Because of their well-documented advantages, randomized controlled trials (RCTs) represent the gold standard for testing hypotheses in medical research. However, RCTs are not ideal for addressing certain research questions (e.g., the risk for renal insufficiency associated with alcohol intake). In addition, conducting a properly designed RCT often requires substantial time and resources. For these reasons, including barriers of cost and need for rapid answers, the majority of clinical research studies in the renal literature use an observational design. Although there clearly is a great need for more RCTs that are conducted in populations with kidney disease, rigorous observational studies are extremely valuable and in many circumstances yield similar results as rigorous RCTs (1). Furthermore, although RCTs are considered the gold standard for examining the efficacy of a therapy, studies of prognosis often are addressed best by cohort studies.

In this first of a multipart series dedicated to reviewing clinical research methods in nephrology, after providing a brief overview of observational studies, we focus on one of the major subtypes of observational designs: The cohort study. A glossary of common terms that are used in this and subsequent sections of this series is included in Table 1. Details on other methods (e.g., case-control studies, RCTs) will be the subject of future reviews.

Brief Overview of Observational Studies

The simplest form of observational study is the case report or case series, which describes the clinical course of individuals with a particular condition or diagnosis. Such studies often highlight a single clinical condition and at times even suggest a potential biologic mechanism. In a case series, the clinician gains an appreciation of the breadth of abnormalities (e.g., range of proteinuria in patients with HIV-associated nephropathy) that may characterize a single disease process. Although these studies may attempt to identify factors or treatments that may have influenced the outcome, by definition they do not include a control or comparison group without the exposure or outcome and hence cannot support strong associations between the two. In rare circumstances, the data from a case series can be sufficiently compelling to change clinical or regulatory decisions (e.g., the relation between phocomelia and thalidomide), but usually conclusions from these studies must be viewed with extreme caution.

In cross-sectional studies such as surveys or chart reviews, exposures and outcomes are ascertained at the same time. Cross-sectional studies have several potential advantages, including low cost, simplicity, and reduced risk for certain types of bias, such what occurs with loss to follow-up (see below). However, systematic differences between those who agree and do not agree to participate (responder bias) can be problematic in cross-sectional studies. In addition, because exposure and outcome are assessed simultaneously, causality cannot be determined conclusively. For example, although elevated serum levels of C-reactive protein (a putative “exposure”) are seen in patients with chronic kidney disease (CKD; the putative “outcome”), it would be misleading to conclude that elevated levels of C-reactive protein predispose patients to CKD when both were ascertained at the same time. Finally, because cross-sectional studies include prevalent rather than incident cases, they may be prone to incidence-prevalence bias. This form of bias occurs because groups of prevalent individuals may be systematically different from incident individuals with the same condition. For example, a high proportion of patients die within 90 days of initiating dialysis, often as a result of serious underlying illness or multiple comorbidities. Because of their shorter survival, there will be less opportunity to enroll these individuals in studies of prevalent patients, although they would be included in a study of incident patients.

Case-control studies begin by identifying participants with and without the condition of interest (“cases” and “controls,” respectively). Exposures then are determined retrospectively, and their frequency is compared between cases and controls. Exposures that are more common in cases may give clues to causal relationships, provided that cases and controls are similar in all respects except for the condition of interest. Indeed, selection of appropriate controls is a major challenge in all case-control studies and is a major determinant of whether the conclusion of the study is valid. Case-control studies also require careful consideration of bias relating to differential recall of exposures. Although these requirements can be difficult to meet, the case-control design is inexpensive and efficient, especially when the outcome of interest is rare. Case-control
studies will be covered in subsequent sections of this continuing series.

**Cohort Studies**

The term *cohort* refers to a group of individuals who have a common feature when they are assembled and who are followed in time. Therefore, cohort studies begin by ascertaining exposure among a group of individuals who are free of the outcome of interest and evaluate participants for incident events that occur with time (Figure 1). Over time, outcomes may occur in both groups, and the analysis therefore examines the frequency of the outcome in the exposed *versus* the unexposed groups. The features of cohort studies allow investigators to define the temporal sequence between exposure and outcome and avoid the risk of recall bias (*i.e.*, asking those with or without the outcome of interest to recall a past exposure). Follow-up time should be sufficient for outcomes to occur. For example, a cohort of incident dialysis patients who are followed for 1 yr would allow analyses that are targeted at determining whether a baseline exposure was associated with mortality, because at the end of 1 yr, mortality is expected in >20% of the cohort (2).

Participants in prospective cohort studies are identified, classified with respect to exposure status at baseline, and then are
followed over time to ascertain outcomes (Figure 1). In historical/retrospective cohort studies (also known as nonconcurrent cohort studies), a group of individuals (the cohort) is identified on the basis of a common feature or features that were determined in the past (e.g., starting dialysis during a particular period; see references [3,4] for examples). Historical cohort studies represent studies in which exposures and outcomes were collected sometime in the past, but the ascertainment of exposures antedated the development of the outcomes, and hence the temporal sequence of events (provided that no subjects have the outcome at the time exposures are measured) is preserved. The concern about historical cohort studies is that exposure data usually are not collected specifically for the study, and, hence, the possibility of missing information or unmeasured confounders is high. The availability of electronic medical records has greatly facilitated the conduct of historical cohort studies, which are efficient, inexpensive, and ideal for less common diseases, especially for those with long latency periods. Both historical and prospective types of cohort studies are well suited to study rare exposures and examining multiple potential effects of a single exposure. Certainly, cohort studies allow testing of multiple hypotheses. However, the possibility of bias relating to multiple comparisons that analyses and results should be hypothesis driven and biologically plausible (5). Finally, cohort studies, in particular prospective cohort studies in which samples are stored when patients enroll, allow the investigator to go back later and test new hypotheses as they arise.

Although recently critical discrepancies between results from observational studies and randomized trials have been highlighted (6), in general, the results from well-conducted cohort studies often are similar to those from randomized trials (1). In fact, in a systematic review of the subject, “well-designed observational studies did not systematically overestimate the magnitude of the associations between exposure and outcome as compared with the results of randomized, controlled trials of the same topic” (1). In examining outcomes among women who received hormone replacement therapy (HRT), although observational studies did not yield similar results in the area of coronary heart disease, the two types of studies yielded almost identical point estimates in the areas of risk for breast and colorectal cancer, hip fracture, stroke, and pulmonary embolism (6). Grodstein et al. (6) systematically reviewed these studies and suggested that methodologic differences may have explained why discrepancies were noted for certain outcomes and not for others. For example, observational data supporting the use of HRT primarily examined women who initiated therapy at the time of menopause, whereas approximately 70% of women in the Women’s Health Initiative (RCT examining HRT and outcomes) were enrolled at the 60 yr or older. Re-analysis of each study, stratified by time of initiation of HRT, are more consistent: A trend toward a benefit with HRT in both types of studies in younger women closer to menopause and a trend toward no benefit and potential harm among older women who initiate HRT several years after onset of menopause (7,8). Therefore, the two types of studies yielded similar results when stratified by timing of HRT initiation. Nevertheless, it is important to remember that in observational studies, exposures among the cohort members are not randomly assigned; therefore, the possibility that “other” differences (e.g., residual confounders) explain differences in exposures is high.

Analysis of Cohort Studies

The major objective in cohort studies is to compare the risk for an outcome or outcomes in groups that are defined by exposure status. Because participants are free of the outcome at baseline, investigators usually are interested in incident (rather than prevalent) cases. For the purposes of health studies, survival time usually is the metric of index. Therefore, incidence rate (the number of cases per unit of time) generally is of greater interest than the crude incidence (the number of cases). For example,

Incidence of ESRD in two general groups, A (n = 10,000) and B (n = 1000)

Incidence of ESRD among group A = 1500 new cases between 2001 and 2003
Incidence rate = 1500/3 yr, or 500 cases per year
Incidence of ESRD among group B = 300 new cases between 2001 and 2003
Incidence rate = 300/3 yr = 100 cases per year

Because age and gender are such fundamental properties of populations (and often are associated with outcomes), incidence rates often are age and gender adjusted. However, for purposes of this example, we assume that age and gender are distributed evenly between the two groups and, thus, that the incidence rates for the two groups are similar with and without adjustment. In this example, the incidence and incidence rate of ESRD both are higher for group A than group B.

Table 2. Hypothetical risk of ESRD in two groups of patients

<table>
<thead>
<tr>
<th>Group</th>
<th>ESRD</th>
<th>No ESRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1500</td>
<td>8500</td>
</tr>
<tr>
<td>B</td>
<td>300</td>
<td>700</td>
</tr>
</tbody>
</table>
RR of ESRD in Group A = \[
\frac{[1500(1500 + 8500)]}{[300(300 + 700)]} = 0.15/03 = 0.5
\]

In this case, the higher incidence rate of ESRD in group A is driven by the larger prevalence of this group in the total population (10,000 in A/11,000 in A + B), because group A actually is at lower risk for ESRD (reflected by the RR of <1). In this case, the period of observation (2001 to 2003) was the same for both groups. However, because this is not always the case, the denominator for the RR can be a measure of person-time (i.e., person-years at risk) rather than the number of people at risk.

This latter point is especially important because each subject may have different follow-up times, each contributing a different number of person-years to the denominator. Person-years of follow-up then can be normalized for each group (e.g., number of cases per 1000 person-years) for interpretation and comparison purposes.

For certain outcomes (e.g., mortality, renal allograft failure), it may be particularly relevant to consider the time until the event occurs, rather than the incidence of the event. To give an absurd example, the incidence of death would be equal in all subgroups of a cohort study after two centuries had elapsed, regardless of any true association between exposure and risk. However, even when the outcome is not inevitable, refining estimates of risk by considering time to event usually results in increased statistical power compared with analyses that simply evaluate whether the event occurred. Two commonly used approaches for analyzing time-to-event data include Kaplan-Meier analysis (which allows univariate comparison of survival times between groups) and Cox proportional hazards analysis (which allows both univariate and multivariate comparisons).

Kaplan-Meier plots are used to display graphically time to event, often comparing survival of groups who have different baseline exposures with frequency of the event on the y axis and time on the x axis (Figure 2). The statistical test to compare one curve with another is the log-rank test, which is a simple modification of the \( \chi^2 \) test. Although not all individuals who enter the study will reach the end point of interest (not all subjects die) during the study period, it also is important to remember that subjects may leave the study for reasons other than the primary outcome before the end of the follow-up period (e.g., leaving the cohort because of recovery, moving to a different state); therefore, the exact survival time or time to the event will be unknown. Nevertheless, these subjects contribute person-time information until they are censored. Unfortunately, if subjects leave the study for reasons related to their exposure or outcome (informative censoring), then the observations may become distorted (e.g., sicker patients leave the study before they die), and the results then may be erroneous.

**Sources of Error in Cohort Studies: Generalizability, Bias, Confounding, and Chance**

Compared with randomized trials (which often study a select group of patients on the basis of inclusion and exclusion criteria), observational studies often include participants with a wider spectrum of disease severity and comorbidity (1). Therefore, results from cohort studies that highlight the effects of treatment may have better external validity (generalizability to the affected population) than those from randomized trials. However, there is a higher risk for drawing incorrect inferences about treatment effects from cohort studies because of the increased likelihood of bias (because treatments are not randomly assigned); therefore, results from observational studies should be confirmed by randomized trials whenever possible. Even in cases in which a randomized trial is not feasible, the possibility of biased results from a cohort study remains.

Bias occurs when results of a study systematically deviate from the truth because of nonrandom factors. Although many specific types of bias have been described, there are three broad categories: Selection bias, information bias, and confounding. Selection bias occurs when study participants are not representative of the broader population at risk for the outcome. This is particularly relevant for case-control studies, in which the criteria for selecting cases and controls must be as similar as possible. However, selection bias also can occur in cohort studies. One example of selection bias that can occur in cohort studies is length-time bias, which occurs when participants with milder disease are preferentially enrolled because their more indolent course allows a longer period for detection (and a higher likelihood of participating). This is relevant because pathophysiology, natural history, and expected response to treatment all may differ substantially in those with mild versus severe forms of disease. Other types of selection bias can occur when loss to follow-up is high because the likelihood of
follow-up may be directly related to the exposure and outcome under study.

Information bias occurs when data on exposure or outcome are systematically incorrect, either when the exposure is measured differently in people with the outcome or when the likelihood of detecting the outcome varies between exposed and unexposed participants. For example, the risk for microalbuminuria associated with diabetes might be overestimated in a cohort study that used data from clinical laboratories (in which urine albumin assays were ordered on everyone as part of routine care) to define the outcome. The risk might seem to be greater if physicians were more likely to screen for microalbuminuria in people with diabetes (detecting additional cases that might be missed in those without diabetes) or if the discovery of microalbuminuria triggered screening for diabetes (increasing the likelihood that the exposure would be identified). Conversely, misclassification of exposure or outcome that occurs at random will tend to bias toward the null (the finding that exposure and outcome are not associated). Although statistical techniques can be used to offset partially some types of selection and exposure bias, these problems are prevented best by careful study design. When the study design cannot be changed, sensitivity analyses (e.g., assuming that all participants who were lost to follow-up actually experienced the outcome; see reference [4] for example) can provide reassurance about the potential impact of these biases.

Confounding occurs when a factor is associated with both exposure and outcome. A classic example of confounding is the observation that coffee drinkers are at higher risk for lung cancer. In this example, cigarette smoking is a confounder, because it is associated with both the exposure (smokers are more likely to drink coffee) and the outcome (cancer). An example in nephrology would be a study of pain medications and risk for kidney failure in which an exposure (e.g., acetaminophen; nonsteroidal anti-inflammatory drug) is found to be associated with an outcome (e.g., kidney failure), yet the exposure may be the result of a predisposing factor (e.g., previous illness causing pain that is linked with kidney disease) that is linked directly with the outcome and, thus, the predisposing factor (e.g., illness related to kidney disease) confounds the relationship between the exposure of interest and outcome. Confounders should not be involved directly in the causal pathway between exposure and outcome. For example, in studies of the relation between air pollution and death, reduced lung diffusion capacity might be associated with both the exposure (polluted air) and the outcome (death) but may not confound the association between the two because reduced lung function might be the underlying mechanism for increased mortality.

Although stratification, matching, and statistical adjustment can correct for confounding, it is important to note that these techniques account only for potential confounders that have been measured. Because such factors may be poorly quantified, unavailable, or unsuspected, considerable potential for residual confounding (i.e., confounding that remains despite attempts to adjust for potential confounders) exists even in the most carefully conducted cohort studies. Traditionally, only associations with very large effect sizes (five- or 10-fold increases in risk) were believed to be at low risk for residual confounding. The increasing frequency of cohort studies with very large sample sizes means that associations may be highly statistically significant despite a small increase in relative risk. In addition to their potentially lower clinical relevance, such associations also are more likely to be spurious as a result of residual confounding. Although cohort studies remain important for generating hypotheses about therapy, they cannot replace randomized trials. This is particularly relevant to the nephrology literature, given several recent examples in which observational data had a significant impact on clinical practice before well-conducted randomized trials negated their findings (9–11).

A final mechanism by which cohort studies may draw incorrect conclusions is by chance alone. When statistical significance is set at the P < 0.05 level, an exposure will be found incorrectly to be associated with the outcome once in every 20 comparisons. Because cohort studies may collect data on hundreds of potential exposures, there can be substantial potential for spurious results because of multiple comparisons. Similar considerations apply to the use of subgroup analysis (showing that an exposure is associated with the outcome only in participants with certain characteristics). These potential limitations can be mitigated by the formulation of a priori hypotheses, statistical correction for multiple comparisons, and formal tests for interaction. Although associations that seem to be biologically plausible might be more likely to be correct, speculative potential mechanisms often are relatively easy to devise even when no relation between exposure and outcome exists. Conversely, insisting on biologic plausibility may lead to rejection of a novel (but correct) association with little supporting evidence.

Establishing and Maintaining Renal Cohorts

Establishing and maintaining a cohort study can be costly, time consuming, and resource intensive. Identifying the specific group of individuals (e.g., incident [12] versus prevalent [13] dialysis patients) to be followed and defining and improving the quality of the exposures to ascertain at baseline and follow-up represent some of the early challenges. An efficient mechanism to capture exposures is to “convert” previous randomized trials into follow-up cohort studies, taking advantage of the enormous amounts of data that were collected in the past (14,15). As is the case with all cohort studies, supporting the maintenance of the cohort, capturing all relevant outcomes, and minimizing loss to follow-up represent some of the ongoing challenges. Widely known prospective cohort studies that were established primarily to study “nonrenal” outcomes but subsequently examined renal end points include the Framingham Heart Study (16) and the Nurses’ Health Study (17). The Framingham Heart Study assembled >5000 individuals who were free of heart disease in 1949, examined each biennially for evidence of coronary heart disease, and reported strong associations between specific baseline exposures (e.g., BP, cholesterol) and coronary heart disease over >30 yr of follow-up. With an abundance of exposure information, including renal function and proteinuria, the Framingham investigators have
examined renal outcomes in both a cross-sectional (18) and a prospective (16) manner. Given the strengths of cohort studies, specific “renal” cohort studies also have emerged in the past decade (e.g., Choices for Healthy Outcomes in Caring for ESRD [CHOICE] [12], African American Study of Kidney Disease and Hypertension [AASK] [14], Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease [CRISP] [19]). In the spirit of emphasizing specific points mentioned above, rather than being exhaustive, we highlight only two prospective cohort studies next.

Chronic Renal Insufficiency Cohort

In response to the rising epidemic of CKD and its associated consequences, namely cardiovascular disease and progression to ESRD, the National Institute of Diabetes and Digestive and Kidney Diseases established the Chronic Renal Insufficiency Cohort (CRIC) Study in 2001 (20). Overarching goals for this prospective cohort study focused on identifying nontraditional risk factors for cardiovascular morbidity and mortality and risk factors for progression toward ESRD. Recruitment for CRIC has been under way since 2003, with the goal of recruitment to reach 3000 racially and ethnically diverse participants who are aged 21 to 74 and have a wide variety of chronic renal diseases. Entry criteria targeted individuals with specific GFR measurements ranging from 20 to 50 or 70 ml/min per 1.73 m² (depending on age), and the follow-up time is 5 yr. Exclusion criteria included patients with polycystic kidney disease, those who are on active immunosuppression for glomerulonephritis, and those who have HIV or malignancy. Data will be collected via yearly in-person study visits as well as telephone contact every 6 mo. Baseline measures include anthropometric parameters, measures of renal function and proteinuria, and measures of quality of life. Outcomes will focus on measures of renal function and cardiovascular disease. Methods for cohort retention and assessment of quality of measures are incorporated. Therefore, for individuals with CKD, CRIC represents a classic prospective cohort study with 5-yr follow-up. Establishment and maintenance of this important cohort study justifiably will require a multitude of resources.

Accelerated Mortality in Renal Replacement

In 2004, Thadhani et al. initiated a national prospective cohort study of incident hemodialysis patients who were receiving chronic hemodialysis in one of >1000 dialysis facilities throughout the United States that are operated by Fresenius Medical Care North America (FMC). The primary goal of the study (funded initially by industry and then by the National Institute of Diabetes and Digestive and Kidney Diseases) is to identify risk factors and potential mediators that are involved in the Accelerated Mortality in Renal Replacement (ArMORR). At baseline (within 14 d of initiating chronic hemodialysis) and every 90 d thereafter, demographic and dialysis-related characteristics, standard laboratory tests, and leftover serum and plasma samples are collected. Usual blood samples are sent to Spectra East in New Jersey, the central laboratory for most specimens that are processed and analyzed by FMC, and remnant specimens are stored for use in future analyses. All incident patients, regardless of age, race, or cause of ESRD, were included, and 10,018 participants with baseline blood samples have enrolled as of July 2005. Each patient will be followed for 1 yr from the initiation of dialysis. The strength of this prospective cohort study is the ability to examine alterations in novel biomarkers that may antedate hard outcomes such as mortality. All exposures and primary outcomes (e.g., mortality) are collected prospectively, and measurement of biomarkers using stored specimens will minimize bias because all samples are collected in a similar manner without knowledge of exposure or outcome status. Because all patients who remain in the FMC system have their data and blood collected on a routine basis and detailed information on patients who leave the FMC system (e.g., transfer, transplant, death) is available, information is expected to be relatively complete. Like all cohort studies, however, this study does not collect information on all potential confounders, is subject to confounding by indication because treatments are not assigned randomly, and is subject to other biases, including loss to follow-up.

Conclusion

Cohort studies, in particular prospective cohort studies, offer several important advantages over other forms of observational studies. Cohort studies remain at the “top” of the hierarchy of observational studies linking exposure to outcome. Indeed, well-executed cohort studies often yield results that resemble those from RCT. Furthermore, for certain hypotheses, RCT may not be possible for ethical, logistical, or economic reasons. Nevertheless, cohort studies remain susceptible to bias and confounding, although these certainly can be mitigated with proper study design. Readers of the medical literature should be familiar with cohort studies because such studies are commonplace. In particular, readers should understand their strengths and limitations and appreciate when cohort studies in general or a specific cohort in particular (because not all cohorts are established and maintained in a similar manner) can or cannot appropriately test the hypothesis in question. Certainly, the support for any link between exposure and outcome rarely rests on a single study, regardless of study design, and requires aggregate support from a multitude of biologic, translational, and clinical studies.

Acknowledgments

This study was supported by grants DK71674 and HD39223 (R.T.) and Alberta Heritage Foundation for Medical Research and Canadian Institutes for Health Research (M.T.).

References

3. Teng M, Wolf M, Lowrie E, Ofsthun N, Lazarus JM,


