Clinical studies of drug effects in humans

WAYNE A. RAY

Many important drug properties cannot be predicted from in vitro or animal studies but must be quantified from studies of clinically relevant endpoints in humans. Clinical study designs include the randomized controlled trial (RCT) and a variety of observational designs. In RCTs, randomization usually ensures treatment group comparability with respect to other factors, so outcome differences reflect treatment differences per se. Properly conducted RCTs thus are the strongest study design, the mainstay of mandatory premarketing studies, and essential for evaluation of therapeutic efficacy. RCT limitations include frequent use of surrogate endpoints, limited power, short-term follow-up, and cross-contamination of study groups. For ethical reasons, some questions cannot be studied with RCTs. The simplest observational design, the case series, has limited value because it lacks a denominator. Limitations of cohort and case-control studies include misclassification, selection bias, and confounding. Despite their limitations, properly conducted experimental and observational clinical studies provide essential data for clinical practice.

INDEXING TERMS: randomized controlled trial • cohort study • case-control study • Reye syndrome • thalidomide • myalgia • nonsteroidal antiinflammatory drugs • L-tryptophan • eosinophilia • cancer, endometrial • prostate resection

Many of the most important effects of drugs in humans cannot be predicted from in vitro or animal studies. Much of the data needed to guide therapy must come from clinical studies of the effects of drugs on clinically relevant endpoints in humans. The objectives of this overview are (a) to illustrate the need for clinical studies, (b) to describe the major types of study design, and (c) to discuss common methodological problems that arise in conducting clinical studies. These topics are complex and have many aspects that remain controversial. In this brief review I present some of the more important issues that clinical researchers encounter in studies of drug effects in humans.

Need for Clinical Studies

The following are examples of the numerous important drug effects discovered or quantified through clinical studies.

Attrition of promising compounds in premarketing testing. In the drug development process, preclinical studies, including in vitro and animal studies, are used to identify promising compounds for marketing. Largely as set forth in the 1962 Kefauver-Harris amendments, these compounds undergo extensive testing before they are approved for marketing. Phase I studies are conducted in small (n = 20–80) [1] numbers of normal subjects, unless the drug is unacceptably toxic, in which case these studies are done in patients. The objectives include elucidation of pharmacokinetics, dose-finding, and establishment of safety. About 70% of compounds pass Phase I testing [1]. Phase II studies consist of therapeutic trials in small numbers (n ≤200) of selected patients. The goals are further understanding of kinetics, dose-finding, and initial assessment of therapeutic efficacy. Parallel long-term studies in animals are conducted to identify potential harmful effects of chronic use. About one-third of the tested compounds pass Phase II testing. Other agents either have unacceptable toxicity or lack efficacy. Phase III studies are larger clinical trials that generally include 500–3000 patients. The goal is to quantify efficacy in more representative populations of patients and to ensure that toxicity is acceptable. About 25–30% of compounds complete Phase III [1].

After Phase III is completed, the manufacturer submits a New Drug Application, which includes the data generated in the clinical testing phases. After review and approval of the New Drug Application by the US Food and Drug Administration (FDA), which usually involves specific advisory committees of independent scientists and may take several years, marketing of the product can begin.1 Marketed products represent ~5–6% of the compounds for which clinical studies were undertaken.

Thalidomide. Thalidomide was introduced in 1958 as a hypnotic. Its primary advantages relative to the barbiturates (the most common hypnotics of that era) were a low rate of residual side effects and a high therapeutic index. The drug was used widely throughout the world, but because of the concerns of an FDA

1 Nonstandard abbreviations: FDA, US Food and Drug Administration; NSAID, nonsteroidal antiinflammatory drug; and RCT, randomized controlled trial.
reviewer over reports of peripheral neuropathy, thalidomide was not approved in the US. Then, in the early 1960s, case reports of phocomelia (a severe birth defect characterized by marked foreshortening of the extremities) in conjunction with use of thalidomide in early pregnancy began to appear. Epidemiological analysis clearly established that thalidomide was responsible for this epidemic; after the drug was withdrawn from the market, the epidemic abated (Fig. 1). An estimated 10,000 children worldwide were born with this birth defect.

Why did this occur? Pregnant women are excluded from nearly all clinical trials. Before this episode, routine premarketing testing did not include teratogenicity studies in animals. Although extensive animal studies are conducted now, the need remains to conduct clinical studies to monitor newly introduced drugs for fetal effects.

Aspirin and Reye syndrome. Reye syndrome is an acute noninflammatory encephalopathy characterized by biochemical abnormalities and pathological liver changes. It was first described in 1963 by Anderson and Reye in Australia from their studies of children, most of whom had had an antecedent febrile illness. The pathological findings led Reye and colleagues to speculate that the syndrome was caused by an exogenous toxin. Although the syndrome was rare (<5 per 100,000 children/year), series of cases studied in other populations suggested an etiological role for salicylates. Ultimately, this hypothesis was confirmed by several epidemiological studies, suggesting that aspirin (or other salicylates) caused the syndrome. The US Public Health Service Study of aspirin and Reye syndrome [2] showed a 24-fold greater risk of contracting Reye syndrome among children with a qualifying febrile illness who were treated with aspirin (Table 1).

Other effects of aspirin and other nonsteroidal antiinflammatory drugs (NSAIDs). The antipyretic and analgesic properties of willow bark have been recognized for centuries [3]. In 1899 the acetylated derivative acetylsalicylic acid was introduced under the name aspirin and became widely used throughout the world. Beginning with phenylbutazone in 1946, a host of other antiinflammatory agents with a similar mechanism of action were introduced. NSAIDs all suppress biosynthesis of the prostaglandins, eicosanoids that mediate the inflammatory response.

NSAIDs are among the most frequently consumed medications in the US. About 75 million prescriptions for NSAIDs are now written annually in the US (4.5% of all prescriptions), accounting for ~$2.5 billion of drug sales [4–6]. Given the recent wide use of these medications, one might speculate that their health effects would be well understood. In fact, a review of epidemiological history records the discovery of multiple major effects (some adverse, some beneficial) of these medications (Table 2) after their marketing.

Thus, when a drug is considered for use in humans, many of its important effects may be unknown. Even after the drug is approved for marketing, ongoing clinical monitoring is still necessary.

---

**Table 1. Aspirin and Reye syndrome: results of US Public Health Service study.**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Salicylate</th>
<th>No salicylate</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of cases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reye syndrome cases (N = 27)</td>
<td>96</td>
<td>4</td>
</tr>
<tr>
<td>Controls (N = 140)</td>
<td>37.9</td>
<td>62.1</td>
</tr>
</tbody>
</table>

**Table 2. Some health effects of aspirin and other NSAIDs discovered after marketing.**

<table>
<thead>
<tr>
<th>Drug</th>
<th>Outcome</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspirin</td>
<td>Reye syndrome</td>
<td>Adverse: RR* &gt; 10</td>
</tr>
<tr>
<td>Aspirin</td>
<td>Acute myocardial infarction</td>
<td>Beneficial: RR ~ 0.5</td>
</tr>
<tr>
<td>Aspirin</td>
<td>Gastric ulcer</td>
<td>Adverse: RR ~ 2</td>
</tr>
<tr>
<td>NSAIDs</td>
<td>Peptic ulcer</td>
<td>Adverse: RR ~ 4</td>
</tr>
<tr>
<td>NSAIDs</td>
<td>Acute renal failure</td>
<td>Adverse</td>
</tr>
<tr>
<td>Aspirin</td>
<td>Cerebral thrombosis in TIA</td>
<td>Beneficial</td>
</tr>
<tr>
<td>Aspirin</td>
<td>Colon cancer</td>
<td>Beneficial?</td>
</tr>
<tr>
<td>NSAIDs</td>
<td>Lower gastrointestinal disease</td>
<td>Adverse?</td>
</tr>
<tr>
<td>Aspirin</td>
<td>Asthma</td>
<td>Adverse?</td>
</tr>
<tr>
<td>Aspirin</td>
<td>Cataracts</td>
<td>Adverse?</td>
</tr>
<tr>
<td>NSAIDs</td>
<td>Antihypertensive efficacy</td>
<td>Adverse?</td>
</tr>
<tr>
<td>NSAIDs</td>
<td>Hepatotoxicity</td>
<td>Adverse?</td>
</tr>
<tr>
<td>NSAIDs</td>
<td>Septic shock</td>
<td>Beneficial?</td>
</tr>
</tbody>
</table>

* RR, relative risk; TIA, transient ischemic attack.
needed. This process involves a clear interplay between the basic and clinical sciences, often with suggestions of hypotheses by the former and confirmation in humans by the latter.

**Study Designs**

The primary study designs available for population-based research include the randomized controlled trial (RCT) and a variety of observational designs (Fig. 2).

**Randomized controlled trials.** In experimental studies the investigator manipulates the factor of interest. In studies of drug effects this process generally involves random assignment of patients to treatment conditions, e.g., RCTs. Randomization produces groups that are comparable with respect to other factors; thus, if the trials are properly conducted, any differences in outcome between the treatment conditions reflect differences between the treatments per se. Accordingly, properly conducted RCTs are considered the strongest study design [7] and provide the best evidence of whether a drug has an effect. RCTs are the mainstay of premarketing studies required by the drug-regulatory authorities [1, 8, 9] and are essential for evaluation of nearly all questions of therapeutic efficacy.

**Cohort studies.** The structure of a cohort study resembles that of an RCT. Groups of subjects with and without the exposure of interest are identified. The cohort at baseline usually is free of the disease under study. The groups are then followed over time to ascertain outcomes, and rates between the two groups are compared.

One of the first studies of the adverse effects of smoking was a cohort study of British physicians conducted by Doll and Peto [10]. British physicians received a questionnaire that sought information on smoking habits. Death certificates were used to identify fatal lung cancer; analysis demonstrated a 10-fold increase in mortality rates among heavy smokers.

A more contemporary study is the Nurses' Health Study, a prospective cohort study of heart disease, cancer, and other serious illness in women [11]. In 1976, a questionnaire was mailed to all 172,000 female, married registered nurses residing in 11 US states. The 121,700 women who responded became the Nurses' Health Study cohort. One objective of this study was to assess the effects of medications on the above diseases; thus, baseline and follow-up data were collected that classified the women by drug exposure. These women have been followed up at regular intervals by questionnaire to determine changes in baseline variables and to detect occurrence of study diseases.

**Case–control studies.** The case–control study identifies groups of subjects with and without the disease of interest and determines the prevalence of exposure before disease onset in subjects who have the disease and exposure prevalence during a comparable time period for the control subjects. For example, when the New Mexico Health Department reported three patients, all of whom used l-tryptophan, with acute-onset eosinophilia and severe generalized myalgia, the Centers for Disease Control and the New Mexico Health Department conducted a case–control study [12]. Eleven patients with idiopathic eosinophilia and incapacitating myalgia were identified from hematology records of state laboratories and review of patients' medical records. Also, 22 neighborhood controls were selected, matched for age and sex. All of the patients (100%) but only two of the control subjects (9%, P ≤0.00002) had used l-tryptophan before the index date. This study, initiated in November 1989, was reported in the *Lancet* in March 1990 [12].

There are two justifications for this design. First, intuition suggests that if exposure is much more frequent in persons with a disease, then this exposure is a cause of the disease. That 100% of patients with acute-onset eosinophilia but only 9% of the control subjects had used l-tryptophan is highly suggestive. Second, under certain circumstances, a case–control study is equivalent to a cohort study in which sampling of nonaffected controls is used to increase efficiency [13].

**LIMITATIONS OF RCTS**

**Surrogate vs clinical endpoints.** For logistic reasons, most RCTs have been of moderate size (<3000 patients) and duration (<1 year). For many drugs this limitation necessitates the study of surrogate rather than clinical endpoints: blood pressure and lipid concentrations instead of myocardial infarction; CD4 concentrations instead of the opportunistic infections themselves; and bone mass instead of fractures. The choice of surrogate endpoints is generally based on clinical or epidemiological evidence linking the surrogate measure with the relevant clinical outcome. However, this reasoning can be misleading. Other drug effects may counterbalance the therapeutic drug effect. For example, clofibrate decreased cholesterol and coronary artery disease; however, total mortality increased because of increased cancer deaths. Also, our understanding of the meaning of surrogate endpoints may be incomplete. For example, in the Cardiac Arrhythmia Suppression Trial [14], drugs that prevented premature ventricular depolarizations (considered to be a risk factor for life-threatening ventricular arrhythmias) actually increased the risk of death. The issue of the
meaning of surrogate endpoints also often arises in the interpretation of animal carcinogenicity studies.

Sample size. Inadequate power resulting from inadequate sample size is a common problem in study design because the high cost of clinical trials discourages large sample sizes. For example, the Cooperative Veterans' Administration study of vaccine efficacy for prevention of pneumococcal pneumonia in the elderly [15] contained a controversial case definition that was dominated by "pneumococcal bronchitis." This probably reflected the sample size (n = 1179 and 1175 for treatment and placebo, respectively), which was inadequate for the study of proven pneumococcal pneumonia (i.e., bacteremia) because only three such cases were noted. Thus, the finding of a lack of vaccine efficacy in this study was not widely accepted. However, case-control studies in elderly subjects, which did include an adequate number of cases of proven infections (n = 1054), estimated a 56% protective efficacy against strains present in the polyvalent vaccine [16].

Infrequency of adverse effects. Another implication of limited premarketing sample sizes is the inability to detect infrequent adverse events, as illustrated in Fig. 3, which shows sample size requirements for clinical trials in relation to frequency of adverse events. Thus, detection of associations between oral contraceptives and vascular disease, NSAIDs and peptic ulcers, and psychotropic drugs and injuries have all required large, careful, observational studies.

Limited follow-up. In RCTs it is difficult to follow subjects for long enough periods to detect long-latency adverse effects. The potential importance of such effects is illustrated by the cohort study of Ron et al. [17] on the association between radiotherapy for tinea capitis (ringworm) and nervous system tumors. Medical records were used to identify 10 384 Israeli children (the E+ group) who had received such treatment between 1948 and 1960. For each child, the dose of radiation to the brain was estimated from information in the records on therapeutic dose, treatment techniques, number of courses of therapy, and simulation studies, performed with a simulated model of a 6-year-old child. The unexposed group (E−), identified from the Central Population Registry, consisted of 10 384 unrelated children who had not been irradiated and 5392 siblings who had not been irradiated. For each child, follow-up began on the date of the first treatment and extended through 1982, the date of death, or the date of first tumor. The total number of person-years of follow-up was 711 739, an average of 26.4 years per child. During the follow-up period, 73 tumors (validated when possible by review of pathology reports) of the brain or nervous system were ascertained from the Israel Cancer Registry, hospital pathology records, and death certificates. The investigators found a striking association between estimated doses to the brain and tumor incidence (Fig. 4).

This study illustrates the advantages of observational cohort studies. Logistic considerations would make it virtually impossible to conduct an experimental study with 26 years of follow-up per subject.

Exposure differential and noncompliance/contamination of controls. To preserve the inferential benefits of randomization, the analysis must be performed on an "intention to treat" basis. That is, during the analysis, patients must remain in the groups to which they were initially assigned. However, the exposure

---

2 This point may seem counterintuitive, because persons who in fact did not take a drug were kept in the active-treatment group for the analysis. This is necessary because the factors associated with compliance may also be associated with outcomes. Studies of lipid-lowering drugs have shown impressive reductions in cardiovascular disease outcomes when patients who complied with the treatment regimen were compared with noncompliers, even when the "treatment" was a placebo. Another common circumstance occurs if a subject changes groups as related to prognosis, e.g., if patients who deteriorate in the control group are switched as a last resort to the active-treatment group. Also, a treatment group loaded with poor-prognosis patients may obscure a beneficial treatment effect. Thus, when this occurs, in essence, analysis of an RCT according to what drug people take turns the RCT into a cohort study, with the necessity to control for numerous possible selection factors. In practice, if RCTs do not achieve high compliance rates, they do not provide a fair test of the study hypothesis. That is why measures to ensure patient compliance are central to executing a good RCT.
difference between groups often is less than intended. This usually occurs when those in the active-treatment group do not take the drug (noncompliance) but can also result if some control subjects receive the active drug. The effect of this behavior is to bias the findings toward the null hypothesis of no difference between exposure groups. This factor is a potential explanation for a negative finding in an RCT. In contrast, observational studies can compare patients on the basis of the drugs that they actually received.

Ethical issues. The RCT design is particularly difficult to use when the primary question is toxicity. For example, it probably would not have been possible to randomize children to receive either aspirin or acetaminophen merely to determine whether the incidence of Reye syndrome was higher in the group who received aspirin.

In summary, RCTs are the strongest type of clinical study design. However, a variety of circumstances commonly encountered in studies of drug effects can often limit the utility of the tests. Under such conditions, a poorly conducted RCT is likely to provide lower-quality evidence than properly conducted observational studies.

LIMITATIONS OF OBSERVATIONAL STUDIES

Denominator-free reasoning. Three physicians practicing in a pediatric clinic at Travis Air Force Base in California reported seeing seven children with marrow failure in an 8-year period [18]. Six had aplastic anemia and one had acute lymphoblastic leukemia. All had been exposed to household insecticides containing organophosphates (Raid Ant and Roach Killer in five cases) in the 28 weeks before diagnosis. From these data, the authors concluded that these agents occasionally cause severe hematological toxicities.

This type of analysis is common in the clinical literature. However, although such case reports may provide valuable clues, in themselves they usually are inconclusive because they lack a control group. In the present example (e.g.), insecticide use may be very common in military housing.

Misclassification. Observational studies require measurement of the drug under study [the exposure (E)], the outcome [the disease (D)], and other factors that influence disease risk. Although this aspect of an observational study is conceptually simple, operationally it can be one of the most difficult [19]. One of the problems encountered in measuring drug use is illustrated by the study of Honkanen et al. [20], a case–control study of motor-vehicle crashes. Cases were identified in the emergency room; controls were selected from drivers at filling stations near the crash scene. The investigators compared serum tests for benzodiazepines with interview data. Of the 17 patients with positive serum tests, only 5 (29%) reported benzodiazepine use. Similarly, Morris et al. [21] reported that 5 of 13 patients with positive urine screens for benzodiazepines, barbiturates, or opiates interviewed at the clinic admitted using these substances. Although problems with interview accuracy can be avoided by biological measurement techniques, these methods have substantial limitations for observational studies, e.g., objections by subjects (particularly those who are disease free), inability in most cases to study past exposures, expense, lack of practical assays for all drugs, reliance on measurement of metabolites that may be common to several drugs in the same class, cross-reactivity of related compounds, and inability to measure drug concentrations at the relevant receptor sites.

Selection bias. Virtually all studies of replacement estrogen use have shown that use of unopposed estrogens in women with an intact uterus is associated with a striking (often 10-fold or greater) increase in the risk of endometrial cancer. However, one case–control study [22] indicated that estrogen users had no increased risk of endometrial cancer: The cases were 56 patients admitted from 1963 through 1965 at the University of Iowa Hospital’s obstetric and gynecology service; the controls were 86 women admitted to the same service for postmenopausal bleeding, but without cancer or hyperplasia. The proportions of estrogen use among the patients and the controls were 28.6% and 27.5%, respectively, suggesting no association. The error in this study, however, is that the control subjects are not a random sample of the population from which the patient cases were drawn. Rather, a selection bias favored estrogen users, because estrogen use predisposes women to vaginal bleeding.

Confounding. Confounding, the distortion of a disease–exposure association brought about by the association of other factors with both the disease and exposure [23], is perhaps the central problem of observational studies. Confounding occurs when other differences between the exposed and unexposed groups are responsible for the observed difference in disease rates. Consider the increased rate of suicide among persons who take antidepressant medications. Is this an effect of the drugs per se? A more plausible explanation is that depression is a confounder. Compared with those who are not depressed, those with depression (X in Fig. 5) have a greater relative risk (R1 >1) for suicide (D). However, persons with depression also are more likely (p >0, where p is the correlation between being depressed and taking antidepressants) to take antidepressants (E), thus creating the association (R2 >1) between drug and disease.

For confounding to be present, the confounder must itself be a risk factor for the disease (R1 ≠ 1). However, the confounder must also be associated with the exposure (p ≠ 0) in the population under study that is free of disease. The spurious association induced by confounding (R2 in Fig. 5) is generally substantially less than the true association between confounder and disease (Breslow and Day [23]). Thus, confounding is most plausible for small effects.
If potential confounders can be measured, then one can use various statistical techniques to control for their influence. Thus, the key to resolving this problem is identifying the possible confounders and measuring them.

The controversy over treatment for benign prostatic hyperplasia illustrates the potential role of confounding. In elderly men, surgical treatment of benign prostatic hyperplasia is common. Since the 1970s, transurethral resection of the prostate has replaced open surgery as the treatment of choice, presumably because the former has lower operative morbidity and permits shorter hospital stays. However, Roos et al. [24] noted a paradoxical increase in mortality associated with transurethral resection of the prostate. Because this study was conducted with the use of large insurance claims databases, the investigators were unable to measure precisely other illnesses in patients who were undergoing this surgery. A subsequent study by Concato et al. [25] suggested that the increased mortality among the resected patients was due to higher comorbidity in this group.

In conclusion, this review provides only a broad overview of the very extensive topic of clinical studies, and many important issues have not been discussed. Although the limitations of clinical studies have been discussed to underscore the complexity of this topic, the reader should not be nihilistic about the utility of such studies. Despite their limitations, properly conducted experimental and observational clinical studies provide data that form the bedrock of clinical practice. Students of drug effects in humans must understand these designs and their appropriate use.

This research was supported in part by a cooperative agreement with the Food and Drug Administration (FD-U-000073).

References