Volume 57, Number 9
OBSTETRICAL AND GYNECOLOGICAL SURVEY

SUPPLEMENT

Clinical Research in Obstetrics and Gynecology: A Baedeker for Busy Clinicians

David A. Grimes, MD and Kenneth F. Schulz, PhD, MBA

From Family Health International, Research Triangle Park, North Carolina

INTRODUCTION

In the early 1800s, Karl Baedeker, a German publisher, launched a series of travel guidebooks. By the twentieth century, the guidebooks had achieved such international fame that his name became synonymous with the genre. As clinicians wade into the flood of clinical research being published, a guidebook can be a handy navigational aid. In this article, we offer a Baedeker for reading the literature, an approach distilled from our three decades of clinical practice and research experience. Interested readers can find more detail in our recent series on research methods in *The Lancet* (1–11).

Reading research is mandatory if a clinician is to keep up. With greater age and experience, clinical practice should improve. Paradoxically, however, greater age and clinical experience often translate into rusty practice. As has been shown for treatment of hypertension, one of the strongest determinants of appropriate practice is number of years since medical school graduation; stated alternatively, practice quality deteriorates over time (12, 13). Keeping current is difficult after leaving formal training, and that difficulty may be greater for those who practice in smaller communities (14). If one cannot (or chooses not to) read, then one's practice is condemned to becoming obsolete. This indirectly hurts patients.

A second benefit of critical reading of clinical research is appropriate adoption (or rejection) of new technologies. Obstetrics and gynecology has a long, blemished record of adoption and dissemination of new tests and procedures without evidence of benefit (15). Episiotomy, one of the most common operations performed on adults in the last century, swept into practice based on DeLee's analogy that childbirth is tantamount to impalement on a pitchfork (16). Urinary estriol measurement to monitor a fetus thought to be in jeopardy has been replaced by an even more expensive and cumbersome test (nonstress testing) for which no evidence of benefit exists either (17). Electronic fetal monitoring took U.S. obstetrics by storm in the absence of demonstrable benefit; a quarter century of study has failed to show any lasting benefit to babies (18), and the poor predictive value of worrisome tracings has needlessly driven up the cesarean delivery rate. Liquid-based cervical cytology screening has not been shown to reduce cervical cancer incidence or mortality, and the cost per case of cancer detected is higher with this approach than with conventional cytology (19). Ironically, poor women at highest risk of this cancer may not be able to afford the screening (20). Reports of new laparoscopy operations have recently been retracted by an editor, because the reported information could not be corroborated (21, 22). This hurt patients as well.

While reading clinical research is clearly important, the task is daunting. First, the volume being published is overwhelming, with an estimated 25,000 biomedical journals in print. One challenge is picking and choosing what to read. In general, most readers should limit themselves to articles that are both relevant to their practices and likely to be of high scientific value. These two criteria will immediately narrow the field.

Once an article is selected, another problem emerges: many clinicians in obstetrics and gynecology report that they cannot critically read the literature (23). Our graduates leave their training full of the

Correspondence to: Family Health International, P.O. Box 13950, Research Triangle Park, NC 27709. Email: dgrimes@fhi.org
DOI: 01.OGX.0000027851.57565.C7

outputs of science (stuffed like overstuffed sofas), yet they are not sufficiently scientific in their approach either to the literature or to their practices. Stated alternatively, scientific illiteracy remains a stubborn problem and a major failing of medical education (24). To address this problem, this guide will outline the major types of clinical research, describe two fundamental questions about validity, outline a simple four-point check list for readers, describe the interpretation of common measures of association, and highlight the rules of conduct for performing and reporting randomized controlled trials. Although our focus will be research involving humans, the same principles apply to animal and laboratory research as well.

TWO APPROACHES TO CLINICAL QUESTIONS

Should the peritoneum at the vaginal cuff be closed as one completes an abdominal hysterectomy? Gynecologists face this question hundreds of thousands of times annually in the U.S. The answer may influence operating time, morbidity, and speed of recovery. Hence, clinicians need to know the answer. One approach is **total enumeration**, study every abdominal hysterectomy nationwide, tally their outcomes, and see which closure technique is preferable. This "census" approach is generally impractical for logistical reasons. Hence, the more common tactic is to select a **sample** of patients having the cuff left open or sutured shut, study their outcomes, and then infer from the sample to the broad population of women. Regarding cuff closure, studies using different types of samples have addressed this question, and the answer remains unclear (25–27).

Are the Findings Valid?

When readers encounter research results, they need to consider two basic questions. First, did the study measure what it set out to measure? This characteristic is termed **internal validity**. It can be undermined by several types of bias described below. Assuming the study results have internal validity, the next question to answer is **external validity**: can the results be extrapolated (generalized) to one's patients?

Herein lies a paradox. Observational studies, which reflect the usual practice of medicine, are vulnerable to bias but are more representative of women in general than are participants in randomized controlled trials. The latter are all volunteers who pass inclusion criteria. In contrast, randomized controlled trials, if properly done, are more immune to bias but are less likely than observational studies to represent rank-and-file patients. Observational research may have poorer internal validity and better external validity than randomized controlled trials; the opposite is true for randomized controlled trials. Obviously, if a study lacks internal validity, generalizing these invalid results is worthless and possibly misleading.

Measuring Dichotomous Outcomes

The terminology used in measuring dichotomous outcomes (e.g., sick or well) is often confusing and inconsistent. This makes reading research reports needlessly difficult. For example, many investigators misuse the simple term **rate**. Probably the most notable example is the misnomer "maternal mortality rate," which has appeared in textbooks and research reports for decades.

As shown in Figure 1, a **ratio** is the product obtained by dividing one number by another. Whether the numerator and denominator are related determines the type of ratio. As shown on the right side of Figure 1, if the numerator is not included in the denominator, then the product remains a ratio. In the maternal mortality example, a woman who dies of complications of a complete hydatidiform mole would be included in the numerator but not the denominator (women with live births). Hence, the venerable maternal mortality "rate" is, in fact, a ratio.

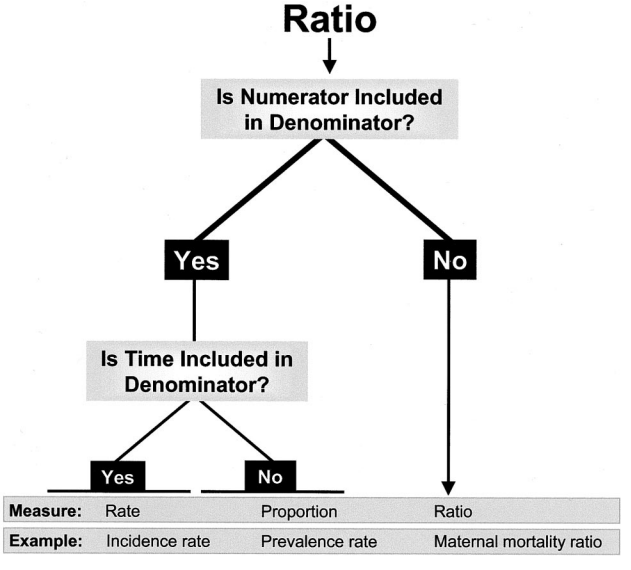

Fig. 1. Algorithm for distinguishing rates, proportions, and ratios. Reprinted with permission from Elsevier Science (Lancet 2002;359:57–61).

Another example would be the abortion ratio: the number of induced abortions divided by the number of live births in a population.

If the numerator is included in the denominator, the product may be either a rate or a proportion. A **rate** indicates the risk of an outcome in a population as a function of time. Rates have two hallmarks: units of time and a multiplier. An example would be the incidence rate of syphilis in 1999: 13.2 cases per 100,000 population per year (28). In contrast, a **proportion** does not have a time element. An example would be the proportion of adult women in the U.S. who were cigarette smokers in 1999: 21.6 per 100 women, or 21.6% (28). For both rates and proportions, all those in the numerator are included in the denominator.

MEASURES OF ASSOCIATION

A less informative way of comparing two groups being studied is **hypothesis testing**. The investigator assumes that no difference exists between the two groups and then sees if the results in the sample studied are consistent with that assumption. If the P value is less than .05, then the null hypothesis is rejected, and a real difference is assumed. This approach of raising (and destroying) a straw man is not intuitive to clinicians. . .or to their patients.

Hence, the preferred way of comparing two groups that have dichotomous (i.e., positive or negative) results is by **interval estimation** (29). Relative risks and odds ratios are the usual measurements reported. The **relative risk** (**RR**) (also termed a rate ratio or risk ratio) is simply the rate of outcome in the exposed group divided by the rate in the unexposed group. Ratios higher than 1.0 imply an increased risk, and vice versa. Because the units (e.g., per 100 patients) divide out in the calculation, relative risks and odds ratios have no units (e.g., 3.2).

The **odds ratio** (**OR**) is the usual measure of association in case-control studies. This term indicates the odds of exposure among the cases (those with the condition) divided by the odds of exposure among the controls (those without the condition). As shown in Figure 2, the odds of exposure among cases is a/b and that among controls is c/d, so the ratio is (a/b)/(c/d). Algebraic division yields the final formula for the odds ratio: ad/bc, which is also termed the **cross products ratio** (30).

The interpretation of the odds ratio is analogous to that of the relative risk: ratios higher than 1.0 imply increased risk, and those below a protective effect. When the condition being studied is uncommon (say, <5%), then the odds ratio becomes a good proxy for the true relative risk. Hence, the two terms are often used interchangeably. In reality, one cannot calculate a true relative risk for a case-control study, since the denominator of persons at risk is not known.

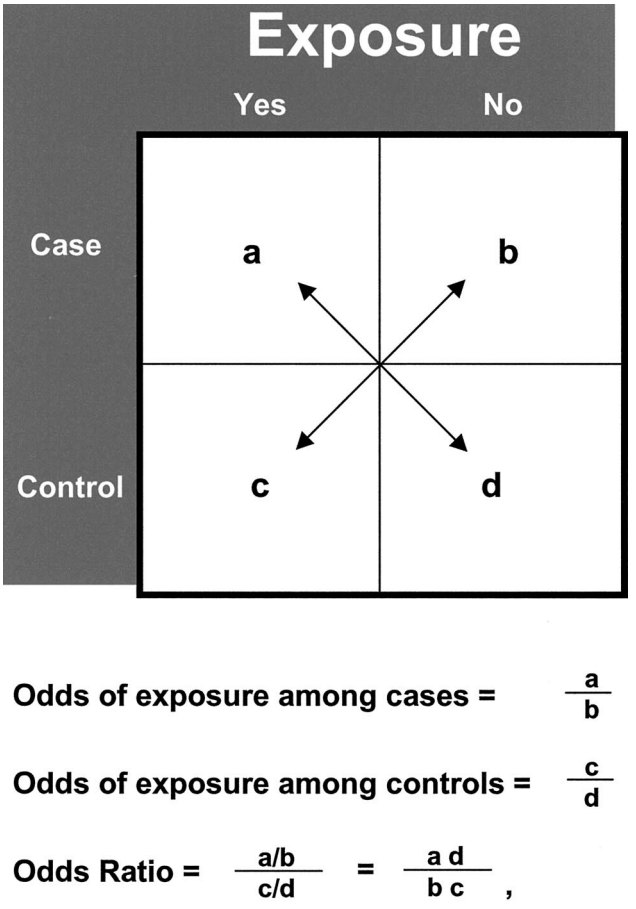

Fig. 2. The cross-products ratio (odds ratio) in a case-control study.

In any study, the result is only guess (often termed a **point estimate**) as to what is happening in the larger population. Accordingly, readers need to know how much precision the estimate has. To answer this question, researchers usually calculate **confidence intervals (CI)** around relative risks and odds ratios. These intervals designate a range of plausible values. The wider the confidence interval, the less precise is a result, and vice versa. If a study is repeated 100 times with the same sample size, then 95 times out of 100 the true value of the relative risk or odds ratio will fall within the confidence interval. In addition, if a 95% confidence interval does not cross 1.0, then the difference observed is statistically significant at

the traditional .05 level. However, using the confidence interval as a sneaky approach to hypothesis testing is inappropriate (31).

TAXONOMY OF RESEARCH TYPES

An important first step to reading a research report is to figure out what type of study has been done. This can be difficult for two reasons. Often, investigators do not explicitly say what they have done, or they use cumbersome verbiage to describe simple study designs. For example, "open-label, single-arm, multicenter, clinical trial" is a recent 15-syllable description of a case-series report. A second, more worrisome obstacle is that investigators sometimes do not know what they have done. The most common error may be calling a **retrospective cohort study** a case-control study (32–34). That such mistakes survive the editorial process is more concerning.

Deducing the study type is fundamentally important. The type of study (its research anatomy) dictates what it can do (its physiology). In addition, for some types of research (the **randomized controlled trial**), a well-established set of rules exists, so the reader can easily check to see that the rules of conduct were obeyed (35, 36).

The universe of all clinical research can be neatly divided into two principal categories: **experimental** or **observational** (Fig. 3). How the patients got their treatments (or other exposures) separates the two. In

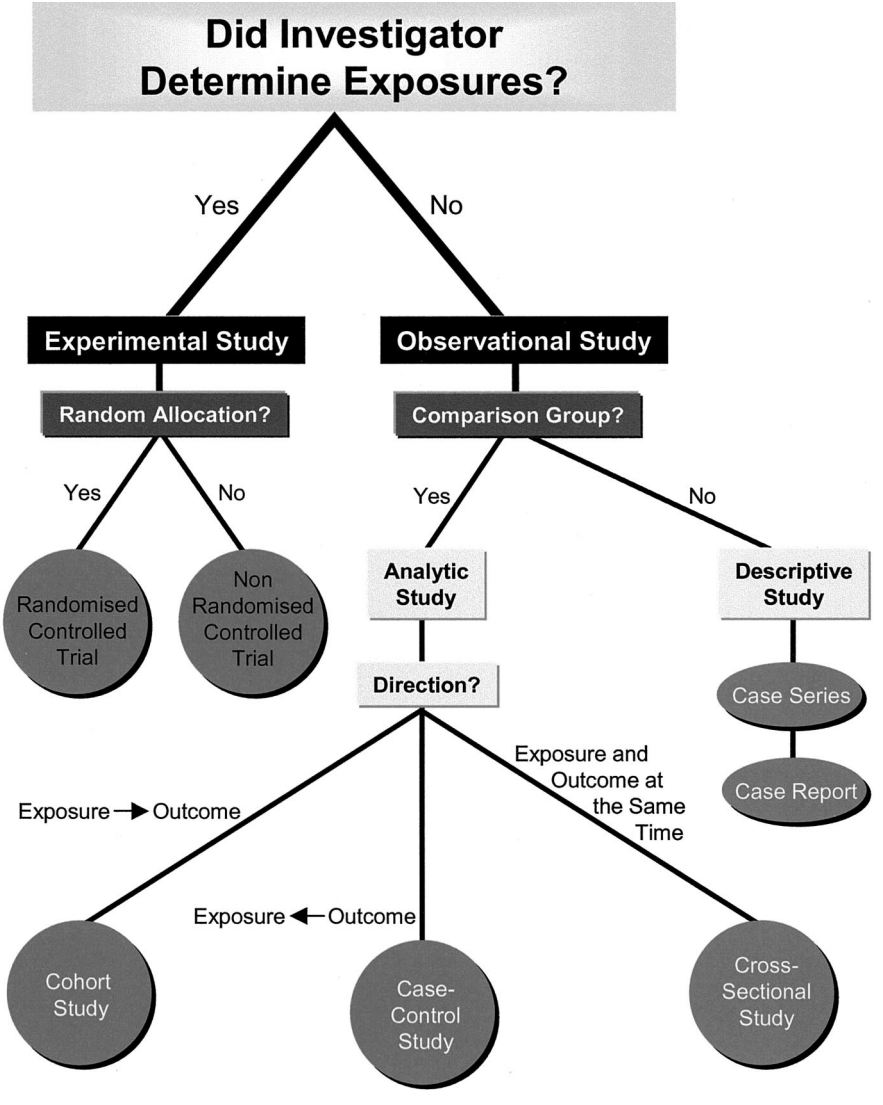

Fig. 3. Algorithm for classification of common types of clinical research. Reprinted with permission from Elsevier Science (Lancet 2002;359:57–61).

observational research, patients receive their treatments in their usual way. Analogous to a football game, the investigator in observational research sits passively on the sidelines and watches the game, noting what plays are chosen and monitoring the score. In experimental research, the investigator intervenes in the clinical setting and actively assigns the treatments (or other exposures) rather than the patient's clinician doing this. In the football analogy, the investigator at the sidelines calls in every play to the quarterback then assesses the results.

Observational research dominates the literature. Indeed, one recent survey revealed that about 80% of the articles published in obstetrics and gynecology are observational (37). Within the world of observational research, two major types exist: studies with or without a comparison group. For simplicity, we will term those with a comparison group **analytic** studies and those without, **descriptive**.

DESCRIPTIVE STUDIES

One definition notes that a descriptive study is "concerned with and designed only to describe the existing distribution of variables, without regard to causal or other hypotheses" (38). The important qualifier here is that causal hypotheses are beyond the purview of this type of study, a feature sometimes forgotten by zealous investigators. Descriptive studies are often the first exploration of a new health problem, such as toxic shock syndrome. They describe the characteristics of those affected and the features of the disease itself.

Good descriptive reporting has been likened to good newspaper reporting; a number of key "W" questions need to be answered (2). **Who** has the disease? Characteristics of those with the disease often provide clues; for example, the incidence of ovarian cancer increases with age, and preeclampsia is more common among primigravidas than among women who have children.

What is the disease in question? A clear, specific, and measurable definition of what is a case of the disease is a prerequisite to its study. For example, the scores of research studies on endometritis are largely uninterpretable, inasmuch as no uniform case definition exists. How tender must a uterus be, and who determines it? Fever, an objective outcome, is a better proxy to study than is endometritis.

Why did the condition arise (what clues about causation might be evident)? Hunches from descriptive studies can be tested in more rigorous analytic or experimental research. For example, early descriptive reports of benign hepatocellular adenomas suggested a link with high-dose oral contraceptives. This hypothesis led to a large case-control study, which confirmed and quantified the association (39).

When does the disease occur? Temporal relationships can often provide insights into etiology. For example, the peak incidence of cervical cancer is several decades earlier than that of ovarian cancer, suggesting different etiologies. A rise in primary and secondary syphilis rates in the 1980s was linked with exchanging sex for drugs in crack houses in U.S. cities (40). An epidemic of endometrial cancer in the 1970s was temporally associated with the use of unopposed estrogen replacement therapy, a common practice in that era (41). The recent epidemic of multiple births in the U.S. has been attributed largely to assisted reproductive technologies, which raises serious medical, financial, and ethical concerns (42).

Where does the disease occur (or not occur)? Syphilis today is rare in the U.S., except in the Southeast (43). Rates of induced abortion are inversely related to distance from a metropolitan area, where most abortion providers are located (44). Sperm counts may decline during the summer in hot, humid locales (45).

The final "W" may be the most important: **so what**? What is the clinical import of the report? The usefulness of descriptive reports varies widely, from first announcements of new and important illnesses to "me too" additions of a few more cases to the world's literature.

Several types of descriptive studies appear in the medical literature. At the bottom of the research hierarchy is the **case report** (Fig. 3). When more than one patient is reported, the study design becomes a **case-series report**. Case-series reports often herald an epidemic. For example, the occurrence of a number of cases of clear-cell carcinoma of the vagina, an uncommon event, suggested the presence of an epidemic in New England. A subsequent case-control study linked the disease to *in utero* exposure to diethylstilbestrol (46), although these data may contain substantial bias (47).

Another type of descriptive research is the **prevalence study**. This can be thought of as a snapshot of a community at one point in time. For example, the federal government periodically conducts the National Survey of Family Growth (48), a household interview of women across the U.S. It provides a rich source of information about family planning, family size, and other demographic information. Another ongoing survey is the National Hospital Discharge Survey, which abstracts data from the face sheet of patients discharged from short-term, nonfederal hos-

pitals across the country. Again, this survey provides valuable information about diseases and treatments in a representative sample of American acute-stay hospitals (49).

Surveillance is yet another important type of descriptive study. This entails watchfulness over a community, with feedback to the community about the results an important element (38). **Active** surveillance makes a determined effort to find cases of interest; **passive** surveillance relies on existing reporting systems, such as reporting of chlamydia cases to state health departments. As might be predicted, the yield of active surveillance is higher than passive. Active surveillance played a pivotal role in the eradication of naturally occuring smallpox from the planet, and the same may occur soon with polio as well (50).

An important limitation of descriptive studies without comparison groups is that they allow no testing of hypotheses about causation. Although they may generate hunches as to causation, only studies with a comparison group (e.g., analytic and experimental) have this ability. A common error is making causal deductions based on case-series reports; substantial harm has resulted (51).

ANALYTIC STUDIES

The second major type of observational study is the **analytic study**. Here, the investigator has a comparison (or control) group, a powerful scientific tool. Comparison groups provide investigators a benchmark against which to compare the study group. For example, in a cohort study, the comparison group indicates how frequent is the **disease** (or other outcome) in the community. If disease is more common among those exposed than among the comparison group not exposed, then a positive association exists. An example would be assisted reproductive technologies and multiple gestations. If less disease occurs among the exposed than among those not exposed, then the exposure is associated with a protective effect. Here, an example would be oral contraceptives and ovarian cancer.

Cross-sectional, case-control, and cohort studies are the most common types of analytic studies. A **cross-sectional study** can be thought of as a snapshot at one point in time (Fig. 3). Both exposure and outcome are determined simultaneously. Although this keeps costs down, it also leads to problems in judging temporal associations. An example would be a study of pica and iron-deficiency anemia in pregnant women in labor. Upon admission, all patients would be queried about pica and all would have their hematocrit determined. Given a positive association between pica and anemia, the temporal (and causal) association is unclear. Did eating starch or clay crowd food out of the diet (or inhibit iron absorption) and thus lead to anemia (52), or did anemia from some other cause lead to cravings for pica? This chicken-and-egg question cannot be resolved in a cross-sectional study, because it focuses on prevalence rather than incidence. An exception would be an exposure which surely preceded the outcome, such as blood type or race.

COHORT STUDIES

In contrast, **cohort** studies are much easier to comprehend. They run forward in time from exposure to outcome (4). If an outcome is more frequent among those exposed than among those not exposed to a factor, then a positive association exists, and vice versa. Because clinical practice flows forward in time from exposure to outcome, this research design is intuitive.

While all cohort studies run forward in time from exposure to outcome, not all are done in real-time (Fig. 4). Stated alternatively, cohort studies can be **concurrent, nonconcurrent**, or **ambidirectional** (4). Regrettably, the terminology often muddies the water here: synonyms for cohort study include longitudinal, incidence, follow-up, forward-looking, and prospective. A **concurrent** cohort study enrolls persons exposed and unexposed and follows them forward contemporaneously to determine who gets the outcome of interest. A **nonconcurrent** cohort study goes back in time to comprise the groups of exposed or unexposed (for example, through hospital charts or employment records) and tracks them forward in time to determine outcomes. Although the forward direction is the same, the data collection process is not contemporaneous. An **ambidirectional** cohort study combines both approaches above; data gathering is done historically and contemporaneously. This can expedite getting results.

Cohort studies have important strengths. They are the best way to determine the incidence of disease (along with relative risks and confidence intervals), and they also portray the natural history of disease. Starting with a single exposure, an investigator can examine several possible outcomes. However, a potential abuse here is examining many outcomes and reporting only those that emerge as statistically significant. Although case-control studies are impractical for rare exposures, cohort studies are more efficient here. For example, one could follow a cohort of women who skydive and a similar group who do not and determine the frequency of fractures in each group.

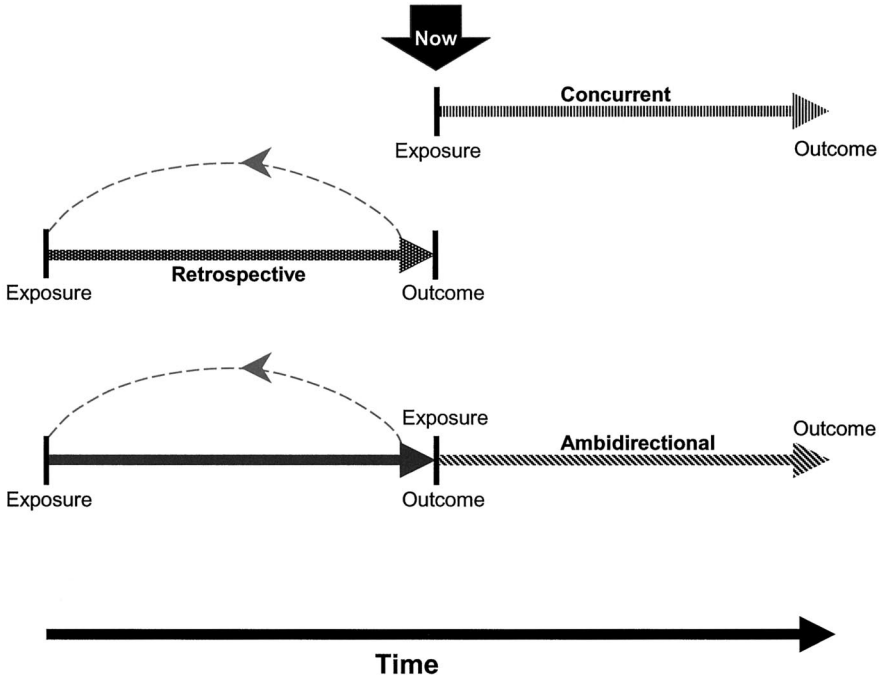

Fig. 4. Schematic diagram of concurrent, retrospective, and ambidirectional cohort studies. Reprinted with permission from Elsevier Science (Lancet 2002;359:341–345).

Cohort studies have weaknesses as well. Selection bias is inherent because the exposed and unexposed are not allocated to their groups at random. For example, women choose to skydive or not. Skydivers may differ in important ways from other women, such as participation in other outdoor sports, not using seatbelts, or other risky behaviors (53). Thus, the groups might not be similar in all important respects except for skydiving or not. In addition, cohort studies can be impractical for examining diseases that are rare or that take decades to develop. Nevertheless, several large cohort studies have contributed important information, albeit at large expense (54).

Reports of cohort studies should meet several reporting requirements. The exposure needs to be clearly defined. Likewise, the outcome should be clear, specific, and measurable. Determination of outcomes should be identical for both the exposed and unexposed; blinding the observer as to the exposure group can help to avoid subtle bias in making this determination. Differential losses to follow-up can lead to bias, and this problem becomes acute when the observation period is long.

CASE-CONTROL STUDIES

Case-control studies work backward (5, 30). They begin with a disease or condition (**cases**) and look back in time at exposures that might be related to the disease. For comparison purposes, a group of persons without the disease (**controls**) undergoes the same scrutiny. If the exposure of interest is more frequent among the cases than among the controls, then a positive association exists, and vice versa. This type of research is especially useful for studying rare events (such as cancer) or diseases that take a long time to develop (such as heart disease); cohort studies are often impractical in these settings. Case-control studies often provide a "quick and dirty" way to study potential associations.

In a case-control study, controls indicate the prevalence of the **exposure** in the community. If the exposure is more common among cases (with the disease or other outcome) than among the controls (who are free of disease), then the exposure is positively associated with the outcome. An example would be cigarette smoking and cervical cancer. In contrast, if the exposure is less common among cases than among control, a protective association is evident. An example here would be tubal sterilization and ovarian cancer (55).

Case-control studies have made important contributions to women's health. For example, by identifying risk factors, case-control studies led to important reductions in the incidence of AIDS well before the responsible virus had even been discovered. In addition, this design has been used to study a wide

variety of potential associations, ranging from tandem trucks and traffic accidents to cat ownership and schizophrenia (5).

Case-control studies have a more sinister side, however. Finding an appropriate control group can be difficult, so selection bias is a common problem. Although case-control studies can be easy to do, they are also easy to do poorly, and the literature is replete with flawed and misleading case-control studies. Because this type of study runs backward in time from outcome to exposure, this design is conceptually difficult for many clinicians.

Choosing appropriate controls is the Achilles heel of this research design. Persons chosen as controls should represent those in the community at risk of the disease in question. For example, assume an investigator plans a study of the potential association between oral contraceptives and venous thromboembolism. Cases are new admissions to a medical service, and controls are a random sample of women admitted to the medical service with conditions other than thromboembolism. Here the controls might have a high prevalence of chronic illness, reduced mobility, and thus increased risk of developing clots. If so, then this hospital-based control group would not be representative of women in the general community, and this would bias the results.

Although case-control studies are usually considered a faster way to answer a research question than a cohort study, this is not always true. In settings where the exposure of interest is rare, conducting a case-control study can be prohibitively expensive. For example, consider a case-control study of the potential relationship between skydiving and fractures. Cases would be women with new fractures identified in a hospital emergency department, controls a sample of those without fractures, and the exposure recent skydiving. Because the proportion of women in the community who engage in this sport is so low, finding sufficient numbers of women who had been exposed to skydiving would take a huge study. This feature of case-control studies is not well known: the prevalence of the **exposure** in the general community drives the sample size in a case-control study, not the frequency of the outcome. In this skydiving example, a cohort study would be more efficient.

Variations of Cohort Studies

Two other variations on the cohort theme are the before-after study and the nested case-control study. The **before-after study** is a common but weak design. An investigator wants to show the effect of an intervention, commonly a drug. The researcher takes baseline measurements, exposes the participants to the intervention, and repeats the initial measurements. An example would be administration of statin drugs to women with high cholesterol values. If cholesterol values decline after treatment, the investigator might infer that the decrease was the result of the drug. However, without a contemporaneous control group, this inference is not secure. **Regression to the mean** (38) may have been responsible, at least in part. The more abnormal a value on first measurement (e.g., cholesterol), the more likely it is to be closer to the mean of the population on repeat measurement. Alternatively, other influences may be involved, such as a change in diet or exercise prompted by alarm over the aberrant values and need for drug therapy.

Nested case-control studies are sometimes built into cohort studies or randomized controlled trials. This approach can be useful when a measurement is considered too difficult or costly to do on all participants. An expensive blood test is a prototype. All participants have a sample of blood taken on enrollment, and the serum is frozen until the study is over. Those in the cohort or trial who develop the illness become cases; a random sample of well participants becomes the control group. At this point, the investigator performs the expensive blood test on the banked serum for only cases and controls (a small fraction of the study participants). This avoids the expense of running the laboratory test on the entire group.

EXPERIMENTAL STUDIES

The **experimental** design in clinical medicine is the closest proxy to the controlled experiment of the basic scientist. Within this category, two principal types exist: **randomized controlled trials** and **nonrandomized controlled trials**. Both are formal trials in which the investigator intervenes and decides which participant gets which exposure. Participants are then followed forward in time to measure outcomes, and the study is analyzed like a cohort study. The important difference between cohort and experimental studies is assignment of exposures by the investigator in the latter.

In **nonrandomized trials**, some method short of true randomization is used to assign the exposures. An example of this approach is **alternate assignment**. This tactic has been used in large studies in which participants are assigned to treatment by alternate months, for example, restricted versus liberal use of electronic fetal monitoring (56). With large

TABLE 1 **Levels of evidence in clinical research**

Quality of evidence	
I	Evidence from at least one properly designed randomized controlled trial.
II-1	Evidence obtained from well-designed controlled trials without randomization.
II-2	Evidence from well-designed cohort or case control studies, preferably from more than one center or research group.
II-3	Evidence from multiple time series with or without the intervention. Important results in uncontrolled experiments (such as the introduction of penicillin treatment in the 1940s) could also be considered as this type of evidence.
III	Opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees.
Strength of recommendation	
A	Good evidence to support the intervention.
B	Fair evidence to support the intervention.
C	Insufficient evidence to recommend for or against the intervention, but recommendation might be made on other grounds.
D	Fair evidence against the intervention.
E	Good evidence against the intervention.

U.S. Preventive Services Task Force (57).

numbers of participants, this approach may approach (but never equal) randomization. This approach has important methodological weaknesses. The U. S. Preventive Services Task Force (Table 1) terms such studies Level II-1, indicating less scientific rigor than randomized controlled trials (Level I) (57). Importantly, allocation concealment is often impossible to achieve in trials that do not use random allocation. Although some researchers refer to these studies as "quasi-randomized," we believe the term is misleading, an example of medical puffery. "Random," like "pregnant," is a dichotomous adjective. A trial cannot be "quasi-randomized" any more than a woman can be "quasi-pregnant."

BIASES

Bias has a different meaning in clinical research than in everyday life. In common parlance, bias means prejudice. In research, it means a systematic distortion or deviation from the truth (3). Biases are pervasive in observational research. The challenge for readers is to detect those biases and figure out how they might have influenced the results presented. We will provide a simple four-point check list to help with this assessment.

Selection bias results from a lack of comparability between the groups and is unmeasured. Stated alternatively, the groups being studied are different at the starting blocks. For example, one group might be older, sicker, or have a higher proportion of cigarette smokers. This "stacks the deck" before the analysis. In a cohort study of failures after two different methods of tubal sterilization, one group might be younger (and hence more fertile) than the other (58). In a case-control study of exposure to ovulation-induction agents and later ovarian cancer, poorly chosen controls might be of higher parity and thus at lower than average risk of exposure to fertility drugs.

Information bias results from a lack of comparability in data gathering between groups. Synonyms for this include ascertainment, measurement, and observation bias. In a cohort study, information about development of the outcome should be determined in exactly the same way for both the exposed and unexposed groups. In a case-control study, information about prior exposures should be collected identically for both cases and controls. Information bias is a common problem in case-control studies, especially those that rely solely on recall of past exposure (without corroboration from other sources such as medical or pharmacy records).

Information bias comes in two varieties: **systematic** misclassification (working preferentially in one direction) or **random** misclassification (noise in the system). The effects differ. Systematic misclassification increases or decreases the measure of association; the odds ratio or relative risk spuriously goes up or down. In contrast, random misclassification obscures real effects; the odds ratio or relative risk moves toward 1.0, wiping out an effect.

Recall bias poses a huge problem in case-control studies that rely on memory of remote events. Cases (who are sick or who have the condition of interest) have likely searched their memories to explain why this misfortune has befallen them. In contrast, healthy controls have no such motivation to rummage through their memories. Thus, cases are more likely to recall distant events than are controls; this relative underreporting of past exposures by controls biases case-control studies that do not have independent validation of exposure.

A case-control study of use of oral contraceptives and limb-reduction defects among children born later illustrates this problem (59). Women who had given birth to children with serious limb-reduction defects were interviewed as cases; controls were women with healthy children of a similar age. At an average of 4.5 years after the birth of the children involved, mothers were asked about prior exposure to oral contraceptives (about 5 years before the interview). As might be expected, a higher proportion of cases

recalled use of oral contraceptives than did controls. **Recall bias** among controls probably accounted for this association; no biological explanation is readily evident.

Although recall bias is commonly blamed for spurious associations in case-control studies, investigators in Sweden have provided compelling documentation of its existence and its potency. The alleged association between induced abortion and later development of breast cancer has become a *cause celebre* among some opponents of legal abortion (www.abortionbreastcancer.com). Indeed, many case-control studies have consistently found a positive association. However, women with breast cancer are likely to have searched their memories for possible causes of this deadly disease, and they are more likely to report honestly to researchers. Healthy controls have no such motivation. In some Scandinavian countries (60, 61), all citizens receive a unique identification number at birth that follows them for life. In addition, health care is provide by the state, offering a unique opportunity to study health through national, comprehensive records with excellent linkage capability.

The investigators performed the same case-control study of abortion and breast cancer using two different approaches to information gathering (61). In the traditional approach, they interviewed cases and controls in person. In the alternative approach, they repeated the study using national health care records. The findings were striking: fewer controls than cases accurately reported prior abortions to the interviewers (abortions documented by government records). The likelihood of this degree of underreporting of abortions by controls having occurred by chance was remote. Using the traditional case-control approach to data collection, the study found no effect of abortion on breast cancer risk; using the more comprehensive national statistics, abortion was associated with a 40% reduction in risk. The case-control literature, which consistently points to an association, is consistently biased (62). Bias, not biology, is at work here.

Many clinicians would agree that **confounding** is aptly named, because it readily confounds understanding. In everyday use, confounding means confusing or bewildering; in epidemiology, it means a mixing or blurring of effects. An investigator sets out to examine the association between an exposure and an outcome but winds up measuring the influence of a third factor (the confounding factor). A confounding factor is associated with both the exposure and the outcome but is not involved in the causal pathway. Even this definition is not much help; confounding is often better grasped through clinical examples. The effects of confounding are not intuitive or logical to most clinicians.

For example, a cohort study might examine the putative association between IUD use and salpingitis. Exposed women are those who choose an IUD; unexposed are women not using contraception. In recent decades, women choosing IUDs for contraception were older than other users of contraception, and the risk of upper genital tract infection is inversely related to age (63). Thus, age would be a potential confounding factor in this study: age is associated with both the exposure (choice of an IUD) and with the outcome (a reduced risk of salpingitis) but not directly involved in the causal pathway to infection. In this example, the older age of IUD users would artificially lower the relative risk (few older women get salpingitis, regardless of contraceptive choice).

Age might have the opposite effect in a case-control study of intrauterine contraception and myocardial infarction. Cases would be women with a confirmed myocardial infarction, and controls a sample of healthy women in the community. The exposure would be contraceptive used in the past 3 months. IUD users tend to be older than other users of contraception, and older women are more likely to have heart attacks. Here, age would be a potential confounding factor: age is associated with both the exposure (choice of an IUD) and the outcome (an increased risk of myocardial infarction) but is not directly involved in the causal pathway to coronary occlusion. In this example, age might lead to a spurious positive association between IUD use and heart attack.

Unlike selection bias and information bias, confounding can be controlled after the fact. Stated alternatively, if selection bias or information bias is present in a study, the investigation is irrevocably corrupted. On the other hand, if confounding is present, the investigator can control for its effects, provided the researcher anticipated it and gathered information about it during the data collection phase. Control of confounding can take place during the data gathering or during the analysis phases of a study.

CONTROLLING FOR CONFOUNDING

Restriction is the simplest way to control for a potential confounding factor. For example, if cigarette smoking is deemed a potential confounding factor in the relationship between oral contraceptives and cervical cancer, simply exclude smokers from the study. This neatly eliminates any effect that smoking might have on the relationship. Although valid, this approach

is not popular for two practical reasons. First, this exclusion limits the external validity of the findings: one can extrapolate the results only to women who do not smoke. Second, this approach inevitably reduces the sample size and thus the ability to detect effects of a given size should they exist (**power**).

Matching is another established approach to controlling for confounding. If smoking is deemed to be a potential confounding factor in a case-control study of oral contraception and cervical cancer, then for every case who smokes, a corresponding control who smokes is enrolled. Similarly, for every nonsmoking case, a nonsmoking control is selected. This maneuver makes the cases and controls homogeneous for smoking status. Although a valid approach to controlling confounding, matching can be cumbersome to do, slowing enrollment. In addition, an investigator cannot examine the effect of any confounding factor for which matching is done.

Stratification is a variation on restriction, a form of *post hoc* restriction done in the analysis phase rather than during the recruitment and accrual phase of a study. If an investigator suspects that smoking may confound the relationship between oral contraceptives and cervical cancer, then the researcher can examine the relationship separately for women who smoke and women who do not smoke to see if the same association is evident.

A mathematical technique, termed the **Mantel-Haenszel procedure** (3), combines the various 2 × 2 tables (here, smokers and nonsmokers) into a single summary estimate of the effect. The tables (or strata) are weighted inversely to their variance in deriving the summary statistic; stated alternatively, 2 × 2 tables with large numbers contribute more to the final statistic than do those with few patients. If the summary estimate of the treatment effect from the Mantel-Haenszel procedure differs substantially from the crude estimate from the overall data set, then confounding is deemed present.

An example may help show how this works. Assume that a large cohort study is following a group of married, parous women. The research question is the potential association between use of an IUD and secondary infertility. The exposed group includes 2000 women using an IUD; the unexposed group consists of 2000 women not using an IUD. Secondary infertility is defined as failure to conceive a planned pregnancy within 12 months after discontinuation of the current contraceptive method. As shown in Figure 5*A*, use of an IUD is associated with an overall three-fold increase in the risk of infertility (RR 3.0; 95% CI, 2.0–4.5). This is the crude estimate for all 4000 women.

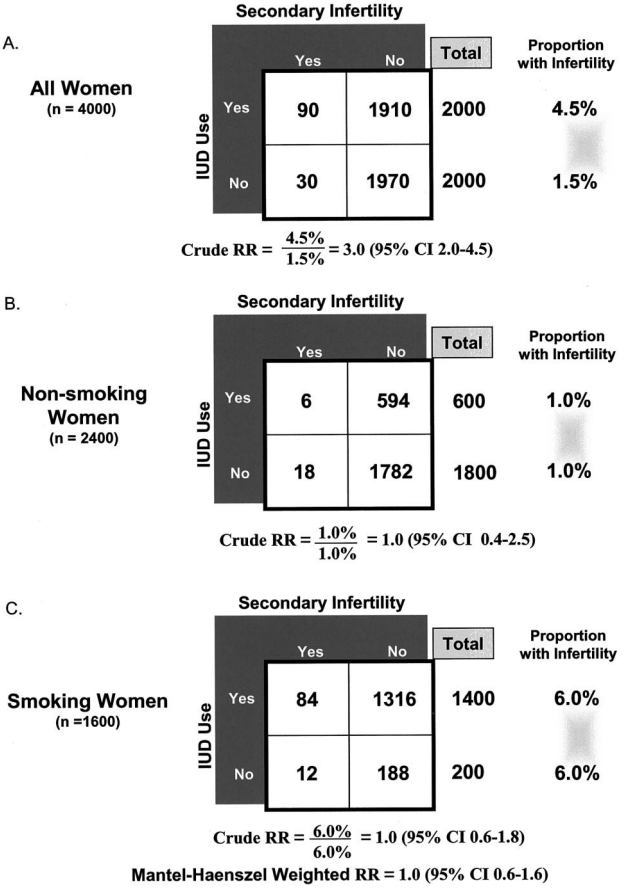

Fig. 5. Hypothetical cohort study of intrauterine device use and secondary infertility. *A*, crude analysis of entire cohort. *B* and *C*, analysis stratified by smoking status. After control for the confounding effect of smoking, the increased risk disappears.

However, this result seemed implausible to the investigator, and she suspected that cigarette smoking may be a confounding factor. With the permission of her Institutional Review Board and the consent of the participants, she collected saliva for cotinine levels on all participants. Figure 5*B* presents the same 4000 participants stratified by whether their saliva indicated recent cigarette smoking. As shown in the top 2 × 2 table, the proportion of nonsmoking women who developed infertility was 1%, independent of IUD use. However, in the bottom table, smoking women had a much higher risk of infertility: 6%. Nevertheless, the risk was identical in both exposed and unexposed. The Mantel-Haenszel adjusted relative risk of infertility is 1.0 (95% CI, 0.6–1.6); the crude relative risk without adjustment for smoking was 3.0.

Because two relative risk estimates (1.0 vs. 3.0) differ widely, confounding is present in this example. Here, the investigator would present the Mantel-

Haenszel adjusted (for smoking) relative risk estimate of 1.0 as the more accurate. The confounding effect of smoking caused a spurious three-fold increase in risk, which disappeared when the confounding was controlled. Smoking was associated with both the exposure (choice of an IUD; note top left cell in table of smokers) and with the outcome (infertility) but was not directly involved in the causal pathway.

Chance is the final item in the four-point checklist. Its listing as the last item is intentional. Many readers (and some editors) jump to the P value to judge if a study is valid – and worth reading. As shown in the example above, bias can easily result in highly significant findings that are completely wrong ($P < .001$ in Figure 5*A*). Hence, one needs to ask this question: could the findings of a study be due to selection bias, information bias, or confounding? If the results cannot be accounted for by these three considerations, could they be due to chance alone? Here, the P value comes into play. By convention, a P value of $<.05$ is used, meaning that a difference this large could be due to chance less than one time out of 20. However, excessive reliance on P values in clinical research has often led to incorrect interpretation (1, 31). If results still cannot be explained away, then they may be real and of clinical importance.

RANDOMIZED CONTROLLED TRIALS

Randomized controlled trials stand at the pinnacle of the research hierarchy. Indeed, when medical historians look back upon the twentieth century, the development of the randomized controlled trial will stand out as one of its sentinel achievements (64). The unique contribution of randomized controlled trials is avoidance of bias. Bias is innate in observational research. The randomized controlled trial is the only known way to avoid bias in clinical research. When one is examining a large effect, such as cigarette smoking and lung cancer, almost any study design will unearth the association. However, the randomized controlled trial really shines when looking for small associations. Here, bias in observational studies could obscure or exaggerate small but real effects. Only by ridding the trial of the distorting effects of bias can one ferret out the truth.

Assignment to treatments by chance, rather than choice, is the defining feature of randomized controlled trials. This simple, yet powerful, tool ensures that the groups under study are similar in all important respects except for the exposure in question. By balancing baseline characteristics, **randomization** levels the playing field at the start of a study; **selection bias** is precluded. By having outcomes measures defined in advance and determined in the same way for the groups under study, **information bias** is avoided or minimized. Potential **confounding** factors should be balanced between the two groups, assuring comparability at the starting blocks, except for chance imbalance. This is important for known confounding factors but even more critical for those factors that are unsuspected. They, too, will be equally distributed among the groups under study.

Methods of Randomization

Despite its importance in avoiding bias, randomization is often not done well. . ..or at all. . .in published trials. Some techniques called "random" are not; this appeared in 5% of a sample of trials examined (65). Using the last digit of hospital chart numbers (odd or even) or year of birth is not a random technique, because patients do not receive chart numbers (or enter the world) at random. An even worse example of a nonrandom technique claiming to be random is use of alternate days or weeks to assign treatments (66).

Some techniques are truly random but suboptimal for other reasons. Flipping a coin to determine the next treatment assignment is random, yet it is vulnerable to tinkering. Investigators or those enrolling participants may find a string of heads unnerving and may, therefore, insert a tail in the sequence (since a tail was "overdue"). Chance, however, has no memory. In addition, flipping a coin multiple times leaves no paper trail than can be audited at a later time. For example, the only randomized controlled trial in the world to have found a significant reduction in perinatal mortality associated with use of electronic fetal monitoring used coin tosses in two Greek hospitals (67). Widely different sample sizes resulted; the likelihood of a disparity this large occurring by chance alone is remote. Stated alternatively, the odds are about 19 to 1 that the authors did not do what they reported, raising serious concerns about the validity of the results.

Preferred approaches to developing a randomization sequence include tables of random numbers or computer-generated random numbers (6). Using odd or even numbers to assign participants to two treatment arms is the simplest approach (**simple randomization**). With small trials, imbalance in the size of the treatment groups can occur by chance alone. However, when trials are larger than about 200 participants, the potential for a big disparity largely disappears.

Restricted randomization is an alternative approach; it controls the likelihood of getting a chance imbalance in the numbers assigned to the treatment groups. A common type of restricted randomization is **blocking. Random permuted blocks** assign participants to treatments by chance but guarantee that equal numbers are assigned to each group as the trial progresses. Instead of randomly allocating all study subjects (simple randomization), this process randomizes participants in a series of blocks of specified sizes, such as 6 or 8 participants. For example, with a block size of 6 participants, 20 different permutations of three assignments to A and three to B exist (e.g., AAABBB, ABBBAA, AABBAB, etc.) Each of these blocks is given a number or range of numbers and the sequence of 6-participant blocks is chosen with a random number table or computer-generated sequence of numbers. Thus, after every sixth participant (e.g., at 6, 12, 18, 24, etc.) equal numbers will have been assigned to A and to B. The longer the block, the less likely inquisitive investigators or study staff are to be able to decipher the block length and thus the upcoming assignment. In addition, randomly varying the block length and inserting a run of simple randomization (a process called **mixed randomization,** described later) can render the allocation sequence nearly impenetrable (11).

Allocation concealment is the second critical bias-eliminating element of randomization. This term means that both participants and trial staff are unaware of the upcoming treatment assignment. Allocation concealment needs to be distinguished from blinding as to treatment. Allocation concealment is mandatory and always possible; treatment blinding is not always possible and may not always be important. The value of allocation concealment in avoiding selection bias has only recently been appreciated. . .and quantified (7). Four separate investigations have concurred: failure to maintain allocation concealment in randomized controlled trials exaggerates treatment effects by as much as 40%. Bias this large may overwhelm real treatment effects.

Given a truly random sequence of treatment assignments, why should knowledge of the upcoming assignment introduce bias? Armed with that foreknowledge, those recruiting participants can steer subjects into or out of the study at will, thus introducing selection bias. Ironically, randomization is designed expressly to avoid selection bias. Assume that a trial is comparing three different operations for genuine stress urinary incontinence (68). The physicians recruiting participants can see all the upcoming assignments, which are posted in the clinic. A physician interviews a patient with severe incontinence, and the physician believes that a Burch procedure is the best option for the patient. If the next treatment to be assigned was an anterior colporrhaphy, the physician could send the patient to the laboratory (or lunch) during which time another participant gets assigned to the vaginal repair. When the initial patient returns, a Burch procedure is, to everyone's delight, the next assignment. The patient with severe incontinence gets a Burch procedure, but not due to chance. If this process were to be repeated, it would overload the Burch procedure group with difficult cases, biasing the comparison against this operation.

As with methods of random-sequence generation, methods of allocation concealment vary widely in their rigor. The following techniques are considered adequate: sequentially numbered, sealed, opaque envelopes with method indicator card inside; pharmacy-controlled distribution of numbered pill bottles; and central randomization (e.g., telephone calls to a central office where the randomization sequence is maintained). Each of these approaches, if properly used, should guard against discovery of the upcoming assignment. Despite the importance of allocation concealment, only a quarter of published trials provide readers sufficient information to confirm that this was done (7).

Each of these methods is vulnerable to tampering, and we have learned of examples of each (69). For example, study staff have taken opaque envelopes to the "hot light" in the radiology department to see the writing on the card inside. A pharmacist ran out of an experimental cephalosporin one weekend, and, to avoid slowing down recruitment, allocated all participants joining the study over the weekend to the comparison antibiotic, cefoxitin, which he had in stock. A busy clinician cajoled a central office into providing the next several treatment assignments, because he anticipated being "too busy" to call back for each patient. . .and the office acceded to this improper request. Randomized controlled trials are "anathema to the human spirit" (69). Clinicians and others involved with trials have a powerful personal interest in figuring out the upcoming assignments. Hence, investigators have a responsibility to build in safeguards against deciphering that are equally powerful. Readers deserve to know about these safeguards and whether they worked.

Reports of randomized controlled trials should document that the randomization yielded groups similar in all important respects, except for the exposure being studied. The first table of such reports usually fulfills this role. This table presents demographic and other relevant clinical features by treatment group,

e.g., age, race, parity. This enables readers to know the types of participants enrolled. Can the results be extrapolated to the readers own patients? In addition, are the groups similar in all important respects except for the exposure being studied?

A common error in many reports of randomized controlled trials is statistical testing of baseline characteristics. Investigators commonly perform tests to show that no significant differences occurred; the far right column of the baseline characteristics table typically consists of P values. Statistical testing here addresses the likelihood that difference seen between the groups could be due to chance. This question is gratuitous for a randomized controlled trial, since *all* differences in baseline characteristics are *known* to be due to chance (the investigator assigned participants to treatment at random) (70). Thus, the likelihood that differences are due to chance is 100%. Instead of performing statistical tests, investigators should describe pertinent baseline characteristics. For example, continuous variables, such as age, should appear as an average with a measure of variability, e.g., mean and standard deviation. Numbers and proportions should be provided for categorical variables.

BLINDING OF TREATMENT

Blinding, also termed **masking**, entails keeping participants, healthcare providers, and outcome assessors (those collecting data) unaware of who got what treatment. Whereas allocation concealment avoids selection bias, blinding is used to prevent information bias and protects the random sequence after allocation (8). If participants know which treatment they are receiving, then this might influence their expectations and their compliance with the treatment. If the clinicians involved are aware of the treatment assignments, they may convey their opinions—subtly or openly—to participants. Alternatively, they might differentially apply co-interventions to one group or the other (e.g., supplementary treatments or care).

Blinding is especially helpful when the outcome measure is subjective. In such cases, knowledge of the treatment group can easily influence determination of the outcome. For example, in a placebo-controlled trial of multiple sclerosis treatment, unblinded neurologists found a benefit of treatment, whereas their blinded colleagues did not. Pain is another outcome for which treatment blinding may be critical. In contrast, for objective outcomes, such as fever or death, blinding may be unnecessary. Moreover, sometimes blinding cannot be accomplished. For example, a trial comparing mini-laparotomy versus laparoscopy for tubal sterilization would be impossible to blind.

The terminology used for blinding is slippery. No consensus exists concerning the terms single- , double- , and triple-blinding (8). Hence, instead of these shorthand terms, we suggest that investigators say explicitly who was blinded (participants, clinicians, data gatherers, etc.) and how the blinding was achieved. In general, reports of randomized controlled trials neglect to provide readers with this information.

Placebos are often used to maintain blinding. A placebo is a pharmacologically inactive drug or agent that is given to the control group in a trial. Administration of a placebo to the control group balances the psychological placebo effect that occurs in the experimental group. In addition, it maintains blinding. Use of a placebo is appropriate when no established therapy exists for a condition; if an effective treatment is available, then use of a placebo is inappropriate and unethical. Instead, the new treatment can be compared with an existing treatment.

When used, placebos should be identical in appearance, smell, and taste to prevent deciphering of the treatment assignments. When disparate interventions are being compared (e.g., an injection versus a tablet), placebos can still be used to maintain blinding. In this situation, a **double-dummy** approach is helpful: each participant receives one active and one inactive treatment. For example, those assigned to the injection group would receive an active injection and a placebo tablet, whereas those allocated to the tablet group would get a placebo injection and an active pill (identical to the placebo tablet).

EXCLUSIONS FROM A TRIAL

Randomized controlled trials commonly have two types of exclusions, those before and those after randomization (9). The impact of these types of exclusions differs. **Exclusions before randomization** are common in most randomized controlled trials. Volunteers are admitted to randomized controlled trials only after having passed eligibility criteria specified in the study protocol. These criteria range from medical (e.g., one treatment would not be safe for the potential participant) to logistical (e.g., the prospective participant is deemed unlikely to return for follow-up visits). Whether capricious or reasonable, none of these exclusion criteria will bias the study: the internal validity should be intact if the trial is properly done. What may suffer from extensive exclusions before randomization is external validity: the trial may include such an eclectic sample that

readers cannot extrapolate the results to their patients.

In contrast, **exclusions after randomization** open the door to bias. Many investigators and readers fail to appreciate that the only unbiased comparison groups are those initially composed by the randomization. Any attrition after that leads to groups that are no longer comparable (unless the erosion is random, which is usually not the case). Hence, the guiding principle is to conduct an **intention-to-treat analysis**. Stated alternatively, "once randomized, always analyzed." Each participant should be analyzed with the group to which he or she was initially assigned. . . .regardless of what occurs thereafter.

A common error is to focus the analysis on only those participants who complied with therapy. Since noncompliance does not occur at random, this distorts the comparison. The primary analysis in any trial should be the intention-to-treat analysis of all participants for whom information is available. Other secondary analyses, such as those with perfect compliance are acceptable, provided researchers plan these in advance and designate them as nonrandomized (cohort) comparisons. Few reports of randomized controlled trials provide readers sufficient detail to confirm that the appropriate unbiased analysis was done (9).

The best way to deal with exclusions after randomization is not to have any. In this regard, an ounce of prevention may be worth a ton of explanations. For example, randomization should be done at the last possible moment to minimize dropouts after randomization but before treatment.

A common cause of exclusions after randomization is **eventual discovery of ineligibility**. Trials commonly enroll participants who, on subsequent inspection, should not have been recruited. They clearly violate the inclusion criteria, but this was overlooked or misunderstood on entry. A common response is to drop these inappropriate participants from the analysis. However, this tactic likely introduces bias, since this discovery after the fact is unlikely to be random. For example, those with side effects or with an unfavorable outcome probably receive greater scrutiny than other participants and thus are more likely to be found to be ineligible. In general, these participants should remain in the trial and be analyzed in the groups to which they were assigned.

Deviations from the protocol are another common (but improper) reason for dropping participants from a trial. Many investigators exclude them to maintain the "purity" of the treatment groups and to provide a "fair" comparison. In reality, those who deviate from one treatment may be so different from those who do not comply with the other treatment that the remaining comparison is highly biased. The bottom line is that, no matter what happens during the course of the trial, all participants should be analyzed with their original treatment groups. All an investigator can test in a trial is the policy of giving the treatment, and not the treatment itself. However, this is precisely what clinicians need to know: how will the policy of giving the treatment work in the real world, in which noncompliance is common? Inclusion of a flow chart (71), which tracks all participants through all phases of a trial, enables readers to see if the appropriate analysis has been done.

Loss to follow-up can also bias a randomized controlled trial. Any losses to follow-up impair the internal validity of the study. However, if that loss is different between the comparison groups, major damage may result. Again, prevention is better than cure. Avoiding enrollment of those considered unlikely to follow-up may help. For those enrolled, additional contact information, such as the telephone number of a friend or relative who is likely to know the whereabouts of the participant, is useful in tracking. Hiring personnel expressly to track participants has led to high rates of follow-up in even developing countries (72, 73). Keeping the data collection questionnaire brief and the follow-up sites convenient can also promote follow-up.

How much loss to follow-up is too much? The answer depends on the trial. For some, in which the participant is under observation and the outcome is quickly determined, losses should be negligible. An example would be immediate morbidity of childbirth in hospital. In contrast, community trials running over several years may encounter loss rates that threaten the internal validity. Some have suggested a **5-and-20** rule of thumb. Trials with less than 5% loss to follow-up are probably secure; those with rates higher than 20% may be unsalvageable. Few trials with losses in excess of 20% would be able to withstand the charge of bias. Another rule of thumb is that the loss-to-follow-up rate should not exceed the outcome event rate (9). Simply put, the only loss-to-follow-up rate that avoids a biased comparison is zero.

UNEQUAL SAMPLE SIZES

Many investigators (and readers) expect treatment groups to have equal sizes in a randomized controlled trial. Indeed, in published reports the numbers assigned to treatment groups are more similar than should occur by random chance alone (11). However, precisely equal

sample sizes contribute little to statistical power and may paradoxically impair the unpredictability of the trial. Sample size requirements are influenced little by unequal group sizes until the disparity approaches a ratio of about 2:1 between them.

Unequal sample sizes are useful in trial in which blinding is not feasible and in which random permuted blocks of a fixed length are used. Those involved with recruiting and enrolling participants may figure out the block size, tally the numbers assigned as the trial progresses, and thus discern the upcoming assignments for some potential participants at the end of each block. For example, if the block length is determined to be six, and three As and two Bs have just been assigned, then the next participant is sure to get B (analogous to counting cards used in a game of blackjack). The sequence becomes transparent, and allocation concealment repeatedly fails.

Unequal sample sizes help to guard against this possibility. For trials with more than 200 participants, simple randomization is an appealing approach. The sequence is completely unpredictable, and large discrepancies in assignments are unlikely. For smaller trials, restricted randomization may be preferred. We have proposed a hybrid approach to restricted randomization that combines the balance afforded by random permuted blocks and the unpredictability conferred by simple randomization. In **mixed randomization**, simple randomization produces a block with specified inequality in the numbers assigned to each group (e.g., block of 10, with 3 As and 7 Bs). If this imbalance is not achieved on the first try, this sequence is discarded and another is chosen until a block of the desired degree of imbalance results, a process termed **replacement randomization**. Thereafter, random permuted blocks continue the allocation sequence. Use of long blocks and blocks that randomly vary in length helps to prevent deciphering the sequence. Another safeguard is the introduction of an unbalanced block generated by replacement randomization from time to time (11).

CONSORT GUIDELINES

The "gold standard" in clinical research now has its own gold standard: the CONSORT guidelines (71). This acronym stands for the Consolidated Standards of Reporting Trials. These guidelines for the conduct and reporting of randomized controlled trials have been adopted by a variety of journals, including some in obstetrics and gynecology. The guidelines, now in their second version, are evidence-based as much as possible. Importantly, early evidence suggests that journals that have adopted these guidelines have had more rapid improvement in the quality of their reports than has a journal that did not adopt them (74). The second CONSORT guidelines, published in 2001 (71), have corrected several glitches in the original version (75). Nevertheless, the guidelines should be considered a work in progress; additional refinement is planned.

The CONSORT guidelines are succinct to a fault. Hence, the authors of CONSORT wrote a longer companion article (36), which provides supporting documentation and examples from the published literature. Both those performing and those reading randomized controlled trials need to be familiar with CONSORT guidelines: they offer a road map for avoiding bias in this type of research.

TYPE OF ASSOCIATIONS

When a reader discovers a statistical association between exposure and outcome in a report, the next job is to figure out what sort of association it might be. Three basic associations exist: **spurious, indirect**, and **causal** (3). The first type reflects bias, such as information and selection bias. Indirect associations are due to confounding; the association is real but due to the effect of the confounding factor and not the exposure itself. The holy grail of clinical research is discovery of the causal association, linking an exposure to an outcome.

Despite their importance, causal associations are often difficult to establish. Whereas in physics, investigators have laws of thermodynamics which are constant and universal, researchers in clinical medicine have only hypotheses with which to work. We have few absolute truths. The more often the hypotheses are tested and hold up to scrutiny, the more likely they are to be true.

In the 1960s, Sir A. Bradford Hill proposed a series of criteria for judging whether associations are causal (76). The criteria have enjoyed wide use over the years. Some criteria are strong, whereas others are weak. For example, **temporal sequence** is a strong criterion. If the outcome preceded the exposure, then the hypothesis that the exposure caused the outcome evaporates.

Strength of the association is another strong criterion. The stronger the association between exposure and outcome (measured by the relative risk or odds ratio), the greater the likelihood of a cause-and-effect relationship. For example, some epidemiologists suggest that a relative risk greater than 3 in a

cohort study or 4 in a case-control study offers compelling evidence of a causal association. Stated alternately, bias would have to be very strong to account for such a large effect.

Consistency is another useful criterion. In general, one usually should not base practice on a single study, no matter how large or well done. Repetition of the finding in different populations, with different study designs and different researchers, supports the association being causal. An example would be the highly consistent literature from around the world indicating that use of combined oral contraceptives protects against endometrial cancer later in life (77). An important caveat is that consistent bias can result in consistent bogus results, as has occurred with the questions of abortion and breast cancer (62) and with IUDs and salpingitis (78).

A **dose-response relationship** also argues for causation. If increasing amounts or duration of exposure are linked with increasing effect, this provides further support for causation. An example would be the increased protection against ovarian cancer with increasing duration of oral contraceptive use. Another would be the increased risk of lung cancer as number of pack-years of smoking increases (79). Hence, researchers commonly try to demonstrate a biological gradient of effect, rather than just reporting a dichotomous exposure, e.g., smoking, yes or no.

Specificity is a weak criterion, inasmuch as many diseases have more than one cause. Rarely in medicine does one exposure lead to one outcome only. **Biological plausibility** is another weak one, limited by our meager understanding of human biology. In 1970, the notion that sanitary hygiene products might be linked with toxic shock syndrome would have been rejected as ridiculous. **Ancillary evidence** can be helpful. For example, the effect of HIV on human lymphocytes in the laboratory is consistent with the hypothesis that the virus causes AIDS in humans. **Reasoning by analogy** is a Pandora's box, opened profitably by plaintiff's lawyers. For example, because prenatal exposure to thalidomide caused birth defects, then prenatal exposure to spermicide can cause defects. This hypothesis, lacking support in the medical literature, led to an award of over $5 million to a plaintiff.

CONCLUSION

Understanding the basic taxonomy of clinical research is a prerequisite to critical reading of the literature. Most published reports in obstetrics and gynecology are observational. The results of these may be more capable of extrapolation to one's practice than those of experimental studies, but biases in observational research compromise internal validity. When reading observational studies, a four-point check list (selection bias, information bias, confounding, and chance) is a useful approach. When reading randomized controlled trials, the CONSORT guidelines (71) should be used as the standard. Published research reports using dichotomous outcomes should provide raw numbers, measures of frequency (such as rates), measures of association (such as relative risks), and 95% confidence intervals (to indicate precision). Hypothesis testing with P values alone is not recommended. As with surgery, critical reading is a learned skill. Proficiency grows with practice. Learning to read critically is both fun and important (for rust-proofing one's practice). We hope this brief Baedeker will help to guide the way.

REFERENCES

1. Grimes DA, Schulz KF. An overview of clinical research: The lay of the land. Lancet 2002;359:57–61.
2. Grimes DA, Schulz KF. Descriptive studies: What they can and cannot do. Lancet 2002;359:145–9.
3. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002;359:248–52.
4. Grimes DA, Schulz KF. Cohort studies: Marching toward outcomes. Lancet 2002;359:341–5.
5. Schulz KF, Grimes DA. Case-control studies: Research in reverse. Lancet 2002;359:431–4.
6. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: Chance, not choice. Lancet 2002;359: 515–9.
7. Schulz KF, Grimes DA. Allocation concealment in randomised trials: Defending against deciphering. Lancet 2002;359:614–8.
8. Schulz KF, Grimes DA. Blinding in randomised trials: Hiding who got what. Lancet 2002;359:696–700.
9. Schulz KF, Grimes DA. Sample size slippages in randomised trials: Exclusions and the lost and wayward. Lancet 2002;359: 781–5.
10. Grimes DA, Schulz KF. Uses and abuses of screening tests. Lancet 2002;359:881–4.
11. Schulz KF, Grimes DA. Unequal group sizes in randomised trials: Guarding against guessing. Lancet 2002;359:966–70.
12. Ramsey PG, Carline JD, Inui TS et al. Changes over time in the knowledge base of practicing internists. JAMA 1991;266: 1103–7.
13. Evans CE, Haynes RB, Birkett NJ et al. Does a mailed continuing education program improve physician performance? Results of a randomized trial in antihypertensive care. JAMA 1986;255:501–4.
14. Grimes DA, Blount JH, Patrick J et al. Antibiotic treatment of pelvic inflammatory disease. Trends among private physicians in the United States, 1966–1983. JAMA 1986;256: 3223–6.
15. Grimes DA. Technology follies. The uncritical acceptance of medical innovation [see comments]. JAMA 1993;269:3030–3.
16. Grimes DA. How can we translate good science into good perinatal care? Birth 1986;13:83–90.
17. Enkin M, Keirse MJNC, Neilson J et al., eds. A Guide to

Effective Care in Pregnancy and Childbirth, 3rd Ed. Oxford: Oxford University Press, 2000.
18. Thacker SB, Stroup DF, Peterson HB. Efficacy and safety of intrapartum electronic fetal monitoring: An update. Obstet Gynecol 1995;86:613–20.
19. Myers ER, McCrory DC, Subramanian S et al. Setting the target for a better cervical screening test: Characteristics of a cost-effective test for cervical neoplasia screening. Obstet Gynecol 2000;96:645–52.
20. Sawaya GF, Grimes DA. New technologies in cervical cytology screening: A word of caution. Obstet Gynecol 1999;94: 307–10.
21. Nezhat C, Pennington E, Nezhat F et al. Laparoscopically assisted anterior rectal wall resection and reanastomosis for deeply infiltrating endometriosis. Surg Laparosc Endosc 1991;1:106–8.
22. Nezhat F, Nezhat C, Pennington E et al. Laparoscopic segmental resection for infiltrating endometriosis of the rectosigmoid colon: A preliminary report. Surg Laparosc Endosc 1992;2:212–6.
23. Olatunbosun OA, Edouard L, Pierson RA. Physicians' attitudes toward evidence based obstetric practice: A questionnaire survey. BMJ 1998;316:365–6.
24. Grimes DA, Bachicha JA, Learman LA. Teaching critical appraisal to medical students in obstetrics and gynecology. Obstet Gynecol 1998;92:877–82.
25. Berman ML, Grosen EA. A new method of continuous vaginal cuff closure at abdominal hysterectomy. Obstet Gynecol 1994;84:478–80.
26. Korn AP, Grullon K, Hessol N et al. Does vaginal cuff closure decrease the infectious morbidity associated with abdominal hysterectomy? J Am Coll Surg 1997;185:404–7.
27. Aharoni A, Kaner E, Levitan Z et al. Prospective randomized comparison between an open and closed vaginal cuff in abdominal hysterectomy. Int J Gynaecol Obstet 1998;63: 29–32.
28. Eberhardt MS, Ingram DD, Makuc DM et al. Urban and Rural Health Chartbook. Health, United States, 2001; Hyattsville, MD: National Center for Health Statistics, 2001, Table 44.
29. Grimes DA. The case for confidence intervals. Obstet Gynecol 1992;80:865–6.
30. Peipert JF, Grimes DA. The case-control study: a primer for the obstetrician-gynecologist. Obstet Gynecol 1994;84: 140–5.
31. Sterne JA, Smith GD. Sifting the evidence-what's wrong with significance tests? BMJ 2001;322:226–31.
32. Silver RK, Helfand BT, Russell TL et al. Multifetal reduction increases the risk of preterm delivery and fetal growth restriction in twins: A case-control study. Fertil Steril 1997; 67:30–3.
33. Garg SK, Chase HP, Marshall G et al. Oral contraceptives and renal and retinal complications in young women with insulin-dependent diabetes mellitus. JAMA 1994;271: 1099–102.
34. Berenson AB, Chacko MR, Wiemann CM et al. A case-control study of anatomic changes resulting from sexual abuse. Am J Obstet Gynecol 2000;182:820–31; discussion 831–4.
35. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA 2001;285:1987–91.
36. Altman DG, Schulz KF, Moher D et al. The revised CONSORT statement for reporting randomized trials. Explanation and elaboration. Ann Intern Med 2001;134:663–94.
37. Funai EF, Rosenbush EJ, Lee MJ et al. Distribution of study designs in four major U.S. journals of obstetrics and gynecology. Gynecol Obstet Invest 2001;51:8–11.
38. Last JM, ed. A Dictionary of Epidemiology, 2nd Ed. New York: Oxford University Press, 1988.
39. Rooks JB, Ory HW, Ishak KG et al. Epidemiology of hepatocellular adenoma. The role of oral contraceptive use. JAMA 1979;242:644–8.
40. Balshem M, Oxman G, van Rooyen D et al. Syphilis, sex and crack cocaine: Images of risk and morality. Soc Sci Med 1992;35:147–60.
41. Ziel HK. Estrogen's role in endometrial cancer. Obstet Gynecol 1982;60:509–15.
42. Fritz MA. Addressing the dramatic rise in multiple pregnancies. Hosp Pract (Off Ed) 2000;35:124–6, 129–30, 133–4.
43. Centers for Disease Control and Prevention. Primary and secondary syphilis–United States, 1999. MMWR Morb Mortal Wkly Rep 2001;50:113–7.
44. Shelton JD, Brann EA, Schulz KF. Abortion utilization: Does travel distance matter? Fam Plann Perspect 1976;8: 260–2.
45. Levine RJ, Mathew RM, Chenault CB et al. Differences in the quality of semen in outdoor workers during summer and winter. N Engl J Med 1990;323:12–6.
46. Herbst AL, Anderson S, Hubby MM et al. Risk factors for the development of diethylstilbestrol-associated clear cell adenocarcinoma: A case-control study. Am J Obstet Gynecol 1986; 154:814–22.
47. Keirse MJ. Evidence-based childbirth only for breech babies? Birth 2002;29:55–9.
48. Trussell J, Vaughan B. Contraceptive failure, method-related discontinuation and resumption of use: Results from the 1995 National Survey of Family Growth. Fam Plann Perspect 1999; 31:64–72, 93.
49. Korn AP, Learman LA. Operations for stress urinary incontinence in the United States, 1988–1992. Urology 1996:48:609–12.
50. Centers for Disease Control and Prevention. Progress toward global eradication of poliomyelitis, 2001. MMWR Morb Mortal Wkly Rep 2002;51:253–6.
51. Caillouette JC, Koehler AL. Phasic contraceptive pills and functional ovarian cysts. Am J Obstet Gynecol 1987;156: 1538–42.
52. Thomas FB, Falko JM, Zuckerman K. Inhibition of intestinal iron absorption by laundry starch. Gastroenterology 1976;71: 1028–32.
53. Bullock MI. Ripcord release capability of female parachutists. Aviat Space Environ Med 1978;49:1177–83.
54. Egan KM, Stampfer MJ, Hunter D et al. Active and passive smoking in breast cancer: Prospective results from the Nurses' Health Study. Epidemiology 2002;13:138–45.
55. Hankinson SE, Hunter DJ, Colditz GA et al. Tubal ligation, hysterectomy, and risk of ovarian cancer. A prospective study. JAMA 1993;270:2813–8.
56. Leveno KJ, Cunningham FG, Nelson S et al. A prospective comparison of selective and universal electronic fetal monitoring in 34,995 pregnancies. N Engl J Med 1986;315: 615–9.
57. U. S. Preventive Services Task Force. Guide to Clinical Preventive Services, 2nd Ed. Baltimore, MD: Williams & Wilkins, 1996.
58. Peterson HB, Xia Z, Hughes JM et al. The risk of pregnancy after tubal sterilization: findings from the U.S. Collaborative Review of Sterilization. Am J Obstet Gynecol 1996;174: 1161–8; discussion 1168–70.
59. Kricker A, Elliott JW, Forrest JM et al. Congenital limb reduction deformities and use of oral contraceptives. Am J Obstet Gynecol 1986;155:1072–8.
60. Melbye M, Wohlfahrt J, Olsen JH et al. Induced abortion and the risk of breast cancer. N Engl J Med 1997;336:81–5.
61. Lindefors-Harris BM, Eklund G, Adami HO et al. Response bias in a case-control study: Analysis utilizing comparative

data concerning legal abortions from two independent Swedish studies. Am J Epidemiol 1991;134:1003–8.
62. Bartholomew LL, Grimes DA. The alleged association between induced abortion and risk of breast cancer: Biology or bias? Obstet Gynecol Surv 1998;53:708–14.
63. Washington AE, Cates W Jr, Zaidi AA. Hospitalizations for pelvic inflammatory disease. Epidemiology and trends in the United States, 1975 to 1981; JAMA 1984:251:2529–33.
64. Hill AB. Suspended judgment. Memories of the British Streptomycin Trial in Tuberculosis. The first randomized clinical trial. Control Clin Trials 1990;11:77–9.
65. Schulz KF, Chalmers I, Grimes DA et al. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals [see comments]. JAMA 1994;272:125–8.
66. Isenberg SJ, Apt L, Wood M. A controlled trial of povidone-iodine as prophylaxis against ophthalmia neonatorum. N Engl J Med 1995;332:562–6.
67. Vintzileos AM, Antsaklis A, Varvarigos I et al. A randomized trial of intrapartum electronic fetal heart rate monitoring versus intermittent auscultation. Obstet Gynecol 1993;81:899–907.
68. Bergman A, Ballard CA, Koonings PP. Comparison of three different surgical procedures for genuine stress incontinence: Prospective randomized study. Am J Obstet Gynecol 1989; 160:1102–6.
69. Schulz KF. Subverting randomization in controlled trials. JAMA 1995;274:1456–8.
70. Senn S. Testing for baseline balance in clinical trials. Stat Med 1994;13:1715–26.
71. Moher D, Schulz KF, Altman DG et al. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001;357: 1191–4.
72. Roddy RE, Zekeng L, Ryan KA et al. Effect of nonoxynol-9 gel on urogenital gonorrhea and chlamydial infection: A randomized controlled trial. JAMA 2002;287:1117–22.
73. Sinei SK, Schulz KF, Lamptey PR et al. Preventing IUCD-related pelvic infection. The efficacy of prophylactic doxycycline at insertion. Br J Obstet Gynaecol 1990;97: 412–9.
74. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: A comparative before-and-after evaluation. JAMA 2001;285:1992–5.
75. Begg C, Cho M, Eastwood S et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637–9.
76. Hill AB. The environment and disease association or causation. Proc Royal Soc Med 1965;58:295–300.
77. Schlesselman JJ. Risk of endometrial cancer in relation to use of combined oral contraceptives. A practitioner's guide to meta-analysis. Hum Reprod 1997;12:1851–63.
78. Grimes DA. Intrauterine device and upper-genital-tract infection. Lancet 2000;356:1013–9.
79. Doll R, Peto R, Wheatley K et al. Mortality in relation to smoking: 40 years' observations on male British doctors. BMJ 1994;309:901–11.