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

Samuel Shapiro

Scope of article

In Part 1 of this review the following principles of causation were considered: time order, specification of the study base, specificity, bias due to random misclassification, and bias due to systematic misclassification.[1] Part 2 continued with confounding, effect modification, and strength of association.[2] Part 3 concludes the consideration of causal principles and will discuss:

3a: Statistical stability
3b: Dose- and duration-response effects
3c: Internal consistency
3d: External consistency
3e: Analogy
3f: Biological plausibility.

3a: Statistical stability

The focus in this review is on exposed and non-exposed, and diseased and non-diseased, individuals ('categorical variables'). The statistical assessment of measurements made on a continuous scale ('continuous variables') (e.g. mean height or weight) is not considered.

Confidence in causality is strengthened if an association is stable: that is, if the numbers (the numerators and denominators of the compared rates) are sufficient to rule out chance ('sampling error') with reasonable confidence. By convention, 'confidence' is estimated by setting a confidence interval (CI) ('confidence limits') around the relative risk (RR) estimate ('point estimate'). CIs are usually set at 95% or, if one wants to be more rigorous, at 99%. A CI that excludes 1.0 is taken to suggest that it is unlikely that the observed association is due to chance. In addition, if the CI is relatively narrow, the extent to which the RR might vary if the same study were to be repeated is correspondingly reduced.

An alternative to estimation of the RR and its CI is to estimate statistical significance. By convention, an association is deemed to be significant if the probability that it could occur by chance is <5% or, if one wants to be especially rigorous, <1% (technically, '$p<0.05$' or '$p<0.01$'). RRs and their CIs, and p values, are derived from the same calculations.

A drawback to p values is that they do not provide direct insight into the magnitude of the association or into the extent that a RR point estimate may vary by chance if the same study were to be repeated. For this reason, description of an association in terms of the RR and its CI is usually preferable since it is more informative. However, when all that is at issue is whether or not any given observation is due to chance, it may be more convenient to give p values, for example, in describing trends (e.g. dose- or duration-response effects: see Part 3b below).

Throughout the 20th century, great advances have been made in statistical methods, among them the ability to control for multiple potentially confounding factors simultaneously by means of multivariate analysis. Indeed, it is virtually impossible to conceive of modern epidemiology without the benefit of those advances. Again, however, the same advances have at the same time engendered certain drawbacks. One important drawback is a lack of transparency if the data are presented in a way that makes it impossible for an independent observer to check the evidence for himself ('black box statistics'). That drawback can be avoided or minimised if the salient evidence from a multivariate analysis is also presented in simple tabular form. Unfortunately this is not always done. The late Bradford Hill once remarked that quantitative findings from any study should not be accepted at face value unless the main results can be checked on the back of an envelope.

When a study is planned it is customary to project the number of people who will need to be enrolled in order to document a hypothesised association, and 'reject the null hypothesis (RR = 1.0)'. If, for example, we wish to test the hypothesis that oral contraceptives increase the risk of venous thromboembolism (VTE) by three-fold, we may estimate the required 'sample size' ('power') based on the following assumptions:

1. That the hypothesised RR is 3.0.
2. That if the RR estimate is 3.0, the 95% CI will exclude a value of 1.0.
3. That we wish to be 80% confident that we will not, by chance, 'miss' a RR estimate of 3.0.

Items 2 and 3 are conventionally the stipulations ('parameters') that are set in estimating statistical power. Item 2 is to ensure that if the same study were to be repeated 100 times, a RR of 3.0, if not present, would only mistakenly be identified less than five times ('alpha error'). Item 3 is to ensure that if there is indeed a RR of 3.0 it would correctly be identified in at least 80 repetitions of the study ('beta error').

Several factors that need not be considered here determine statistical power. What is relevant is that in follow-up studies important determinants include the projected numbers of cases among exposed and non-exposed persons, or in case-control studies the numbers of exposed cases and controls. In both follow-up and case-control studies the single most important determinant of power is usually the number of exposed cases (because that number is usually the smallest).

J Fam Plann Reprod Health Care 2008; **34**(4): 261–264
(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

Sometimes the hypothesis in a projected study is that there is no association (RR = 1.0). In that case it is customary to set the upper 95% confidence limit one wishes to exclude at 2.0 (but occasionally at 1.5) and power is calculated accordingly. As a practical matter, the exclusion of a RR of much less than 2.0 usually calls for massive numbers and is seldom feasible.

A further consideration that bears on the estimation of statistical power is the interrelationship between the strength of any given association (see Part 2c: strength of association[2]) and its statistical stability: the greater the magnitude of the RR the smaller is the number of exposures and outcomes needed to document it. For high RR estimates (say ≥5.0) small numbers may be sufficient to set confidence limits that exclude 1.0. However, estimates based on small numbers, even when significant, are not robust, they have extremely wide CIs, and they are uniquely susceptible to the vagaries of misclassification, whether random (see Part 1d: bias due to random misclassification[1]) or systematic.

Assume, for example, that in a hypothetical case-control study there are 30 cases, of which three (10%) are exposed, that the RR is 5.0, and that the 95% CI is 1.1–12.8. If only a single case has misreported his or her exposure, and in fact was not exposed, the association would no longer be statistically significant. In this instance the misclassification necessarily affects one (33%) of the three exposed cases. By contrast, in a much larger study the confidence limits would be narrower, and any misclassification, if present, has the 'opportunity' to be much less than 33%, and to affect a smaller proportion of the cases.

The possibility of confounding (see Part 2a: confounding[2]) further limits the interpretability of small numbers. Assume, in the present example, that it is necessary to adjust simultaneously for the effects of age, sex, years of education, and smoking. The number of factors exceeds the number of exposed persons, and multivariate adjustment ('black box adjustment'), if attempted, becomes meaningless: 'the model (one hopes) fails to converge'. By contrast, if there are, say, 30 exposed individuals, the data are likely to be sufficiently robust in statistical terms to accomplish meaningful multivariate adjustment for confounding.

A further limitation to small numbers is that it is impossible to evaluate whether the data are internally consistent (see Part 3c: internal consistency). Thus, in the above example it would be impossible to determine whether, given only three exposed persons, the association is present among men and women, smokers and non-smokers, rich and poor, and so on.

It is for these reasons that associations based on small numbers, even when statistically significant, are considered fragile. Such associations can only be considered valid when there are strong grounds to assume that there is virtually no error. Even then the fragility intrinsic to small numbers necessarily limits their interpretability. As a rough rule of thumb, concern about fragility only recedes when the numbers of exposed cases and non-cases each exceed 10 or, preferably, 20 or 30 individuals.

At this point it is necessary to draw attention to a common fallacy, which is to interpret 'statistically significant' associations as 'causal'. As already pointed out, they may not be causal if the data are biased or confounded. Statistical methods are essential to good epidemiological practice. However, while they enable us to estimate the magnitude of any given association and the confidence with which chance can be excluded, and to adjust for confounding, they do not eliminate bias or residual confounding, if present.

Another important issue is the interpretability of significant associations derived from multiple comparisons conducted within a single set of data (usually a large database). Probability theory dictates that if any body of data is sampled thousands of times, as is sometimes done, 'significant' associations will inevitably arise by chance, in much the same way as exceptional hands are sometimes dealt in a game of cards.

For these reasons, any association identified *post hoc ('a posteriori')*, either in the course of multiple comparisons, or sometimes serendipitously ('hypothesis generation'), must be regarded as tentative, and in need of independent confirmation (see Part 3d: external consistency). If the association is 'real' it should be confirmed; if it is an artefact it should not be ('regression to the mean'). Conversely, any hypothesis proposed in advance *('a priori')*, and then confirmed in a study specifically designed to test it, carries much greater weight – although any study, no matter how good, always benefits from independent confirmation.

Example: Fragile data and failure to recruit targeted cases and controls. Appetite suppressants containing phenylpropanolamine and the risk of haemorrhagic stroke. In a case-control study, based on six exposed cases of haemorrhagic stroke and one control exposed to appetite suppressants containing phenylpropanolamine, the RR was 16.58 (95% CI 1.51–182.21).[3] Given such exceptionally fragile data, no inferences were justified. Only 41% of the targeted cases were enrolled, and those who were not may well have differed in their use of appetite suppressants. The controls were recruited using a technique known (unfortunately) as 'random digit dialling' in which telephone calls are made to determine whether there is a potentially eligible control in the household. The proportion of targeted controls not enrolled (for reasons such as telephone not answered, call forwarding, non-cooperation) could not be determined, but it was undoubtedly high. Both because of fragility and failure to enrol 59% of the targeted cases, and an unknown proportion of the targeted controls, this was an uninformative study.

Example: A spuriously elevated relative risk identified in multiple comparisons. Reserpine and breast cancer. In the course of multiple comparisons conducted within a large database, based on a comparison of 150 cases of breast cancer and 600 controls, the RR of breast cancer for the use of reserpine (an antihypertensive drug) was 3.5 ($p = 0.00007$).[4] Several further studies failed to confirm the association.[5] Finally, in a study that compared 1881 cases of breast cancer and 1523 controls, the RR was 0.8 (95% CI 0.5–1.1).[6] The methods used in the hypothesis-generating and final study were essentially the same, and the initial RR estimate of 3.5 was an artefact that arose by chance in the course of multiple comparisons.

3b: Dose-response and duration-response effects

As a general rule, confidence in causality is strengthened when there is evidence of a dose- or duration-response relationship, and when there are sound clinical or biological grounds to anticipate that there should be such a relationship. The steeper the slope ('gradient'; 'linear gradient'; 'monotonic gradient') of the response curve, the more plausible is it that it may be causal. However, evidence of a dose- or duration-response effect by no means excludes the possibility that it may be due to bias (e.g. if, in a case-control study, the cases tend to overestimate the duration of exposure) or confounding (e.g. if the most severely ill patients take the highest doses).

Causation can also occur in the absence of a dose- or duration-response effect (e.g. allergy to beestings), and sometimes the RR may even decline with increasing duration of exposure, as already illustrated in the example of the modifying effect of the duration of oral contraceptive use on the risk of VTE.[7–9]
Example: See below.

3c: Internal consistency ('coherence')

In any given study, confidence in causality is strengthened if an association is consistently evident ('coherent') within subgroups in which it might reasonably be expected to be. For example, if smoking causes lung cancer in men, it ought also to do so in women, as well as in those who completed primary school, high school, university, and so on. If internal consistency is not demonstrable, confidence in causality is weakened, unless an effect is demonstrably confined to a subgroup (see Part 2b: effect modification[2]).

Ideally, in order to evaluate internal consistency, numbers should be statistically stable not only for estimation of the overall association, but also for relevant subgroups. In practice that ideal may not always be realistic. However, as pointed out above, if numbers are very small, it is not possible to show any degree of consistency at all, and confidence in causality is correspondingly weakened.
Example: See below.

3d: External consistency

Confidence in causality is strengthened when evidence derived from different epidemiological studies, preferably based on different research strategies (e.g. follow-up and case-control studies) consistently converges on the same association. A causal inference may be further strengthened if the magnitude of the observed association is broadly consistent among the studies. Ideally, no association should be accepted as causal until it has been repeatedly and independently confirmed.

3e: Analogy

Analogy refers to indirect information that may support causality, and it is a relatively weak causal criterion. The evidence that oral contraceptives cause VTE and myocardial infarction (in smokers), for example, might be invoked as evidence to support the claim that they also cause strokes, but the reasoning is weak.

3f: Biological plausibility

An association identified in epidemiological research gains in credibility when other lines of scientific evidence (e.g. animal experiments; laboratory tests) suggest that it is biologically plausible. A drawback, however, is that sometimes the experimental evidence may not be applicable to human beings. In addition, different effects are sometimes evident in different animal species or in different experiments. Quite commonly there is copious, and contradictory, experimental evidence for or against causality, leaving the biased epidemiologist free to quote whatever evidence fits his or her preconceptions. Sometimes it is even the case that the epidemiological evidence is correct, and the experimental evidence is wrong [e.g. thalidomide and phocomelia: the association was not demonstrated in animal experiments until the 'right' animal (the rabbit) was selected].[10]

Despite these limitations to the criterion of biological plausibility, in principle it is only when there is full accordance in the totality of the evidence, and when the biological mechanisms are fully understood, that causation can be said to be established beyond any reasonable doubt – an ideal that is seldom, if ever, fully achieved. Short of that ideal, solid biological evidence strongly supports epidemiological evidence of causality.
Example: Proper specification of time order, strong associations, statistical stability, duration- and dose-response effects, internal consistency, external consistency, and biological plausibility. Conjugated estrogen use and uterine cancer. In 1975 two studies, published back-to-back, reported an increased risk of endometrial cancer among users of supplemental estrogens.[11,12] The findings were criticised on the grounds that estrogens could commonly have precipitated endometrial cancer that would otherwise have remained 'clinically silent'.[13] That criticism was rebutted in a case-control study that demonstrated that the risk of endometrial cancer was increased among women who had last used estrogens as much as ≥5 years previously.[14] The overall associations for durations of ≥5 years of use were strong (>3.0), both among current and past users. Duration-response effects ($p<0.01$) were evident. The findings were consistent for early and advanced cancer. Other studies have confirmed the overall association, and have demonstrated dose-response effects.[15] In multiple studies, experimental evidence has shown that estrogens induce endometrial hyperplasia, an established precursor of neoplasia.[16] It is established that supplemental estrogens cause endometrial cancer.

Conclusions

In this series of articles, many of the examples selected to illustrate various causal principles have been fallacies. Fallacies are by no means unique to epidemiology: they burden all disciplines. So it is worth re-emphasising that epidemiology has to its credit a great many non-fallacious and major achievements in causal research, and that there will many more to come. Here, I hope my interpretation of causal principles for clinicians will assist them in distinguishing some of the wheat from the chaff. Unfortunately, as illustrated by several examples in this series, there is now no shortage of chaff.

The central limitation to causal thinking in modern epidemiology is the overarching belief that technical advances now enable us to interpret the marginal evidence intrinsic to small RRs as causal, whereas previously we were unable to do so. Despite the lack of evidence to support that belief, and despite the need for scepticism, that belief has come more and more to be sustained by dependence on 'black box' statistics, large or massive studies, meta-analyses, and sensitivity analyses. If they are to justify themselves, some epidemiologists need to convince themselves that at least some of the small associations they identify can be interpreted, and that they can be interpreted as causal. An analogous belief is that blinded, long-duration, randomised trials that cease to be blinded or randomised can nevertheless be analysed and interpreted as if they are, and that they can document small risks as causal: they can only do so when they do not, in effect, become observational studies. Since, for common diseases, small RR increments can translate into large absolute risks, many epidemiologists are simply unable to accept that causal judgments based on such small increments must usually remain uncertain.

Not only has the belief in the interpretability of small risks become dominant, but it has become politically entrenched. The epidemiological sections of the monographs on the causes of cancer published by the International Agency for Research on Cancer, for example,

now consist almost entirely of meta-analyses of publications from the world literature. Similarly, the *ex-cathedra* prescriptive activities ('systematic reviews'; 'Cochrane reviews') of 'evidence-based medicine' and 'Cochrane centres' (see the website home page: http://www.Cochrane.org) are largely based on the belief that there is a hierarchy of valid evidence in which controlled trials most closely approximate 'the truth', followed by cohort studies, followed by case-control studies (all or some of which can be melded in meta-analyses), followed by the rest, with anecdotal evidence at the bottom of the heap.[17] There is no hierarchy: each of the research strategies described here have strengths and weaknesses, and it is the best evidence, however derived, that must be given the greatest weight in deciding on causality.

The late Alvan Feinstein once remarked that if some insuperable scientific obstacle interferes with one's preconceptions, the temptation to ignore it and pretend it does not exist may be irresistible. Can this state of affairs be remedied? If it is to be, an essential requirement is that experienced clinical insight must be restored to the leadership in causal research. The associations at issue are usually subtle, and clinical judgment is essential if they are to be properly interpreted. In the absence of clinical judgment, epidemiology runs the risk of becoming stupid epidemiology.

Elsewhere I have stated that: "If we can move away from the paradigm of the randomised controlled trial as the most superior methodology under all circumstances, and if we can learn to accept that some questions cannot be answered, we also need to reassert the ascendancy of clinical medicine, in its broadest sense, in causal thinking within epidemiology".[17] That need has become urgent, and if this article helps to fulfil it then it will have served its purpose.

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 1. Principles of causality in epidemiological research: time order, specification of the study base and specificity. *J Fam Plan Reprod Health Care* 2008; **34**: 83–87.
2 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**: 185–190.
3 Kernan WN, Viscoli CM, Brass LM, Broderick JP, Brott T, Feldmann E, *et al.* Phenylpropanolamine and the risk of hemorrhagic stroke. *N Engl J Med* 2000; **343**: 1826–1832.
4 Boston Collaborative Drug Surveillance Program. Reserpine and breast cancer. *Lancet* 1974; **ii**: 669–671.
5 Shapiro S. Meta-analysis/shmeta-analysis. *Am J Epidemiol* 1994; **140**: 771–778.
6 Shapiro S, Parsells JL, Rosenberg L, Kaufman DW, Stolley PD, Schottenfeld D. Risk of breast cancer in relation to the use of rauwolfia alkaloids. *Eur J Clin Pharmacol* 1984; **26**: 143–146.
7 Suissa S, Blais L, Spitzer WO, Cusson J, Lewis M, Heinemann L. First-time use of newer oral contraceptives and the risk of venous thromboembolism. *Contraception* 1997; **56**: 141–146.
8 Suissa S, Spitzer WO. Oral contraceptives and the risk of venous thromboembolism: impact of duration of use. *Contraception* 1998; **57**: 64–68.
9 Suissa S, Spitzer WO, Rainville B, Cusson J, Lewis M, Heinemann L. Recurrent use of newer oral contraceptives and the risk of venous thromboembolism. *Hum Reprod* 2000; **15**: 817–821.
10 Heinonen OP, Slone D, Shapiro S. *Birth Defects and Drugs in Pregnancy* (8th edn) Littleton, MA: Publishing Sciences Group, 1977.
11 Smith DC, Prentice R, Thompson DJ, Herrmann WL. Association of exogenous estrogens and endometrial carcinoma. *N Engl J Med* 1975; **293**: 1164–1167.
12 Ziel HK, Finkle WD. Increased risk of endometrial carcinoma among users of conjugated estrogens. *N Engl J Med* 1975; **293**: 1167–1170.
13 Horwitz RI, Feinstein AR. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. *N Engl J Med* 1978; **299**: 1089–1094.
14 Shapiro S, Kaufman DW, Slone D, Rosenberg L, Miettinen OS, Stolley PD, *et al.* Recent and past use of conjugated estrogens in relation to adenocarcinoma of the endometrium. *N Engl J Med* 1980; **303**: 485–489.
15 International Agency for Research on Cancer (IARC). H*ormonal Contraception and Post-Menopausal Hormone Therapy (IARC Monographs on the Evaluation of Carcinogenic Risk to Humans Vol. 72).* Lyon, France: IARC, 1999.
16 International Agency for Research on Cancer (IARC). *Combined Oestrogen-Progestin Contraceptives and Combined Oestrogen-Progestin Menopausal Therapy (IARC Monographs on the Evaluation of Carcinogenic Risk to Humans Vol. 91).* Lyon, France: IARC, 2008. http://monographs.iarc.fr/ENG/recentpub/mono91.pdf [Accessed 4 February 2008] OR IARC 1989?
17 Shapiro S. Looking to the 21st century: have we learned from our mistakes, or are we doomed to compound them? *Pharmacoepidemiol Drug Saf* 2004; **13**: 257–265.

UN wall charts

The United Nations (UN) has produced two new wall charts – World Contraceptive Use 2007 and World Abortion Policies 2007 – that might be of interest to health professionals. The website also includes a number of very useful articles on sexual and reproductive health. Visit the UN website for further information.

Source: www.unpopulation.org

HPV immunisation programme in Scotland

From September 2008 to June 2009, around 90 000 girls in Scotland will receive three separate injections over a 6-month period as part of Scotland's Human Papilloma Virus (HPV) National Immunisation Programme to help protect teenage girls from the future risk of cervical cancer. Over 15 000 information packs are being issued by Health Protection Scotland (HPS) to a range of health professionals across Scotland from June 2008. The pack, which has been developed by HPS and NHS Health Scotland to help health professionals implement and deliver the immunisation programme from 1 September this year, will include examples of the campaign's marketing materials, Q&As for parents and carers and their daughters, and detailed medical information including a fact sheet and a copy of the Green Book Chapter on HPV.

Source: www.hps.scot.nhs.uk

Pro-life' pharmacies and birth control

Previously in News Roundup it was reported that certain UK pharmacists were unwilling to sell emergency contraception.[1] News from the USA reveals that a pharmacy that opened in the state of Virginia this summer will not sell condoms, birth control pills or emergency contraception. R Alta Charo, a University of Wisconsin lawyer and bioethicist, told the *Washington Post*: "We may find ourselves with whole regions of the country where virtually every pharmacy follows these limiting, discriminatory policies and women are unable to access legal, physician-prescribed medications. We're talking about creating a separate universe of pharmacies that puts women at a disadvantage."

Reference

1 Pharmacist refuses to sell emergency contraception [News Roundup]. *J Fam Plan Reprod Health Care* 2005; **31**: 324.

Source: http://www.washingtonpost.com/wp-dyn/content/article/2008/06/15/AR2008061502180_pf.html

Reviewed by **Henrietta Hughes**, MRCGP, DFSRH
General Practitioner, London, UK