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

Samuel Shapiro

Scope of article

In Part 1 of this series the following principles of causation were considered: time order, specification of the study base, and specificity (bias due to random misclassification and bias due to systematic misclassification).¹ Part 2 will consider:

2a: Confounding

2b: Effect modification

2c: Strength of association.

Part 3 concludes the consideration of causal principles and will discuss statistical stability, dose- and duration-response effects, internal consistency, external consistency, analogy and biological plausibility.²

2a: Confounding

Confounding exists when a risk factor other than the exposure under study is associated, independently, both with the exposure and with the outcome. Confidence in causality is increased if an association can be judged to be reasonably unconfounded. When present, confounding is the same in a follow-up or case-control study carried out in the same study base. To the degree that confounding can be precisely measured, its effect can be removed or minimised.

To illustrate, consider the effect of age on the association of oral contraceptive use with venous thromboembolism (VTE). With advancing age, oral contraceptive use declines, and the risk of VTE increases. If in any given study the compared groups are of different ages, that difference will confound the association. For example, in a case-control study, if the cases are older than the controls, and if age is not allowed for, the magnitude of the association will be underestimated.

Age is an example of a risk factor that can be precisely measured, and its potentially confounding effect can therefore be eliminated ('controlled') by appropriate adjustment: for example, by setting age limits ('exclusion'); by comparing women of similar age ('comparisons within strata of age'); by making comparisons after applying percentages from strata of age to a single standard age distribution ('standardisation'); or by multivariate analysis (broadly, the more or less simultaneous statistical adjustment for the confounding effects of several factors, including age). In case-control studies age can also be allowed for by selecting controls matched to be of the same age as each individual case

('matching') or by ensuring that there are similar ratios of cases to controls within strata of age ('series-matching', 'frequency-matching').

Some epidemiologists consider confounding to be a type of systematic bias (see Part 1e: bias due to systematic misclassification¹). But there is an important distinction. In principle, provided a confounder can be precisely measured, its effect can be eliminated by adjustment. By contrast, once any random or systematic bias is present in any study, its effect cannot be eliminated.

Unfortunately, many confounding factors cannot be measured precisely, and our inability to do so limits the interpretability of observational data. When a confounder is imprecisely measured it is only possible to adjust for that imprecise measurement, and not for the 'truth'. Examples of incompletely measurable factors include wealth or poverty ('socioeconomic status'), exercise, diet and health behaviours, to mention just a few. In addition, random bias (see Part 1d: bias due to random misclassification¹), as well as systematic bias (see Part 1e: bias due to systematic misclassification¹), may affect not only the ascertainment of exposures and outcomes, but also the ascertainment and measurement of confounders.

A further hindrance to causal inference is that in any observational study it is always possible that there may be confounding factors that are not recorded or even suspected. That possibility places an absolute limit on the certainty with which any association can be interpreted. The designation 'residual confounding' is used to denote the effect of confounding that may still be present in a given study, despite ostensible 'control', either because a confounder has been imprecisely measured, or because one or more confounders have not been recorded or suspected.

A major advantage of controlled trials, relative to observational research, is that randomisation usually results in more or less equal distributions of all confounders, known and unknown. But even then, unequal distribution of one or two confounders may occur by chance ('luck of the draw'). In that case, however, it is unlikely that, in two or more trials, there will be the same unequal distribution of the same confounders. A proviso, however, is that if there is low adherence to the assigned treatment, or if there are 'crossovers' from treatment to no treatment (or vice versa), or if losses to follow-up are high, confounding can occur, as in any observational study. The reason is that persons who do not adhere, or who 'cross over', or who are lost to follow-up, may differ in their risk of the outcome from those who adhere.

One form of confounding, 'confounding by indication', if present, further limits the interpretability of observational data. Such confounding is present when the underlying reason for exposure is also a determinant of the outcome. To illustrate, consider the hypothesis that the administration of insulin to diabetic patients in a way that keeps blood glucose levels normal at all times reduces the risk of blindness, relative to less stringent control. The underlying severity of the diabetes (the 'indication') is likely to be a determinant of how rigorously insulin is used,

J Fam Plann Reprod Health Care 2008; **34**(3): 185–190

(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

and independently, a determinant of the risk of becoming blind. It is impossible to measure the severity of the diabetes with sufficient precision to fully control its confounding effect. Even if rigorous treatment in fact *reduces* the risk of blindness, that benefit may be missed – and if the confounding is sufficiently great, such treatment may even appear, spuriously, to *increase* it. Consequently, when the indication for treatment is itself a determinant of the outcome, confounding by indication may only confidently be avoided by randomisation.³

To get around the difficulty that we can never be certain that all confounding (or bias) has been eliminated in observational research, some epidemiologists argue that in any reasonably well-conducted study, although residual confounding (or bias) is always possible, the onus is on those who question any given association to provide evidence of such confounding (or bias).⁴ Some epidemiologists even argue that in a well-designed study residual confounding (or bias) is unlikely to invalidate any but the smallest of associations (see Part 2c: strength of association). As shown in the second example below, that argument is not tenable.

Finally, it is important to appreciate when any given factor does not confound. First, a factor does not confound if it is not *independently* associated with the exposure and the outcome, but is instead part of the causal pathway. ‘Control’ for such a factor is not valid, and the effect of doing so would be to underestimate or eliminate a ‘true’ association. For example, unsterile needles are part of the causal pathway through which injections ‘cause’ HIV. If, in any given study the data were to be stratified according to the use or non-use of unsterile needles, the association would be missed.

Second, if a factor is associated *only* with the exposure, or *only* with the outcome, but it is not associated *independently*, with *both* the exposure *and* the outcome, it does not confound. Adjustment for such a factor is unnecessary, and it would be without any effect on the relative risk (RR) estimate. For example, brown fingers are associated with smoking, but not with lung cancer: ‘adjusting’ for brown fingers would make no difference.

Example: Residual confounding and systematic misclassification (detection bias) in studies of the risk of cervical cancer among oral contraceptive users. Based on a review of the world literature, the World Health Organization (WHO) International Agency for Research on Cancer (IARC) has decreed that oral contraceptives cause cervical cancer.^{5,6} One reason for reaching that conclusion is that in multiple studies adjustment for the confounding effect of sexual activity did not eliminate the association. [What is really meant by the IARC decree is that oral contraceptives potentiate the carcinogenicity of high-risk human papillomavirus infection – the established cause of cervical cancer – which, if true, would be an example of effect modification (see Part 2b: effect modification)].

In fact, full adjustment for sexual activity was not made in any of the reviewed studies, and could not be. Most of the studies adjusted only for the reported age at first intercourse, and for the number of sexual partners, and not for the ‘truth’. It is likely that women who had many partners systematically underreported the true number, while women who had one partner, or no partners, could hardly have done so. Moreover, none of the studies had information on the sexual activity of the male partners. The summary RR estimates derived from a meta-analysis of 28 studies⁷ cited in the IARC monograph (Table 2.3.1)⁶ were 1.1, 1.6, and 2.2, respectively, for durations of use of <5, 5–9 and ≥10 years. Residual confounding could readily

have accounted for such small associations (see Part 2c: strength of association). In addition, of course, sexual activity would likely have been the greatest, and the most underreported (see Part 1e: systematic misclassification¹ and Part 3b: dose- and duration-response effects²), among women who used oral contraceptives for the longest time. Detection bias due to screening with Papanicolaou (Pap) smears would also have been the greatest among long-term users.

Example: Residual confounding. Smoking and suicide.

In the observational component of the Multiple Risk Factor Intervention Trial, 361 662 subjects were followed to determine whether changes in risk behaviour (e.g. smoking cessation) reduces the risk of coronary heart disease.⁸ A coincidental finding was an increased risk of suicide among smokers. After adjustment for a large number of confounding factors, compared with non-smokers, the RRs of suicide for those who smoked 1–19, 20–39, 40–59 and ≥60 cigarettes per day were 1.4, 1.9, 2.3 and 3.4, respectively. The trend (‘trend P’) was statistically stable (see Part 3a: statistical stability²), as were the individual RR estimates. As pointed out by the authors, it is absurd to propose that cigarettes ‘cause’ suicide. The obvious explanation is that depressed people tend to smoke and also to commit suicide. That is, depression confounded the association; it produced a statistically significant, but nevertheless spurious, dose-response relationship (see Part 3b: dose-response effects²), and among heavy smokers a large and spuriously elevated RR (see Part 2c: strength of association). The confounding effect of depression could not be eliminated because it was not adequately defined or measured.

2b: Effect modification

Effect modification (‘interaction’) exists when one factor modifies the effect of another. For example, heavy cigarette smoking without heavy alcohol consumption, and heavy alcohol consumption without heavy smoking, each increase risk of oral cancer by some five-fold.^{9,10} Yet the combination of heavy smoking and drinking increases the risk by some 100-fold. In this example the evidence to show that cigarettes modify the effect of alcohol (or vice versa) is unambiguous since the RRs are not simply ‘additive’ (5.0 + 5.0 or less) but ‘multiplicative’ (some 10-fold more than 5.0 + 5.0). Things may be less clear, however, when two RR estimates are different, but within the ‘additive’ range.^{11,12} Nevertheless, as a practical matter effect modification can also exist without the RRs being multiplicative, as illustrated in the next example.

Example: Effect modification, confounding by indication, and violation of time order. Risk of VTE among ‘third’- and ‘second’-generation oral contraceptive users. It has been shown that oral contraceptives increase the risk of VTE maximally during the first year of use, and then to a lesser degree following more prolonged use^{13–15} [a phenomenon sometimes labelled (unfortunately) as ‘depletion of susceptibles’ or (also unfortunately) as a ‘healthy user effect’, the idea being that early on people at the highest risk are ‘weeded out’]. In addition, the pattern of an initial high risk, followed by a lower but still elevated risk, appears to repeat itself each time oral contraceptive use is stopped and then restarted. However, each time women recommence use, the risk elevations appear to be somewhat lower than during the first episode of use (although this is uncertain). Thus, duration of use modifies the effect of oral contraceptives on the risk of VTE; and on each occasion that oral contraceptive use is restarted, the effect modification appears to be the same, but perhaps less marked.

Consider a comparison of two groups of oral contraceptive users with different durations of use. The overall RR estimate would be higher in the group with the higher proportion of short-term users, but the estimates within strata of duration (i.e. strata of the effect modifier) would be the same. In addition, for successive episodes of use (first, second, third, and so on) the duration-specific estimates would again be the same.

In a recent controversy concerning the effects of so-called 'second'- and 'third'-generation oral contraceptives on the risk of VTE, based on findings from a study conducted by the WHO,¹⁶ it was claimed that relative to the use of 'second-generation' oral contraceptives, 'third-generation' products increased the risk by more than two-fold.

The WHO study did not take the modifying effects of duration of use during all lifetime episodes of oral contraceptive use fully into account. Yet in order to do so it was mandatory to compare overall use, and duration of use, among cases and controls who commenced use for the first time ('starters'), and then to do the same comparisons among repeat users ('restarters'), for the second episode of use, the third episode, and so on.

An additional limitation to the WHO study was that allowance was not made for oral contraceptive users who switched from one product to another ('switchers'). Women who switched from a 'second'- to a 'third'-generation pill may selectively have done so if they or their doctors became concerned about a possible increased risk of VTE (e.g. because of recent trauma or a swollen leg), and if they believed, mistakenly, that the newest products were the safest. That is, among switchers there may have been confounding by indication (e.g. a sprained ankle) or violation of time order (e.g. if a swollen leg was due to undiagnosed deep vein thrombosis). As explained above, confounding by indication cannot confidently be eliminated in observational data; and violation of time nullifies any association. However, the risks could legitimately have been evaluated for starters and for women who restarted the first, second and subsequent times.

Almost contemporaneously with the WHO study, an additional study, the Transnational Study, also evaluated the risk of VTE among oral contraceptive users.¹⁷ When the WHO findings were published, the Transnational data were initially subjected to much the same analysis, and relative to the use of 'second'-generation products, the RR for 'third'-generation products was 1.5. That association was interpreted as independent evidence confirming the WHO findings. However, in later multivariate analyses ('multiple logistic regression' and 'Cox regression with time-dependent covariates') in which duration of use was fully taken into account among starters and restarters, there was no longer any evidence of an increased risk.^{13–15,18}

With regard to possible confounding by indication, one of the 'third'-generation progestogens (desogestrel) had been marketed twice, at the same dose (150 µg), once in 1981 together with 30 µg ethinylestradiol (EE), and then again in 1992 together with 20 µg EE. Yet, relative to the receipt of oral contraceptives containing levonorgestrel (a 'second'-generation progestogen) the RR was higher (RR 2.8) for desogestrel combined with the lower dose of EE (20 µg) than for desogestrel combined with the higher dose (30 µg) marketed earlier (RR 1.5).¹⁹ In addition, for all oral contraceptives introduced after 1981, the more recent the date of marketing, the higher was the RR, a trend that was statistically significant (see Part 3a: statistical stability²). These phenomena were clear evidence of a tendency for

the women at highest risk to be prescribed the most recently marketed products. That is, there was confounding by indication.

2c: Strength of association

The strength of an association refers to the magnitude of the RR estimate. Assuming any given study is well designed and executed, the higher the RR, the more confident we can be that the association may be causal. No study is entirely free of bias or confounding. Nevertheless, in a reasonably well-conducted study that yields a high and statistically stable RR (see Part 3a: statistical stability²), we may judge that an association might perhaps be weakened somewhat, but not be obliterated, even if it were possible to entirely eliminate all sources of bias and confounding. Or put another way, we may judge it unlikely that the residual effect of any biases or uncontrolled sources of confounding, if present, could exceed or even come close to the magnitude of the observed association. Thus, despite the imperfections, we might cautiously infer that the evidence supports causation (or protection).

By contrast, we can seldom, if ever, confidently make that inference when associations, even if statistically stable, are of low magnitude ('weak', 'small').²⁰ It takes relatively little bias or confounding, sometimes less than can be eliminated in the best-designed studies, to produce weak associations, and in order to render a weak and non-causal association 'statistically significant' all that is needed is to make the study large enough, or if necessary, massive enough. In that case all that is accomplished is to set narrow confidence limits (see Part 3a: statistical stability²) around a biased or confounded RR estimate.

As a rough rule of thumb, RR estimates of the order of 2.0 or less are generally considered 'weak' or 'small', and any value of less than 3.0 should be considered tentative. Moreover, if a study is egregiously biased, or confounded, even a 'strong' association (e.g. a RR of 5.0) may not be causal and, as demonstrated above, the 'world record' for an association that has been shown to be non-causal is 27 (fertility drugs and ovarian cancer).^{21,22}

The reasoning that RR estimates of <3.0 should be considered tentative is based on empirical experience, as shown in one non-causal example already mentioned (a three-fold increase in the risk of suicide among heavy smokers.⁸ Or to give another example, the use of supplemental estrogens has been associated with a five-fold reduction in the risk of fatal accidents or homicide.²³ The latter associations were obviously confounded by prudent behaviour, and by high socioeconomic status, among estrogen users.

In the past, the limits to causal inference imposed by small associations were widely accepted, but some public health advocates chafed against them because for common diseases low RRs translate into high absolute risks. In that circumstance a small RR, *if causal*, may be of major public health importance. To illustrate: in a hypothetical study of a common disease, if in the exposed and non-exposed groups the incidence rates are 2% and 1.5% per year, respectively, the RR is 1.3 (2.0/1.5) (i.e. small) but the absolute risk is 0.5% (2.0%–1.5%), or 5 per 1000 per year (i.e. large). By contrast, if the disease is rare, and the respective incidence rates are 2.0 and 1.5 per million per year, the RR is again small, 1.3 (2.0/1.5) and the absolute risk is also small, 0.5 (2.0–1.5) per million per year, or 5 per 10 million per year.

The snag is in the words, *if causal*, and regardless of the public health importance of small elevations in the RRs when the diseases are common, discriminating among bias, confounding and causation is simply beyond the resolving

power of the epidemiological microscope. We are always left with uncertainty. The topic of public health policy is beyond the scope of this essay, but if a great deal hangs on whether or not an association is causal, and if a policy decision must be made, that decision can only be based on judgment made in the face of uncertainty.

Perhaps because of a desire for concrete answers, or perhaps because causal thinking in epidemiology has increasingly come to be dominated by a statistical rather than a clinical perspective, it has been argued that improved methods now make it possible to interpret small RR estimates as causal with greater reassurance than was previously the case. The literature is now replete, even dominated, with RR estimates, sometimes as low as 1.2 or less, that are interpreted as definitive evidence of causation (e.g. passive smoking and lung cancer).²⁴

Three ostensible advances in methodology have been invoked to support that argument. The first is the advent of very large studies (e.g. the Nurses Health Study with over 120 000 participants²⁵), in which small RRs of the order of 1.5 are commonly statistically stable. In such studies, once chance has been disposed of as a possible explanation, it becomes tempting to believe that bias and confounding may not be terribly important after all. More recently large studies have been superseded by truly massive ones. In the Million Women Study (MWS),²⁶ for example, virtually any RR that deviates from 1.0 is 'statistically significant'. Nevertheless, despite the more ready attainment of statistical significance in large or massive studies, the same limitations apply.

The second argument is that the advent of meta-analysis has made it possible to produce statistically stable results for small RRs by synthesising information across multiple studies. Meta-analysis is beset by multiple deficiencies, a detailed consideration of which is beyond the scope of this essay, and the interested reader is referred to two other articles.^{22,27} Briefly, as is demonstrated in several of the examples throughout this review, one major deficiency is that when bias is present in one study, it is likely to be present in more than one. Aggregation of the information simply serves to render the collective biases more 'statistically significant', not to eliminate them.

A further deficiency of meta-analysis is that when a series of studies yield conflicting ('heterogeneous') results, statistical tricks ('random effects models') are sometimes used in order to produce a single aggregated RR. Yet heterogeneity is evidence against causality (see Part 3c: internal consistency² or 'coherence'). In addition, the methods used to test for heterogeneity are not statistically robust.²⁸ Consequently, study results can be discordant without being 'significantly' heterogeneous when tested. Random effects models make no clinical or biological sense.

A further deficiency of meta-analysis is that the investigator can usually predict the result in advance by simply 'eyeballing' the individual studies, and the decision as to whether or not to perform a meta-analysis may be influenced by the anticipated result. And still further deficiencies include variation in quality among studies; inability on the part of those not engaged in the meta-analysis to check and verify the findings for themselves (i.e. lack of transparency); and variation in methodology, including variable measurement and control of confounding.

The third argument is that the advent of 'sensitivity analysis' has made it possible to check the validity of small RR estimates. In such an analysis various assumptions are made about the magnitude of plausible sources of bias or confounding. Those assumptions are then simulated

('modelled') and the association is re-estimated. If it survives the sensitivity analysis it is argued that the evidence favours causation.

In practice there has been little evidence to support that argument. The main accomplishment of sensitivity analysis has been to show how little bias or confounding it takes to nullify small associations,^{20,29,30} but it has not proven useful in demonstrating the converse. The reason is that the investigators may already have decided about the validity of their evidence in advance. That is, they may be biased (unconsciously biased, to be sure, but biased all the same), and if they are, they may specify the terms of a sensitivity analysis so that it conforms to their bias.

One advantage of randomised controlled trials is that since randomisation can eliminate confounding, and blinding can eliminate detection bias, small associations can more readily be interpreted as causal. As already mentioned, however, an important proviso is that such studies must remain randomised and blinded. Additional provisos are that there must be good adherence to the assigned treatment and few crossovers. If, as is commonly the case in trials conducted on an epidemiological timescale, these requirements are not met, randomised trials can take on the properties of observational research, with all of its intrinsic limitations, and they should be analysed using the methods of observational research.^{30,31}

To sum up, given adequate study design, when RRs are large, epidemiology excels at the identification of causal relationships (e.g. asbestos and mesothelioma; HIV and AIDS). But when RRs are small, it may be difficult or impossible to discriminate among causation, bias and confounding.

Example: A small and spuriously elevated relative risk in a meta-analysis. Supplemental estrogens and breast cancer. In a meta-analysis ('collaborative re-analysis') of 51 studies involving 52 705 women with breast cancer and 108 411 controls, the RR for the current receipt for ≥ 5 years of supplemental hormone therapy, over 80% of which was conjugated estrogens without progestogens, was 1.35.³² The association, although small, was 'highly significant', and it was interpreted as evidence to support causality. Yet, as was the case in the meta-analysis of oral contraceptive use and breast cancer,³³ systematic bias could readily have accounted for it.³⁴

By contrast, in the Women's Health Initiative (WHI) randomised controlled trial of conjugated estrogens versus placebo, based on an intention-to-treat analysis, the RR of breast cancer was 0.82 (not statistically significant).³⁵ However, intention-to-treat analysis was improper because the discontinuation rates of the assigned treatments exceeded 50%. In that respect, although the WHI study started out as a controlled trial, it became an observational study, and the data should have been analysed accordingly. When this was done, among women who adhered to their treatments, the RR was 0.71, statistically significant, and compatible with a *reduced* risk of breast cancer among conjugated estrogen users. The latter estimate must be considered the best estimate in the WHI study.

In another respect the WHI study of conjugated estrogen use and breast cancer remained a controlled trial. In the arm of the WHI study of combined estrogen/progestogen therapy, unblinding was common³⁶ because postmenopausal bleeding occurred commonly, and endometrial cancer had to be excluded. By contrast, the WHI study of conjugated estrogens was confined to hysterectomised women, bleeding did not occur, and <2% of the subjects were 'unblinded'.³⁷ That is, unblinding was extremely uncommon, and biased ascertainment of the outcome was for all practical purposes avoided.

Despite some limitations due to poor adherence to treatment, the WHI study was unequivocally superior to the 51 studies included in the meta-analysis. At minimum, the WHI findings suggested that estrogens do not increase the risk of breast cancer. The absence of an increased risk has since been independently confirmed.³⁸ Based on the evidence of no increase in the risk of breast cancer among estrogen recipients in the WHI study, and possibly even a decrease, it is likely that bias accounted for the increased risk observed in the meta-analysis.

Example: Biased specification of a sensitivity analysis. Supplemental hormone use and ovarian cancer. In the MWS, among hysterectomised women almost all RRs of ovarian cancer among hormone users were small, 1.53 or less, but because of its massive size, statistically stable.³⁹ Losses to follow-up were substantial (36%), and the investigators performed a sensitivity analysis to determine the extent to which those losses may have affected the findings. They concluded that the losses had no material effect.

Sensitivity analyses were not performed to test how much misclassification of women not at risk for ovarian cancer (because of bilateral oophorectomy at the time of hysterectomy), how much biased ascertainment of the outcome among exposed women, or how much violation of time order (because early symptoms of as yet undiagnosed ovarian cancer 'caused' hormone use) it would have taken to nullify the observed RRs. Given the low magnitude of the RRs, it is doubtful whether they could have survived those analyses.

Future article

In the October 2008 issue of the Journal, the final article in this series (Part 3) will cover additional principles of causality: statistical stability, dose- and duration-response effects, internal consistency, external consistency, analogy, and biological plausibility.²

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

- 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.
- 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).
- Miettinen O. The need for randomization in the study of intended effects. *Stat Med* 1983; **2**: 267–271.
- Hertz-Picciotto I. Invited commentary. Shifting the burden of proof regarding biases and low magnitude associations. *Am J Epidemiol* 2000; **151**: 946–948.
- Cogliano V, Grosse Y, Baan R, Straif K, Secretan B, El Ghissassi F; WHO International Agency for Research on Cancer. Carcinogenicity of combined oestrogen-progestogen contraceptives and menopausal treatment. *Lancet Oncol* 2005; **6**: 552–553.
- 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].
- Smith JS, Green J, Berrington de Gonzalez A, Appleby P, Peto J, Plummer M, et al. Cervical cancer and use of hormonal contraceptives: a systematic review. *Lancet* 2003; **361**: 1159–1167.
- Davey Smith G, Philips AN, Eaton JD. Smoking as an "independent" risk factor for suicide: illustration of an artifact from observational epidemiology. *Lancet* 1992; **340**: 708–712.
- International Agency for Research on Cancer (IARC). *Alcoholic Drinking (IARC Monographs on the Evaluation of Carcinogenic Risk to Humans Vol. 83)*. Lyon, France: IARC, 1988.
- Day GL, Blot WJ, Austin DF, Bernstein L, Greenberg RS, Preston-Martin S, et al. Racial differences in the risk of oral and pharyngeal cancer: alcohol, tobacco and other determinants. *J Natl Cancer Inst* 1993; **85**: 465–473.
- Kupper L, Hogan MD. Interaction in epidemiologic studies. *Am J Epidemiol* 1978; **108**: 447–453.
- Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol* 1980; **112**: 467–470.
- 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.
- Suissa S, Spitzer WO. Oral contraceptives and the risk of venous thromboembolism: impact of duration of use. *Contraception* 1998; **57**: 64–68.
- 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.
- World Health Organization Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception. Effect of different progestogens in low oestrogen oral contraceptives on venous thromboembolic disease. *Lancet* 1995; **346**: 1582–1588.
- Spitzer WO, Lewis MA, Heinemann LA, Thorogood M, MacRae KD. Third generation oral contraceptives and the risk of venous thromboembolic disorders: an international case-control study. Transnational Research Group on Oral Contraceptives and the Health of Young Women. *BMJ* 1996; **312**: 83–88.
- Lewis MA, MacRae KD, Kühl-Habichl D, Bruppacher R, Heinemann LA, Spitzer WO. The differential risk of oral contraceptives: the impact of full exposure history. *Hum Reprod* 1999; **14**: 1493–1499.
- Lewis MA, Heinemann LA, MacRae KD, Bruppacher R, Spitzer WO. The increased risk of venous thromboembolism and the use of third generation progestogens: role of bias in observational research. The Transnational Research Group on Oral Contraceptives and the Health of Young Women. *Contraception* 1996; **54**: 5–13.
- Shapiro S. Bias in the evaluation of low-magnitude associations: an empirical perspective. *Am J Epidemiol* 2000; **151**: 939–945.
- Whittemore AS, Haris R, Intyre J. Characteristics relating to ovarian cancer risk: collaborative analysis of 12 US case-control studies II. Invasive epithelial cancer in white women. *Am J Epidemiol* 1992; **135**: 1184–1203.
- Shapiro S. Is meta-analysis a valid approach to the evaluation of small effects in observational studies? *J Clin Epidemiol* 1997; **50**: 223–229.
- Petitti DB, Perlman JA, Sidney S. Noncontraceptive estrogens and mortality: long-term follow-up of women in the Walnut Creek Study. *Obstet Gynecol* 1987; **70**: 289–293.
- 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.
- Belanger CF, Hennekens CH, Rosner B, Speizer FE. The Nurses Health Study. *Am J Nurs* 1978; **78**: 1039–1040.
- Million Women Study Collaborators. Breast cancer and hormone replacement therapy in the Million Women Study. *Lancet* 2003; **362**: 419–427.
- Shapiro S. Meta-analysis/shmeta-analysis. *Am J Epidemiol* 1994; **140**: 771–778.
- Greenland C. Can meta-analysis be salvaged? *Am J Epidemiol* 1994; **140**: 783–787.

- 29 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.
- 30 Shapiro S. The Million Women Study: potential biases do not allow uncritical acceptance of the data [Editorial]. *Climacteric* 2004; **7**: 3–7.
- 31 Garbe E, Suissa S. Issues to debate on the Women's Health Initiative (WHI) study. Hormone replacement therapy and acute coronary outcomes: methodological issues between randomized and observational studies. *Hum Reprod* 2004; **19**: 8–13.
- 32 Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52 705 women with breast cancer and 108 411 women without breast cancer. *Lancet* 1997; **350**: 1047–1059.
- 33 Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 10 239 women without breast cancer from 54 epidemiological studies. *Lancet* 1996; **347**: 1713–1727.
- 34 Shapiro S. Adverse neoplastic and cardiovascular outcomes of HRT: the validity of the evidence. *Endocrine* 2004; **24**: 203–210.
- 35 Stefanick ML, Anderson GL, Margolis KL, Hendrix SL, Rodabough RJ, Paskett ED, *et al*; WHI Investigators. Effects of conjugated equine estrogens on breast cancer mammography screening in postmenopausal women with hysterectomy. *JAMA* 2006; **295**: 1647–1657.
- 36 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.
- 37 The Women's Health Initiative Steering Committee. Effects of conjugated estrogen in postmenopausal women with hysterectomy. The Women's Health Initiative Randomized Controlled Trial. *JAMA* 2004; **291**: 1701–1712.
- 38 Zhang SM, Manson JE, Rexrode KM, Cook NR, Buring JE, Lee IM. Use of oral conjugated estrogen alone and risk of breast cancer. *Am J Epidemiol* 2007; **165**: 524–529.
- 39 Million Women Study Collaborators. Ovarian cancer and hormone replacement therapy in the Million Women Study. *Lancet* 2007; **369**: 1703–1710.

OBITUARY

Professor Norman Morris

Professor Norman Morris, who died on 29 February 2008, will be remembered primarily as the obstetrician responsible for re-evaluating and reforming the care of women in pregnancy and labour. He was appointed Professor at Charing Cross Hospital in London in 1958. The appointment was made on the basis of his academic research, at Hammersmith and University College Hospitals, into blood pressure in pregnancy and placental transport. He retained these interests and his department made valuable contributions in these areas.

It is, however, in his humane approach to women's health for which Norman will be particularly remembered. He stated his position clearly and courageously in his address on the inauguration of his Chair. He spoke of the complacency of his obstetric colleagues who had reduced maternal and perinatal mortality dramatically, but in so doing had ignored women's feelings and emotions. This address was published in *The Lancet* in 1960 and re-published recently in that journal as one of the most influential papers of recent times. As a result of his championing, with the support of midwives and patient groups, antenatal classes became routine, husbands were encouraged to attend classes and to be with their

wives in labour, and labour wards were furnished and altered to be as welcoming as possible. The routine shave and enema on admission in labour were done away with. The role of the episiotomy was reviewed.

It was not only in the area of obstetrics that Norman's passionate concern for the well-being of women manifested itself. He addressed the problems of unwanted pregnancy and of contraception vigorously and with great empathy. His was one of the first academic departments to run specific family planning and termination clinics. These services were pioneered with the help of Geraldine Howard and Margaret Blair, and were seen as having an essential role in the department as a whole, not only for patient care but also for the training of students and doctors.

His concern for these aspects of women's health led to him writing a book on sterilisation together with his colleague Humphrey Arthure. He was interested in trying to develop a reversible method of sterilisation and devised an ingenious operation for burying the ovaries in the broad ligament. Unfortunately, in those pre-laparoscopy days, reversal required a second laparotomy.

Norman also instituted the Wyeth Symposia held annually at the Royal College of Obstetricians and Gynaecologists (RCOG), where the whole day was dedicated to family planning and sexual health. They were always oversubscribed and covered a broad spectrum, with speakers from all over the country. I remember when I spoke at one, being rather daunted by the almost exclusively female audience, but the atmosphere was very different from the usual RCOG meetings. The audience were warm, receptive, appreciative and enthusiastic. Norman rightly set great value by these symposia and they were a feature in the recognition of family planning and sexual health becoming a specialty in its own right.

Norman Morris was not a conventional man and his very considerable contribution was not conventional. Women's health has lost one of its greatest friends but he leaves a permanent and invaluable legacy.

The Morris family are planning a memorial service in September 2008 which Norman's professional colleagues and friends are invited to attend. Further details about this memorial service are available from Nicholas Morris (nicholashmorris@hotmail.com).

Submitted by **Michael E Pawson**, MRCS, LRCP, FRCOG, *Consultant Gynaecologist (Retired), Chelsea & Westminster Hospital NHS Trust, London, UK and Honorary President, British Society of Psychosomatic Obstetrics, Gynaecology and Andrology (BSPOGA)*