PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2004; 13: 257–265
Published online 3 February 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/pds.903

INVITED LECTURE

Looking to the 21st century: have we learned from our mistakes, or are we doomed to compound them?[‡]

Samuel Shapiro[*,†]

Emeritus Director, Slone Epidemiology Center, Boston University, Boston, MA, USA

First I would like to dedicate this talk to the memory of my beloved friend, Dennis Slone, who died much too young, and much too unrecognized, in 1980. I missed him then, I miss him now and I will miss him for the rest of my life, as I know will a handful of other people present today.

Next, I would like to thank the organizers, particularly Sam Lesko, a colleague and friend with whom I had the pleasure of working for many years, for inviting me to give one of the keynote addresses at this conference. The invitation was particularly gracious, since it is no secret that at its inception I opposed the establishment of a pharmacoepidemiological society, and that I have always questioned the need for pharmacoepidemiology as a discrete discipline, on esthetic, political and conceptual grounds. To be gracious in return, I will not discuss my esthetic or political objections until the coffee break; the conceptual objection I will return to at the end of this talk.

I must confess that when I selected my topic, I did not appreciate that I had bitten off more than I could chew. When I began to think about it, I realized that the subject matter is exceedingly varied and complex. In Figure 1, I list some of the domains in which I believe there is a need to come to grips with serious limitations to the way in which some approaches have been applied, or misapplied, to the evaluation of drug effects – and that list is by no means exhaustive. If one is to be serious about covering everything, a monograph, rather than a single lecture, is needed. In 1967, Richard Doll published a sponsored monograph, introduced by a public lecture, on the prevention of cancer.[1] I suggest that an analogous model may be needed, and this, if you like, is my public lecture in which I will concentrate only on Item 1, skepticism as a scientific principle.

A bit of history first. This conference is taking place just short of 37 years since I arrived in the United States in September 1967 and switched from clinical medicine to epidemiology, more specifically to the exploration of epidemiological methods that could be applied to drug surveillance, then an ‘almost virgin’ subject.

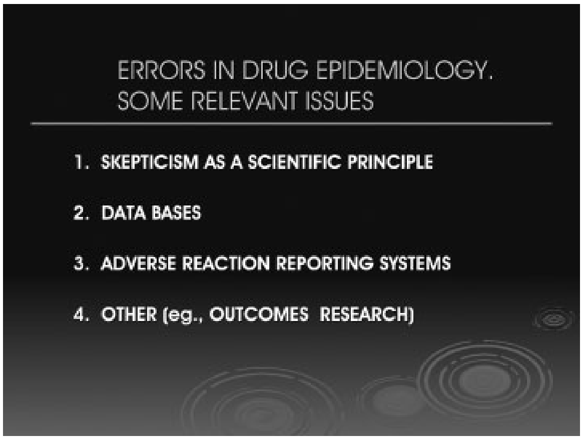
When I arrived, the first colleagues with whom I was associated in that effort, Herschel Jick and Dennis Slone, had only been at it for a year or two and at the time that I joined them all three of us were still learning epidemiology, ‘on the road’, as it were. And when it came to the application of what we were learning to drug surveillance, we were for the most part not only groping in the dark, but also groping in the wrong place. We were studying hospitalized patients while the main public health issues concerned drug safety in the population at large. Like the fabled drunkard, we were not looking for the penny where it dropped but where the street-lamp cast some light. Then, to labor the metaphor, as we sobered up and as dawn approached, we began to look further by adding case-control methods to our armamentarium and we succeeded in finding a penny or two, after all. I must quickly add that we also found a slug or two.

I have described that experience at length and in rather less florid language elsewhere.[2] But what is relevant here is that as we gained experience, Jick on one hand and Slone and I on the other came to differ more and more about how epidemiology should

[*]Correspondence to: Dr S. Shapiro, Emeritus Director, Slone Epidemiology Center, Boston University, Boston, MA, USA.
E-mail: sydzeev@aol.com
[†]Visiting professor of Epidemiology, Mailman School Of Public Health. Columbia University.

[‡]The author has been paid as a consultant (or in a similar capacity) by a company with a vested interest in the product being studied on issues related and unrelated to the product being studied.

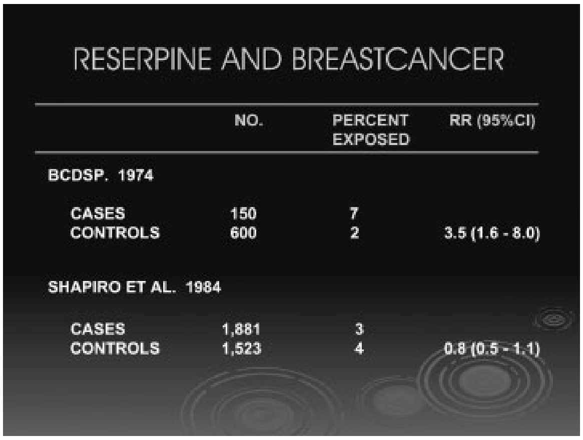
Received 28 August 2003
Accepted 18 September 2003



ERRORS IN DRUG EPIDEMIOLOGY.
SOME RELEVANT ISSUES

1. SKEPTICISM AS A SCIENTIFIC PRINCIPLE
2. DATA BASES
3. ADVERSE REACTION REPORTING SYSTEMS
4. OTHER (eg., OUTCOMES RESEARCH)

Figure 1. Errors in drug epidemiology. Some relevant issues



RESERPINE AND BREASTCANCER

	NO.	PERCENT EXPOSED	RR (95%CI)
BCDSP. 1974			
CASES	150	7	
CONTROLS	600	2	3.5 (1.6 - 8.0)
SHAPIRO ET AL. 1984			
CASES	1,881	3	
CONTROLS	1,523	4	0.8 (0.5 - 1.1)

Figure 2. Reserpine and breast cancer

properly be used as a tool in drug surveillance. Eventually, in 1974, our group, which had come to be known as the Boston Collaborative Drug Surveillance Program (BCDSP), underwent binary fission—some might say nuclear fission. Jick remained as head of the BCDSP, while Slone and I co-founded the Drug Epidemiology Unit (after his death renamed the Slone Epidemiology Unit and finally, the Slone Epidemiology Center).

Personality clashes were one reason for the split but the scientific reasons had largely to do with an experience we had undergone, which I have since described as the reserpine/breast cancer disaster.[2] We (i.e. Jick, Slone, I and others) had reported a hypothesized causal association between the use of reserpine (then a popular antihypertensive) and breast cancer. After considerable hoopla, that association was eventually shown in multiple studies to be spurious. In retrospect and after the split, Slone and I came to realize that our initial hypothesis-generating study was sloppily designed and inadequately performed. In addition, we had carried out, quite literally, thousands of comparisons involving hundreds of outcomes and hundreds (if not thousands) of exposures. As a matter of probability theory, 'statistically significant' associations were bound to pop up and what we had described as a possibly causal association was really a chance finding.

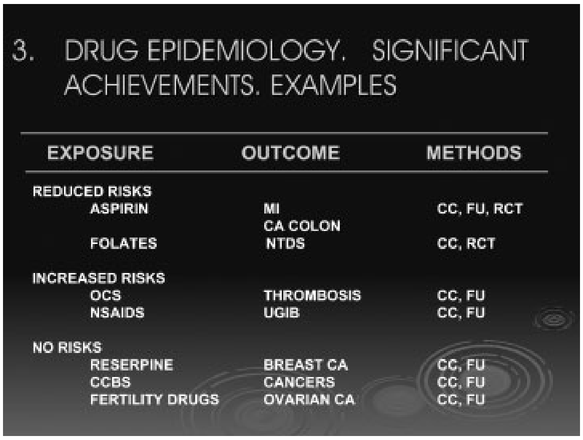
As soon as the error became clear to us, Slone and I acknowledged it.[3] Indeed, we (and our colleagues) went further (Figure 2): in due course we published findings from a better executed and much larger study that effectively ruled out an increased risk of breast cancer among reserpine recipients.[4,5] After a full decade had elapsed, that publication finally closed the chapter on an unhappy but educational episode. Provided they can acknowledge their mistakes and learn from them, I think all epidemiologists stand to benefit from a disaster or two.

I have dwelled on that disaster because Slone and I drew lessons from it that were to influence the remainder of our academic careers. The principal lesson was this: we should never forget that good science is skeptical science. Or to put the matter more formally, one way in which science proceeds is by attempting to falsify hypotheses; and further, we can embrace any given hypothesis, and then only tentatively, only for as long as we are unable to falsify it. That principle is not new. Karl Popper elaborated on it at length, even if he did go too far in rejecting other, inductive, lines of scientific reasoning. For us, however, the reserpine disaster embedded the falsification principle in our psyches, making it deeply personal and difficult to forget even in moments of excessive enthusiasm. Here I argue that the need for skepticism and for rigorous attempts to falsify hypotheses, especially when investigators interpret their own findings, is increasingly being lost sight of and I will explore some of the reasons.

But before I get to that subject, I must first lay some groundwork by considering the respective roles played by clinical medicine, clinical pharmacology and epidemiology in the evaluation of drug effects in humans. To deliberately oversimplify: in clinical medicine, the focus is on the cross-sectional or longitudinal observation of individual patients. In clinical pharmacology, the focus is on human experiments, such as pharmacodynamic studies or clinical trials, usually relatively short-term trials. In epidemiology, the focus is on populations and usually over much longer time spans than in clinical pharmacology.

Of course, the borderlines between the disciplines are indistinct and in places, they overlap, as they should, since they complement and reinforce each other. I would go even further: if valid inferences concerning drug effects are to be drawn, such inferences will commonly have to be based on evidence drawn from all three disciplines and that evidence should broadly converge on the same conclusion. Still further, causal inferences must also be informed by background evidence drawn from other related fields, such as pathology, microbiology and so on. Epidemiology, in isolation, runs the risk of being stupid epidemiology.

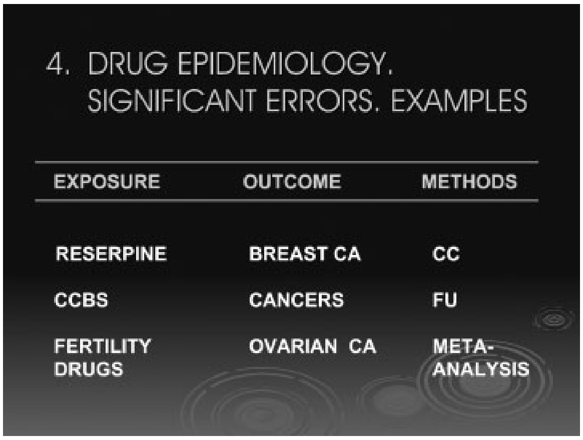
Secondly, since I will later on be concentrating on mistakes as a further piece of groundwork I want, early on, to make it clear that I do not contend that scientific skepticism equates with nihilism. Evidence that, for now, is not falsifiable, is what makes the world go round. More or less idiosyncratically, in Figure 3, I set out a few selected examples of triumphs in drug epidemiology—all of them of major clinical and public health importance. Epidemiology has documented reduced as well as increased risks, attributable to the use of drugs—and what is perhaps even more important is that it has demonstrated safety in the face of allegations to the contrary. Moreover, in this audience it would not be difficult to agree on a considerable expansion of the list, under all three headings. Note also that a wide range of methodologies have contributed to these achievements. I believe we can justly be proud of our accomplishments.



3. DRUG EPIDEMIOLOGY. SIGNIFICANT ACHIEVEMENTS. EXAMPLES

EXPOSURE	OUTCOME	METHODS
REDUCED RISKS		
ASPIRIN	MI CA COLON	CC, FU, RCT
FOLATES	NTDS	CC, RCT
INCREASED RISKS		
OCS	THROMBOSIS	CC, FU
NSAIDS	UGIB	CC, FU
NO RISKS		
RESERPINE	BREAST CA	CC, FU
CCBS	CANCERS	CC, FU
FERTILITY DRUGS	OVARIAN CA	CC, FU

Figure 3. Drug epidemiology. Significant achievements. Examples



4. DRUG EPIDEMIOLOGY. SIGNIFICANT ERRORS. EXAMPLES

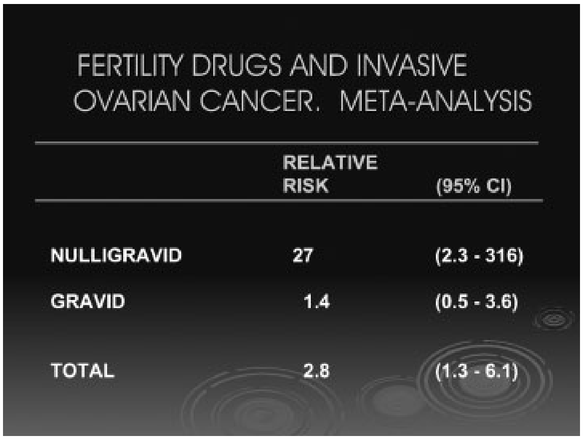
EXPOSURE	OUTCOME	METHODS
RESERPINE	BREAST CA	CC
CCBS	CANCERS	FU
FERTILITY DRUGS	OVARIAN CA	META-ANALYSIS

Figure 4. Drug epidemiology. Significant errors. Examples

So now, I move on to errors committed in the 20th century, and Figure 4 gives some egregious examples. Notice that they are the same examples as those given in the previous slide under the heading, 'no risks', although the methods used to commit the errors were not quite the same as those used to correct them. What can we learn from these examples? I have already considered the reserpine/breast cancer disaster. The claims made for the calcium channel blocker/cancer association[6] were based on a poorly designed follow-up study, which identified, for a biologically and clinically absurd outcome (*all* cancers—an outcome for which not even tobacco can be blamed), a low-magnitude relative risk estimate of 1.72.

But the real beauty in this array of spurious findings is the fertility drug/ovarian cancer association,[7] with meta-analysis, a disreputable methodology in search of statistically significant associations, no matter what, as the culprit. Based on this meta-analysis, it was suggested that this association may be causal. There was considerable publicity and the number of infertile women who denied themselves a potential opportunity to become pregnant will never be known. Ironically, it was subsequently shown that fertility drugs could not account for the association because the drugs at issue were not fertility drugs:[8] they were substances like estrogens with or without progestogens, thyroid hormone or even 'speed' (dextroamphetamine plus amobarbital).

For illustrative purposes, some of the data from the meta-analysis are given in Figure 5. For invasive ovarian cancer, the overall relative risk estimate was 2.8 and almost the entire contribution to that association came from the stratum of nulligravidae, in which the relative risk was a whopping 27. That estimate, although significant, was based on small numbers and it was exceedingly fragile as indicated by the extraordinarily wide confidence limits. In the archives of spurious associations, I suspect that a relative risk



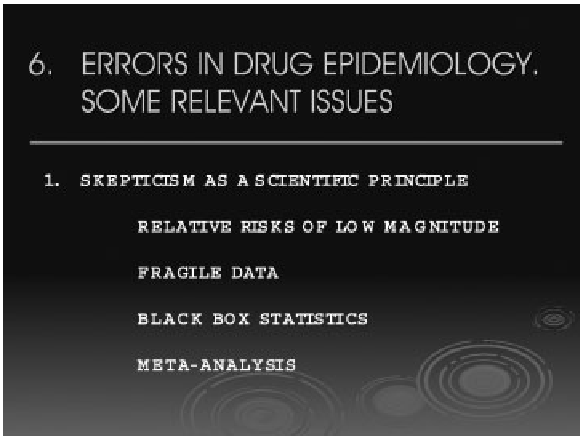
FERTILITY DRUGS AND INVASIVE OVARIAN CANCER. META-ANALYSIS

	RELATIVE RISK	(95% CI)
NULLIGRAVID	27	(2.3 - 316)
GRAVID	1.4	(0.5 - 3.6)
TOTAL	2.8	(1.3 - 6.1)

Figure 5. Fertility drugs and invasive ovarian cancer. Meta-analysis

point estimate of 27 may be a world record, which raises a question. This study was carried out collaboratively and with access to the 'raw' data by some of the most experienced epidemiologists in the United States: if they could screw up a relative risk estimate of 27, how can we ever hope to interpret estimates of 1.27?

Clearly, while we have enjoyed some major successes, we have also have to accept responsibility for having committed some major blunders. How has this state of affairs come about and can we do better? Obviously there are many possible answers to those questions, but I suggest that there are also some overarching systemic problems. It is time to examine the issues from a broader perspective and I return to item 1 in my earlier list of generic errors, the tendency toward the abandonment of skepticism (Figure 6) and to what I conceive to be some of the components of that tendency.



6. ERRORS IN DRUG EPIDEMIOLOGY. SOME RELEVANT ISSUES

1. SKEPTICISM AS A SCIENTIFIC PRINCIPLE
   - RELATIVE RISKS OF LOW MAGNITUDE
   - FRAGILE DATA
   - BLACK BOX STATISTICS
   - META-ANALYSIS

Figure 6. Errors in drug epidemiology. Some relevant issues

Figure 7. An epidemiologist struggles with relative risk estimates

Firstly, small relative risks. In 1968, when I attended a course in epidemiology 101, Dick Monson was fond of pointing out that when it comes to relative risk estimates, epidemiologists are not intellectually superior to apes. Like them, we can count only three numbers: 1, 2 and BIG (I am indebted to Allen Mitchell for Figure 7). In adequately designed studies we can be reasonably confident about BIG relative risks, sometimes; we can be only guardedly confident about relative risk estimates of the order of 2.0, occasionally; we can hardly ever be confident about estimates of less than 2.0, and when estimates are much below 2.0, we are quite simply out of business. Epidemiologists have only primitive tools, which for small relative risks are too crude to enable us to distinguish between bias, confounding and causation.

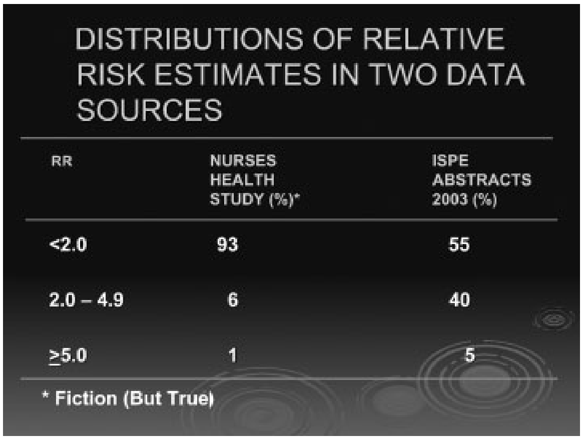
The problem, as Lynn Rosenberg pointed out several years ago,[10] is that we are running out of large estimates. In addition, despite refinements in our tools based on techniques such as sensitivity analysis, we have at the most shifted the boundary only slightly, if at all. For the most part we remain unable to interpret small risk increments[11] and the main value of sensitivity analysis is in testing the validity of relatively large ones, or alternatively, in demonstrating the vulnerability of small risk increments to bias and confounding.

Several years ago, Alvan Feinstein made the point that if some scientific fallacy is demonstrated and if it cannot be rebutted, a convenient way around the problem is simply to pretend that it does not exist and to ignore it.[12] No one has shown that small relative risks are interpretable but in the absence of large ones they have nevertheless become the rage. Without any new rationale being offered, they are being interpreted as causal and to make things worse, thanks to the advent of massive databases as well as massive studies, we are now able to identify more statistically significant but small relative risks than in the good old days and the temptation to interpret them as causal has become difficult to resist.

To illustrate that point, I have to allude to a problem that is usually avoided because to mention it in public is considered impolite: I refer to bias (unconscious, to be sure, but bias all the same) on the part of the investigator. And in order not to obscure the issue by considering studies of questionable quality, I have chosen the example of putatively causal (or preventive) associations published by the Nurses Health Study (NHS). For that study, the investigators have repeatedly claimed that their methods are almost perfect.

Over the years, the NHS investigators have published a torrent of papers and Figure 8 gives an entirely fictitious but nonetheless valid distribution of the relative risk estimates derived from them (for relative risk estimates of less than unity, assume the inverse values). The overwhelming majority of the estimates have been less than 2 and mostly less than 1.5, and the great majority have been interpreted as causal (or preventive). Well, perhaps they are and perhaps they are not: we cannot tell. But, perhaps as a matter of quasi-religious faith, the investigators have to believe that the small risk increments they have observed can be interpreted and that they can be interpreted as causal (or preventive). Otherwise they can hardly justify their own existence. They have no choice but to ignore Feinstein's dictum.

Figure 8 also presents a non-fictitious distribution of what I have judged, arbitrarily, to be the most representative relative risks in the published abstracts of this conference. The ISPE abstracts, you will be glad to know, have done somewhat better than the NHS: some 45% of the relative risks are above 2, and about 5% of them are respectably higher. However, some 55% of the estimates are below 2. To the credit of the investigators, a small number of the latter estimates are interpreted as suggesting no increased risk. However, a great majority of the associations are described in language such as this: 'Our findings suggest that x may increase (or decrease) the risk of y', when what should be said is: 'We have identified a small association, and have not the foggiest notion of what it means'. It appears that a substantial proportion of the membership of ISPE also stands charged with the over-interpretation of small relative risks.



DISTRIBUTIONS OF RELATIVE RISK ESTIMATES IN TWO DATA SOURCES

RR	NURSES HEALTH STUDY (%)*	ISPE ABSTRACTS 2003 (%)
<2.0	93	55
2.0 – 4.9	6	40
≥5.0	1	5

* Fiction (But True)

Figure 8. Distributions of relative risk estimates in two data sources (percent)

Next is fragile data. Not much needs to be said here except to note that very large relative risks are now commonly interpreted as causal, even though based on small numbers, simply because they are statistically significant. Of course, if relative risks based on small numbers were not large, they would not be significant. However, the requirement that under those circumstances there have to be strong grounds to justify the assumption that the data are error-free (or virtually so) is commonly ignored as illustrated in the meta-analysis of fertility drugs. Again the temptation to translate statistical significance into causality becomes irresistible.

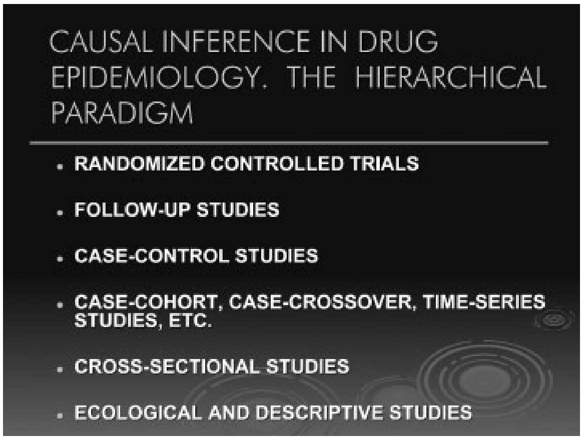
Next is black box statistics. Properly used, multivariate analysis has become a powerful and indeed indispensable tool in modern epidemiology. However, when it comes to small or fragile associations, it can be misused as a multivariate meat grinder to control simultaneously more factors than can be counted and hence to produce non-transparent, statistically significant relative risk estimates, the arithmetic of which neither you nor I can check for ourselves. Bradford Hill once remarked that no scientific paper that presents quantitative data is satisfactory if an independent reader is unable to check at least the main results on the back of an envelope.

Lastly is meta-analysis. Apart from what I have already said, I have on three occasions published detailed and (I believe) rigorous critiques of this approach.[8,9,11] Almost no one has responded and the few who have, have acknowledged the criticisms. However, they have also gone on to make the pie in the sky prediction that those deficiencies will be avoided in some future heaven. As yet, no heaven, and no pie in the sky. Not only is this state of affairs another instance of ignoring Feinstein's dictum but meta-analysis is now represented as the jewel in the crown of evidence-based medicine, a relatively recent fad which also calls, urgently I think, for its own monograph chapter.

So, as I see it, the overarching and systemic problem can be expressed as follows: when we consider the limits to causal inference in epidemiology imposed by factors such as low magnitude associations and fragile data as well as the smokescreens raised by techniques, such as the improper use of black box statistics or by meta-analysis, why has drug epidemiology (and indeed epidemiology in general) progressively moved away from a posture of skepticism?

I suggest that the answer to that question may be that we are operating under a false paradigm. But since the term has been much abused lately, I must first explain what I mean by a 'paradigm'. Kuhn originally used it to denote a set of governing and generally agreed upon scientific principles under which we operate and he used the term 'paradigm shift' to denote any radical transformation of that set of principles. Here (I hope) I use those terms more or less in the way he intended.

Figure 9 gives a hierarchical paradigm which depicts how the validity of the various approaches, used in causal research, have conventionally been ranked for many years. In this scheme (which for present purposes is knowingly oversimplified), randomized controlled trials (RCTs) rank as the 'gold standard' because randomization is supposed to take care of confounding, while 'blinding' takes care of bias. Follow-up studies rank next because they are deemed most closely to approximate the ideal of RCTs, even though they are susceptible to confounding and bias. Case-control studies rank next because they are represented as being more susceptible to bias than follow-up studies. Then comes a hodge-podge of approaches, the exact ranking order of which can be argued about but which is unimportant for present purposes.



CAUSAL INFERENCE IN DRUG EPIDEMIOLOGY. THE HIERARCHICAL PARADIGM

- RANDOMIZED CONTROLLED TRIALS
- FOLLOW-UP STUDIES
- CASE-CONTROL STUDIES
- CASE-COHORT, CASE-CROSSOVER, TIME-SERIES STUDIES, ETC.
- CROSS-SECTIONAL STUDIES
- ECOLOGICAL AND DESCRIPTIVE STUDIES

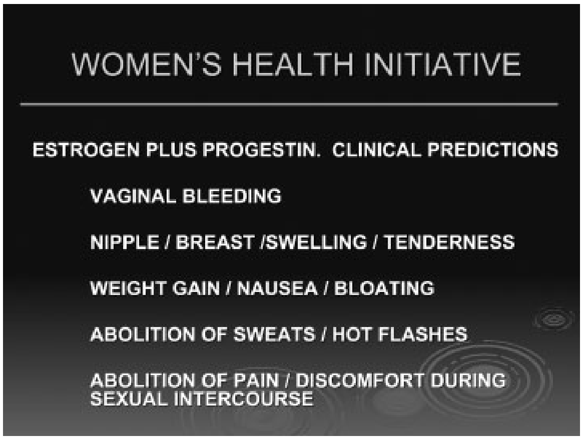
Figure 9. Casual inference in drugs epidemiology. The hierarchial paradigm

That paradigm is already being challenged. For example, today it is appreciated that it makes no sense to rank follow-up studies as superior to case-control studies or *vice-versa*: the two approaches simply constitute alternative methods of sampling exposed and non-exposed, and diseased and non-diseased persons, from a study base and in terms of confounding and bias, each approach has certain strengths and certain weaknesses. However, the paradigm has not shifted in that RCTs continue to be regarded as the undisputed 'gold standard'.

It is time to challenge the assertion that RCTs always constitute the 'gold standard'. On what can broadly be termed the clinical pharmacological time scale of days, weeks or months, when 'blinding' can truly be maintained and when adherence to the assigned treatment is high, that assertion may sometimes be defensible. However, on the epidemiological time scale, when exposed and non-exposed populations must be followed for years, rather than days, weeks or months, it is seldom possible to maintain 'blinding' or to accomplish acceptably high levels of adherence. In exceptional cases, those objectives may be achieved (e.g. aspirin prophylaxis against ischemic heart disease). For the most part, however, although studies of this type start out as RCTs, they soon become 'unblinded' and consequently, potentially biased and confounded. That is, they become encumbered with all of the limitations inherent in observational research.

A further consequence of having a hierarchical paradigm with RCTs occupying the throne is that statistical insight rather than clinical or biological insight, tends to assume the dominant role in causal inference with consequences that can be unfortunate.

These points need to be driven home and to do so, I turn to the risk of breast cancer in relation to the use of estrogen plus progestin, as reported in the Women's Health Initiative (WHI) RCT, the most ambitious, multi-dimensional and expensive RCT in the history of the universe. The estrogen plus progestin component of the study was stopped after a mean follow-up of 5.2 years for two reasons, one of them being a relative risk estimate of 1.26 for breast cancer, which '. . .almost reached nominal statistical significance'.[13] That association was interpreted as follows: the WHI was '. . .the first RCT to confirm that combined estrogen plus progestin *does* (my emphasis) increase the risk of breast cancer'. The possibility that the data may have been biased or confounded was not considered. Yet, bias and confounding were plausible explanations of that finding and were not ruled out.



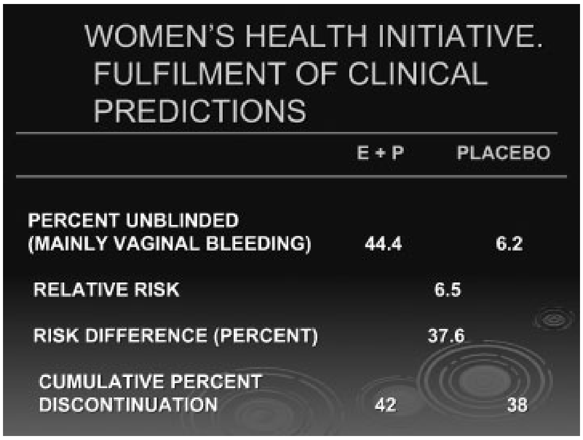
WOMEN'S HEALTH INITIATIVE

ESTROGEN PLUS PROGESTIN. CLINICAL PREDICTIONS

- VAGINAL BLEEDING
- NIPPLE / BREAST /SWELLING / TENDERNESS
- WEIGHT GAIN / NAUSEA / BLOATING
- ABOLITION OF SWEATS / HOT FLASHES
- ABOLITION OF PAIN / DISCOMFORT DURING SEXUAL INTERCOURSE

Figure 10. Women's health initiative

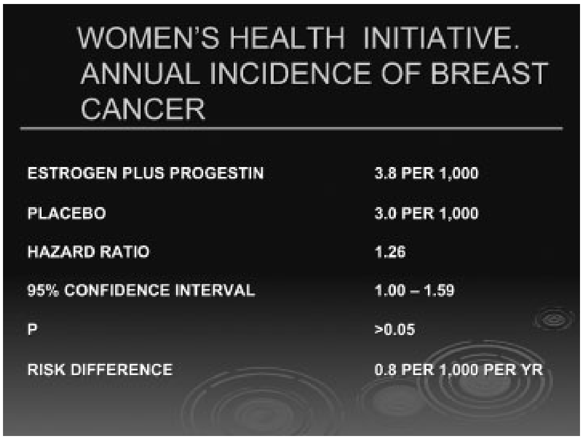
Firstly bias: any gynecologist could have predicted that, inevitably, a large proportion of women given estrogen plus progestin would not remain 'blind' for a multiplicity of reasons (Figure 10). And in fact (Figure 11), in practice it turned out that 44.4% of the exposed women, as against 6.2% of the placebo recipients, had to be 'unblinded', mainly because of break-through bleeding: a relative risk of 6.5 and a risk difference of 37.6%. These differences are BIG by any standard and conservative, since many women on estrogen plus progestin who remained 'blinded' would nevertheless have become aware that they were taking hormones because of symptoms such as breast tenderness.



WOMEN'S HEALTH INITIATIVE. FULFILMENT OF CLINICAL PREDICTIONS

	E + P	PLACEBO
PERCENT UNBLINDED (MAINLY VAGINAL BLEEDING)	44.4	6.2
RELATIVE RISK	6.5	
RISK DIFFERENCE (PERCENT)	37.6	
CUMULATIVE PERCENT DISCONTINUATION	42	38

Figure 11. Women's health initiative. Fulfilment of clinical predictions



WOMEN'S HEALTH INITIATIVE. ANNUAL INCIDENCE OF BREAST CANCER

ESTROGEN PLUS PROGESTIN	3.8 PER 1,000
PLACEBO	3.0 PER 1,000
HAZARD RATIO	1.26
95% CONFIDENCE INTERVAL	1.00 – 1.59
P	>0.05
RISK DIFFERENCE	0.8 PER 1,000 PER YR

Figure 12. Women's health initiative. Annual incidence of breast cancer

It is well established that at any given time there is a large pool of undiagnosed but potentially detectable breast cancers among postmenopausal women and indeed it is for that reason that screening appears to be effective.[14] Figure 12 shows that if knowledge of estrogen plus progestin use augments the detection of breast cancer that would otherwise remain clinically silent by as little as 0.8 per 1000 per year, it would entirely account for the association. In addition, for a lower 95% confidence limit of 1 which, in any case, is only 'almost significant' (investigator bias?), any amount of detection bias would be sufficient to account for it.

Secondly, confounding: in a 5.2 year follow-up study, with at least 44.4% of exposed women being aware that they received estrogen plus progestin and with a discontinuation rate among them of 42%, a large proportion of which occurred years before the diagnosis of breast cancer, confounding (e.g. because of more frequent breast examinations) not only becomes possible, but is even likely. In RCTs, analysis according to intention to treat is advocated as the appropriate way to deal with confounding in circumstances when 'unblinding' is not a major problem and when losses to follow up are modest and of short duration. However, when they are not, that approach becomes pointless and misleading. There is no option but to consider and to allow for confounding, to the extent possible, as in any observational study.

Do female hormones increase the risk of breast cancer? Perhaps yes, perhaps no, but it certainly cannot be claimed that the WHI study has '. . .[confirmed] that combined estrogen plus progestin *does* (my emphasis) increase the risk of breast cancer'. What started out as an RCT became an observational study, as was clinically predictable and as was quantitatively confirmed.

And to generalize, many, perhaps most, other trials conducted on an epidemiological time scale clearly become observational studies. For example, in a RCT of cholesterol-lowering agents in relation ischemic heart disease risk, the exposed patients inevitably tend to become aware of the fact because their cholesterol levels improve. To this it must be repeated that, as in any observational study, patients start and stop the medication for a host of reasons, get lost to follow-up and so on.

In short, on the epidemiological scale, the concept that RCTs are the 'gold standard' and superior to observational studies, is not defensible. Certainly there are circumstances when they may offer unique advantages but so may other methods. What is needed is a paradigm shift under which it is recognized that there is no hierarchy, at least with regard to the ranking of RCTs, follow-up studies and case-control studies: each approach has advantages and drawbacks and no method is intrinsically superior or inferior to the other.

I return to a consideration of the limits to causal inference in epidemiology: when we are confronted by small or fragile risks which are further compounded by black box statistics and techniques such as meta-analysis, epidemiology will remain an unsatisfactory tool for making the distinction among bias, confounding and causality. Our talent is in the identification of large risks. Such risks are scarce right now, so how do we get out of the dilemma? I suggest that the way out is through more intimate collaboration between clinical medicine, biology and epidemiology: under those conditions well formulated hypotheses, if they are indeed causal, should yield large relative risk estimates.

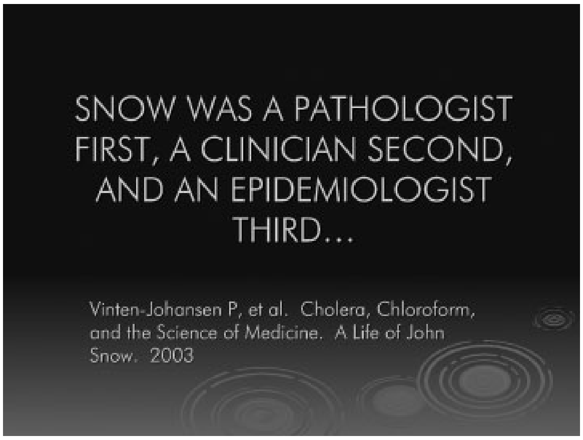
Some may argue that it is of public health importance to identify and evaluate possible causal implications of small relative risks because for common diseases these can translate into large absolute risks. That is a complex subject that also calls for its own monograph chapter. Unfortunately however, not all questions are answerable even if we desperately want answers and public health importance does not equate with scientific validity. The answers to some questions, if they can be answered, may have to depend on methods other than those used in epidemiology.

If we are to move away from the paradigm of the RCT 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. For several decades, clinical medicine has been in retreat under the onslaught of the paradigm of the superiority of RCTs and that paradigm in its turn has given pride of place to statistics. Since we are in the business of measurement and counting, statistics of course, is vital to epidemiology but under the paradigm of the RCT as the 'gold standard' it has become the master when it should be the servant.

I promised to return to my conceptual objections to the organization of pharmacoepidemiology as an entity separate from the general community of epidemiologists. I do not believe there is a separated discipline of pharmacoepidemiology and as illustrated in several of the examples I have given, a great deal of the research was carried out by epidemiologists without the 'pharmaco-' prefix. I subscribe to Occam's razor: there is only epidemiology and different concentrations such as pediatric, drug or cancer epidemiology are simply part of it. In recognition of that unity, other groups, for example pediatric epidemiologists operate within the framework of existing epidemiological organizations. I suggest that ISPE should do the same.

In the 21st century, are we doomed to compound our errors? I would not even hazard a guess in answer to that question: it all depends. However, in my dealings with colleagues in other sectors of medicine, I have the impression that epidemiology is increasingly coming into disrepute because bad epidemiology is more and more tending to obscure good epidemiology. If we are to become serious again, it may depend on a reaffirmation of scientific skepticism and on a paradigm shift.

Recently, a remarkable book on the life of John Snow, the father of modern epidemiology, has been published[15] (Figure 13). The authors observe that 'Snow was a pathologist first, a clinician second, and an epidemiologist third'. I think Snow got it exactly right. And last of all, I believe epidemiology must also be informed by everyday experience, including



SNOW WAS A PATHOLOGIST FIRST, A CLINICIAN SECOND, AND AN EPIDEMIOLOGIST THIRD...

Vinten-Johansen P, et al. Cholera, Chloroform, and the Science of Medicine. A Life of John Snow. 2003

Figure 13. Snow got it right

non-quantitative experience. Many years ago, CP Snow wrote about the divorce between science and art and its detrimental consequences. Good art make us think in novel ways and in a recent harrowing novel Gunter Grass made a comment that all of us would do well to remember. He was not talking about epidemiology but about drowned refugees. This is what he said: 'But what do numbers tell us? Numbers are never accurate. In the end you always have to guess'.[16] I agree.

REFERENCES

1. Doll R. *Prevention of Cancer*. The Nuffield Provincial Hospitals Trust 1967.
2. Shapiro S. Case-control surveillance. In *Pharmacoepidemiology* (2nd edn), Strom BL (ed.). John Wiley and Sons: Chichester, UK; 1994. 301–322.
3. Shapiro S, Slone D. Case-control study: consensus and controversy. *Comment J Chronic Dis* 1979; **32**: 105–107.
4. Boston Collaborative Drug Surveillance Program: reserpine and breast cancer. *Lancet* 1974; **ii**: 669–671.
5. Shapiro S, Parsells J, Rosenberg L, *et al*. Risk of breast cancer in relation to the use of Rauwolfia alkaloids. *Eur J Clin Pharmacol* 1984; **26**: 143–146.
6. Pahor M, Guralnik JM, Corti M-C, *et al*. Calcium channel blockade and incidence of cancer in aged populations. *Lancet* 1996; **348**: 493–497.
7. Whittemore AS, Harris R, Intyre J, *et al*. Characteristics relating to ovarian cancer risk; collaborative analysis of 12 US case-control studies. II. Invasive epithelial ovarian cancers in white women. *Am J Epidemiol* 1992; **136**: 1184–1203.
8. Shapiro S. Commentary: Is meta-analysis a valid approach to the evaluation of small effects in observational studies? *J Clin Epidemiol* 1997; **50**: 223–229.
9. Shapiro S. Meta-analysis/shmeta-analysis. *Am J Epidemiol* 1994; **140**: 771–777.
10. Rosenberg L. Presidential address, 26th annual meeting of the Society for Epidemiologic Research, Keystone, Colorado, 1993.
11. Shapiro S. Bias in the evaluation of low-magnitude associations: an empirical perspective. *Am J Epidemiol* 2000; 939–945.
12. Feinstein AR. *Clinical Biostatistics*. The CV Mosby Company: St. Louis, Missouri, 1977.
13. Writing group for the Women's Health Initiative investigators: risks and benefits of estrogen plus progestin in healthy postmenopausal women. *JAMA* 2002; **288**: 321–333.
14. Shapiro S, Venet W, Strax P, *et al*. Ten-to fourteen-year effect of screening on breast cancer mortality. *J Natl Cancer Inst* 1982; **69**: 349–355.
15. Vinten-Johansen P, Brody H, Paneth N, Rachman S, Rip M. *Cholera, Chloroform and the Science of Medicine: A Life of John Snow*. Oxford University Press: Oxford, UK; 2003.
16. Grass G. *Crabwalk: RCTs, Follow-up Studies, and Case-Control Studies*. Harcourt Inc: New York, 2003.