Clinical Research Methodology 2: Observational Clinical Research

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

Case-control and cohort studies are invaluable research tools and provide the strongest feasible research designs for addressing some questions. Case-control studies usually involve retrospective data collection. Cohort studies can involve retrospective, ambidirectional, or prospective data collection. Observational studies are subject to errors attributable to selection bias, confounding, measurement bias, and reverse causation—in addition to errors of chance. Confounding can be statistically controlled to the extent that potential factors are known and accurately measured, but, in practice, bias and unknown confounders usually remain additional potential sources of error, often of unknown magnitude and clinical impact. Causality—the most clinically useful relation between exposure and outcome—can rarely be definitively determined from observational studies because intentional, controlled manipulations of exposures are not involved. In this article, we review several types of observational clinical research: case series, comparative case-control and cohort studies, and hybrid designs in which case-control analyses are performed on selected members of cohorts. We also discuss the analytic issues that arise when groups to be compared in an observational study, such as patients receiving different therapies, are not comparable in other respects. (Anesth Analg 2015;121:1043–51)

Observational clinical studies are attractive because they are relatively inexpensive and, perhaps more importantly, can be performed quickly if the required data are already available. Epidemiologic and health services investigators have used such approaches for decades. But until recently, surgical and perioperative retrospective studies were too often "100-patient chart reviews," which rarely produced valid conclusions. Increasingly, though, the availability of electronic data and sophisticated statistical techniques makes retrospective observational studies of surgical and perioperative practices a valuable tool.

Especially in anesthesia, there has been a revolution in data quality and availability because of electronic anesthesia and hospital records. Consequently, some institutions have large and dense (i.e., minute-to-minute) registries of anesthesia care. Often they are linked to related hospital databases, such as those from clinical laboratories and blood banks, and to outcomes such as duration-of-hospitalization and date-of-death. Billing codes also provide valuable information about diagnoses and procedures although the intricacies and requirements of reimbursement mechanisms, and of administrative record systems more generally, can sometimes distort clinical realities. To glean the most information from such registries, there are now several national registries that pool data from various institutions, notably the National Surgical Quality Improvement Project, Multicenter Perioperative Outcomes Group, and the American Society of Anesthesiologists Anesthesia Quality Institute.

Recent retrospective perioperative studies include data from tens-of-thousands to tens-of-millions of patients.[1,2] Large numbers per se limit research errors attributable to chance, but do not prevent bias and confounding, and can exacerbate their effects by inducing overconfidence in biased results. But large samples do support more in depth and effective application of statistical techniques to compensate for known confounding factors than is possible with small studies.

Although analyses of observational data should generally be considered exploratory rather than definitive, they are nonetheless often a relatively inexpensive and quick way to evaluate the plausibility of hypotheses and build support for subsequent experimental studies. Done well, retrospective analyses can provide good estimates of treatment effect[3–7] although they tend to underestimate harms associated with interventions.[8] Excellent guidelines have been published to encourage uniform and complete reporting of observational studies, including full acknowledgement of limitations.[9,10]

CASE SERIES

A case series—a description of what happened to a series of patients with a particular diagnosis, perhaps treated with a particular strategy—is certainly an improvement over anecdotal experience and case reports because compiled data from a group are less likely to be idiosyncratic than results from 1 or a few individuals. There are many examples in which case series, and even case reports, have provided critical advances.[9] Malignant hyperthermia, for example, was initially reported as a case series and, because it is so rare, has never been subject to randomized trial.[10,11]

In a typical case series, physicians might report that 79 of their patients given a particular treatment for heart failure had a median survival of 37 months. If you are comparable to their patients and get exactly the same treatment, it is

From the *Department of Outcomes Research, Anesthesiology Institute, Cleveland Clinic, Cleveland, Ohio; and †Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio.

Accepted for publication May 1, 2015.

Funding: Supported by internal sources.

The authors declare no conflicts of interest.

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

DOI: 10.1213/ANE.0000000000000861

reasonable to expect to have about a 50-50 chance of living more or less than 37 months.

The trouble is that this result, although useful for considering the prognosis of an individual already committed to the particular treatment, offers no comparative context. Median survival for comparable patients is not all that we want to know; what we really need to understand is whether this survival (or any other outcome) is better or worse with this treatment than with alternatives, and that is where the logic gets tricky. The danger is that in assessing alternative treatment plans, the results of a case series of a new treatment are almost always implicitly or explicitly compared with previous results, that is, to a "historical control."

Historically controlled studies tend to falsely generate a conclusion that new treatment or local management is superior because the comparative effects of the treatment tend to get mixed up with, or confounded by, other time-dependent changes. For instance, recent patients may have been diagnosed earlier in their disease courses than historical patients because of the improvement in diagnostic imaging or other technology. Patients diagnosed earlier—the recent ones—would thus be expected to survive longer from the time of diagnoses whether or not the new treatment they receive is an improvement. This problem is known as "lead time" bias in the context of studies of medical screening. Because there is no way to recover when diagnoses would have occurred had the historical reference subjects been seen under present conditions, it is difficult or impossible to retroactively correct for this problem. Alternatively, therapy for some disease-related complications may have improved over time. Improved treatment of these complications, rather than the new therapy, may have improved survival.

Conclusions based on historical comparisons are, startlingly often shown to be misleading, usually exaggerated, in subsequent randomized trials.[12,13] Comparisons to historical controls are generally invalid because: (1) the comparison patients differ from those in the case series in important ways that have not been accounted for and may even be unknown (confounding); and/or (2) outcome measurement accuracy differs nonrandomly (measurement bias).

That said, population-based observational research—following a group of subjects over time—is the proper tool for determining the natural history and prognosis of various diseases, and the conditions (i.e., "risk factors") that frequently precede their development. For example, observational studies would be used to evaluate weight gain or development of hypertension over time in a population. We will return later to observational designs involving more rigorous interperiod comparisons than the historically controlled case series. But it is essential to avoid unwarranted causal inference such as the conclusion that health will be improved by preventing obesity or hypertension—which may or may not prove to be the case.

CASE-CONTROL STUDIES

Case-control studies are usually the best approach—and often the only practical one—to study rare diseases or outcomes. Consider unanticipated tracheal reintubation in the postanesthesia care unit. This is a potentially severe complication that is often used as a quality metric; fortunately, reintubations are also rare—too rare, in fact—to practically evaluate in prospective trials.

An alternative approach is to find patients who required reintubation in the postanesthesia care unit and compare them to a similar group of patients who did not. Investigators can then look backward in time and determine, for example, which patients in each group had their neuromuscular block reversed. If patients who required reintubation were significantly less likely to have been given reversal agents, one can conclude that there is an association between inadequate block reversal and emergent reintubation. The basic approach to such case-control studies is shown in Figure 1.

For case-control studies to be valid, the case and control groups must be chosen or statistically adjusted to be comparable with respect to potential confounding factors, and the assessment of exposures and potential confounding factors must be equally complete and accurate for both. A danger with clinical records is that important potential confounders may not have been evaluated or may have been inaccurately recorded. They may also be nonrandomly erroneous: for example, patients who have complications may have a more complete list of preexisting conditions. A further danger is that important confounding factors may be unknown and thus not even considered by investigators.

But because investigators must look back in time—sometimes quite far back—it is difficult to determine the extent to which the groups might meaningfully differ. Equivalent exposure assessment for both groups may also be hard to ensure. For example, data about cases often must be obtained directly from patients or their families, whose memories may be stimulated by the symptoms, diagnostic processes, and concerns about the disease. In contrast, the memories of and about control patients may not undergo such reinforcement. The tendency for stronger memories in 1 group to distort exposure comparisons between case and control subjects is known as recall bias.

It is possible to conduct case-control studies within well-defined cohorts and thus ensure that both cases and controls represent the same defined population. For instance, disease registries that centrally record all cases within a defined geographic area allow investigators to conduct

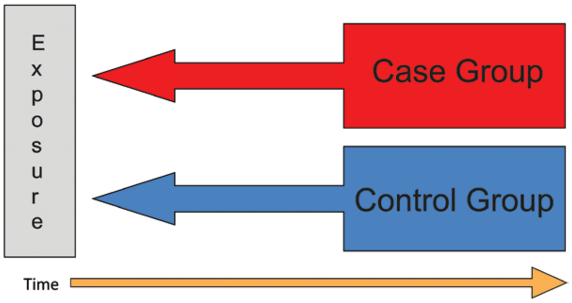

Figure 1. The cardinal feature of case-control studies is that investigators start with a group of people who have already experienced the outcome-of-interest (reintubation) and an appropriate group of controls who have not. They then look backward in time to compare the frequency and the extent of exposures (in this case, reversal of muscle relaxation) between the groups. Case-control studies are most often retrospective, in that investigators collect the data about exposure, often many different exposures, backward in time, after the cases have occurred.

"population-based" studies with cases drawn from the registry and controls drawn from the general population in the area covered by the registry.

"Case-cohort" and "nested case-control" studies describe hybrid designs combining elements of conventional cohort and case-control research. In each, an underlying cohort study defines the context within which a case-control study is conducted. In case-cohort studies, the exposures of cases that develop within the course of the cohort study are compared with those of a random sample of cohort members, which may be selected before disease develops. In nested case-control studies, controls are matched to each case individually by random selection from other cohort members observed and remaining disease free for at least as long as the case. This matches the observation times and data collection periods of controls with those of the cases.

Case-cohort and nested case-control designs are more assuredly valid than conventional case-control or retrospective cohort studies. But they require surveillance of a cohort to obtain the cases and controls and thus may take longer and inevitably cost more than fully retrospective approaches. Savings nonetheless accrue because exposure and outcome data need not be ascertained for all cohort members. For instance, if exposures are genetic or otherwise available from preserved biosamples, only those from cases and selected controls, who together may constitute only a small subset of the cohort, will require analysis.

COHORT STUDIES

Cohort studies differ from case-control studies in that they look forward in time from exposure to disease/outcome. As above, the term "exposure" is used broadly and can refer to a patient's genetics, environmental exposure, or treatment. The term "disease" is equally broad and includes complications, progression, and death. The general logic of cohort studies is shown in Figure 2.

In cohort studies, investigators start with groups differing in their exposures or levels of exposure and look forward in time, comparing subsequent disease incidence between these groups. But that does not mean that data collection needs to be prospective. For example, investigators can "look forward" from exposure to disease within the confines of an existing database. This latter approach is termed a "retrospective cohort study" because the research is conducted after the period for which disease development is to be compared. Usually, and preferably, the exposure groups are subsets of a single natural cohort although on occasion they may arise from different sources.

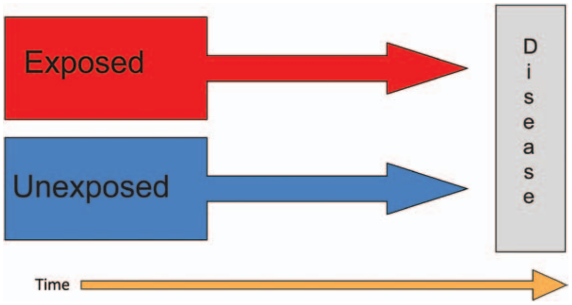

Figure 2. The cardinal feature of cohort studies is that investigators start with groups of people who differ in their exposure and analytically look forward in time to compare the development of disease between these groups.

Cohort studies can also start after exposure but before development of disease, an approach termed an "ambidirectional cohort study." Ambidirectional studies are used in public health crises when the exposure was unexpected but when its effects on health are potentially important (e.g., a leaking nuclear power facility, chemical plant explosion, or widespread exposure to contaminated air or water). Exposed subjects are then followed prospectively for ill effects, even though the exposure occurred before the study started. Studies of smoking, air pollution, diet, blood pressure, etc. are also considered ambidirectional because the (ongoing) exposure is present at the time the study starts.

Finally, cohort studies can be completely prospective. Some are observational, which is appropriate when exposure cannot be controlled by the investigators (i.e., employment in a chemical plant, a genetic factor, or a lifestyle factor or maintenance medication that cannot ethically or practically be manipulated for a sufficient time frame). But a completely prospective cohort approach, in principle, allows the investigators to take an experimental approach by controlling the exposure. This can vastly increase the validity of the study.

A special case of a prospective experimental cohort study is a "randomized clinical trial," in which the exposure (in this case, a preventive strategy or therapeutic treatment for a disease) is randomly assigned. The 3 "flavors" of cohort studies are shown in Figure 3. (Also, see the third article in this series that focuses on randomized trials.)

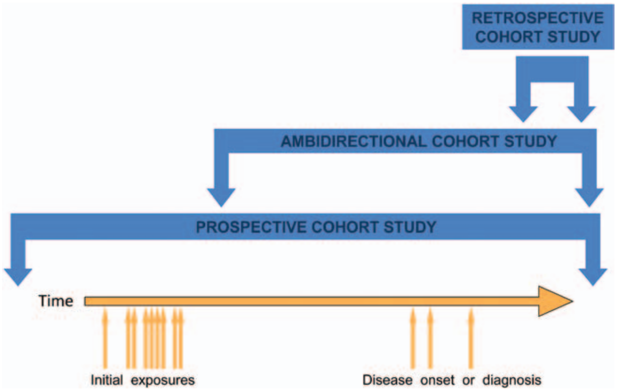
RETROSPECTIVE COHORT STUDY
AMBIDIRECTIONAL COHORT STUDY
PROSPECTIVE COHORT STUDY
Time
Initial exposures
Disease onset or diagnosis

Figure 3. In cohort studies, investigators start with exposure and look forward in time, comparing development or progression of disease between groups differing in their exposures. In a retrospective cohort study, investigators assemble the groups and make these comparisons within the confines of existing records. An ambidirectional approach is used when the exposure has occurred (i.e., a chemical plant exploding) or is ongoing (i.e., cholesterol concentration or diet), and its effects on health are potentially important and thus worth following. Prospective cohort studies, in which investigators collect their data as the exposures and subsequent disease events occur, can be especially powerful because this approach may allow investigators to control measurements and manipulate the exposure, which vastly increases the study reliability. A special case of a prospective cohort study is a randomized clinical trial, in which exposure (treatment) is randomly assigned.

INTERPERIOD COMPARATIVE DESIGNS

"Before-and-after" studies, which compare outcomes of a case series after initiation of an intervention with those of historical controls in an immediately preceding period, use temporally proximal historical controls who are also highly similar in other respects to those coming after and thus exposed to the intervention. Such studies often allow planned, uniform measurements throughout the full observation period. For these reasons, they are clearly less vulnerable to error from extraneous influences than more loosely historically controlled case series. Before-and-after studies are frequently used in health services research because implementations of complex practice changes, such as electronic records or multispecialty enhanced recovery pathways, tend to occur at discrete points in time and are essentially irreversible.

Observational analogs of experimental crossover designs and "n-of-1" clinical studies,[14,15] in which individuals or groups are exposed to planned sequences of 1 or more experimental and control conditions in discrete periods, are more resistant to errors than simple before-and-after comparisons. Such designs are often called "quasiexperimental" and are most straightforward for studies of exposures with short-term biological effects and of people whose health conditions are chronic and relatively stable; more restrictive assumptions and complex analyses are required when responses vary systematically among periods because of disease progression or persistent effects of therapies in previous periods. For example, at the population level, cardiac arrest incidence rates might be compared among periods of high and low air pollution, as determined by a relevant air quality metric.

Interrupted time series studies involve measurements over multiple time periods, preferably a considerable number, bracketing an intervention, or other exposure of interest. This elaboration of a conventional before-and-after design allows observation of trends coincident with the intervention. The technique allows differentiation of a response to intervention from stable secular trends caused by extraneous factors, such as steady increases or decreases or perhaps cyclical effects because of seasonal variation. Close temporal association between exposure and outcome, in a long series of stable or predictably changing outcomes, can compellingly suggest causality in the context of observational research.

Elaborations of this design, where possible, involve observations of similar time series across multiple groups for which the timing of the exposure has varied. Controlled interrupted time series studies include a second series of observations, at the same times as the first, in a group that remains unexposed throughout the study. Stepped wedge designs observe multiple time series at similar points, from groups initially exposed at different times in the sequence of observations. The additional time series improve investigators' ability to distinguish the effects of exposure from those of extraneous, irregular influences on the outcome.

Similarly, in what are known as "case-crossover studies" at the individual patient level, measures of chronic pain might be compared between periods in which patients used different analgesics or doses; childhood accidents might be associated with amount of sleep on the preceding night; and episodes of bleeding and thrombosis for patients on ventricular assist devices might be compared across periods of differing preventive medication regimens.

BIAS IN CASE-CONTROL, COHORT, AND INTERPERIOD COMPARISON STUDIES

The "Hawthorne effect," initially identified in the 1940s,[16] is a subtle type of bias favoring the intervention in before-and-after studies. It refers to the fact that simply being in a study, and the associated attention from investigators, alters responses of participants. For example, patients in a placebo-controlled clinical trial of *Ginkgo biloba* for treatment of mild-moderate dementia exhibited more improvement when intensively monitored than with less intense evaluation.[17]

Two difficulties with before-and-after studies and more elaborate interperiod comparisons are that (1) the presence of the exposure may influence the measurements of the outcome unless the measurement process is under tight control of the researcher throughout the study period; and (2) medical improvements, or secular trends concurrent with but *irrelevant to the intervention*, may make an intervention appear more (or less) effective than it really is. The second of these can be greatly mitigated by multiple series and multiperiod studies because such studies can effectively exclude many simple alternative explanations for data, leaving causality as the sole remaining plausible scenario.

Combined concerns about concurrent time-dependent improvements, the Hawthorne effect, and the potential for measurement biases, make simple before-and-after studies an especially weak design. Where feasible, multiperiod crossover studies, including stepped wedge[18] and interrupted time series[19,20] comparisons, are generally superior alternatives and even more so when they include a parallel control series.

Case-control studies look back in time from disease/outcome to exposure, whereas cohort studies look forward in time from exposure to disease/outcome. A consequence is that selection and measurement biases apply differently. In case-control studies, selection bias applies to selection of the case and control groups. To the extent that the 2 groups differ on variables extraneous to the focus of the study, the effects of such variables may contaminate inferences about the exposure(s) of interest, compromising the validity of the conclusions.

The opposite is true for cohort studies. There, selection bias applies to selection of the exposed and unexposed groups—which need to be otherwise comparable. An example from anesthesia is that patients who get neuraxial anesthesia usually differ from those who do not; a simple comparison of outcomes between patients who do and do not have epidural or spinal anesthesia will thus be confounded by such differences to the extent that they influence the outcomes of interest.

Such issues emerged in the controversy over risks and benefits of hormone replacement therapy for postmenopausal women. Hormone replacement therapy users tend to be more affluent, better educated, have greater access to care and treatment of comorbidities, and tend to be more medically compliant than nonusers.[21] To the extent that these factors predispose to better health outcomes, direct observational comparisons between users and nonusers randomly sampled from the population were potentially biased in favor of hormone replacement therapy.

Simple comparisons of such an outcome in cohort studies would thus compare groups of women collectively at different risk levels. Case-control studies would similarly find fewer using hormone replacement among cases than controls because more affluent and educated women with better medical care would be less frequent among the case group, which would thus be relatively more populated with women less likely to have access to and use hormone replacement therapy. Thus, confounding in both types of studies could arise from failure to compensate from the selective manner in which hormone replacement therapy becomes available to, and is used by, perimenopausal and postmenopausal women.

For case-control studies, measurement bias applies primarily to estimating exposure, which might be genetic, environmental, or a treatment. Again, the opposite is true for cohort studies; there, measurement bias applies primarily to assessing disease or outcome. For instance, good clinical practice typically requires greater surveillance for, and higher sensitivity to, disease signs in patients known or suspected to be at increased risk, but such differences in relation to a suspect exposure produce ascertainment bias in cohort studies. Figure 4 shows the 2 types of studies and where selection and measurement bias are of most concern in each.

Case-control and cohort studies are both important research tools. Case-control studies usually involve retrospective data collection although a case-cohort or nested case-control analysis is sometimes planned within a randomized trial. Cohort studies can involve retrospective, ambidirectional, or prospective data collection. More complex observational studies may combine features of both the case-control and cohort approaches. Comparisons between observational studies and subsequent large randomized trials suggest that the observational studies usually correctly determine the direction of the effect being evaluated but often overestimate the magnitude of the treatment effect.[6]

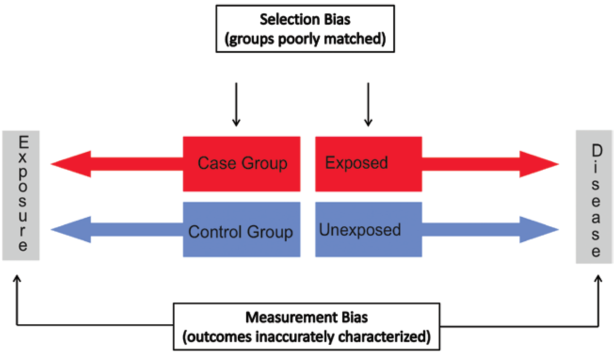

Figure 4. In case-control studies (left), selection bias applies to selection of the case and control groups, and measurement bias applies to determination of exposure. In cohort studies (right), selection bias applies to selection of exposed and unexposed groups, and measurement bias applies to determination of disease or outcome.

LINKS BETWEEN CASE-CONTROL AND COHORT STUDIES

Although cohort and case-control studies differ substantially, they share an identical basic purpose: to shed light on how current status and exposures, such as medical treatments, predict and affect future outcomes.

It is not intuitively obvious that comparisons of statistics about past exposures in case-control studies can be used to describe the statistical patterns of evolution of disease outcomes in current and future patients. In fact, precise mathematical reasoning, including a bit of algebraic jujitsu, is required, and this reasoning does not apply to arbitrarily assembled case and control groups. For the logic to work, both cases and controls must be sampled from a common underlying cohort wherein the relevant pathogenic processes are generally stable. In other words, the control sample in a case-control study must be representative of the larger group of all those who, had they developed the disease under study, would have been detected and included in the case group. In practice, it is often difficult to establish such groups, especially when studying rare diseases and potential links to distant occupational and environmental exposures.

Consider, for example, a case-control study in which hospitalized patients with 1 disease are compared with similarly hospitalized controls with other diagnoses. The difficulty with this approach is that hospitalized patients are special by virtue of their need for hospitalization and thus may differ greatly from the cases in characteristics and past exposures related to other causes of hospitalization. So the selection processes that result in hospitalization for various conditions can introduce serious confounding that is often hard to anticipate—and even harder to eradicate—a problem known as Berkson bias.

With hindsight, we can well appreciate an example of this in 2 seminal 1950s studies of smoking and health conducted in England by Doll and Hill, who were both subsequently knighted for their contributions. One was a prospective study of smoking and mortality from various causes in a cohort of physicians.[21] The other was a case-control study of smoking and lung cancer incidence, in which hospitalized lung cancer patients were compared with controls selected from inpatients of the same hospitals with other diagnoses.[22] Because lung cancer was almost always fatal, and there is no particular reason to expect its relation with smoking to differ in physicians from its relation with smoking in others, one might expect estimates of the smoker-to-nonsmoker lung cancer incidence ratio from the hospital-based case-control study and the smoker-to-nonsmoker lung cancer mortality ratio from the prospectively studied physician cohort to be similar.

However, the hospital-based case-control study incidence ratios for categories of cigarette smoking at levels above half-a-pack per day were only half the corresponding mortality ratios observed in the physician cohort, whereas the incidence and mortality ratios for less frequent smokers were similar. Subsequent research has clearly supported the ratios found from the physician cohort, and it is now accepted that the case-control study substantially underestimated the strength of the smoking and lung cancer linkage.

We know now what was not known then, and why such a discrepancy between the results of these studies at higher smoking levels was virtually inevitable. Cigarette smoking causes many diseases, including myocardial infarctions and ischemic stroke; collectively, cardiovascular and cerebrovascular diseases linked to smoking were then, and are now, much more common than lung cancer. In the 1950s, heart attack patients were typically hospitalized for 3 weeks. Hence, the British hospitals from which controls were selected in the early 1950s were likely filled with heart attack survivors, whose past smoking exposures had contributed to pathogenesis of their heart disease and thus selected them to be available as controls.

Control patients in this study were thus bound to overestimate the smoking levels of those without lung cancer in the population, leading to underestimation of the disparity between smoking levels of cases and controls and consequently of the inferred disparities between lung cancer incidences of nonsmokers and smokers—but only at smoking levels sufficient to noticeably affect heart attack risk. Thus, with current knowledge, it is unsurprising that the effect was not seen in Doll and Hill's lowest smoking dose category, where effects of smoking on other diseases were weakest.

CONTROLLING CONFOUNDING AND COMPLEXITY IN OBSERVATIONAL STUDIES

Observational clinical researchers attempt to isolate the causal effects of 1 factor from the biological and statistical influences of others. Fortunately, human environments and behaviors are not subject to the sorts of manipulations and controls that are routine for laboratory samples and animal models. But a consequence is that confounding is always possible. Specific features of study design and analysis are thus used to reduce the risk of confounding errors by fostering like-with-like comparisons. These methods are powerful but imperfect.

Matching methods assemble the groups to be compared in a manner that forces them to be similarly constituted with respect to characteristics that otherwise might confound the comparisons of greatest interest. Group compositions might be harmonized as a whole in what is known as "frequency matching." Alternatively, and more aggressively, individuals in each group might be linked to specific similar individuals in the other group(s) in "matched set" studies.

Analytical strategies strive for similarly equitable comparisons by restructuring how observations are organized and weighted when comparisons are made. Stratification methods such as classical direct rate adjustment subdivide the data into strata whose members are similar with respect to confounding variables, obtain like-with-like subcomparisons between the relevant groups within each of these strata, and then reassemble these subcomparisons into an overall summary of group differences. Because the subcomparisons are protected against confounding, their aggregation is also. For example, Florida's mortality rate is higher than Alaska's because Floridians tend to be older. But we would be foolish to move from 1 state to the other to live longer because mortalities for Floridians and Alaskans of comparable ages are similar. By stratifying each state's population and averaging the differences in the resulting age-specific mortalities, we obtain a clearer picture of the effect of state of residence, separated from the known effect of age. But unless a very large sample is available, it is difficult with this method to simultaneously protect against confounding by more than a few variables, and it only protects against known confounders for which reliable data are available.

"Multivariable modeling" refers to more comprehensive and sophisticated statistical multiple regression methods for simultaneously controlling multiple confounders and potentially also evaluating effect modification. One use of such models is to develop an overall propensity score to summarize the multiple characteristics correlated with exposure. The score may then be used to develop individually matched or stratified comparisons, more or less as just described, in relatively simple statistical analyses as if the propensity score itself had been a single prognostic characteristic known at the study's outset. Thus, the modeling effort is directed toward accounting for preceding correlates of the exposure and is conducted entirely without reference to clinical outcomes. When the exposure is a treatment, propensity modeling then attempts to represent physician behavior, about which much may be known a priori and used to inform the modeling process.

Most multivariable modeling, however, is directed at exposure-disease linkages and attempting the more difficult task of representing complex biology and resulting disease behaviors. These models attempt to achieve the benefits of finely stratified analyses by mathematically incorporating the relations between pairs of variables, assuming others are held constant, and then correcting for differences in groups being compared that may generate confounding by mathematical adjustments of outcomes. Essentially, this involves sliding the outcome distributions for each group along scales corresponding to individually predictive variables, until averages of all such variables are closely aligned across all groups. Direct multivariable modeling and propensity matching are both useful, and which is best for a given analysis depends on the questions being asked, the type of data, and the characteristics of the data source.[23]

A basic approach to adjusting for confounding is common to the various types of regression analysis used for modeling continuous measurements, dichotomous outcomes, rates of recurrent disease episodes, and survival times. At the core of most statistical modeling is an index of patient factors known technically as a "linear predictor" and consisting of a weighted sum of patient variables $a_1x_1 + a_2x_2 + ...+ a_kx_k$. The x_i are numbers, or numerical codes for values of ordinal or qualitative patient factors, and the a_i are weights intended to reflect the statistical relations of these variables to the outcome. When enough data are available and relations in nature are simple, confounding can be statistically removed simply by adding 1 or more terms a_jx_j representing confounders into the linear predictor. This approach can, in principle, remove confounding for even a large set of variables.

Effect modification can also be represented in regression analysis by including interaction terms in the linear predictor for which the x portion is constructed as the product of numerical representations of the exposure variable and variables that may modify its effect. When interaction terms are statistically significant, exposure effects should usually be described separately for different values or ranges of the effect modifier.

For instance if, in an observational study, the variable x_1 is respectively 1 or 0 for patients receiving drug A or drug B, and physicians tend to somewhat preferentially prescribe drug A for older and drug B for younger patients—and the average outcome of treatment depends linearly on age—then confounding of the treatment effect by age might be removed by simply including a multiplier of patient age in the linear predictor or including separate variables taking values 0 or 1 to indicate membership in each of several age strata. Effect modification can also be investigated by including an additional variable, say x_6, constructed as the product of x_1 with age: 1 × age = age for patients receiving drug A and 0 × age = 0 for patients receiving drug B.

Occam's Razor, attributed to William of Occam (1285-c. 1347) and sometimes known as the law of parsimony or the KISS (keep it simple stupid) principle, is the practical, and to some extent aesthetic, philosophy of choosing, among competing explanations, the simpler over the more complex.

Occam's Razor requires hypotheses involving effect modification to be held somewhat at arms length until essentially required by data that cannot be explained more simply. Keeping this approach in mind when conducting multivariable modeling is wise because a search for interactions among many variables involves large numbers of combinations, promoting false-positive results.

The apparent simplicity of these powerful methods can be deceptive, especially when the methods are applied in conjunction with automatic or semiautomatic methods for variable selection, which itself is among the most difficult problems in statistical methodology. Specifically:

1. Because confounding is a consequence of sample composition regardless of statistical significance, methods that select variables solely by statistical significance criteria, such as significance of unadjusted associations with either outcome or exposure or by forward, backward, or stepwise variable selection, may fail to recognize and adjust for important confounding. Thus, the common practice of using statistical significance to determine which variables to adjust for in comparisons of outcomes in an observational study is ad hoc rather than justified by statistical principles.
2. Sliding variables along their scales individually to achieve comparability of averages across groups, as is done in the basic modeling approach to confounder adjustment described above, may implicitly assume biologically impossible combinations of variables. In such cases, adjustment models effectively extrapolate beyond the range of the data.
3. A characteristic that cannot itself be measured directly is termed "latent." Pain, for instance, is latent but can be assessed by its effects on multiple behaviors and measurement instruments such as, for young children, the Faces Pain Scale (Faces, Legs, Activity, Cry, Consolability)[24] and numerical visual analog scale.[25] When such multiple measures of the same latent entity are available for analysis, inclusion of alternative highly correlated (technically known as "collinear") measurements in the same linear predictor may conceal important relations. This occurs because the test of each variable individually treats the counterpart variable as a potential confounder and, by doing so, largely adjusts its own effects out of the test.
4. Intermediate variables in a causal pathway between an exposure and outcome, technically known as "mediators," are not confounders. Treating them as if they were, by matching or inclusion in a linear predictor, conceals such portion of the exposure's causal effect that acts through the mediator. For example, anesthesia handovers are associated with, and may be causal for, major in-hospitality morbidity or mortality, potentially because of errors resulting from the loss of critical information during the transfer.[1] If this hypothesis is correct, then if analyses of the association between handovers and outcomes were somehow adjusted for loss of critical information, the adjustment would greatly reduce or even eliminate the association between handoffs and outcomes.
5. Controlling the relations of 2 variables for a mutual consequence, known as a "collider," or for a collider's near relative, can distort the relation unpredictably and yield paradoxical results. To see this, consider the classic example of the relation between 2 causes of wet lawns, rain, and automated lawn sprinkling only on Mondays, Wednesdays, and Fridays. If the association is examined after controlling for lawn status by fixing the lawn as wet, then absence of either cause fully determines—indeed, logically implies—that the other must be present. But of course this is nonsense because rain and automatic sprinkling are statistically independent, despite the conviction of many homeowners that watering brings on rain.
6. Accurate determination of timing, and hence sequencing of exposures, outcomes, and possible mediators, confounders, and effect modifiers may be difficult in observational studies of chronic diseases and/or exposures, leading to inadvertent adjustment for mediators or consequences rather than confounders.
7. Adjustments for confounding in statistical models are themselves influenced by random variation of the estimated relations between the outcome and the confounder, random variation which can be substantial in modestly sized studies.

These issues clarify why (1) a hands-on statistical approach is generally preferable to "by the numbers" handling of confounder control in observational studies; (2) data from some such studies may not have a unique coherent interpretation, even when large amounts of data are available; and (3) data from controlled clinical experiments are preferable to observational data when obtainable and affordable.

STROBE REPORTING GUIDELINES

Various checklists and guidelines have been proposed to enhance completeness and consistency of reporting in observational studies. In practice, the reporting guidelines also serve as guidelines for study conduct because many required elements will only be available if designed into the original protocol. The best known and most commonly used checklist is from the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement.[26]

The STROBE checklist includes 22 items, half-a-dozen of which have subitems. Among these, perhaps the most critical are (1) specific objectives and hypotheses; (2) eligibility criteria, sources, and methods of participant selection and sample size rationale; (3) definition of all outcomes, exposure, predictors, potential confounders, and effect modifiers; (4) data sources; (5) methods of statistical analysis, including matching, adjustments for confounding, subgroups, and interactions; (6) relevant participant characteristics; and (7) results including both unadjusted and (if used) adjusted/matched outcomes with 95% confidence intervals.

CONCLUSIONS

Large registries have revolutionized retrospective perioperative studies by giving investigators access to large amounts of high-quality data. High-density databases including tens-of-thousands to tens-of-millions of patients, combined with sophisticated statistical techniques, have markedly improved the reliability of retrospective studies.

Case series are often explicitly or implicitly compared with historical controls, an approach that is subject to numerous biases. Retrospective case-control studies may be the only way to evaluate rare conditions, and retrospective cohort studies are a quick and inexpensive way to evaluate hypotheses. Long-running prospective cohort studies accumulate invaluable data and biosamples that inform on the natural history of disease and provide a basis for subsequent case-cohort and nested case-control studies.

Done properly, case-control and cohort studies are powerful research tools and may provide the strongest feasible research designs for addressing some questions. Case-control studies usually involve retrospective data collection. Cohort studies can involve retrospective, ambidirectional, or prospective data collection. More complex observational studies may combine features of both the case-control and cohort approaches.

However, observational studies are subject to errors attributable to selection bias, confounding, measurement bias, and reverse causation—in addition to errors of chance. Confounding can be statistically controlled to the extent that potential factors are known and accurately measured but, for the reasons enumerated above, adjustment may not be straightforward in practice, and bias and unknown confounders remain additional potential sources of error, often of unknown magnitude and clinical impact. Causality—the most clinically useful outcome—can rarely be definitively determined from observational data, because intentional, controlled manipulations of exposures are not involved, and clear exclusion of competing noncausal interpretations is logically impossible and exceptionally difficult, even to a practical standard of assuaging reasonable doubt.

Experimentation, however, provides more powerful tools for preventing clinical research errors. Randomized assignment of treatment prevents reverse causation errors and selection bias and, in sufficiently large studies, strongly protects against confounding. Blinding minimizes measurement bias. The third article in this series discusses randomized clinical trials.

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 design the study and write the manuscript.
Attestation: Peter B. Imrey approved the final manuscript.
This manuscript was handled by: Steven L. Shafer, MD.

REFERENCES

1. Saager L, Hesler BD, You J, Turan A, Mascha EJ, Sessler DI, Kurz A. Intraoperative transitions of anesthesia care and postoperative adverse outcomes. Anesthesiology 2014;121:695–706
2. Dalton JE, Glance LG, Mascha EJ, Ehrlinger J, Chamoun N, Sessler DI. Impact of present-on-admission indicators on risk-adjusted hospital mortality measurement. Anesthesiology 2013;118:1298–306
3. Abraham NS, Byrne CJ, Young JM, Solomon MJ. Meta-analysis of well-designed nonrandomized comparative studies of surgical procedures is as good as randomized controlled trials. J Clin Epidemiol 2010;63:238–45
4. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342:1878–86
5. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887–92
6. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, Contopoulos-Ioannidis DG, Lau J. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001;286:821–30
7. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol 2005;58:550–9
8. Papanikolaou PN, Christidi GD, Ioannidis JP. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. CMAJ 2006;174:635–41
9. Albrecht J, Meves A, Bigby M. Case reports and case series from Lancet had significant impact on medical literature. J Clin Epidemiol 2005;58:1227–32
10. Denborough M, Lovell R. Anaesthetic deaths in a family (letter). Lancet 1960;ii45
11. Denborough MA, Forster JF, Lovell RR, Maplestone PA, Villiers JD. Anaesthetic deaths in a family. Br J Anaesth 1962;34:395–6
12. Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. Am J Med 1982;72:233–40
13. Bhansali MS, Vaidya JS, Bhatt RG, Patil PK, Badwe RA, Desai PB. Chemotherapy for carcinoma of the esophagus: a comparison of evidence from meta-analyses of randomized trials and of historical control studies. Ann Oncol 1996;7:355–9
14. Guyatt G, Sackett D, Adachi J, Roberts R, Chong J, Rosenbloom D, Keller J. A clinician's guide for conducting randomized trials in individual patients. CMAJ 1988;139:497–503
15. Guyatt G, Sackett D, Taylor DW, Chong J, Roberts R, Pugsley S. Determining optimal therapy—randomized trials in individual patients. N Engl J Med 1986;314:889–92
16. Mayo E. The Human Problems of an Industrial Civilization. 2nd ed. New York, NY, MacMillan, 1946
17. McCarney R, Warner J, Iliffe S, van Haselen R, Griffin M, Fisher P. The Hawthorne effect: a randomised, controlled trial. BMC Med Res Methodol 2007;7:30
18. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. BMC Med Res Methodol 2006;6:54
19. Ma ZQ, Kuller LH, Fisher MA, Ostroff SM. Use of interrupted time-series method to evaluate the impact of cigarette excise tax increases in Pennsylvania, 2000–2009. Prev Chronic Dis 2013;10:120268

20. Terner Z, Carrol T, Brown DE. Time series forecasts and volatility measures as predictors of post-surgical death and kidney injury. Healthcare Innovation Conference (HIC), 2014 IEEE, 319–22
21. Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. Br Med J 1950;2:739–48
22. Doll R, Hill AB. The mortality of doctors in relation to their smoking habits; a preliminary report. Br Med J 1954;1:1451–5
23. Blackstone EH. Comparing apples and oranges. J Thorac Cardiovasc Surg 2002;123:8–15
24. Gomez RJ, Barrowman N, Elia S, Manias E, Royle J, Harrison D. Establishing intra- and inter-rater agreement of the Face, Legs, Activity, Cry, Consolability scale for evaluating pain in toddlers during immunization. Pain Res Manag 2013;18:e124–8
25. Myles PS, Troedel S, Boquest M, Reeves M. The pain visual analog scale: is it linear or nonlinear? Anesth Analg 1999;89:1517–20
26. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370:1453–7