Doi: 10.1111/j.1742-7843.2008.00207.x
 Basic & Clinical Pharmacology & Toxicology, **102**, 257–266

MiniReview

Establishing Evidence for Early Action: the Prevention of Reproductive and Developmental Harm

David Gee

Strategic Knowledge and Innovation, European Environment Agency, Copenhagen, Denmark

(Received July 20, 2007; Accepted November 12, 2007)

Abstract: Developmental and reproductive toxicants that cause serious disease and dysfunction, either lifelong or occurring late in life, can be initiated in the early life stages of human beings and other species. It is often the timing of the dose more than the dose itself that distinguishes harmful from harmless exposures to such toxicants. As much of the harm is irreversible, and sometimes multigenerational, the timing of actions to prevent such harm is also critical. In determining when there is a sufficiency of evidence to justify early prevention of harm, decision-makers need to take account of the implications of multicausality, the methodological biases within environmental sciences, and the need to take precautionary, as well as preventive actions to eliminate or reduce exposures. The widely used Bradford Hill causal 'criteria' are briefly reviewed in light of multicausality. Reaching agreement between stakeholders on a sufficiency of evidence for early action to reduce exposures to toxicants requires the consistent use of transparent definitions of the concepts and terms used to characterize the strength of evidence between causes and effects. Proposals are made to improve those in current use, including a definition of the precautionary principle.

Serious damage to health can be initiated in the early life stages of man and other species, but it may not become apparent until much later in their development. Harm that originates at the foetal stage can also appear in early life, causing birth defects, life-time dysfunctions, such as lower IQ, or diseases of early adulthood, such as testicular and vaginal cancer and other damage to reproductive organs.

This insight into the foetal origins of much adult reproductive and developmental disease has emerged from environment and health research, particularly over the last 10 years. The research builds on what was known about earlier experiences of foetal toxicity for human beings and wildlife that arose from the cases of thalidomide, diethylstilbestrol (DES), dichloro-diphenyl-trichloroethane (DDT) and tributyltin (TBT) [1–7]. The mechanisms of action in each of these earlier experiences are not yet well established, despite decades of research. However, it seems clear that it is more the timing of the dose, rather than the dose itself, which, *inter alia*, distinguishes harmful from harmless exposures to reproductive and developmental toxicants [8].

Such harm is often irreversible and sometimes multigenerational, causing life-time personal and societal costs that cannot be offset by any benefits to the individual from intrauterine exposures. Thus, biology, economics, equity and morals all justify early actions to prevent developmental and reproductive harm. But how can a sufficiency of evidence for action be established, then communicated, so as to gain sufficient consensus for timely prevention?

This article examines some of the main barriers to achieving such consensus. These barriers include multicausality; its implications for statistical analysis and the application of the Bradford Hill 'criteria' on causality; the methodological biases within environmental sciences against establishing causality; the different strengths of evidence needed to achieve the conflicting goals of science and public policy; and the use of opaque and inconsistent terms to characterise cause–effect relationships. Suggestions are made that may help overcome these barriers.

From monocausality to multicausality

The biological processes that lead to chronic diseases such as breast or prostate cancer, or to other reproductive or developmental harm, appear to involve some, or all, of at least eight main events in the disease process: preparation within the host; initiation; promotion; retardation; progression; disease onset; the strengthening or weakening of the severity; and prevalence of the disease. These steps in the causal chain of the disease process seem to be linked by interdependent, cocausal risk factors. Some factors, including the play of chance, may operate at one or several stages of the same disease process, and/or within other causal chains that can lead to the same disease or disorder.

Author for correspondence: David Gee, European Environment Agency, Kongens Nytorv 6, DK-1050 Copenhagen K, Denmark (fax +45 33 36 71 42, e-mail david.gee@eea.europa.eu).

Timing of exposures is usually critical for reproductive and developmental harm. For example, the harm caused by thalidomide exposure occurred only in a window of susceptibility, which in this case was from days 35 to 49 after conception [9]. And it is often relatively low doses that are sufficient to initiate harm during these critical windows of vulnerability: larger doses outside these sensitive periods seem to have little harmful impact.

It is, therefore, a challenging task to identify the 'causal' and often cocausal factors needed to prevent or reduce the population burden of such ill health, given that exposures occur at different developmental stages; are often interactive, mixed and usually low level; and affect people with specific environmental histories and susceptibilities. Such complex disease processes can be investigated in several ways.

Within the history of the public health sciences, there has long been, and still is, a tension between a monocausal, reductionist approach to investigating disease aetiology, and multicausal, more holistic approaches. Some scientists frame their studies around the view that it is the germ, or the gene, the oncogene, or a single risk factor, which is mainly 'responsible' for disease. Others look to the overall environmental history of the host for the many factors and influences that, if taken together, may explain disease causation [10–12].

This often creative tension is partly due to the early successes of the germ theory of disease in helping to both identify and eliminate the micro-organisms that caused infectious diseases such as cholera [13]. The tradition has continued with the current focus on genes as the 'main' cause of many diseases, despite the wide acceptance that most common diseases arise from the interplay among gene expression, host and environmental influences [14–19].

A monocausal approach to complex biological systems and processes will generate large numbers of individual risk factors, which though perhaps important, shed little light on how they interact with other factors. For example, by 1980 some 246 'risk factors' for heart disease had been identified [20]. And by 2002 there were over 100 oncogenes and 15 suppressor genes that had been identified as possibly playing a key and context-dependent role in cancer causation, with, for example, the over-expression of a specific gene causing cell growth in one cell type but cell inhibition, or cell death, in another [21].

Concentrating research on particular parts of the puzzle, rather than on the causal puzzle itself, may inhibit the clarification of causality. 'The focus on specific factors, with the corresponding neglect of complex causal processes, is a fundamental weakness' in population health research [22].

For example, some 4000 chemical substances have been identified in tobacco smoke, of which 167 are currently classified as toxic. However, the disease process that leads to cancer or heart disease in some smokers, but not in others, is still largely unknown after more than 40 years of research.

Despite the absence of knowledge about the specifics of the disease process within mixtures, it has still been possible to prevent some harm, albeit belatedly, by reducing exposures to the whole mixture, such as tobacco and combustion smoke, welding and rubber fume, and fine particles of air pollution.

The practical difficulties of studying and understanding complex multicausal biological processes have meant that the attraction of a monocausal approach remains strong, despite the need to be more realistic about biological processes [23,24]. Reductionism, and the metaphor of the body as a machine, are powerful paradigms that continue to support the idea of linear relationships between specific causes, long after knowledge about irreducible uncertainties, emergent properties and non-linear dynamics have became available [25,26].

From confounders to cocausal factors?

The tools available to unravel multicausal, complex and dynamic disease processes, are in their infancy [27]. As a consequence, most epidemiologists try to identify specific risk factors while eliminating possible confounding factors via various statistical techniques. Such 'statistical surgery', or context stripping [28,29] may remove many confounders from the analysis that are really cocausal factors. If the focus is on just one toxicant, then other 'environmental properties tend to be regarded as marginal and designated as covariates or confounders: treating such environmental conditions as confounders is equivalent to defining genetic differences as confounders' [30].

Even with a well-studied phenomenon, such as lead poisoning, there is a growing realization that lead exposure, environmental deprivation and enrichment, and neurotoxicity are complex and 'perhaps bidirectional' [31]. For example, an enriched and intellectually stimulating home environment seems to reduce the harmful effects of a toxicant such as lead, while lead exposure can reduce the benefits of such enriched environments. Similarly, a deprived socio-economic environment can increase the harmful effects of lead while reducing the beneficial effects of a reduction in lead exposure. In a study comparing socio-economic variables, such as poverty and the percentage of immigrants, with individual variables, such as sex, and low birth weight, it was found that the community risk factors were significant predictors of reading scores even after the individual risk factors were accounted for [32].

More fundamentally, scientists have also noted that bidirectional relationships, such as cell signalling and cross-talk, implying that causality may be circular [33]. Similar scientific challenges emerge from the field of endocrine disruption in wildlife, as well as within ecotoxicology more generally [34]. These arise from having to investigate and draw inference across biological scales, from population level to lower levels of biological organization and back again, in order to show, for example, whether harm to individual fish can cause fish population decline [35].

It would seem then that the 'key to understanding these causal processes is clearly the ability to elaborate and understand complexity: the interacting systems involved will always overwhelm predictions of independent effects of any single

Table 1.

The Bradford Hill 'criteria' (Bradford Hill 1965) for helping to move from association to causation, with some illustrative examples from the European Environmental Agency report 'Late Lessons from Early Warnings' (European Environmental Agency, 2001).

1. Strength of the association?	John Snow found 71 cholera deaths per 1000 houses served by polluted water but only 5 per 1000 houses served with sewage-free water (London, 1854).
2. Consistent results?	The US Surgeon General Report in 1964 found 36 studies linking smoking with lung cancer.
3. Specific effects?	In 1959, the then rare cancer, mesothelioma, was observed to kill children in South Africa who played on asbestos waste tips without there being increases in other causes of their death.
4. Temporality?	'Is the cart coming before the horse'? The DES exposure of mothers occurred before rare cancers in their daughters were observed (USA, 1970).
5. Biological gradients?	Does the effect increase with dose, if such exposure measurements are available? For example, TBT exposure from boats and imposex in snails (UK, 1986).
6. Biological plausibility?	Depends on the 'knowledge of the day', cannot be robust, as the observation may be new. For example, PCB contamination of eagles, (Sweden, 1966).
7. Coherence?	Is the evidence coherent with general known factors? For example, radiation damage from X-rays (USA, 1904). Also dependent on the knowledge of the day.
8. Experiment (reversibility)?	Does prevention prevent? For example, a reduction of SO_2 eventually leads to less lake/forest acidification (Sweden, 1998).
9. Analogy?	For example, collapsing fish stocks from over-fishing in different areas (e.g. California sardine collapse, 1942, was a useful lesson for other fish stocks).

factor, reducing them to very limited and uncertain information' [36]. These examples illustrate the need for a scientific approach that is based more on relationships between interdependent variables operating within non-linear and complex systems than on the reductionism of monocausality. It also follows that in complex systems very small changes in key variables can have profound effects: 'small' can be very significant in finely balanced non-linear systems, where, as Heraclitus observed centuries ago, there is a 'harmony of opposites'. Removing even the 'smallest' link in an interdependent causal chain can break at least one chain of disease causation.

But how can we escape from this monocausal 'prison of the proximate' [36] in the environmental health sciences without losing the precision and relevance required of good science? And how can we identify possible or probable causality from observed associations in complex biological and ecological systems, so that priorities for public health protection can be agreed?

Multicausality and the Bradford Hill 'criteria' for causality

> 'With preventive medicine in mind the decisive question is whether the frequency of the undesirable event B will be influenced by a change in the environmental feature A'. [37]

Bradford Hill began his classic 1965 article on causation in environmental health by asking how '*the*' environmental feature seen to be *associated* with harm could be reliably identified as *the cause* of that harm. He described nine characteristics ('features' or 'viewpoints') of scientific evidence that, if taken together, could help scientists to move with some confidence from association to causation (table 1). His subsequently misnamed 'criteria' are still widely used [2,38].

Bradford Hill's explicit approach to deriving causation from association was essentially based on monocausality, that is, on finding *the* specific cause of a specific disease. However, there was tension between monocausality and multicausality in his article. The 'decisive' question for him was whether event A 'influenced' the 'frequency' of event B. This suggests that he was aware that several factors would be implicated in disease but that removing one of them may reduce its frequency, or incidence, without necessarily eliminating it entirely. He also acknowledged the other, simpler type of multicausality, which is where one disease can have several different causes, noting that: 'diseases may have more than one cause. It has always been possible to acquire a cancer of the scrotum without sweeping chimneys or taking to mule spinning in Lancashire' [37].

His views on multicausality did not prevent him from feeling that there may be one ultimate cause of disease: 'if we knew all the answers we might get back to a single factor' [37]. In light of multicausality and complexity and after some 40 years of their use and of expanding knowledge, the criteria need to be reviewed [39]. For example, two of the strongest criteria, temporality and consistency of research results, seem less robust now than they did in the essentially monocausal world of Bradford Hill.

Temporality asserts that the cause must precede the effect. This is obviously so: except possibly under conditions of reciprocal relationships where, as noted above, causality may be circular [32]. Temporality then becomes less clear.

In addition, where there are multiple causes, an overall trend in a biological end-point can be established under the influence of some causal factors well before the emergence of other 'component causes' [17] that may reverse, stabilize, or accelerate this trend, depending on their relative strengths. If this feature of reality is not taken into account, then simplistic interpretations of temporality can lead to shaky conclusions. For example, in a review of the evidence on falling sperm counts and endocrine-disrupting chemicals, it was concluded that, as overall sperm counts began to fall in some countries in advance of the rise of chlorine-based

chemistry, such chemical exposures could not be a cause of change in the overall trend [2]. In the context of multicausality, where the combined effects of several causes together determine the overall time trend of a biological end-point, such a conclusion is not soundly based, whatever the true role of chemicals in causing falling sperm counts.

The criterion of *consistency* between the results of different studies into the same phenomena, when present, clearly adds much confidence to assertions of causality. However, consistency in nature does not require that all or even a majority of studies on the same issue find the same effect. 'If all studies of lead showed the same relationship between variables, one would be startled, perhaps justifiably suspicious' [41]. The sources of variability arise both from the study and the investigator, such as the framing and initial assumptions; the models, methods and statistical analyses used; the choice of population group; and the data selected for collection and then analysis. Other sources of variability and bias have been noted by Bailar [42]. In addition, there are the sources of variability arising from the 'sociomics' of environments and the epigenetics of individuals.

It is hardly surprising therefore that, after decades of research, most lead studies can still only 'explain' 30–40% of the variance in most lead linked biological end-points, and sometimes far less [31]. As inconsistency from complex biological and ecological systems is to be expected, the *absence* of consistency between studies will not provide much weight to support assertions about the absence of causality.

Meanwhile, another less used criterion, *analogy*, should perhaps be given greater weight in today's circumstances. Bradford Hill described the use of analogy thus: 'with the effects of thalidomide and rubella before us we would surely be ready to accept slighter but similar evidence with another drug or another viral disease in pregnancy'.

The case for using analogies more often, in conjunction with another criterion, *biological plausibility*, based on the effects of polychlorinated biphenyls (PCBs), chlorofleurocarbons (CFCs), DDT, TBT , DES and other better-known substances, is stronger now than it was in the 1960s because of the knowledge now available from experiences with these substances, despite large gaps in knowledge about their mechanisms of action. Applying this knowledge to 'similar' substances from among the 70,000 in common use would help save some of the animal and human resources needed to fill the very large information gaps on their toxicities, and could well prevent much harm. An illustration of this use of analogy and plausibility is provided by a recent assessment of the cancer hazard of some chemicals that are similar to vinyl chloride [43]. However, there is also much evidence about the major differences in biological effects that minor differences in chemical structure can cause. This invites caution when making greater use of analogy and of 'group toxicity' classifications, similar to the toxic equivalents for dioxins.

Two other well-used criteria need to be given less weight in light of today's knowledge. The criterion of *recovery* (or 'experiment', as Bradford Hill rather confusingly called it), by which he meant evidence that removing a cause should lead to less disease, is not very robust in a complex world. This is due to delayed impact recovery times (e.g. due to secondary exposures that can arise from contaminated soils and sediments long after persistent substances, such as lead, PCBs and DDT, have been banned); to the generational effects of developmental toxicants; and to the difficulties of identifying specific causes of successful prevention in multicausal systems, except when one toxicant dominates causality, for example, tobacco and lung cancer, leaded petrol and blood lead, and the coal smoke that causes respiratory disease.

Bradford Hill included *a linear dose response relationship* as another important criterion. However, where the timing of exposure is more important than the dose itself, and where non-monotonic, 'low-dose' effects are present, then the absence of a linear dose–response relationship does not provide robust evidence against causality.

Less weight should also be placed on *specificity* as a criterion, given the widespread prevalence of 'many to many' cause and effect relationships, and the capacity of many substances, such as PCBs, asbestos, lead, mercury, etc., to cause many types of harm.

Finally, the *strength of association*, which Bradford Hill put first in his list of features, is clearly still very relevant, but with caveats that arise from multicausality. A low relative risk of, say, 1.5, if replicated in several studies, can be very robust for a multicausal disease such as heart disease, such as the case with smoking and heart disease. Such a 'low' relative risk will also represent much harm if the disease is widespread.

In judging strength of association Bradford Hill also warned against the overuse and misuse of statistical significance testing: 'we waste a deal of time, we grasp the shadow and lose the substance, we weaken our capacity to interpret data and to take reasonable decisions whatever the value of P. And far too often we deduce "no difference" from "no significance" '. Despite similar cautions being regularly repeated since then, the misinterpretation of statistical significance, and the relative neglect of confidence intervals, continues [44]. As Bradford Hill intimated, a statistical significant relationship with wide confidence intervals is generally not as precise or robust as a non-significant result with narrow confidence intervals.

The weight that Bradford Hill gave to each of his 'criteria', as well as to their totality, was nuanced. He recognized that biological complexity, and the practicalities of prevention, invited caution in the use of his criteria: 'What I do not believe – and this has been suggested – is that we can usefully lay down some hard-and-fast rules of evidence that must be obeyed before we can accept cause and effect. None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non'.

The complexity and multicausality of biological systems, and the time lag between critical exposures and health outcomes, especially with reproductive and developmental harm, means that the task of linking harm to sometimes decade's old exposures is very difficult. The observed

Table 2.

Some methodological features of environmental health science and their main directions of error.*

Scientific studies	Some methodological features	Main† directions of error-increases chances of detecting a
Experimental studies (animals)	• High doses‡	• False positive
	• Short (in biological terms) range of doses	• False negative
	• Low genetic variability	• False negative
	• Few exposures to mixtures	• False negative
	• Few foetal life-time exposures	• False negative
	• High fertility strains	• False negative (developmental/reproductive end-points)
Observational studies (wildlife and human beings)	• Confounders	• False positive§
	• Inappropriate controls	• False positive/negative
	• Non-differential exposures misclassification	• False negative
	• Inadequate follow-up	• False negative
	• Lost cases	• False negative
	• Simple models that do not reflect complexity	• False negative
Experimental and observational studies; and scientific culture	• Publication bias towards positive studies	• False positive
	• Low statistical power (e.g. from small studies)	• False negative
	• Use of 5% probability level	• False negative
	• Scientific cultural pressure to avoid	• False negative
	• Much re-analysis of 'positive' studies	• False negative
	• Little replication of 'negative' studies	• False negative

*'Direction' of error only: size of error will vary.
†All features can err in either direction; some features err equally in either direction (e.g. inappropriate controls) but most of the features mainly err in the direction shown in the table.
‡When there are 'low-dose' effects, this feature will err towards false negative.
§Maybe false negative if there are cocausal factors.

associations will, in most cases, be multiple, weak, confounded, interactive and often inconsistent.

In such circumstances, the asymmetry that characterizes the Bradford Hill criteria is even more pronounced, that is, the *presence* of the criteria can be robust evidence *for* a causal association, while the *absence* of the criteria is not robust evidence *against* a causal association. Bradford Hill drew attention to this asymmetry, but some of his followers have forgotten this [39].

The use, and sometimes misuse, of the Bradford Hill criteria in the difficult context of multicausality is one barrier to the timely establishment of sufficient evidence for preventive action. But is there another barrier from the systemic biases that arise from some of the common methods used in the epidemiology and toxicology that generate the evidence?

Main directions of error in environmental health sciences

Table 2 illustrates the main directions of error for a number of methodologies used in toxicology, epidemiology and statistical analysis, and from the culture of scientific publishing. Most features are biased towards generating false negatives (i.e. falsely attributing safety when harm is the reality). These biases help to produce robust science by avoiding false positives at the expense of false negatives, but they can contribute to poor public policymaking, where 'false negatives', such as asbestos, DES, TBT, PCBs, etc., can lead to much harm and costs [45]. A key goal of public policy is to avoid such false negatives, sometimes, though rarely, at the expense of false positives.

Both policymakers and scientists need to acknowledge and take account of these main directions of error when they evaluate the methods and results of research. Many scientists do [46,47] but awareness of these biases among many stakeholders appears to be low.

A promising line of research would be to investigate whether there could be improved scientific methods for the environmental health sciences that would help to achieve a more ethically acceptable, and economically efficient, balance between the generation of 'false negatives' and 'false positives', but which did not compromise good science. Meanwhile, when do we know enough to take timely action to reduce threats of harm?

A sufficiency of evidence for action?

> Public health decisions about moving from 'evidence to action' are a balancing act between what needs to be known and what ought to be done. [48]

It took more than 40 years of much scientific endeavour and debate between the 1940s and the 1980s, before what was known about smoking and lung cancer was applied to protect public health, following sustained opposition from economic and political interests. In this case, the opportunity for *precautionary* action on a likely hazard in the 1960s was lost: by the 1990s only the *prevention* of known harm was possible.

Both the prevention and precautionary principles, as well as the proportionality principle, which prevents the unreasonable use of precaution, are now explicit parts of European Union

(EU) law so the opportunities to implement more timely prevention are greater than in the 1960s.

Applying the precautionary principle to achieve the timely protection of public health from reproductive and developmental harm, therefore, requires action on credible early warnings, even though there may not be proof of causality, or knowledge of mechanisms of action. Beneficent actions without known mechanisms of action are common: much medical practice achieves 'recovery' of the patient in the absence of knowledge about how they recover.

An example of an early warning, which illustrates several key issues involved in the application of the precautionary principle, is provided by the 1969 recommendation from the UK Medical Research Council Swan Committee to severely restrict antibiotics in animal feed. It was based only on 'a sufficiently sound basis for action': causality had not been 'established', and 'mechanisms of action' were unknown: and largely still are. But the further research that was clearly needed did not provide sufficient grounds for deferring action to reduce such potentially serious and possibly widespread harm from antibiotic resistance.

Some countries, particularly Sweden, later heeded this warning and took action in 1985 to stop the use of antibiotics as animal growth promoters, but most did not, despite the clarity and weight of this early warning. It took 30 years since the 1969 early warning before the EU banned antibiotics as growth promoters [44]. The supplier of the growth promoter did not believe that there was a sufficiency of evidence for such action and challenged the European Commission's decision. The European Court of Justice upheld the Commission's use of the precautionary principle to justify its action: the likely costs of inaction greatly outweighed the likely costs of action in this case, even though the evidence for causation was not very strong [48].

Different strengths of evidence for different purposes

Bradford Hill saw the need for: 'differential standards before we convict'. He recognized that an appropriate choice of strength of evidence for each case was essentially dependent on the likely costs (and, critically, on their distribution between different parties), of being wrong in acting, or not acting, to eliminate or reduce exposures.

He illustrated this approach with three strengths of evidence that he judged would be appropriate for three very different circumstances: 'relatively slight evidence' for a ban on a widely used pregnancy pill; 'fair evidence' for eliminating exposure to a probably carcinogenic mineral oil used at work; and 'very strong evidence' for public restrictions on smoking, or on diets of fats and sugars. 'If we are wrong in deducing causation from association no great harm or injustice (in these cases) will be done'.

Societies regularly use different strengths of evidence for different purposes. For example, a high level is used in criminal courts, where the costs of a wrong conviction fall heavily on the innocent victim; while a lower level is used in civil courts in compensation cases, where the costs of failing to compensate the victims of negligence fall on already injured and usually financially constrained people. In both cases, the costs of being wrong in the other direction, that is, failing to convict a guilty person, and awarding compensation to someone who was not the victim of negligence, fall on the broader shoulders of society, or of insurance companies, and are, therefore, deemed by most people to be more acceptable.

Choosing an appropriate strength of evidence for specific cases of potential hazards necessarily involves these kinds of trade offs between the consequences of being wrong in 'establishing' or 'not establishing' cause and effect.

The most well-known classification of strengths of evidence for public health purposes is that of the International Agency for Research on Cancer (IARC). The IARC uses four main categories to characterize the strengths of evidence for carcinogens: from 'human', to 'probable', 'possible' and 'unlikely' carcinogens. In the field of air pollution, the World Health Organization (WHO) has used four categories to characterize strength of evidence: evidence sufficient to infer causality; evidence suggestive of causality; evidence insufficient to infer causality; and evidence showing no association [49]. Within ecological sciences such as climate change both the strength of evidence, and how it is moving in time with advances in scientific knowledge, have also been explicitly addressed.

For example, the United Nations Intergovernmental Panel on Climate Change (IPCC) concluded in 2001 that on the 'balance of evidence' (but not 'beyond all reasonable doubt'), mankind was having a discernible influence on global climate. Many scientists, policymakers and businesses regarded this as sufficient evidence for action on greenhouse gases. The evidence has, for most hypotheses, significantly strengthened since then [50].

Table 3 summarizes five of the IPCC seven strengths of evidence. I have illustrated their use for different purposes in US and EU societies. It would be useful if other areas of the environmental and health sciences, such as those concerning endocrine-disrupting substances, developed similar schemes for classifying strengths of evidence. In addition, there needs to be explicit agreement on the rules for choosing between them: 'weight of evidence', for example, is usually either not defined or defined very differently by different users [51]. A novel approach to the controversial evidence on electromagnetic fields used a qualitative Bayesian approach and more transparent rules for estimating degrees of confidence in the evidence, ideas that could perhaps be taken up more widely [52].

In the three examples above, from IARC, WHO and IPCC, the different strengths of evidence are scientific judgements that are made without regard for different purposes, consequences or contexts. Bradford Hill would have agreed with this: 'on scientific grounds . . . the evidence is there to be judged on its merits and the judgement . . . should be utterly independent of what hangs upon it – or who hangs because of it'.

However, he also noted that 'in real life we shall have to consider what flows from that decision . . . In occupational

Table 3.

Different strengths of evidence for different purposes: some examples and illustrations.

Strength of evidence	Illustrative terms	Examples of use
Very strong (90–99%)	• Statistical significance	• Part of strong scientific evidence of 'causation'
	• Beyond all reasonable doubt	• Most criminal law, and the Swedish Chemical Law 1973, for evidence of 'safety' of substances under suspicion-burden of proof on manufacturers
Strong (65–90%)	• Reasonably certain	• Food Quality Protection Act, 1996 (USA)
	• Sufficient scientific evidence	• To justify a trade restriction designed to protect human, animal or plant health under World Trade Organization Sanitary and Phytosanitary Agreement, Art. 2, 1995
Moderate (33–65%)	• Balance of evidence	• Intergovernmental Panel on Climate Change 1995 and 2001
	• Balance of probabilities	• Much civil and some administrative law
	• Reasonable grounds of concern	• European Commission Communication on the Precautionary Principle 2000
	• Strong possibility	• British Nuclear Fuels occupational radiation compensation scheme, 1984 (20–50% probabilities triggering different awards up to 50% + which triggers full compensation)
Weak (10–33%)	• Scientific suspicion of risk	• Swedish Chemical Law 1973, for sufficient evidence to take precautionary action on potential harm from substances-burden of proof on regulators
	• Available pertinent information	• To justify a provisional trade restriction under World Trade Organization Sanitary and Phytosanitary Agreement, Art. 5.7, where 'scientific information is insufficient'
Very weak (1–10%)	• Low risk	• Household fire insurance
	• Negligible and insignificant	• Food Quality Protection Act, 1996 (USA)

medicine our object is usually to take action . . . to intervene to abolish or reduce death or disease'. In other words, in real life, the separation of 'risk assessment' and 'risk management' is artificial and unreal, and differential strengths of evidence, depending on purposes and consequences, are necessary.

The case for action

Bradford Hill ended his article with the 'case for action': 'All scientific work is incomplete – whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not confer on us a freedom to ignore the knowledge we already have, or to postpone the action that it appears to demand at a given time'.

Today's knowledge is often seen as static, with just a few troublesome gaps in knowledge that further research will remove. Such 'further research' can then become an excuse to postpone precautionary, or even preventative, actions.

Figure 1 illustrates the dynamic nature of knowledge, where today's certainties can become tomorrow's mistakes

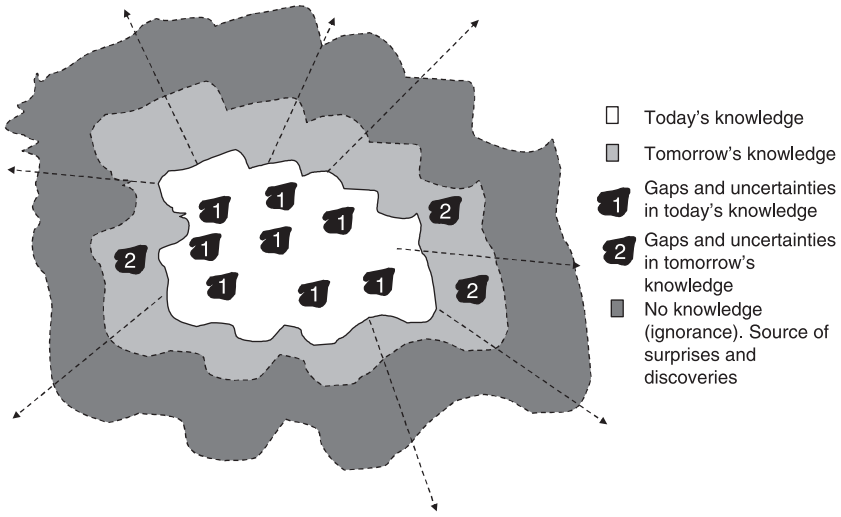

Fig. 1. Knowing and not knowing: a dynamic expansion.

Table 4.

Some examples of terms often used to characterise varying strengths of evidence for cause and effect relationships.

Effects are mostly obscure
Suspicion that
Seems to be linked to
Tempting to suggest that
Might be associated with
Implies that
Increasing evidence that
Points to their origin in
Considered to be associated with
Has been shown to contribute to
Exposure/outcome associations observed
A statistical association has been observed
The association is not causal
Substantially contributed to
The association is causal
No overall evidence
No convincing evidence
Not enough evidence

Table 5.

A proposal for consistent cause and effect terminology linked to different strengths of evidence.

Terminology	Strength of evidence
Causally linked to	Very strong (>95%)
Strongly associated with	Strong (65–95%)
Associated with	Moderate (35–65%)
Little evidence that*	Weak (10–35%)
Unlikely to be*	Very weak (<10%)

*As 'no evidence of harm' is not the same as 'evidence of no harm' these statements need to refer to the relevant research base.
The probability bands are illustrative only.
European Environmental Agency; stimulated by Bradford Hill; Intergovernmental Panel on Climate Change (IPCC) report, Summary for Policymaker's, 2007 Working Group I; and Guidance Notes for Lead Authors of IPCC 4th Assessment on Addressing Uncertainties, IPCC July 2005.

and where new uncertainties, knowledge gaps and areas of ignorance open up as others are closed down. It follows that waiting for 'full' risk assessments, or for the elimination of uncertainties, before actions to prevent harm is unrealistic.

The decision about when there is a sufficiency of evidence to justify preventive action clearly involves more inputs to decision-making than from science alone. The strength of evidence that is deemed appropriate depends on such non-scientific criteria as the costs of being wrong with actions or inactions (including their nature and distribution between different groups and generations); the justification for, and benefits of, the agents or activities that pose potential threats to health; and the availability of feasible alternatives.

There is an obvious need for stakeholder participation in taking such value laden decisions. However, a necessary condition for good communications between stakeholders is the use of clear and consistent terms to characterize evidence about cause and effect. Unfortunately, this is often not the case.

Table 4 illustrates the wide variations in terminology used in summaries of scientific evidence about causality in environmental health. There is clearly scope for improvement.

Table 5, inspired by the IPCC and IARC work, provides some examples of concepts and terminology that, if widely adopted, would help improve dialogue between stakeholders.

It would be helpful if the EU agencies responsible for environment, food, pharmaceuticals, chemicals, occupational health and infectious diseases could agree to use common, transparent and consistent concepts and terminology on cause/effect relationships. Similarly, the rules about how stocks of scientific evidence get translated into conclusions about strengths of evidence, via expert judgement, need to become more transparent and more consistently applied across the different risk domains addressed by the EU agencies.

Conclusions

Over the last 10 years, the evidence on the potential for some chemicals and other environmental stressors, usually in combination, to produce reproductive and developmental harm has increased, particularly in laboratory animals and wildlife. At the same time, evidence from human beings is beginning to accumulate, along with knowledge about mixtures of substances and about environmentally low, but sometimes harmful, doses. Exposure to these stressors via consumer products, water and food, is widespread. Meanwhile, the types of harm that one would expect to see from reproductive and development risks, such as breast, testicular and prostate cancers, reproductive problems such as some birth defects, low birth rates, infertility and early puberty, and some neurodevelopmental disorders, are generally increasing, particularly in Europe and the USA.

The evidence linking these particular disorders with specific endocrine-disrupting substances and other environmental stressors is, overall, not very strong. However, this is to be expected from the use of current scientific methods on complex, multicausal and often reciprocal systems and disease processes.

These are difficult circumstances for policymakers. However, the case for some precautionary action to reduce exposures is compelling, given widespread exposures, particularly to vulnerable groups such as foetuses and children. It invites the judicious and case specific use of the precautionary principle in order to help ensure that the timing of preventive action, and its cost-effectiveness, both in terms of quantitative and qualitative costs, is optimal from the point of view of society as a whole.

Unfortunately, there is no widely acceptable definition of the precautionary principle that indicates how it could be implemented and in what sorts of circumstances. In order to improve understanding, communication and debates on the principle the European Environment Agency has produced a working definition that is proving to be useful in clarifying the main issues:

> The Precautionary Principle provides justification for public policy actions in situations of scientific complexity, uncertainty and ignorance, where there may be a need to act in order to avoid, or reduce, potentially serious or irreversible threats to health or the environment, using an appropriate level of scientific evidence, and taking into account the likely pros and cons of action and inaction.

The limitations of scientific knowledge imply moral courage in taking precautionary action in time to avert harm. As Lewontin has observed:

> Saying that our lives are the consequence of a complex and variable interaction between internal and external causes does not concentrate the mind nearly so well as a simplistic claim; nor does it promise anything in the way of relief for individual and social miseries. It takes a certain moral courage to accept the message of scientific ignorance and all that it implies. [53]

Multi-causality implies that bolder and more timely precautionary measures are needed without waiting for high levels or proof of causality.

Mistakes will be made, surprises will occur. But if the quality of the scientific and stakeholder processes used to arrive at decisions such are sound, and the best of science is used, then living with the consequences of such decisions will be more acceptable. Earlier prevention will especially help those who, being most exposed, are likely to bear the health costs of poor timing in the avoidance of potential harm from reproductive and developmental hazards.

Acknowledgements

John Bailar; EEA colleagues Pernille Folkmann, Dorota Jarosinska, Gabi Schoning, Markus Ehhard; and three anonymous peer reviewers whose comments contributed significantly to the structure and argumentation of the article.

References

1 European Commission. Weybridge Report, European Workshop on the Impact of Endocrine Disrupters on Human Health and Wildlife. European Commission, EUR 17549, Brussels, Belgium, 1997.
2 World Health Organization/International Programme on Chemical Safety. Global Assessment of the State-of-the-Science of Endocrine Disruptors. WHO, Geneva, Switzerland, 2002, http://www.who.int/ipcs/publications/new_issues/endocrine_disruptors/en/
3 Segner H. Developmental, reproductive, and demographic alterations in aquatic wildlife: establishing causality between exposure to endocrine-active compounds (EACs) and effects. Acta Hydrochim Hydrobiol 2005;**33**:17–26.
4 CREDO; http://www.credocluster.info.
5 Schwartz DA, Korach KS. Emerging research on endocrine disruptors, NIEHS director's perspective. Environ Health Perspect 2007;**115**:A13.
6 EDEN Endocrine Research, The Prague Declaration; 2006, http://www.edenresearch.info/declaration.
7 Skakkebæk NE, Andersson AM. Endocrine Disrupters and Consumer products: Possible Effects on Human Populations. Abstracts from the 4th Copenhagen Workshop on Endocrine Disrupters, University Department of Growth and Reproduction, Rigshospitalet, Copenhagen, Denmark, 2007.
8 The Faroes Statement: Human Health Effects of Developmental Exposure to Chemicals in Our Environment. Tórshavn, Faroe Islands, Thursday, 24 May 2007. Basic Clin Pharm Toxicol 2008;**102**:73–75.
9 Thalidomide Victims Association of Canada, 'The History of Thalidomide', Dr. Widukind Lenz. http://www.thalidomide.ca/en/information/history_of_thalidomide.html/ September 2007.
10 Dubos R. Man Adapting. Yale University Press, New Haven, CT, 1965;164–65.
11 Duncan DF. Epidemiology. Macmillan, New York, 1988.
12 Sing CF, Stengard JH, Kardi SLR. Dynamic relationships between the genome and exposures to environments as causes of common human diseases. In: Simopoulos AP, Ordovas JM (eds). Nutrigenetics and Nutrigenomics World Review of Nutrition and Diet. Karger, Basel, Switzerland, 2004;77–91.
13 Rosen G. A History of Public Health. JHU Press, New York, 1958.
14 Gordon D. Tenacious assumptions in Western medicine. In: Lock M, Gordon D (eds). Biomedicine Examined. Kluwer, Dordrecht, The Netherlands, 1988.
15 Baird PA. Identification of Genetic Susceptibility to Common Diseases: the case for regulation. Perspect Biol Med 2002;**45**:516–28.
16 Rothman KJ. Introduction to Epidemiology. Oxford University Press, Oxford, UK, 2002.
17 Collins FF, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. Nature 2003;**422**:835–47.
18 Edwards TM, Myers JP. Environmental exposures and gene regulation in disease etiology. Environ Health Perspect 2007;**115**:1264–70.
19 Hopkins PN, Williams RR. A survey of 246 suggested coronary risk factors. Atherosclerosis 1981;**40**:1–52.
20 Weinstein B. Addiction to oncogenes – the achilles heal of cancer. Science 2002;**297**:63–4.
21 Dean K. Integrating theory and methods in population health research. In: Dean K (ed.). Population Health Research: Linking Theory and Methods. SAGE Publications, London, 1993.
22 Soto AM, Sonnenschein C. Emergentism as a default: cancer as a problem of tissue organisation. Bioscience 2005;**30**:101–16.
23 Gee D. Late lessons from early warnings: towards realism and precaution with endocrine disrupting substances. Environ Health Perspect 2006;**114**:152–60.
24 May R. Simple mathematical models with very complicated dynamics. Nature 1970;**261**:459–67.
25 Khaimovich L. Recent Developments in Nonlinear Dynamics: Unfolding the Meaning of Sociologically Relevant Concepts. Paper prepared for the 1992 American Statistical Association Meeting, 1992.
26 Di Guiliio RT, Benson WH. Interconnections between Human Health and Ecological Integrity. SETAC Press, Brussels, Belgium, 2002.
27 Guba G, Lincoln YS. Fourth Generation Evaluation. Sage Publications, Newbury Park, CA, 1989.
28 Cory-Slechta D. Studying toxicants as single chemicals: does this strategy adequately identify neurotoxic risk? Neurotoxicology 2005;**26**:491–510.
29 Weiss B, Bellinger DC. Social ecology of children's vulnerability to environmental pollutants. Environ Health Perspect 2006;**114**:1479–85.
30 Bellinger D. Lead neurotoxicity in children: decomposing the variability in dose–effect relationships. Am J Ind Med 2007;**50**:720–8.
31 Rauh VA, Parker FL, Garfinkel RS. Biological, social, and community influences on third-grade reading levels of minority Head start children: a multilevel approach. J Comp Psychol 2003;**31**:255–78.
32 Soto AM, Sonnenschein C. Emergentism by default: a view from the bench. Synthese 2006;**151**:361–76.

33 Newman MC. Population Ecotoxicology. Wiley & Son, Chichester, UK, 2001.
34 Kidd AK, Blanchfield PJ, Mills KH et al. Collapse of a Fish Population after Exposure to a Synthetic Estrogen. Proceedings of the National Academy of Sciences, http://www.pnas.org/cgi/doi/10.1073/pnas.0609568104 21 May 2007.
35 Bellinger D, Leviton A, Waternaux C, Allred E. Methodological issues in modelling the relationship between low-level lead exposure and infant development: examples from the Boston lead study. Environ Res 1985;**38**:119–29.
36 McMichael AJ. Prisoners of the proximate: loosening the constraints on epidemiology in an age of change. Am J Epidemiol 1999;**149**:887–8.
37 Hill B. Environment and disease: association or causation? Proc R Soc Med 1965;**58**:295–300.
38 Ashby J, Houthoff E, Kennedy SJ et al. The challenge posed by endocrine-disrupting chemicals. Environ Health Perspect 1997;**105**:164–9.
39 European Environmental Agency. Association and Causation: the Bradford Hill Criteria Revisited. European Environment Agency, Copenhagen, Denmark, in press.
40 Needleman H. Making models of real world events: the use and abuse of inference. Neurotoxicol Teratol 1995;**17**:241–2.
41 Bailar J. How to distort the scientific record without actually lying: truth and the arts of science. Eur J Oncol 2007;**11**:217–24.
42 Grosse Y, Baan R, Straif K et al. Carcinogenicity of 1,3-butadiene, ethylene oxide, vinyl chloride, vinyl fluoride, and vinyl bromide. Lancet Oncol 2007;**8**:679–80. http://oncology.thelancet.com 29 September 2007.
43 Poole, C. Low P-values or narrow confidence intervals: which are more durable? Epidemiology 2000;**12**:291–4.
44 European Environmental Agency. Late Lessons from Early Warnings: the Precautionary Principle 1896–2000. European Environment Agency, Copenhagen, Denmark, 2001.
45 Grandjean P. Underestimation of risk due to exposure misclassification. Int J Occup Med Environ Health 2004;**17**:131–6.
46 Grandjean P. Non-precautionary aspects of toxicology. Toxicol Appl Pharmacol 2005;**207**:652–7.
47 Weed DL. Precaution, prevention and public health ethics. J Med Philos 2004;**29**:313–32.
48 Case T-13/99, Pfizer Animal Health v. Council, 2002, ECRII-3305.
49 World Health Organisation. Effects of Air Pollution on Children's Health and Development. A Review of the Evidence. WHO, European Office, Bonn, Germany, 2005.
50 Intergovernmental Panel on Climate Change. Summary for policy makers. Climate Change: the Physical Science Basis. Contribution of Working Group 1 to the 4th Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK, 2007.
51 Weed, DL. Weight of evidence: a review of concept and methods. Risk Anal 2005;**25**:1545–57.
52 Neutra R, DelPizzo P, Geraldine M. An evaluation of the possible risks from electric and magnetic fields (EMFs) from power lines, internal wiring, electrical occupations, and appliances. California EMF Programme, Oakland, CA, 2002.
53 Orrell D. The Future of everything: the Science of Prediction. Lewontin in Thunder's Mouth Press, New York, 2007;379.

