT: Yale University Press, 1987
8. Sackett DL, Haynes RB, Tugwell P, Guyatt G. Clinical Epidemiology: A Basic Science for Clinical Medicine, 2nd edition. Philadelphia, PA: Lippincott, Williams and Wilkins, 1991
9. Medical Research Council Investigation, Streptomycin in Tuberculosis Trials Committee. Streptomycin treatment of pulmonary tuberculosis. Br Med J 1948;2:769–82
10. Medical Research Council Investigation. The prevention of whooping-cough by vaccination. Br Med J 1951;1:1463–71
11. Chalmers TC. Randomization and coronary artery surgery. Ann Thorac Surg 1972;14:323–7
12. Chalmers TC, Eckhardt RD, Reynolds WE, Cigarroa JG Jr, Deane N, Reifenstein RW, Smith CW, Davidson CS. The treatment of acute infectious hepatitis. Controlled studies of the effects of diet, rest, and physical reconditioning on the acute course of the disease and on the incidence of relapses and residual abnormalities. J Clin Invest 1955;34:1163–235
13. Cochrane AL. Effectiveness and Efficiency: Random Reflections on Health Services. London: Nuffield Provincial Hospitals Trust, 1973
14. Sackett DL, Richardson SW, Rosenberg W, Haynes RB. Evidence-Based Medicine: How to Practice and Teach. Edinburgh: Churchill Livingstone, 1996
15. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA 1992;268:240–8
16. Chalmers I, Haynes B. Reporting, updating, and correcting systematic reviews of the effects of health care. BMJ 1994;309:862–5
17. Ware JH. Statistical practice and statistical education in cardiology. Circulation 1987;75:307–10
18. Ivanova A, Rosner GL, Marchenko O, Parke T, Perevozskaya I, Wang Y. Advances in statistical approaches oncology drug development. Ther Innov Regul Sci 2014;48:81–9
19. Sessler DI, Kurz A. Departmental and institutional strategies for reducing fraud in clinical research. Anesth Analg 2012;115:474–6
20. Kopyeva T, Sessler DI, Weiss S, Dalton JE, Mascha EJ, Lee JH, Kiran RP, Udeh B, Kurz A. Effects of volatile anesthetic choice on hospital length-of-stay: a retrospective study and a prospective trial. Anesthesiology 2013;119:61–70
21. Chavalarias D, Ioannidis JP. Science mapping analysis characterizes 235 biases in biomedical research. J Clin Epidemiol 2010;63:1205–15
22. Freedman M, Osicka T. Reader variability: what we can learn from computer-aided detection experiments. J Am Coll Radiol 2006;3:446–55
23. Whiteside JL, Hijaz A, Imrey PB, Barber MD, Paraiso MF, Rackley RR, Vasavada SP, Walters MD, Daneshgari F. Reliability and agreement of urodynamics interpretations in a female pelvic medicine center. Obstet Gynecol 2006;108:315–23
24. Quinn TJ, Dawson J, Walters MR, Lees KR. Reliability of the modified Rankin Scale: a systematic review. Stroke 2009;40:3393–5
25. Price DD, Finniss DG, Benedetti F. A comprehensive review of the placebo effect: recent advances and current thought. Annu Rev Psychol 2008;59:565–90
26. Benedetti F. Placebo and the new physiology of the doctor-patient relationship. Physiol Rev 2013;93:1207–46
27. Kaptchuk TJ, Friedlander E, Kelley JM, Sanchez MN, Kokkotou E, Singer JP, Kowalczykowski M, Miller FG, Kirsch I, Lembo AJ. Placebos without deception: a randomized controlled trial in irritable bowel syndrome. PLoS One 2010;5:e15591