Causation, bias and confounding: a hitchhiker's guide to the epidemiological galaxy

Part 1. Principles of causality in epidemiological research: time order, specification of the study base and specificity

Samuel Shapiro

Definitions and methods

In this series of articles the focus is on epidemiology as a tool in the exploration of causation. Other functions (e.g. administrative functions such as monitoring birth or death rates, or life expectancy, or the use of epidemiological data for guidance in health policy) are not considered.

In epidemiological research possible causation (or prevention) is explored by determining and comparing the incidence ('occurrence') of diseases ('outcomes', 'events'), as these occur among exposed and non-exposed persons in defined populations, over specified time intervals. Incidence per unit time is known as an incidence rate. A defined or implied population/time experience is known as a study base, and it can be analysed in two ways: in a follow-up ('cohort') study or in a case-control ('case-referent') study.

Follow-up studies

In a follow-up study, incidence rates of disease are compared among groups of exposed and non-exposed persons. Quite commonly, but not always, the rates are based on units of person-time (e.g. person-years). If the rates are similar (and unbiased and unconfounded: see Parts 1d, 1e and 2a¹) the evidence does not support causation; if the rate is higher among exposed persons it does; if it is lower it supports protection. The incidence rate in the exposed, divided by the rate in the non-exposed, is the estimated relative risk (RR) (the terms 'rate ratio' and 'incidence rate ratio' are synonyms; the terms 'hazard ratio' and 'hazard rate ratio' are slightly different, but for practical purposes the differences can be disregarded). Any RR estimate is only a rough approximation, and the practice of representing RRs to two decimal places is pseudo-precision;^{2,3} it has nevertheless become popular.

A RR estimate of 1.0 denotes no association ('the null'), and hence no evidence of causation or protection; >1.0 denotes a positive association (which may be causal) and <1.0 denotes a negative (and possibly protective) association. If an association is causal, the incidence rate in the non-exposed group, subtracted from the rate in the exposed group, roughly represents the excess incidence attributable to the exposure, or the absolute risk ('attributable risk', 'risk difference', 'rate difference' or, if an association is protective, 'risk reduction', 'risk difference', 'rate difference'). To confuse matters, however, the term 'attributable risk' is sometimes used, erroneously,

to denote the incidence of the disease in the *total* population, exposed *plus* non-exposed, that can be attributed to exposure. The correct designation is the 'population-attributable risk'.

One important type of follow-up study is a randomised, double-blind, controlled trial, in which persons ('subjects') are randomly allocated to two or more treatments, one of which is often, but not always, a placebo. Both the subjects and the observers are kept unaware of the assigned treatment. If virtually everyone is kept unaware it may (unfortunately) be called a 'triple-blind' trial. If only the subjects are kept unaware, it may (again unfortunately) be called a 'single-blind' trial, and if everyone knows the treatment it may simply be called a randomised trial.

Randomised trials, whether or not 'blinded', are said to be controlled ('experimental') whereas follow-up or case-control studies (described next) are said to be observational ('non-experimental'). Again to confuse matters, however, the adjective 'controlled' is sometimes used to denote quantitative observational evidence, as against clinical ('anecdotal') evidence, and 'experimental' means different things to different people.

Case-control studies

In a case-control study, either all cases of a disease occurring in a study base, or a sample of cases selected to be representative of the base, are compared with a representative sample of 'non-cases' ('controls'). If all cases identified in a study base (i.e. all cases occurring over a defined time interval in a defined population) are selected, the study is said to be 'population-based'. Sometimes, for efficiency, the data in a follow-up study (usually a massive one) are used to carry out a case-control analysis in which cases that occur during follow-up are compared with a sample of non-cases ('nested case-control study').

Rates of exposure before the onset of the disease ('index date') are compared among the cases and the controls. If the rates are similar (and again, unbiased and unconfounded) the evidence does not suggest causation; if the rate is higher among the cases it does; if it is lower it suggests protection.

The ratio of exposed to non-exposed cases (i.e. the odds of the cases being exposed) is divided by the corresponding ratio among the controls (i.e. the odds of controls being exposed) to yield a further ratio, the odds ratio (OR). Except in extreme conditions the OR is a close approximation of the RR, and it may simply be called a RR. With simple arithmetic, based on the OR, and on the percentage of exposed cases, the percentage of cases attributable to the exposure ('aetiological fraction') can be estimated. Then, if the total incidence rate of the disease is known, either from the data (as in a population-based or nested case-control study), or otherwise from some external source (e.g. a cancer registry), the aetiological fraction can be applied to that rate in order to estimate the absolute risk. If the incidence rate is not known, the absolute risk cannot be estimated in a case-control study.

J Fam Plann Reprod Health Care 2008; **34**(2): 83–87

(Accepted 12 February 2008)

Department of Epidemiology, University of Cape Town, Cape Town, South Africa

Samuel Shapiro, FCP(SA), FRCP(E), Visiting Professor of Epidemiology (also Emeritus Director, Slone Epidemiology Center, Boston University, Boston, MA, USA)

Correspondence to: Professor Samuel Shapiro, Department of Public Health and Family Medicine, University of Cape Town, Cape Town, South Africa. E-mail: samshap@mweb.co.za

As should be clear from the above description, both the follow-up and case-control approaches simply constitute alternative methods of identifying, selecting ('sampling') and comparing exposed and non-exposed, and diseased and non-diseased individuals, in a study base. Therefore, if a follow-up or a case-control study is carried out in the same study base, the association should be the same.

The decision as to which approach to use is governed principally by practical considerations. If an outcome is relatively common (e.g. pregnancy) and the exposure is rare (e.g. *in vitro* fertilisation), a follow-up study is usually the best choice. Conversely, for the evaluation of the risk of a relatively uncommon disease (e.g. breast cancer in premenopausal women) in relation to a common exposure (e.g. oral contraceptives), the case-control approach is generally the more feasible one. However, when there are massive databases (e.g. national registries of exposures and outcomes, or so-called 'automated record linkage systems' or 'automated databases'), or when massive resources are invested in an *ad hoc* study (e.g. the observational component of the Multiple Risk Factor Intervention Trial),⁴ follow-up studies can occasionally be conducted even when an outcome is relatively uncommon. Sometimes, given sufficient resources, both approaches can be used (e.g. oral contraceptive use and the risk of venous thromboembolism).⁵

The terms 'prospective' and 'retrospective' are commonly used to denote follow-up and case-control studies. However, those terms can be confusing, since follow-up studies can be 'retrospective' and case-control studies can be 'prospective'. For example, a follow-up study may be called a 'retrospective cohort study' if the exposures and diseases have all occurred in the past (e.g. in an automated record linkage system containing data on prescriptions and diagnoses). Conversely, a case-control study may be called 'prospective' if newly occurring cases of disease are prospectively identified only after the study has commenced. And in the silly season, designations such as 'retrospective prospective' or 'prospective retrospective', or worse, may be encountered.

Data from a series of follow-up or case-control studies (or both) are sometimes synthesised ('aggregated') in a meta-analysis ('combined analysis', 'pooled analysis', 'synthetic meta-analysis'). To confuse matters, however, the term 'pooled analysis' is sometimes used to denote the synthesis of the 'raw data' from the individual studies, as opposed to the synthesis of only the published data. In such studies (also sometimes called 'collaborative analyses' or 'collaborative re-analyses') the 'raw data' from the individual studies are reclassified, standardised and then synthesised.

In a meta-analysis statistical techniques that make allowance for the sizes of the individual studies, and for other factors, are used to synthesise RRs across a range of studies, in order to produce a 'summary RR' ('aggregated RR', 'pooled RR') which is more statistically robust (see Part 3a: statistical stability⁶) than the RRs in the individual studies. If the RRs among the studies are statistically heterogeneous (i.e. if the variability is unlikely to be due to chance) they are sometimes synthesised by means of a 'random effects model' that purportedly takes the heterogeneity into account.

For completeness some variants of the follow-up and case-control approaches also need to be mentioned. In general they tend to be less rigorous and more restricted as tools in causal research.

A cross-sectional study is basically a case-control study in which cases present in a study base either at a single instant, or over a specified time interval, usually a short

one, are included. That is, a cross-sectional study represents a hodgepodge of pre-existing and newly occurring cases ('prevalent cases'). The more long-lasting the disease, the higher is the proportion of pre-existing cases, and the greater is the possibility that the onset of the illness preceded the date of the first exposure (see Part 1a: time order).

A case-cohort ('case-base') study is a form of case-control study in which the cases are compared with controls selected at random (or equivalent) from the *total* study base (i.e. the potential cases *plus* the potential controls). Thus some of the 'controls' may actually be cases. The distortion thereby introduced is deemed to be minor if the cases constitute only a small proportion of the study base.

In a case-only study, another form of case-control study, the exposure rate in the cases is compared with an estimated 'control' exposure rate, based on theoretical considerations. Except when there are secure grounds to assume that the latter rate is exceedingly low, that estimate may be suspect.

A case-crossover study is one in which the hypothesised causal exposure is of short duration, and the disease has a rapid onset. Assumed 'induction' and 'non-induction' periods within the same individuals are compared. The two periods are deemed, as it were, to be the 'case' and the 'control' periods. Case-crossover designs have been advocated for exploring whether acute 'exposures', such as sexual intercourse, or alcoholic binges, precipitate acute events such as heart attacks⁷ (one wonders about the combination).

In ecological ('correlational') studies independent trends over time ('secular trends') are compared, whereas in observational studies each individual is classified as exposed or non-exposed, and diseased or non-diseased. The term 'analytical' is sometimes used to distinguish observational from ecological studies. Once again to confuse matters, however, that term is sometimes used to distinguish epidemiological evidence, whether observational or ecological, from anecdotal evidence.

Ecological studies can confirm what is otherwise already obvious (e.g. correlated declines in the incidence of lung cancer and tobacco sales),⁸ but occasionally even seemingly 'obvious' correlations may be misleading. Recently, for example, a decline in the use of menopausal hormones in the USA has been correlated with a decline in the incidence of breast cancer,⁹ even though the data indicated that the incidence began to drop before the decline in hormone use commenced (see Part 1a: time order). That correlation was nevertheless interpreted as evidence to support causality. A correlated decline in hormone use and breast cancer incidence has not been observed elsewhere.¹⁰⁻¹²

Since ecological studies do not compare incidence rates in exposed and non-exposed persons, but simply compare trends over time, they are subject to what has been called the 'ecological fallacy'. During much of the 20th century, for example, the incidence of heart attacks correlated with the sales of cars.

Standardised incidence or mortality ratios ('SIRs' or 'SMRs') are incidence or mortality rates, as determined in exposed cohorts, divided by rates recorded in general population statistics. The ratios are usually adjusted ('standardised') for the confounding effects (see Part 2: confounding¹) of age, and sometimes for a few other factors, such as sex ('gender') or race ('ethnic group'), but it is seldom possible to adjust for much more. ['Mortality rates' and 'fatality rates' should not be confused: mortality rates are incidence rates of deaths, with the study base as the denominator; fatality rates are incidence rates of deaths, with diseased persons as the denominator.]

Principles of causality in epidemiological research

In Part 1 of this series, five principles are defined, and considered in turn:

1a: Time order

1b: Specification of the study base

1c: Specificity

1d: Bias due to random misclassification

1e: Bias due to systematic misclassification.

Additional principles will be considered in Parts 2 and 3.

Whatever the type of study undertaken, whether follow-up or case-control, once the data have been assembled and analysed, if an association is shown, the next step is to evaluate whether or not the evidence supports a causal (or protective) inference – and if it does, how strongly. The same evaluation must be made if there is no association, and causality is to be rejected. Any evaluation is a matter of judgment, and to assist that judgment certain principles ('criteria') of causality have been developed by authorities such as Bradford Hill in the UK¹³ and Susser^{14,15} in the USA. Apart from the self-evident principle that for an exposure to be a cause it *must* commence before the onset of the illness (see Part 1a: time order), none of the principles are absolute, and most of them are interrelated.

Different epidemiologists conceive of causal criteria in different ways, and apply them differently, but to justify an inference of causality (or protection or no effect) the evidence should satisfy a broad array of generally accepted criteria. But since the criteria are not absolute, judgments can vary. The following is a distillation of the criteria that I use in making judgments about causality.

1a: Time order ('temporality')

Even though the requirement that the exposure must occur before the onset of the illness is obvious, it is not always met. Violation of time order may occur in cross-sectional or ecological studies, or when early symptoms of a disease are already present, but as yet undiagnosed. In the latter instance a seeming association may arise, not because the exposure causes the disease, but because the disease 'causes' the exposure (so-called 'protopathic bias' – which is not a bias).

Things can become especially tricky for diseases with a gradual onset. For example, in a study of end-stage renal disease it may be impossible to determine when the disease actually commenced, and antecedent symptoms of renal insufficiency may provoke the use of analgesics (the alleged cause of analgesic nephropathy).¹⁶ The evaluation of cancer risk may also be problematic since the onset is commonly insidious. To deal with these difficulties, 'time-lag' analyses, in which exposures during specified periods immediately before the date of diagnosis are censored, are sometimes carried out on the assumption that exposures that took place in the more distant past are more likely to have preceded the 'true' onset of the illness. Even then, however, such analyses can sometimes be frustrating and difficult to interpret.

Example: Violation of time order. Menopausal hormones and ovarian cancer. In a recent report from the Million Women Study it was claimed that the use of menopausal hormones causes ovarian cancer.¹⁷ That study had several defects, considered elsewhere.¹⁸ What is relevant here, however, is that the risk of ovarian cancer was increased only while the women were still using menopausal hormones. As soon as they stopped the risk was no longer increased. On clinical and pathological grounds, under a causal hypothesis an instantaneous drop in the incidence as soon as hormone use ceases is impossible.¹⁸ The only plausible explanation is that women

with as yet undiagnosed ovarian cancer commonly experienced symptoms such as pelvic discomfort, pain during sexual intercourse or urinary difficulties that 'caused' the use of menopausal hormones, not the reverse.

1b: Specification of the study base

For a study to be valid the base should include persons who are at risk both for the exposure and for the outcome. Surprisingly, this requirement is not always met.

Example: Women not at risk for the exposure. Fertility drugs and ovarian cancer. In a meta-analysis ('collaborative analysis') of three case-control studies, among women who had never been pregnant ('nulligravidae'), fertility drugs were alleged to increase the risk of ovarian cancer by some 27-fold.¹⁹ At the time when the infertile women were treated, fertility drugs either were under investigation and not yet generally available, or they did not yet exist.²⁰

Example: Women not at risk for the outcome. Hormone replacement therapy and ovarian cancer. In the above-mentioned study of ovarian cancer,¹⁷ hysterectomised women who were not at risk because their ovaries had also been removed, were included.¹⁸

1c: Specificity ('precision')

Epidemiological studies can never be totally precise, and the degree of imprecision ('misclassification', 'error', 'misclassification error', 'measurement error') can vary. Given imprecision there is always an opportunity for the occurrence of bias, and the greater the imprecision the greater is that opportunity. Bias may either be random or systematic.

1d: Bias due to random misclassification

If there is roughly equal imprecision in the specification of exposures or outcomes in the groups compared, such misclassification tends to introduce 'random noise', the effect of which is almost always to bias elevated (or reduced) RR estimates toward 1.0 ('the null'). Conversely, a causal inference gains in credibility if an association becomes stronger as the data become more precise.

In follow-up studies, misclassification may occur if the exposures or outcomes are poorly defined. Misclassification may also occur if there are losses to follow-up, or in a randomised trial, if there is non-adherence to treatment. In short-term clinical trials carried out in a matter of days or weeks (e.g. a comparison of a hypnotic and placebo in the treatment of insomnia) the degree of misclassification may be minor. However, on the epidemiological time scale, when trials last months or years considerable misclassification may occur.

In case-control studies, misclassification may occur if a large proportion of targeted cases or controls (e.g. drug addicts) are not enrolled, or if the exposures and outcomes are poorly defined. It may also occur if the exposures are misremembered (e.g. if menopausal women cannot remember how old they were when they began to menstruate).

When misclassification occurs, if it is extensive it may not be justifiable to assume that it is indeed random, rather than systematically biased or confounded (see Part 1e: bias due to systematic misclassification; and Part 2a: confounding¹). For this reason, ideally, if any epidemiological study is to be demonstrably valid, the aim must be to obtain precise data: in cohort studies follow-up should be virtually complete; in randomised trials, follow-up and adherence to treatment should also be virtually complete; and in case-control studies enrolment of targeted subjects should be virtually complete.

Despite these considerations, somewhat perversely, when a small but statistically significant association is observed (see Part 2c: strength of association¹; and Part 3a: statistical stability⁶), the tendency for random misclassification to bias it toward the null is sometimes invoked as evidence to support causality. It is argued that in the absence of misclassification the association would be stronger, and even more 'significant'. Hence, the estimated magnitude of the observed association is 'conservative', and the observed association would have been stronger if only the data had been more precise.

As should be intuitively obvious, except in instances in which the misclassification is minor, and is indeed likely to be random, that argument is untenable, for two reasons. First, as already mentioned, the greater the imprecision, the greater is the opportunity for some or all of the bias not to be random, but to be systematic, or for the data to be confounded (see Part 2a: confounding¹).

Second, consider a hypothetical study and assume that all the data are misclassified. In such a study any seeming association would of course be nonsensical. If we now move away from that extreme, it is nevertheless obvious that any study in which there is more than minor misclassification is suspect. Imprecise evidence is not converted into valid evidence by means of questionable assumptions. Indeed, a logical extension of those assumptions is that we should always design imprecise studies since any observed associations will then be 'conservative'.

Example: Purportedly random misclassification in a randomised controlled trial. Estrogen plus progestogen therapy and breast cancer. In the initial report from the Women's Health Initiative (WHI) randomised controlled trial of estrogen plus progestogen versus placebo, the RR of breast cancer was 1.26, a finding that "... almost reached statistical significance".²¹ The WHI study of combined estrogen-progestogen therapy had multiple defects: among other things, it soon ceased to be either 'blinded' or randomised, and it took on the characteristics of an observational study.²² What is relevant here, however, is that the RR was based on an 'analysis according to intention-to-treat'. In such an analysis it is assumed that all participants have adhered to their assigned treatments, when in fact some have not. Hence, under the assumption of random misclassification, analysis of the data as if the all subjects adhered to treatment gives rise to RR estimates that are, if anything, 'conservative'. In addition, if the reason for stopping ('discontinuation') (e.g. severe illness) is a determinant in the outcome (e.g. non-recovery) (i.e. if the reason for stopping is a confounding factor), confounding is reduced or avoided (see Part 2a: confounding¹).

This reasoning, although counterintuitive, may have some merit when the rates of non-adherence to the assigned treatments are low. However, in the WHI study the rates were not low: 42% and 38% of the estrogen plus progestogen and placebo recipients, respectively, stopped their treatments, and the assumption of random misclassification was not defensible. In addition, 11% of the placebo recipients were prescribed female hormones by their own doctors ('drop-ins', 'crossovers'), and analysis according to intention-to-treat would not have eliminated any confounding thereby introduced (see Part 2a: confounding¹).

1e: Bias due to systematic misclassification

Systematic bias exists if the ascertainment of the exposure or the outcome is asymmetrical in the compared groups.

Depending on the direction of the asymmetry, such bias may result in overestimation or underestimation of an association. If the bias is severe enough it may even produce an apparent relationship, when in an unbiased comparison there would be no association at all – or at the extreme, if the bias is still more powerful, it may even reverse the direction of a 'true' association. To minimise systematic bias it is essential to obtain the information on the exposure and outcome in as precise a manner as possible, and in exactly symmetrical ways.

In the epidemiological literature different types of systematic bias have been described using a confusing and inconsistent terminology. For example, in case-control studies the terms 'information bias' and 'selection bias', respectively, are commonly used to denote asymmetrical ascertainment of the exposure and the outcome; in follow-up studies the designation may be reversed. And to further confuse matters, in follow-up studies, at recruitment ('time of enrolment', 'baseline'), differences in the distribution of risk factors, such as age or sex, are sometimes mistakenly designated as 'selection bias' when in fact those differences are not instances of bias at all, but of confounding (see Part 2a: confounding¹).

And to confuse matters still further, a bewildering number of biases, with a correspondingly bewildering terminology, have been described. Several years ago one study identified 39 biases,²³ each of which was given its own label. And to make things still worse, over the years that number has increased.

This needless multiplication of labels is a pity, since fundamentally there are only two forms of systematic bias: bias in the ascertainment of the exposure and bias in the ascertainment of the outcome. If that distinction is kept in mind, the inconsistent and sometimes illogical terminology that appears in the literature can be deciphered.

In follow-up studies, since exposed and non-exposed persons do not yet have the disease under study, biased ascertainment of the exposure is seldom a problem. By contrast, in interview- or questionnaire-based case-control studies, if the cases and controls remember differently, biased ascertainment of the exposure may occur. However, if the data are obtained from some 'objective' source (e.g. automated prescription data) that bias may be avoided.

In both follow-up and case-control studies, bias in the ascertainment of the outcome (sometimes called 'detection bias' in follow-up studies, and commonly called 'selection bias' in case-control studies) can occur. For example, awareness that oral contraceptives cause venous thromboembolism may result in a systematic tendency to (correctly) diagnose otherwise 'silent' cases more commonly among exposed than among non-exposed women – or put another way, among non-exposed women 'true' cases may more commonly be 'missed'. One major advantage of controlled trials, relative to observational research, is that 'blinding', if maintained, usually minimises detection bias.

Example: Biased ascertainment of the exposure and the outcome in a case-control study. In a case-control study of oral contraceptive use and venous thromboembolism, relative to non-exposure, the RR for 'first plus second generation' oral contraceptive users was 5.2; for 'third-generation' users it was 48.6.²⁴ In that study the ascertainment of the exposure was asymmetrical ('information bias'): among the cases the information on oral contraceptive use was obtained by abstraction of medical records, augmented by telephone interviews; among the controls it was obtained by a mailed questionnaire. The ascertainment of the outcome was also

asymmetrical ('selection bias'): the cases were identified in hospital discharge registries; the controls were blood donors.

Example: Systematic bias in a meta-analysis. In a meta-analysis ('collaborative re-analysis') of 54 studies (53 297 cases and 100 239 controls) of the risk of breast cancer in relation to the use of oral contraceptives the overall RR for current users was 1.07. Among women last exposed <1, 1–4, 5–9 and ≥10 years previously the RRs were 1.24, 1.16, 1.07 and 1.01, respectively (trend $p < 0.00001$) (see Part 3a: statistical stability and Part 3b: dose- and duration-response effects⁶).

Anxiety about the possibility that oral contraceptives may increase the risk of breast cancer has existed from the time they were introduced, and most of the data were from interview- or questionnaire-based case-control studies. That anxiety would have been shared across studies. The possibility that there was a systematic tendency for the cases to remember their oral contraceptive use more completely than the controls could not be ruled out; nor could the possibility be ruled out that the most recent use was the most completely remembered by the cases ('information bias').

Breast cancers can sometimes be slow growing and go undiagnosed unless actively searched for (e.g. breast examinations, mammography). Across both the follow-up and the case-control studies included in the meta-analysis, such cancers could more commonly have been diagnosed among current oral contraceptive users, and more commonly have been diagnosed the more recent the use ('selection or detection bias').

A sensitivity analysis (see Part 2c: strength of association¹) showed that it would have taken minimal bias in the ascertainment of the exposure, together with minimal bias in the ascertainment of the outcome, to fully account for the findings – and far less bias than could conceivably be ruled out as an alternative explanation to causation.¹

In short, x times bias is still bias, and the only effect of increasing the size of x by synthesising RRs derived from multiple studies is to set narrower confidence limits (see Part 3a: statistical stability⁶) around the magnitude of that bias – or put another way, to make the bias more 'statistically significant'.

Future articles

The next article in this series (Part 2) will cover additional principles of causality: confounding, effect modification and strength of association.¹ Part 3 will cover further principles of causality: statistical stability, dose- and duration-response effects, internal consistency, external consistency, analogy and biological plausibility.⁶

Acknowledgements

Apologies to the late Douglas Adams, author of *The Hitchhiker's Guide to the Galaxy* (Pan Books, 1979). Parts of this essay are taken verbatim, or in modified form, from an expert report (An Overview of Recent Evidence Concerning the Risk of Venous Thromboembolism Among Women Using "Third Generation" and "Second Generation" Oral Contraceptives) submitted by the author as testimony in a trial before a British court [High Court of Justice, Queens Bench Division, Case No. 0002638, Neutral Citation No. [2002] EWFC 1420 (QB), before The Honourable Mr Justice Mackay]. Among other things, that report included a section on principles of causality in epidemiology.

Statements on funding and competing interests

Funding None identified.

Competing interests The author presently consults, and in the past has consulted, with manufacturers of products discussed in this article.

References

1 Shapiro S. Causation, bias and confounding: a hitchhiker's

- guide to the epidemiological galaxy. Part 2. Principles of causality in epidemiological research: confounding, effect modification and strength of association. *J Fam Plan Reprod Health Care* 2008; **34**: (in press).
- 2 Shapiro S. Bias in the evaluation of low-magnitude associations: an empirical perspective. *Am J Epidemiol* 2000; **151**: 939–945.
- 3 Hertz-Picciotto I. Invited commentary. Shifting the burden of proof regarding biases and low magnitude associations. *Am J Epidemiol* 2000; **151**: 946–948.
- 4 Davey Smith G, Phillips AN, Eaton JD. Smoking as an "independent" risk factor for suicide: illustration of an artifact from observational epidemiology. *Lancet* 1992; **340**: 708–712.
- 5 Farmer RD, Lawrenson RA, Todd JC, Williams TI, MacRae K. Oral contraceptives and venous thromboembolic disease. Analyses of UK General Practice Research Database and the UK Mediplex Database. *Hum Reprod Update* 1999; **5**: 688–706.
- 6 Shapiro S. Causation, bias and confounding: a hitchhiker's guide to the epidemiological galaxy. Part 3: Principles of causality in epidemiological research: statistical stability, dose- and duration-response effects, internal and external consistency, analogy and biological plausibility. *J Fam Plan Reprod Health Care* 2008; **34**: (in press).
- 7 Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol* 1991; **133**: 144–153.
- 8 International Agency for Research on Cancer (IARC). *Tobacco Smoke and Involuntary Smoking (IARC Monographs on the Evaluation of Carcinogenic Risk to Humans Vol. 44)*. Lyon, France: IARC, 2004.
- 9 Ravdin PM, Cronin KA, Howlader N, et al. The decrease in breast-cancer incidence in 2003 in the United States. *N Engl J Med* 2007; **356**: 1670–1674.
- 10 Kliever EV, Demers AA, Nugent ZJ. A decline in breast cancer incidence [Letter to the Editor]. *N Engl J Med* 2007; **357**: 509–510.
- 11 Mueck AL, Wallwiener D. Brustkrebsrate und HRT-verordnungen: differierende Daten aus USA und Europa. *Frauenarzt* 2007; **48**: 812–816.
- 12 Zahl P-H. A decline in breast cancer incidence [Letter to the Editor]. *N Engl J Med* 2007; **357**: 510–511.
- 13 Hill AB. The environment and disease: association or causation? *Proc Roy Soc Med* 1965; **58**: 295–300.
- 14 Susser M. *Causal Thinking in the Health Sciences*. New York, NY: Oxford University Press, 1979.
- 15 Susser M. What is a cause and how do we know one? A grammar for pragmatic epidemiology. *Am J Epidemiol* 1991; **133**: 635–648.
- 16 Delzell E, Shapiro S. A review of epidemiologic studies of non-narcotic analgesics and chronic renal disease. *Medicine* 1998; **77**: 102–121.
- 17 Million Women Study Collaborators. Ovarian cancer and hormone replacement therapy in the Million Women Study. *Lancet* 2007; **369**: 1703–1710.
- 18 Shapiro S. False alarm: postmenopausal hormone therapy and ovarian cancer. *Climacteric* 2007; **10**: 469–470.
- 19 Whittemore AS, Harris R, Intyre J. Characteristics relating to ovarian cancer risk: collaborative analysis of 12 US case-control studies II. Invasive epithelial cancer in white women. Collaborative Ovarian Cancer Group. *Am J Epidemiol* 1992; **135**: 1184–1203.
- 20 Shapiro S. Is meta-analysis a valid approach to the evaluation of small effects in observational studies? *J Clin Epidemiol* 1997; **50**: 223–229.
- 21 Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. Principal results from the Women's Health Initiative randomized controlled trial. *JAMA* 2002; **28**: 321–333.
- 22 Shapiro S. Risks of estrogen plus progestin therapy: a sensitivity analysis of findings in the Women's Health Initiative randomized controlled trial. *Climacteric* 2003; **6**: 302–310.
- 23 Sackett DL. Bias in analytic research. *J Chronic Dis* 1979; **32**: 51–63.
- 24 Andersen BS, Olsen J, Nielsen GL, et al. Third generation oral contraceptives and heritable thrombophilia as risk factors for non-fatal venous thromboembolism. *Thromb Haemost* 1998; **79**: 28–31.