Epidemiology series

Bias and causal associations in observational research

David A Grimes, Kenneth F Schulz

Readers of medical literature need to consider two types of validity, internal and external. Internal validity means that the study measured what it set out to; external validity is the ability to generalise from the study to the reader's patients. With respect to internal validity, selection bias, information bias, and confounding are present to some degree in all observational research. Selection bias stems from an absence of comparability between groups being studied. Information bias results from incorrect determination of exposure, outcome, or both. The effect of information bias depends on its type. If information is gathered differently for one group than for another, bias results. By contrast, non-differential misclassification tends to obscure real differences. Confounding is a mixing or blurring of effects: a researcher attempts to relate an exposure to an outcome but actually measures the effect of a third factor (the confounding variable). Confounding can be controlled in several ways: restriction, matching, stratification, and more sophisticated multivariate techniques. If a reader cannot explain away study results on the basis of selection, information, or confounding bias, then chance might be another explanation. Chance should be examined last, however, since these biases can account for highly significant, though bogus results. Differentiation between spurious, indirect, and causal associations can be difficult. Criteria such as temporal sequence, strength and consistency of an association, and evidence of a dose-response effect lend support to a causal link.

Clinicians face two important questions as they read medical research: is the report believable, and, if so, is it relevant to my practice? Uncritical acceptance of published research has led to serious errors and squandered resources.[1] Here, we will frame these two questions in terms of study validity, describe a simple checklist for readers, and offer some criteria by which to judge reported associations.

Internal and external validity

Analogous to a laboratory test, a study should have internal validity—ie, the ability to measure what it sets out to measure.[2] The inference from participants in a study should be accurate. In other words, a research study should avoid bias or systematic error.[3] Internal validity is the sine qua non of clinical research; extrapolation of invalid results to the broader population is not only worthless but potentially dangerous.

A second important concern is external validity; can results from study participants be extrapolated to the reader's patients? Since a total enumeration or census approach to medical research is usually impossible, the customary tactic is to choose a sample, study it, and, hopefully, extrapolate the result to one's practice. Gauging external validity is necessarily more subjective than is assessment of internal validity.

Internal and external validity entail important trade-offs. For example, randomised controlled trials are more likely than observational studies to be free of bias,[4] but, because they usually enrol selected participants, external validity can suffer. This problem of unsuitable participants is also termed distorted assembly.[5] Participants in randomised controlled trials tend to be different (including being healthier[6–8]) from those who choose not to take part, a function of the restricted entry criteria. The filtering process for admission to randomised trials might, therefore, result in "a type of hothouse flower, which cannot bloom or be successfully removed beyond its special greenery."[5]

Lancet 2002; **359:** 248–52

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (D A Grimes MD, K F Schulz PhD)

Correspondence to: Dr David A Grimes
(e-mail: dgrimes@fhi.org)

Bias

Bias undermines the internal validity of research. Unlike the conventional meaning of bias—ie, prejudice—bias in research denotes deviation from the truth. All observational studies (and, regrettably, many badly done randomised controlled trials)[9,10] have built-in bias; the challenge for investigators, editors, and readers is to ferret these out and judge how they might have affected results. A simple checklist, such as that shown in panel 1, can be helpful.[11–14]

Several taxonomies exist for classification of biases in clinical research. Sackett's landmark compilation,[15] for example, included 35 different biases. By contrast Feinstein[5] consolidated biases into four categories that arise sequentially during research: susceptibility, performance, detection, and transfer. Susceptibility bias refers to differences in baseline characteristics, performance bias to different proficiencies of treatment, detection bias to different measurement of outcomes, and transfer bias to differential losses to follow-up. Another approach,[3,11,16,17] which is often used, is to group all biases into three general categories: selection, information, and confounding. The leitmotif for all three is "different".[17] Something "different" distorts the planned comparison.

Selection bias

Are the groups similar in all important respects?

Selection bias stems from an absence of comparability between groups being studied. For example, in a cohort study, the exposed and unexposed groups differ in some important respect aside from the exposure. Membership bias is a type of selection bias: people who choose to be members of a group—eg, joggers—might differ in important respects from others. For instance, both cohort and case-control studies initially suggested that jogging after myocardial infarction prevented repeat

Panel 1: **What to look for in observational studies**

Is selection bias present?

In a cohort study, are participants in the exposed and unexposed groups similar in all important respects except for the exposure?

In a case-control study, are cases and controls similar in all important respects except for the disease in question?

Is information bias present?

In a cohort study, is information about outcome obtained in the same way for those exposed and unexposed?

In a case-control study, is information about exposure gathered in the same way for cases and controls?

Is confounding present?

Could the results be accounted for by the presence of a factor—eg, age, smoking, sexual behaviour, diet—associated with both the exposure and the outcome but not directly involved in the causal pathway?

If the results cannot be explained by these three biases, could they be the result of chance?

What are the relative risk or odds ratio and 95% CI?[11,12]

Is the difference statistically significant, and, if not, did the study have adequate power to find a clinically important difference?[13,14]

If the results still cannot be explained away, then (and only then) might the findings be real and worthy of note.

infarction. However, a randomised controlled trial failed to confirm this benefit.[15] Those who chose to exercise might have differed in other important ways from those who did not exercise, such as diet, smoking, and presence of angina.

In case-control studies, selection bias implies that cases and controls differ importantly aside from the disease in question. Two types of selection bias have earned eponyms: Berkson and Neyman bias. Also known as an admission-rate bias, Berkson bias (or paradox) results from differential rates of hospital admission for cases and controls. Berkson initially thought that this phenomenon was due to presence of a simultaneous disease.[5] Alternatively, knowledge of the exposure of interest might lead to an increased rate of admission to hospital. For example, doctors who care for women with salpingitis were more likely to recommend hospital admission for those using an intrauterine device (IUD) than for those using a hormonal method of contraception.[18,19] In a hospital-based case-control study, this would stack the deck (or gynaecology ward) with a high proportion of IUD-exposed cases, spuriously increasing the odds ratio.

Neyman bias is an incidence-prevalence bias. It arises when a gap in time occurs between exposure and selection of study participants. This bias crops up in studies of diseases that are quickly fatal, transient, or subclinical. Neyman bias creates a case group not representative of cases in the community. For example, a hospital-based case-control study of myocardial infarction and snow shovelling (the exposure of interest) would miss individuals who died in their driveways and thus never reached a hospital; this eventuality might greatly lower the odds ratio of infarction associated with this strenuous activity.

Other types of selection bias include unmasking (detection signal) and non-respondent bias. An exposure might lead to a search for an outcome, as well as the outcome itself. For example, oestrogen replacement therapy might cause symptomless endometrial cancer patients to bleed, resulting in initiation of diagnostic tests.[20] In this instance, the exposure unmasked the subclinical cancer, leading to a spurious increase in the odds ratio. In observational studies, non-respondents are different from respondents. Cigarette smokers are a case in point: smokers are less likely to return questionnaires than are non-smokers or pipe and cigar smokers.[21]

Information bias

Has information been gathered in the same way?

Information bias, also known as observation, classification, or measurement bias, results from incorrect determination of exposure or outcome, or both. In a cohort study or randomised controlled trial, information about outcomes should be obtained the same way for those exposed and unexposed. In a case-control study, information about exposure should be gathered in the same way for cases and controls.

Information bias can arise in many ways. Some use the term ascertainment to describe gathering information in different ways. For example, an investigator might gather information about an exposure at bedside for a case but by telephone from a community control. Diagnostic suspicion bias implies that knowledge of a putative cause of disease might launch a more intensive search for the disease among those exposed, for example, preferentially searching for infection by HIV-1 in intravenous drug users. Conversely, the presence of a disease might prompt a search for the putative exposure of interest. Another type of bias is family history bias, in which medical information flows differently to affected and unaffected family members, as has been shown for rheumatoid arthritis.[22] To minimise information bias, detail about exposures in case-control studies should be gathered by people who are unaware of whether the respondent is a case or a control. Similarly, in a cohort study with subjective outcomes, the observer should be unaware of the exposure status of each participant.

In case-control studies that rely on memory of remote exposures, recall bias is pervasive. Cases tend to search their memories to identify what might have caused their disease; healthy controls have no such motivation. Thus, better recall among cases is common. For example, the putative association between induced abortion and subsequent development of breast cancer has emerged as a hot medical and political issue. Many case-control studies have reported an increase in cancer risk after abortion.[23] However, when investigators compared histories of prior abortions, obtained by personal interview, against centralised medical records, they documented systematic underreporting of abortions among controls (but not among cases) that accounted for a spurious association.[24] In Swedish and Danish cohort studies,[25,26] free from recall bias, induced abortion has had either a protective effect or no effect on risk of breast cancer.

Is the information bias random or in one direction?

The effect of information bias depends on its type. If information is gathered differentially for one group than for another, then bias results, raising or lowering the relative risk or odds ratio dependent on the direction of the bias. By contrast, non-differential misclassification—ie, noise in the system—tends to obscure real differences. For example, an ambiguous questionnaire might lead to errors in data collection among cases and controls, shifting the odds ratio toward unity, meaning no association.

Confounding

Is an extraneous factor blurring the effect?

Confounding is a mixing or blurring of effects. A researcher attempts to relate an exposure to an outcome, but actually measures the effect of a third factor, termed a confounding variable. A confounding variable is associated with the exposure and it affects the outcome, but it is not an intermediate link in the chain of causation between exposure and outcome.[27,28] More simply, confounding is a methodological fly in the ointment. Confounding is often easier to understand from examples than from definitions.

Oral contraceptives and myocardial infarction, and smoking

Early studies of the safety of oral contraceptives reported a pronounced increased risk of myocardial infarction. This association later proved to be spurious, because of the high proportion of cigarette smokers among users of birth control pills.[29–31] Here, cigarette smoking confounded the relation between oral contraceptives and infarction. Women who chose to use birth control pills also chose, in large numbers, to smoke cigarettes, and cigarettes, in turn, increased the risk of myocardial infarction. Although investigators thought they were measuring an effect of birth control pills, they were in fact measuring the hidden effect of smoking among pill users.

IUD insertion and salpingitis, and exposure to sexually transmitted disease

Results of a large case-control study of IUDs indicated a significant increase in salpingitis soon after insertion.[32] However, among married or cohabiting women with only one reported sex partner in the past 6 months, no significant increase in risk was evident.[33] In the study, exposure to sexually transmitted diseases apparently confounded the association. Even among women at low risk of salpingitis, frequent coitus might increase risk of infection,[34] and few studies have controlled for this variable.

Oral contraceptives and cervical cancer, and smoking

Reported associations between oral contraceptives and squamous cervical cancer[35] might be due to unsuspected confounding by cigarette smoking and human papillomavirus infection.[36] Control of confounding is inevitably limited by our meagre understanding of human biology; unsuspected confounding factors evade control in observational studies.[37]

Control for confounding

When selection bias or information bias exist in a study, irreparable damage results. Internal validity is doomed. By contrast, when confounding is present, this bias can be corrected, provided that confounding was anticipated and the requisite information gathered. Confounding can be controlled for before or after a study is done. The purpose of these approaches is to achieve homogeneity between study groups.

Restriction

The simplest approach is restriction (also called exclusion or specification).[28] For example, if cigarette smoking is suspected to be a confounding factor, a study can enrol only non-smokers. Although this tactic avoids confounding, it also hinders recruitment (and thus power) and precludes extrapolation to smokers. Restriction might increase the internal validity of a study at the cost of poorer external validity.

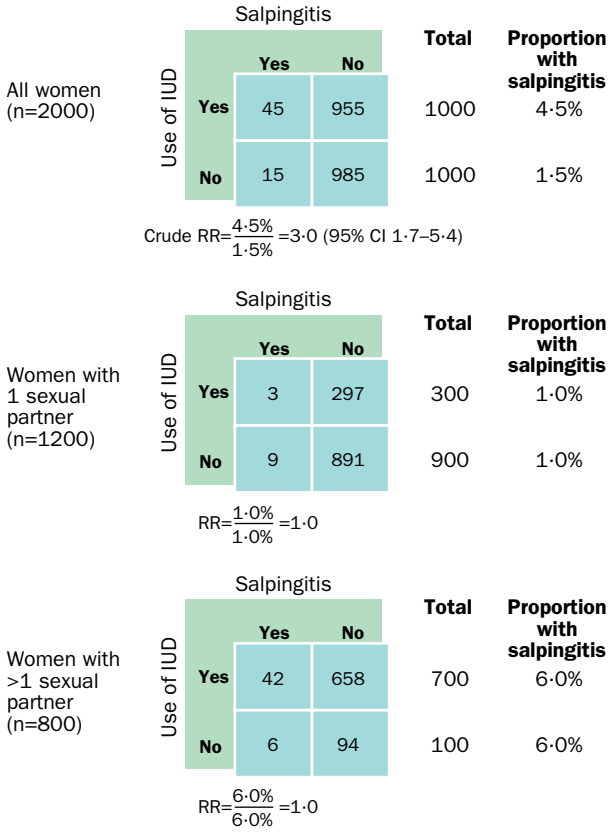
Matching

Another way to control for confounding is pairwise matching. In a case-control study in which smoking is deemed a confounding factor, cases and controls can be matched by smoking status. For each case who smokes, a control who smokes is found. This approach, although often used by investigators, has two drawbacks. If matching is done on several potential confounding factors, the recruitment process can be cumbersome, and, by definition, one cannot examine the effect of a matched variable.[28]

Stratification

Investigators can also control for confounding after a study has been completed. One approach is stratification. Stratification can be considered a form of post hoc restriction, done during the analysis rather than during the accrual phase of a study. For example, results can be stratified by levels of the confounding factor. In the smoking example, results are calculated separately for smokers and non-smokers to see if the same effect arises independent of smoking. The Mantel-Haenszel procedure[38] combines the various strata into a summary statistic that describes the effect. The strata are weighted inversely to their variance—ie, strata with larger numbers count more than those with smaller numbers. If the Mantel-Haenszel adjusted effect differs substantially from the crude effect, then confounding is deemed present. In this instance, the adjusted estimate of effect is considered the better estimate to use.

Confounding is not always intuitive, as shown by the fictitious example in the figure. In this hypothetical



Example of confounding in a hypothetical cohort study of intrauterine device use and salpingitis

When the crude relative risk is controlled for the confounding effect of number of sexual partners, the raised risk disappears.

cohort of 2000 women, use of an IUD was strongly related to development of salpingitis (relative risk 3·0; 95% CI 1·7–5·4). However, the number of sexual partners was related to women's choice of contraception and to their risk of upper-genital-tract infection. Here, a disproportionate number of women with more than one sexual partner chose to use an IUD (700 *vs* 300 women with only one partner). The number of partners was also related to the risk of infection (6% among those with >1 partner *vs* 1% among those with only one partner). In each stratum by number of partners, the relative risk is 1·0, indicating no association between the IUD and salpingitis. The Mantel-Haenszel weighted relative risk, which controls for this confounding effect, is 1·0 (95% CI 0·5–2·0). In this fictitious example, the apparent three-fold increase in risk associated with IUD use was all due to confounding bias.

Multivariate techniques

In multivariate techniques, mathematical modelling examines the potential effect of one variable while simultaneously controlling for the effect of many other factors. A major advantage of these approaches is that they can control for more factors that can stratification. For example, an investigator might use multivariate logistic regression to study the effect of oral contraceptives on ovarian cancer risk. In this way, they could simultaneously control for age, race, family history, parity, &c. Another example would be use of a proportional hazards regression analysis for time to death; this method could control simultaneously for age, blood pressure, smoking history, serum lipids, and other risk factors.[39] Disadvantages of multivariate approaches, for some researchers, include greater difficulty in understanding the results, and loss of hands-on feel for the data.[28]

Chance

If a reader cannot explain results on the basis of selection, information, or confounding bias, then chance might be another explanation. The reason for examination of bias before chance is that biases can easily cause highly significant (though bogus) results. Regrettably, many readers use the p value as the arbiter of validity, without considering these other, more important, factors.

The venerable p value measures chance. It advises the reader of the likelihood of a false-positive conclusion: a difference was seen in the study, although it does not exist in the broader population (type I error). Many clinicians are surprised to learn, however, that the p value of 0·05 as a threshold has no basis in medicine. Rather, it stems from agricultural and industrial experiments early in the 20th century.[40,41] Should a study not achieve significance at this level, one needs to see if the study had adequate power to find a clinically important difference. Many "negative" studies simply have too few participants to do the job.[13,14] Better yet, investigators should present measures of association with confidence intervals[41] in preference to hypothesis tests.

Judgment of associations

Bogus, indirect, or real?

When statistical associations emerge from clinical research, the next step is to judge what type of association exists. Statistical associations do not necessarily imply causal associations.[17] Although several classifications are available,[28] a simple approach includes just three types: spurious, indirect, and causal. Spurious associations are the result of selection bias, information bias, and chance. By contrast, indirect associations (which stem from confounding) are real but not causal.

Judgment of cause-effect relations can be tough. Few rules apply, though criteria first suggested by Hill have received the most attention (panel 2).[17,42,43] The only iron-clad criterion is temporality: the cause must antedate the effect. However, in many studies, especially with chronic diseases, answering this chicken-egg question can be daunting. Strong associations argue for causation. Whereas weak associations in observational studies can easily be due to bias, large amounts of bias would be necessary to produce strong associations. (This large bias is evident in reports that link IUD use with salpingitis.) Some suggest that relative risks more than 3 in cohort studies, or odds ratios greater than 4 in case-control studies, provide strong support for causation.[44] Consistent observation of an association in different populations and with different study designs also lends support to a real effect. For example, results of studies done around the world have consistently shown that oral contraceptives protect against ovarian cancer; a causal relation can, therefore, be argued. Evidence of a biological gradient supports a causal association too. For instance, protection against ovarian cancer is directly related to duration of use of oral contraceptives.[45] The risk of death from lung cancer is linearly related to years of cigarette smoking. In both of these examples, increasing exposure is associated with an increasing biological effect.

Other criteria of Hill's are less useful. Specificity is a weak criterion. With a few exceptions, such as the rabies virus, few exposures lead to only one outcome. Should an association be highly specific, this provides support for causality. However, since many exposures—eg, cigarette smoke—lead to numerous outcomes, lack of specificity does not argue against causation. Biological plausibility is another weak criterion, limited by our lack of knowledge. 300 years ago, clinicians would have rejected the suggestion that citrus fruits could prevent scurvy or that mosquitoes were linked with blackwater fever. Ancillary biological evidence that is coherent with the association might be helpful. For example, the effect of cigarette

Panel 2: **Criteria for judgment of causal associations**[17,42,43]

Temporal sequence
Did exposure precede outcome?

Strength of association
How strong is the effect, measured as relative risk or odds ratio?

Consistency of association
Has effect been seen by others?

Biological gradient (dose-response relation)
Does increased exposure result in more of the outcome?

Specificity of association
Does exposure lead only to outcome?

Biological plausibility
Does the association make sense?

Coherence with existing knowledge
Is the association consistent with available evidence?

Experimental evidence
Has a randomised controlled trial been done?

Analogy
Is the association similar to others?

smoke on the bronchial epithelium of animals is coherent with an increased risk of cancer in human beings. Finally, experimental evidence is seldom available, and reasoning by analogy has sometimes caused harm. Since thalidomide can cause birth defects, for instance, some lawyers (successfully) argued by analogy that Bendectin (an antiemetic widely used for nausea and vomiting in pregnancy) could also cause birth defects, despite evidence to the contrary.[46]

Conclusion

Studies need to have both internal and external validity: the results should be both correct and capable of extrapolation to the population. A simple checklist for bias (selection, information, and confounding) then chance can help readers decipher research reports. When a statistical association appears in research, guidelines for judgment of associations can help a reader decide whether the association is bogus, indirect, or real.

We thank Willard Cates and David L Sackett for their helpful comments on an earlier version of this report. Much of this material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

1 Grimes DA. Technology follies: the uncritical acceptance of medical innovation. *JAMA* 1993; **269:** 3030–33.
2 Last JM, ed. A dictionary of epidemiology, 2nd edn. New York: Oxford University Press, 1988.
3 Ahlbom A, Norell S. Introduction to modern epidemiology, 2nd edn. Chestnut Hill, Massachusetts: Epidemiology Resources, 1990.
4 Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983; **309:** 1358–61.
5 Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: WB Saunders Company, 1985.
6 Anon. The National Diet-Heart Study Final Report. *Circulation* 1968; **37:** I1–428.
7 Moinpour CM, Lovato LC, Thompson IM Jr, et al. Profile of men randomized to the prostate cancer prevention trial: baseline health-related quality of life, urinary and sexual functioning, and health behaviors. *J Clin Oncol* 2000; **18:** 1942–53.
8 Halbert JA, Silagy CA, Finucane P, Withers RT, Hamdorf PA. Recruitment of older adults for a randomized, controlled trial of exercise advice in a general practice setting. *J Am Geriatr Soc* 1999; **47:** 477–81.
9 Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998; **352:** 609–13.
10 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; **273:** 408–12.
11 Rothman KJ. Modern epidemiology. Boston: Little, Brown and Company, 1986.
12 Grimes DA. The case for confidence intervals. *Obstet Gynecol* 1992; **80:** 865–66.
13 Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 "negative" trials. *N Engl J Med* 1978; **299:** 690–94.
14 Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994; **272:** 122–24.
15 Sackett DL. Bias in analytic research. *J Chronic Dis* 1979; **32:** 51–63.
16 Wingo PA, Higgins JE, Rubin GL, Zahniser SC, eds. An epidemiologic approach to reproductive health. Geneva: WHO, 1994.
17 Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown and Company, 1987.
18 Burkman RT. Association between intrauterine device and pelvic inflammatory disease. *Obstet Gynecol* 1981; **57:** 269–76.
19 Kronmal RA, Whitney CW, Mumford SD. The intrauterine device and pelvic inflammatory disease: the Women's Health Study reanalyzed. *J Clin Epidemiol* 1991; **44:** 109–22.
20 Feinstein AR, Horwitz RI. Oestrogen treatment and endometrial carcinoma. *BMJ* 1977; **2:** 766–67.
21 Seltzer CC, Bosse R, Garvey AJ. Mail survey response by smoking status. *Am J Epidemiol* 1974; **100:** 453–57.
22 Schull WJ, Cobb S. The intrafamilial transmission of rheumatoid arthritis: 3, the lack of support for a genetic hypothesis. *J Chronic Dis* 1969; **22:** 217–22.
23 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.
24 Lindefors-Harris BM, Eklund G, Adami HO, Meirik O. Response bias in a case-control study: analysis utilizing comparative data concerning legal abortions from two independent Swedish studies. *Am J Epidemiol* 1991; **134:** 1003–08.
25 Harris BM, Eklund G, Meirik O, Rutqvist LE, Wiklund K. Risk of cancer of the breast after legal abortion during first trimester: a Swedish register study. *BMJ* 1989; **299:** 1430–32.
26 Melbye M, Wohlfahrt J, Olsen JH, et al. Induced abortion and the risk of breast cancer. *N Engl J Med* 1997; **336:** 81–85.
27 Abramson JH. Making sense of data. New York: Oxford University Press, 1988.
28 Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman RB, eds. Designing clinical research: an epidemiologic approach, 2nd edn. Baltimore: Lippincott Williams and Wilkins, 2001.
29 Ory HW. Association between oral contraceptives and myocardial infarction: a review. *JAMA* 1977; **237:** 2619–22.
30 Schwingl PJ, Ory HW, Visness CM. Estimates of the risk of cardiovascular death attributable to low-dose oral contraceptives in the United States. *Am J Obstet Gynecol* 1999; **180:** 241–49.
31 Jain AK. Cigarette smoking, use of oral contraceptives, and myocardial infarction. *Am J Obstet Gynecol* 1976; **126:** 301–07.
32 Lee NC, Rubin GL, Ory HW, Burkman RT. Type of intrauterine device and the risk of pelvic inflammatory disease. *Obstet Gynecol* 1983; **62:** 1–6.
33 Lee NC, Rubin GL, Borucki R. The intrauterine device and pelvic inflammatory disease revisited: new results from the Women's Health Study. *Obstet Gynecol* 1988; **72:** 1–6.
34 Lee NC, Rubin GL, Grimes DA. Measures of sexual behavior and the risk of pelvic inflammatory disease. *Obstet Gynecol* 1991; **77:** 425–30.
35 Schlesselman JJ. Cancer of the breast and reproductive tract in relation to use of oral contraceptives. *Contraception* 1989; **40:** 1–38.
36 Lacey JV Jr, Brinton LA, Abbas FM, et al. Oral contraceptives as risk factors for cervical adenocarcinomas and squamous cell carcinomas. *Cancer Epidemiol Biomarkers Prev* 1999; **8:** 1079–85.
37 Kjellberg L, Hallmans G, Ahren AM, et al. Smoking, diet, pregnancy and oral contraceptive use as risk factors for cervical intra-epithelial neoplasia in relation to human papillomavirus infection. *Br J Cancer* 2000; **82:** 1332–38.
38 Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959; **22:** 719–48.
39 Lang TA, Secic M. How to report statistics in medicine. Philadelphia: American College of Physicians, 1997.
40 Rothman KJ. A show of confidence. *N Engl J Med* 1978; **299:** 1362–63.
41 Sterne JA, Smith GD. Sifting the evidence: what's wrong with significance tests? *BMJ* 2001; **322:** 226–31.
42 Hill AB. The environment and disease association or causation. *Proc R Soc Med* 1965; **58:** 295–300.
43 Streiner DL, Norman GR, Munroe Blum H. PDQ epidemiology. Toronto: BC Decker, 1989.
44 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine, 2nd edn. Boston: Little, Brown and Company, 1991.
45 Grimes DA, Economy KE. Primary prevention of gynecologic cancers. *Am J Obstet Gynecol* 1995; **172:** 227–35.
46 McKeigue PM, Lamm SH, Linn S, Kutcher JS. Bendectin and birth defects: 1, a meta-analysis of the epidemiologic studies. *Teratology* 1994; **50:** 27–37.