

- Flegal, K. M., Graubard, B. I., Williamson, D. F., and Gail, M. H. (2005). Excess deaths associated with overweight, underweight, and obesity. *Jama*. Am Med Assoc 2005 Apr 20; 293(15): 186, 1–7.
- Gallison, P. (1987). *How Experiments End*. Chicago: Chicago University Press.
- Glass, D. C., Gray, C. N., Jolley, D. J., Gibbons, C., and Sim, M. R. (2006). *The Story So Far. Annals of the New York Academy of Sciences, 1076* (1 Living in a Chemical World: Framing the Future in Light of the Past), pp. 80–89.
- Hafeman, D. M., and Schwartz, S. (2009). Opening the black box: A motivation for the assessment of mediation. *Int J Epidemiol*. 38(3): 838–845.
- Hedström, P. and Swedberg, R. (1998). *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge: Cambridge University Press.
- Jensen, C. D., Block, G., Buffer, P., Ma, X., Selvin, S., and Month, S. (2004). Maternal dietary risk factors in childhood acute lymphoblastic leukemia (united states). *Cancer Causes and Control*, 15(6): 559–570.
- Kincaid, H. (2002). Scientific Realism and the Empirical Nature of Methodology. In S. Clarke and T. Lyons, *Recent Themes in the Philosophy of Science* (Dordrecht: Kluwer, 2002), pp. 39–62.
- Kincaid, H. (2008). Do we need theory to study disease? *Perspectives in Biology and Medicine*, 51(3): 367–378.
- Kuhn, T. S. (1977). *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago: University of Chicago Press.
- Levine, B. J. (2008). The other causality question: Estimating attributable fractions for obesity as a cause of mortality. *International Journal of Obesity*, Aug 2008 Supplement 3, Vol. 32, p. S4–S7, 4p.
- Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton: Princeton University Press.
- Morgan, S. L., and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.
- Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.
- Pearson, K. (1900). *The Grammar of Science*. Cambridge: MIT Press.
- Quine, W. V., and Ullian, J. S. (1970). *The Web of Belief*. New York: Random House.
- Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern Epidemiology*. Philadelphia: Lippincott Williams & Wilkins.
- Russo, F. and Williamson, J. (2007). Interpreting probability in causal models for cancer. In F. Russo and J. Williamson (eds), *Causality and Probability in the Sciences*. College Publications, pp. 217–242.
- Schisterman, E. F., Cole, S. R., and Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20(4): 488.
- Sober, E. (1988). *Reconstructing the Past: Parsimony, Evolution, and Inference*. Cambridge: MIT Press.
- VanderWeele, T. J. and Robins, J. M. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology*. 18(5): 561–568.
- Wilson, M. (2006). *Wandering Significance: An Essay on Conceptual Behavior*. New York: Oxford University Press.

5.

The IARC and mechanistic evidence

Bert Leuridan and Erik Weber

Abstract

The International Agency for Research on Cancer (IARC) is an organization which seeks to identify the causes of human cancer. For each agent, such as betel quid or Human Papillomaviruses, they review the available evidence deriving from epidemiological studies, animal experiments and information about mechanisms (and other data). The evidence of the different groups is combined such that an overall assessment of the carcinogenicity of the agent in question is obtained.

In this paper, we critically review the IARC's carcinogenicity evaluations. First we show that serious objections can be raised against their criteria and procedures – more specifically regarding the role of mechanistic knowledge in establishing causal claims. Our arguments are based on the problem of confounders, of the assessment of the temporal stability of carcinogenic relations, and of the extrapolation from animal experiments. Then we address a very important question, viz. how we should treat the carcinogenicity evaluations that were based on the current procedures. After showing that this question is important, we argue that an overall dismissal of the current evaluations would be too radical. Instead, we argue in favour of a stepwise re-evaluation of the current findings.

5.1 Introduction

The IARC, the International Agency for Research on Cancer, is a division of the World Health Organization. In the Preamble to the IARC *Monographs* we read:¹

Through the *Monographs* programme, IARC seeks to identify the causes of human cancer. (IARC 2006, p. 1)

¹ The Preamble to the IARC *Monographs* can be found at the beginning of each *Monograph* (for the current preamble, see e.g. in IARC 2007, p. 9–31). A slightly different version of the current preamble can be found on the web (IARC 2006). Throughout the history of the IARC, in the past four decades, the procedures and the criteria that are listed in the Preamble have repeatedly been changed (Hopkins 1994, pp. 194–196; Cogliano 2006).

More specifically, the objective of the programme is ...

... to prepare, with the help of international Working Groups of experts, and to publish in the form of *Monographs*, critical reviews and evaluations of evidence on the carcinogenicity of a wide range of human exposures. (IARC 2006, p. 2)

The term 'agent' is used to refer to 'any entity or circumstance that is subject to evaluation in a *Monograph*' (IARC 2006, p. 2). The exposures or agents include individual chemicals, but also ...

... groups of related chemicals, complex mixtures, occupational exposures, physical and biological agents and lifestyle factors. (IARC 2006, p. 1)

How is the carcinogenic risk of exposures assessed by the IARC?² The available evidence consists of epidemiological studies (field experiments with humans), animal experiments and information about mechanisms (and other data). The available studies are first evaluated separately. Then their conclusions are combined per group. Finally, the evidence of the different groups is combined into one final assessment.

In this chapter, we will critically review the IARC's carcinogenicity evaluations. In the first part, we will briefly present their procedures and criteria (Section 5.2) and show that serious objections can be raised against them – more specifically regarding the role of mechanistic knowledge in establishing causal claims (Sections 5.3–5.5). This means we will only focus on the evidential role of mechanisms, not on their possible explanatory roles (cf. Glennan 2002, Bechtel and Abrahamsen 2005). Then we will address a very important question, viz. how we should treat the conclusions of all *Monographs* that were based on the current procedures (Section 5.6). We will show that this question is important (given the possible economic and social consequences of the IARC assessments), but that an overall dismissal of the current evaluations would be too radical. Instead, we argue in favour of a stepwise re-evaluation of the current findings.

5.2 Relevant features of the IARC procedures and criteria

The IARC's evaluation procedure consists of three phases and involves three kinds of studies: epidemiological studies, experimental studies on animals, and mechanistic information and other data. (We will often use the labels 'epidemiological', 'experimental' and 'mechanistic' to refer to these respective

² Despite the title of the *Monographs*, the goal of the IARC is to evaluate cancer hazard (i.e. the potential of agents to cause cancer), not cancer risk (i.e. the probability that this potential is realized for a particular section of the population in a defined set of circumstances). (Hopkins 1994, pp. 193–194)

kinds of studies or evidence.) In the first phase, all studies are evaluated separately. In the second phase, assessments are made of the epidemiological, the experimental and the mechanistic group respectively. In the third phase, the evidence of the different groups is combined such that an overall assessment of the carcinogenicity of the agent in question is obtained.

1. Let us first look at the epidemiological studies. For ethical reasons, these studies with humans are almost without exception prospective or retrospective; randomized experiments are very rare. In the first phase, each study is assessed according to three criteria: viz. whether they are plagued by *bias*, *confounding* or *chance*. Bias is defined as

... the operation of factors in study design or execution that lead erroneously to a stronger or weaker association than in fact exists between an agent and disease. (IARC 2006, p. 9)

In order to exclude bias it is required that

... the study population, disease (or diseases) and exposure should have been well defined by the authors. Cases of disease in the study population should have been identified in a way that was independent of the exposure of interest, and exposure should have been assessed in a way that was not related to disease status. (IARC 2006, p. 9)

Confounding occurs when

... the relationship with disease is made to appear stronger or to appear weaker than it truly is as a result of an association between the apparent causal factor and another factor that is associated with either an increase or decrease in the incidence of the disease. (IARC 2006, p. 9)

In order to rule out confounding it is required that

... the authors should have taken into account – in the study design and analysis – other variables that can influence the risk of disease and may have been related to the exposure of interest. Potential confounding by such variables should have been dealt with either in the design of the study, such as by matching, or in the analysis, by statistical adjustment. (IARC 2006, p. 9)

In order to exclude chance the authors must report the basic data on which their conclusions are based, but also their statistical methods:

Finally, the statistical methods used to obtain estimates of relative risk, absolute rates of cancer, confidence intervals and significance tests, and to adjust for confounding should have been clearly stated by the authors. (IARC 2006, p. 10)

Studies that score badly on these criteria have a low credibility, so their weight in the final evaluation is very low.

After the individual screening, the epidemiological studies are compared with each other. The aim of this second phase is to arrive at one of the following conclusions (IARC 2006, pp. 19–20):

- (1) There is *sufficient* epidemiological evidence of carcinogenicity.
- (2) There is *limited* epidemiological evidence of carcinogenicity.
- (3) The epidemiological evidence of carcinogenicity is *inadequate*.
- (4) There is epidemiological evidence suggesting *lack* of carcinogenicity.

Conclusion (1) is drawn if 'a positive relationship has been observed between exposure and cancer in studies in which chance, bias and confounding could be ruled out with reasonable confidence' (IARC 2006, p. 19). Conclusion (2) is drawn if a positive association is observed for which a causal interpretation is credible, but chance, bias or confounding could not be ruled out with reasonable confidence. Conclusion (3) is drawn if there are no studies available, or if the available studies are of insufficient quality or consistency (for the first two conclusions it is required that the positive association occurs in a large majority of the studies). Conclusion (4) is drawn if there are several adequate studies which consistently show no positive association.

2. The experiments with animals are also screened individually in the first phase. One of the considerations is of course whether the animals were allocated randomly to the experimental or the control group: if that condition is not satisfied, the main possible advantage of animal experiments (viz. that they can be randomized trials) is not exploited. Another consideration is whether both male and female animals were used (this prevents a possible *bias*). And of course the data (number of animals studied, number of tumours, length of survival, etc.) should be reported and analysed adequately (elimination of *chance*).

After the individual screening, the results of animal experiments are combined in the second phase. The possible conclusions are (IARC 2006, pp. 20–21):

- (1) There is *sufficient* evidence of carcinogenicity in experimental animals.
- (2) There is *limited* evidence of carcinogenicity in experimental animals.
- (3) The evidence of carcinogenicity in experimental animals is *inadequate*.
- (4) There is evidence suggesting *lack* of carcinogenicity in experimental animals.

Conclusion (1) is drawn if there are high quality studies (randomised, elimination of chance) for two or more species, consistently showing an increased incidence of tumours. Increased incidence in a well-conducted study in both sexes of a single species, can also provide sufficient evidence. Conclusion (2) is drawn if the data suggest a carcinogenic effect but are limited for making a definitive evaluation (e.g. because only one sex of a single species is investigated). The criteria for conclusions (3) and (4) are similar to those for epidemiological studies.

3. The mechanistic data³ include information about toxicokinetics (absorption, distribution, metabolism, and elimination of agents) and mechanisms of carcinogenesis (How does the agent affect the organs, tissues or cells? Does it e.g. lead to genetic mutations?). For carcinogenic effects that have been observed in experimental animals, an evaluation is made of the strength of the evidence that it is due to a particular mechanism. The second-phase categories that are used here are 'weak', 'moderate' and 'strong', but these labels are less clearly defined than those of the epidemiological and experimental studies (IARC 2006, pp. 21–22). For instance, experimental studies which show that suppressing key elements of a mechanism prevents the development of tumours, provide strong evidence for the conclusion that the mechanism operates in the type of experimental animal that is studied. There is also an assessment of how likely it is that a particular mechanism operates in humans. And much attention is paid to the questions whether 'multiple mechanisms might contribute to tumour development, whether different mechanisms might operate in different dose ranges, whether separate mechanisms might operate in humans and experimental animals and whether a unique mechanism might operate in a susceptible group'. (IARC 2006, p. 21)
4. Finally, in the third phase, the three types of evidence are brought together. The agent under investigation is put into one of the following groups (IARC 2006, pp. 22–23):⁴

Group 1: The agent is *carcinogenic to humans*.

Group 2A: The agent is *probably carcinogenic to humans*.

Group 2B: The agent is *possibly carcinogenic to humans*.

Group 3: The agent is *not classifiable as to its carcinogenicity to humans*.

Group 4: The agent is *probably not carcinogenic to humans*.

The Preamble presents a set of rules governing this overall assessment. It is important to note that these are not treated as rigorous rules and that past decisions have resulted in apparent exceptions (as is indicated by formulations starting with 'exceptionally' or 'in some cases'). We will not list all exceptional rules. Thus the following presentation is somewhat simplified.

An agent is placed in Group 1 if there is *sufficient* epidemiological evidence of carcinogenicity. Exceptionally, an agent may also be labelled carcinogenic

³ Originally, mechanistic data were not taken into account. In 1982, the first kind of mechanistic information (viz. genotoxicity evaluation) was taken into account, and in 1991 several other types of mechanistic data (e.g. concerning gene-expression) were incorporated in the IARC criteria (Hopkins 1994, pp. 194–195). In the current preamble (i.e. IARC 2006), even more weight is attached to mechanistic information (Cogliano 2006).

⁴ At this moment, nearly 1,000 agents have been classified. Of these, 107 are in Group 1, 58 are in Group 2A, 249 are in Group 2B, 512 are in Group 3 and, finally, 1 agent (Caprolactam) is in Group 4.

to humans if there is *sufficient* experimental evidence and *strong* mechanistic evidence. But normally, if the epidemiological evidence is less than *sufficient*, it is combined with the evidence from experimental animals and/or with the mechanistic evidence and results in a classification lower than Group 1. For instance, an agent is classified in Group 2A in the following cases:

- (a) If there is *limited* epidemiological evidence and *sufficient* experimental evidence.
- (b) (In some cases) if there is *inadequate* epidemiological evidence, but *sufficient* experimental evidence and *strong* evidence that the carcinogenesis is mediated by a mechanism that also operates in humans.
- (c) (Exceptionally) if there merely is *limited* epidemiological evidence of carcinogenicity in humans.

In condition (b) the mechanistic evidence is used to warrant extrapolation from animals to humans. If this warrant is absent – case (a) – stronger epidemiological evidence is required than in cases where there is such a warrant.

This role of mechanistic evidence, which relates to the extrapolation from animal experiments to humans, can be further clarified by means of the difference between the following rules:

- (d) If there is *inadequate* epidemiological evidence and *sufficient* experimental evidence, but *strong* evidence that the mechanism of carcinogenicity in experimental animals *does not* operate in humans, the agent is classified in Group 3.
- (e) If there is *inadequate* epidemiological evidence and *sufficient* experimental evidence, but no such negative mechanistic evidence, then the agent is classified in Group 2B.

5.3 Mechanisms and the problem of confounders

1. Biomedical scientists investigating the causes of diseases face a fundamental ethical problem. Randomized experiments with the target population (i.e. humans) provide the most reliable method for establishing causal relations in the biomedical sciences:

A decisive test of whether smoking causes heart disease, then, would be to take a large sample of human infants randomly selected from the human population, divide them into two equal groups, and force one group to smoke for the rest of their – no doubt abbreviated – lives. (Dupré 1993, pp. 202–203)

However, these randomized experiments are usually impossible for ethical reasons: they may cause physical harm to the experimental subjects, as in

Dupré's example. Biomedical scientists can avoid the unethical experiments by doing merely observational studies on humans (prospective or retrospective designs) and by doing randomised experiments with animals.⁵

From Section 5.2.4 it is clear that the IARC procedures do take into account the role that information about mechanisms can play in extrapolating results from animal experiments to humans. However, mechanisms can have at least two other evidential roles that are neglected in the IARC procedures. The first role relates to the problem of confounders and is discussed in this section. The second relates to extrapolation over time and is discussed in Section 5.4.

2. The problem of confounders originates from the fact that in a prospective or retrospective design the individuals 'put themselves' into the experimental or the control group by the way they act.⁶ For instance, in a prospective design set up to investigate the relation between smoking and heart disease, people that decided to smoke end up in the experimental group, non-smokers in the control group. Because of this non-random selection, there may be disturbing factors. For instance, if there are more heart diseases among the smokers, this may be due to the fact that both smoking and heart disease are positively influenced by coffee drinking. Randomized experiments avoid this problem by the random division into experimental and control group.

The standard solution to this problem is 'conditioning on potential confounders'. But this solution has its limitations, as Dan Steel points out with respect to the social sciences:

I agree that there are cases in which one can draw reasonable conclusions about what causes what without the aid of experiment or substantial knowledge of underlying mechanisms. However, the usefulness of conditioning on potential causes does not undermine the proposal that mechanisms significantly aid causal inference in the social sciences, since social scientists are rarely able to measure all potential common causes. Indeed, the inability to exhaustively consider all potential common causes is a basic element of the problem of confounders, to which mechanisms are being considered as a partial solution. (Steel 2004, p. 63)

This problem is not limited to the social sciences. Potential disturbing factors (confounders) can be eliminated by means of statistical methods on a one-by-one basis. But we can never be sure that no untested variables will ever turn out to be confounders, and we cannot test all possible variables. For instance,

⁵ As an anonymous referee rightly pointed out, randomized experimental designs also have other shortcomings than those cited above. For example, they can show us *that* two variables are causally linked but not *how* they are linked. It follows that even where randomized experimental designs are feasible, mechanistic information may still add to our knowledge.

⁶ This terminology should not be taken literally. It is not required that individuals are actively responsible for ending up in the experimental or the control group (due to their behaviour). Only where they end up does not depend on any manipulation by the researcher.

we can exclude the possibility that coffee drinking is a common cause in the above example, but we cannot be sure that there is no other variable which causes both smoking and heart disease and is responsible for the correlation. We cannot exclude the possibility that smoking and heart disease have a common cause; we can only test individual variables and exclude them as common causes. More generally, it seems impossible to rule out confounding 'with reasonable confidence' by means of conditioning alone.⁷

How can causal mechanisms help here? Steel (2004) distinguishes two possible roles for mechanisms relating to the problem of confounders. The first possible role is negative: if we don't find a plausible mechanism linking two variables, we can conclude that the correlation between them is spurious (i.e. there is a common cause). Steel argues that this negative role does not work, because we can always find plausible mechanisms. This argument is a bit too strong, however. At least it does not apply straightforwardly to the present context. In biomedical research, it does not suffice to just come up with a plausible mechanism. Mechanistic hypotheses have to be justified empirically. Moreover, the IARC itself sometimes explicitly uses strong evidence suggesting lack of a mechanism (and other relevant data), in tandem with inadequate epidemiological evidence and experimental evidence suggesting lack of carcinogenicity, as a reason to classify an agent in Group 4. (Yet this decision is not taken frivolously.)⁸ Hence we suggest that the negative role of mechanisms does bear on carcinogenicity studies (even if, as Steel argues, it does not work in the social sciences).

The second possible role is positive: if we find a mechanism for which we have good evidence of, we can conclude that there is a causal relation between the two correlated variables. In the case of carcinogenesis, the description of the mechanism would contain claims about how the presence of certain chemical substances (e.g. in the blood) leads to the presence of other chemical substances in cells and to changes in properties of cells (e.g. genetic mutation). These processes can be investigated *in vitro*. This is important, because *in vitro* it is possible to do randomized experiments, in which the problem of confounders is unlikely to occur. An ideal mechanistic argument for a claim about carcinogenicity (or other hazard) consists of a chain of lower-level causal

⁷ The picture is somewhat more complicated. In the epidemiological literature, two general methods for dealing with the problem of confounders in observational studies are distinguished. 'The first is to consider them in the design of the study by matching on the potential confounder or by restricting the sample to limited levels of the potential confounder. The other method is to evaluate confounding in the analysis by stratification [...] or by using multivariate analysis techniques such as multiple logistic regression.' (Greenberg *et al.* 2004, chapter 10, 'Confounding') (See also the quote from the IARC preamble in Section 5.2.1.) But these complications do not affect the main problem addressed by Dan Steel, a problem which is explicitly recognized in the epidemiological literature: 'Only known confounders can be addressed in observational research.' (Greenberg *et al.* 2004, chapter 10, 'Summary')

⁸ See also case (d) in Section 5.2.

claims of which each element is supported by a randomized experiment. Similarly, in the social sciences, an ideal mechanistic argument uses causal claims with respect to the behaviour of individuals which have been tested in randomized trials. Both in the social and in the biomedical sciences, the usefulness of mechanistic evidence 'relies upon causal relationships among components being more directly accessible than those at the macrolevel' (Steel 2008, p. 195).⁹

Looking back at Section 5.2.1 we see that this evidential role of mechanisms is largely neglected in the IARC procedures. It is assumed that one may attempt to exclude the possibility of confounding with reasonable confidence without invoking mechanisms. A sceptic, following Steel's line of reasoning, might argue that confounding can never be excluded with reasonable confidence in this way. So there never is *sufficient* epidemiological evidence for carcinogenicity: that is an empty category. This is an argument for suggesting a change in the IARC procedures. Mechanistic evidence should also be used to better exclude the possibility of confounding in individual epidemiological studies.

5.4 Mechanisms and temporal stability

Consider the following statements, that have an identical logical form:

No gold sphere has a mass greater than 100,000 kg.

No enriched uranium sphere has a mass greater than 100,000 kg.

The second statement is deemed temporally much more stable than the first. The critical mass for enriched uranium is just a few kilograms, so the second statement is not only true at this moment, but will remain true unless some principles that govern the universe change. The truth of the first statement seems to be more contingent (it just happens to be the case that no one did produce such a sphere yet). Examples of even less stable generalizations are 'All screws in Smith's car are rusty' and 'All coins in my pocket are made of copper'.

Likewise, probabilistic causal claims that are true at this moment can also differ with respect to their temporal stability. Consider first an example from

⁹ Let us briefly discuss some doubts raised by an anonymous referee who states that mechanistic evidence is as undetermined as any other and that it is trivial to formulate multiple, contradictory plausible mechanisms for any pathogenic process. We already stated that in biomedical research it does not suffice to come up with just a plausible mechanism. It may be trivial to formulate plausible mechanistic hypotheses, but it takes a lot of work to support them empirically. What we need to rule out confounders is a well-justified model of a mechanism, not just a *mechanism sketch*, i.e. a description of a (possible) mechanism containing missing pieces which we do not yet know how to fill in – cf. Machamer *et al.* (2000, p. 18).

the social sciences. In a book on ethical problems in the social sciences, Paul Davidson Reynolds discusses an experiment which investigates the effects of negative income tax (his source is Kershaw 1972):

The research involved the examination of the effects of different negative income tax plans (direct cash payments) to 'guarantee' a predetermined minimum household income: partial reductions in payments occurred if household earnings increased. The basic question was the extent of labor-force participation of individuals in households with a guaranteed income - i.e. would they work less? The study also estimated the costs of a guaranteed income program if adopted as the major welfare strategy for the nation. The initial study involved 1400 families in five cities in the New Jersey-Pennsylvania area randomly assigned to one of the eight plans (negative income tax schedules) or to a control group (families receiving no guaranteed income). (1982, p. 36)

The eight plans differed in the amount of money that was given if there was no other income, and in the reductions in payment that occurred when there was another income. But in each plan, the reductions were only partial. The aim of the study was to determine whether the advantages of a guaranteed income plan (administrative simplicity, dignity, equity, ...) were or were not outweighed by a possible disadvantage, viz. reduced labour-force participation.

The experiments reported by Reynolds were performed to evaluate guaranteed income as a nationwide welfare strategy. In the early 1970s (when these experiments were performed) the result was that the effect of guaranteed income on labour-force participation was small. Suppose now that the US Government would have taken this result as a basis for adopting negative income tax as the major welfare strategy. Then 35 years later they might have found out that the effect has changed. Causal relations can become weaker or stronger over the years. For instance, if people become less materialistic, they might attach more value to free time and less to extra consumption, so the effect of a guaranteed income on labour-force participation might increase. Causal relations can even be reversed (from positive to negative, or the other way around).

No matter what the nature of our evidence is (random experiments, prospective or retrospective designs) we face the challenge of extrapolating our results to the future. Without such extrapolation, the results have no policy relevance (see Section 5.6 for a more elaborate discussion of policy relevance). How far the extrapolation must go, depends on what we use the causal relation for. Since no government wants to change its welfare strategy fundamentally too often, extrapolation is required for quite a large period. Regardless of how far one should try to extrapolate, it is clear that extrapolation is impossible without insight into the stability of the underlying social mechanism. Once we know the mechanism, we can investigate how changes at the micro-level (people's beliefs, desires and individual decisions) may affect the macro-level

(the relation between negative income tax and labour-force participation). If there is a change at the micro-level that is likely to occur and that has an effect on the causal relation at the macro-level, extrapolation is a risky business. If such changes are unlikely, extrapolation is quite safe.

Let us now go back to the biomedical sciences. There are three ways in which causal generalizations can be unstable in time. First, evolution in the age structure of the population of interest may have an effect on the strength of a causal relation, or even result in new causal relations. Compare a population where everyone dies before the age of 80 with a population in which a substantial part reaches the age of 100 years. It is possible that a compound constitutes a hazard in the second population but not in the first simply because the effects of the compound are very slow and only manifest themselves above the age of 80.

Second, there is risk-instability. Consider the following example. Various risk factors of breast cancer are being explored. In a paper arguing for a causal link between exposure to electromagnetic fields and breast cancer, McElroy *et al.* (2007, p. 266) claim that about half of the variation in breast cancer rate is still unexplained by well explored risk factors such as ionizing radiation, abortion, alcohol consumption, hormone use, etc. That is why they investigate exposure to electromagnetic fields. However, there is also research into the effect of maternal diet on breast cancer. The hypothesis is that maternal diet may increase the risk of breast cancer by inducing changes in the foetus, which alter the susceptibility of the daughter to risk factors that occur later in her life, such as the ones mentioned above (see Hilakivi-Clarke and Clarke 2006, p. 340). This example shows that it is possible that our susceptibility to factors that initiate cancer can vary quite quickly, under the influence of changes in (maternal) diet. This is a typical example of risk-instability. Maternal diet may change quickly in a population. In general, it is possible that a compound does not constitute a hazard at time x (because everyone in the population has a certain property P which neutralizes the effect of the compound) while it does constitute a hazard at time y (because e.g. only half of the population has property P).

A third way in which causal relations can be unstable over time is mechanism instability.¹⁰ Ye *et al.* (2009) study the *evolutionary* mechanism of cancer progression (as opposed to molecular mechanisms). They note that a large number of molecular mechanisms and pathways are known that underlie tumorigenicity. However, no common molecular mechanism underlying all kinds of cancer is known. According to Ye *et al.* genome instability (more specifically, the increased frequency of non-clonal chromosome aberrations) is the common mechanism:

¹⁰ We thank one of the referees for drawing our attention to the difference between risk-instability and mechanism-instability.

Increasing evidence illustrates that the somatic evolution of cancer is similar to natural evolution with system stability mediated genetic heterogeneity playing a key role [...]. [...] An emerging genome-centric concept on cancer evolution states that overall genome level variation coupled with stochastic gene mutations serve as a driving force of cancer evolution by increasing the cell population diversity [...]. (2009, p. 288)

Genome level variation or instability raises population heterogeneity qua molecular mechanisms,¹¹ which in turn raises the probability of a specific pathway leading to cancer (this, together with natural selection at the somatic cell level, constitutes the evolutionary mechanism of cancer; 2009, p. 296). It may itself be caused by genetic, metabolic and environmental (cf. the agents reviewed by the IARC elements (2009, p. 295)).

Thus the picture is as follows: for some or other reason (e.g. the presence of some carcinogenic agent), genome level instability is induced. This raises the number of potential molecular pathways or mechanisms, some of which may lead to cancer. But these 'mechanisms are constantly changing during cancer evolution' (2009, p. 289), which means that different cells may contain different pathways. Here's the crux: if the hallmark of cancer progression is molecular heterogeneity, what reason would we have to presuppose these mechanisms remain stable over time in the human population?

Mechanism-instability is also present in the case of breast cancer susceptibility we used to illustrate risk-instability (cf. supra). Hilakivi-Clarke and de Assis write that '[a]lterations in the fetal hormonal environment, caused by either maternal diet or exposure to environmental factors with endocrine activities, can modify the epigenome, and these modifications are inherited in somatic daughter cells and maintained throughout life.' (2006, p. 340) The fetal hormonal environment induces changes to the mechanism underlying carcinogenesis.

These examples show that there is yet another reason for changing the IARC procedures. The stability over time of carcinogenicity evaluations should be explicitly addressed and mechanistic evidence should be included as much as possible in this assessment. Moreover, carcinogenicity conclusions that have little stability (or whose stability is unknown) should be re-evaluated more frequently than more stable conclusions.¹²

¹¹ The population here is composed of cells, not individuals.

¹² In the history of the IARC evaluations, certain agents have been re-evaluated in different *Monographs*. For example, the possible carcinogenicity of tobacco smoking has been evaluated in IARC (1986) and re-evaluated in IARC (1987) and in IARC (2004). Likewise, tobacco habits other than smoking have been evaluated in IARC (1985) and again in IARC (1987) and in IARC (2007). The most recent evaluations (2004, 2007) did not only rely on more recent studies, they also invoked the most recent criteria (cf. footnote 3).

5.5 Extrapolation from animals to humans

The use of mechanistic evidence in the IARC procedures is lacking in still another way. We mentioned that mechanistic data are used to guide the extrapolation of experimental evidence to humans. At the moment, however, this use is open to improvement.

In Section 5.2.4 we discussed the rules governing the attribution of agents to Group 2A. We found that an agent may be placed in this group if there is *limited* epidemiological evidence of carcinogenicity and *sufficient* evidence of carcinogenicity in experimental animals (this was labelled 'case (a)'). In some cases, an agent may also be placed in this group in case there is *inadequate* epidemiological evidence of carcinogenicity, *sufficient* evidence of carcinogenicity in experimental animals and *strong* evidence that the carcinogenesis is mediated by a mechanism that also operates in humans (this was labelled 'case (b)'). The difference between the cases (a) and (b) is remarkable. Once it is acknowledged that there is no *sufficient* epidemiological evidence available (i.e. when it is either *limited* or *inadequate*) and that hence we need to rely on experimental evidence in animals, we should, strictly speaking, be prepared to show that this experimental evidence is relevant to humans.¹³ But if mechanistic evidence is needed for extrapolation in case (b), why isn't it needed in case (a)? In short, we recommend that the use of mechanistic evidence in extrapolation is treated as consistently as possible in the IARC procedures. Whatever the experimental evidence in animals may be, its relevance for humans should be made clear. (Note that this role for mechanistic evidence is independent of its possible use to rule out confounders with reasonable confidence.)

5.6 The status of the current IARC conclusions

Let us briefly recapitulate the conclusions of the last three sections. First, we argued that epidemiological studies, given their non-experimental nature, may always fall victim to the problem of confounders. Therefore we suggested a change in the IARC procedures: mechanistic evidence should also be used to better exclude the possibility of confounding in individual epidemiological studies. Secondly, we showed that mechanistic evidence is needed in order to assess the temporal stability of causal claims – even in the biomedical sciences – and we argued that this was yet another reason to change the procedures of the IARC. Finally, we drew attention to the unequal role played by mechanistic evidence regarding the extrapolation of experimental evidence on laboratory animals to human beings.

¹³ The need for such a warrant becomes very clear if we realize that different animal species may suggest different causal relations. For instance, aflatoxin B₁ causes liver cancer in rats but not in mice (Steel 2008, p. 82).

These findings and recommendations raise an important question, viz. how should we treat the conclusions of all *Monographs* that were based on the current procedures? Should we dismiss them completely? Our answer to this question consists of three parts. First we further motivate the above question. *Prima facie*, there are good reasons to dismiss the current evaluations (Section 5.6.1). Then we will analyse the possible consequences of such a decision, leading to the conclusion that an overall dismissal of the current evaluations would be too radical (Section 5.6.2). Finally, we will argue for a stepwise re-evaluation of the current IARC conclusions (Section 5.6.3).

1. *Prima facie*, there are good reasons to dismiss the current evaluations. We may fear that the current procedures result in *false positives*: some chemical substances, biological agents, ... are declared carcinogenic while they are not. If there can be *sufficient* epidemiological evidence without any mechanistic backing (see Section 5.3), some agents may erroneously end up in Group 1 (see Section 5.2). And if positive experimental evidence on animals may be considered relevant for human beings without any mechanistic warrant, some agents may erroneously end up in Groups 2A or 2B. We may also fear that the carcinogenicity relations that were discovered in the old *Monographs* have changed (cf. the problem of stability over time we discussed in Section 5.4).¹⁴

False positives would be problematic given that the IARC conclusions serve as the basis for regulation and legislation in large parts of the world. As such, they indirectly have huge financial and economic consequences. (Note that the IARC itself does not directly engage in regulation or legislation, cf. *infra*.)

The *Monographs* are used by national and international authorities to make risk assessments, formulate decisions concerning preventive measures, provide effective cancer control programmes and decide among alternative options for public health decisions. (IARC 2006, p. 3)

For example, one of the tasks of the California Environmental Protection Agency (Cal/EPA) is to 'publish a list of chemicals known to the State of California to cause cancer, birth defects or other reproductive harm'.¹⁵ One of the reasons why a chemical may be listed is that the IARC (or a similar 'authoritative body') has identified it as carcinogenic (Coghiano *et al.* 2004, p. 1269). This list has direct regulatory consequences.

Proposition 65 imposes certain requirements that apply to chemicals that appear on this list. These requirements are designed to protect California's drinking water sources from contamination by these chemicals, to allow California consumers to make informed choices about the products they purchase, and to enable residents or

¹⁴ The problem of the stability of carcinogenicity claims may also result in false negatives.

¹⁵ Quoted from <http://www.calepa.ca.gov/publications/factsheets/1997/prop65fs.htm>

workers to take whatever action they deem appropriate to protect themselves from exposures to harmful chemicals.¹⁶

The legislation of the European Commission provides a second example. For example, in the Commission Directive 2009/2/EC of 15 January 2009 on the classification, packaging and labelling of dangerous substances, it is stated that attention should be paid to the outcome of future discussions within the IARC on the carcinogenicity of nickel substances.¹⁷

In the past, the use of the IARC's conclusions as a basis for regulation and legislation has been criticized by the industry. For example, twenty years ago Barnard *et al.* (1989, p. 85) condemned the fact that the then IARC procedures 'made no attempt to evaluate whether animal evidence is predictively relevant to human cancer risk'.¹⁸ Furthermore, they regretted that

[b]ecause of a misunderstanding of the limited scope of the analysis involved, the IARC [...] lists have recently been used as a basis for legislative and regulatory decisions. (Barnard *et al.* 1989, p. 81)

It should be noted, however, that the IARC itself does not take part in regulation or legislation. Quite the reverse, the preamble explicitly states that

The evaluations of IARC Working Groups are scientific, qualitative judgements on the evidence for or against carcinogenicity provided by the available data. These evaluations represent only one part of the body of information on which public health decisions may be based. Public health options vary from one situation to another and from country to country and relate to many factors, including different socioeconomic and national priorities. Therefore, *no recommendation is given with regard to regulation or legislation, which are the responsibility of individual governments or other international organizations.* (IARC 2006, p. 3, our emphasis)

That the IARC evaluations represent only part of the body of information on which public health decisions are based emerges in two ways. The first way is alluded to in the last quote: one agent (whatever is the group it is attributed to) may be treated differently in different countries. Secondly, the weight of the IARC's verdict on the carcinogenicity to humans of some agent X is not proportional to the issuing regulatory decisions and in many cases a wide range of possible decisions is open for consideration. Hence agents with the same IARC classification may be treated differently in one and the same country. For example, the sale and use of alcoholic beverages, which are carcinogenic to humans (Group 1) is permitted throughout the European Union (of course, in some countries they are more heavily taxed than in

¹⁶ Quoted from <http://www.calepa.ca.gov/publications/factsheets/1997/prop65fs.htm>

¹⁷ Commission Directive 2009/2/EC in *Official Journal of the European Union*, 16.1.2009, L11/7.

¹⁸ At the time of publication, the authors were all linked with chemical companies (Barnard with Cleary, Gottlieb, Steen and Hamilton; Moolenaar with Dow Chemical Co.; and Stevenson with Shell Chemical Co.). Barnard and Stevenson were also members of the American Industrial Health Council.

others). By contrast, the importation, supply and new use of asbestos (which are also in Group 1) is strictly prohibited throughout the European Union. In both cases (alcoholic beverages and asbestos), there are threats of serious damage or harm to human health (cancer!), and in both cases the same level of scientific certainty is attributed (Group 1). Yet strong precautionary measures are taken in the case of asbestos, but not in the case of alcoholic beverages.

To conclude, given the insufficient use of mechanistic evidence by the IARC, one may fear that the current procedures result in false positives. These may have huge economic and social consequences given that the IARC conclusions serve as the basis for regulation and legislation in large parts of the world (even though the IARC itself does not engage in regulation or legislation). It follows that the above question is important and that *prima facie* there are good reasons to dismiss the current evaluations.

2. However, a dismissal would be unjustified. In general, rejecting the best available knowledge solely because it is not the best possible knowledge is counterproductive (provided this best available knowledge is reasonably reliable). Scientific knowledge is rarely sought after for intellectual reasons only. It is aspired for its possible use: as a basis for policy. Although the use of the IARC conclusions in policy (regulation and legislation; cancer prevention – IARC 2006, p. 1) provides reasons to deem the flaws in the IARC procedures problematic, we will see that this very same use safeguards them from an overall dismissal. Given what is at stake (*viz.* the life and the quality of life of thousands of people), we should prefer false positives over total ignorance.

In Sections 5.3–5.5 we showed that the IARC procedures are open to improvement. We did not show that they are completely flawed. Quite the contrary, they incorporate several protocols to provide as sound a scientific basis for evaluation as possible.

Firstly, it is clear from Section 5.2 that the IARC bases its findings on a broad empirical basis, reviewing ‘all pertinent epidemiological studies and cancer bioassays in experimental animals’ (IARC 2006, p. 3) plus part of the mechanistic and other relevant data on the condition that they are ‘published or accepted for publication in the openly available scientific literature’ or stem from ‘government agency reports that are publicly available’ (IARC 2006, p. 4).

Secondly, all participants of the IARC working groups need be qualified and impartial. It is the working groups that are responsible for developing the *Monographs*. Their members are selected by IARC staff together with other experts (IARC 2006, p. 5). The goal of the IARC is to invite the best-qualified experts (Cogliano *et al.* 2004, p. 1273). Study summaries may not be written by or reviewed by someone associated with the study being considered (IARC 2006, p. 6). Potential participants also have to declare, in confidence,

any interests that could constitute a real, potential or apparent conflict of interest, with respect to his/her involvement in the meeting or work between (a) commercial entities and the participant personally, and (b) commercial entities and the administrative

unit with which the participant has an employment relationship. (quoted in Cogliano *et al.* 2004, p. 1273)

In line with the WHO procedures, an apparent conflict of interest exists when the expert’s objectivity could be questioned by others, even if the interest does not necessarily influence the expert (Cogliano *et al.* 2004, p. 1273).¹⁹

Finally, the working groups strive after consensus evaluation (or otherwise majority vote) and the working group members engage in peer-review.

IARC Working groups strive to achieve a consensus evaluation. Consensus reflects broad agreement among Working Group Members, but not necessarily unanimity. The chair may elect to poll Working Group Members to determine the diversity of scientific opinion on issues where consensus is not readily apparent. (IARC 2006, p. 6)

Together these protocols (broad empirical basis, qualified and impartial experts, and peer review and consensus) ensure that the current IARC conclusions are reasonably reliable for policy. For the sake of prudence, we should not opt for an overall dismissal of the current IARC evaluations and our methodological criticisms should not be used by industrial lobbies to undermine the role of the IARC as providing the scientific basis for regulation and legislation.

3. Instead of dismissal, we would argue for a stepwise re-evaluation. We should stick to the IARC conclusions unless and until they are contradicted by more recent *Monographs* and updates.²⁰ In this way, we do not lose the pragmatic value of the current body of knowledge. At the same time, we strive for increasingly reliable conclusions.

Here the industry can play a role. It can suggest which agents are eligible for re-evaluation. Yet such a suggestion cannot by itself undermine our confidence in the current findings. Until a re-evaluation is finished and published, the current conclusions remain our best available knowledge and should serve as a basis for policy.

5.7 Conclusions

The procedures of the IARC should be improved by making more appropriate use of mechanistic evidence. It may be feared that the current evaluations result in false positives that can be avoided. In particular we recommend that

¹⁹ It may be the case that the best-qualified experts have real or apparent conflicts of interest and hence may not serve as working group members. In that case they may act as invited specialists. Invited specialists take part in subgroup and plenary discussions but they may not serve as meeting or subgroup chairs, draft text that discusses cancer data or contribute to the evaluations. (Cogliano *et al.* 2004, p. 1273)

²⁰ In footnote 12 we already mentioned that in the history of the IARC, certain agents have been re-evaluated in different *Monographs*.

mechanistic evidence is used more consistently with regard to extrapolation of experimental findings on animals to cancer in humans, that it is used to better rule out the possibility of confounding in epidemiological studies and that it is used to assess the temporal stability of carcinogenicity claims.

But from this it does not follow that the current evaluations have to be dismissed – at least not until they are contradicted by more recent *Monographs* or updates. Given what is at stake (viz. the life and the quality of life of thousands of people), we should prefer false positives over total ignorance.

Acknowledgements

We thank two anonymous referees, Leen De Vreese, Isabelle Drouet, Anton Froeyman, Federica Russo and Rafal Urbaniak for their comments on earlier drafts of this paper. We also thank the audiences at the First Biennial Conference of the Society for Philosophy of Science in Practice (Enschede) and at CaPitS2008 (especially Nancy Cartwright, Stathis Psillos and Paolo Vineis) for their comments. The research for this paper was supported by the Fund for Scientific Research – Flanders through project nr. G.0651.07. Bert Leuridan is Postdoctoral Fellow of the Research Foundation – Flanders (FWO).

References

- Barnard, R.C., Moolenaar, R.J. and Stevenson, D.E. (1989). IARC and HHS lists of carcinogens: Regulatory use based on misunderstanding of the scope and purpose of the lists. *Regulatory Toxicology and Pharmacology*, 9: 81–97.
- Bechtel, W. and Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421–441.
- Cogliano, V.J. (2006). Use of carcinogenicity bioassays in the IARC *Monographs, Annals of the New York Academy of Sciences*, 1076: 592–600.
- Cogliano, V.J. (2007). The IARC *Monographs*: a resource for precaution and prevention. *Occupational and Environmental Medicine* 64 (9): 572.
- Cogliano, V.J. et al. (2004). The science and practice of carcinogen identification and evaluation. *Environmental Health Perspectives* 112, (13): 1269–1274.
- Cogliano, V.J. et al. (2008). Use of mechanistic data in IARC evaluations. *Environmental and Molecular Mutagenesis* 49: 100–109.
- Dupré, J. (1993). *The Disorder of Things*. Cambridge & London: Harvard University Press.
- Glennan, S.S. (2002). Rethinking Mechanistic Explanation. *Philosophy of Science* 69, 3: S342–S353.
- Greenberg, R.S. et al. (2004, e-book). *Medical Epidemiology*. 4th edition. McGraw-Hill Professional, <http://www.accessmedicine.com/content.aspx?aid=546168>.

Hilakivi-Clarke, L. and de Assis, S. (2006). Fetal origins of breast cancer. *Trends in Endocrinology and Metabolism* 17: 340–348.

Hopkins, J. (1994). The role of cancer mechanism in IARC carcinogen classification. *Food and Chemical Toxicology*, 32, 2: 193–198.

IARC (1985). *Tobacco Habits Other than Smoking; Betel-Quid and Areca-Nut Chewing, and Some Related Nitrosamines*. IARC *Monographs on the Evaluation of Carcinogenic Risks to Humans*, volume 37, IARC, Lyon.

IARC (1986). *Tobacco Smoking*. IARC *Monographs on the Evaluation of Carcinogenic Risks to Humans*, volume 38, IARC, Lyon.

IARC (1987). *Overall Evaluations of Carcinogenicity: An Updating of IARC Monographs Volumes 1 to 42*. IARC *Monographs on the Evaluation of Carcinogenic Risks to Humans*, supplement 7, IARC, Lyon.

IARC (2004). *Tobacco Smoke and Involuntary Smoking*. IARC *Monographs on the Evaluation of Carcinogenic Risks to Humans*, volume 83, IARC, Lyon.

IARC (2006). *Preamble to the IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. <http://monographs.iarc.fr/ENG/Preamble/CurrentPreamble.pdf>

IARC (2007). *Smokeless Tobacco and Some Tobacco-specific N-nitrosamines*. IARC *Monographs on the Evaluation of Carcinogenic Risks to Humans*, volume 89, IARC, Lyon.

Kershaw, D.N. (1972). A negative income tax experiment. *Scientific American* 227, (4): 19–25.

Machamer, P., Darden, L., and Craver, C.F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67: 1–25.

McElroy J. et al. (2007). Occupational exposure to electromagnetic field and breast cancer risk in a large, population-based, case-control study in the United States. *Journal for Occupational and Environmental Medicine* 49: 266–274.

Reynolds, P.D. (1982). *Ethics and Social Science Research*. Englewood Cliffs, New Jersey: Prentice-Hall.

Steel, D. (2004). Social mechanisms and causal inference. *Philosophy of the Social Sciences* 34: 55–78.

Steel, D. (2008). *Across the Boundaries. Extrapolation in Biology and Social Science*. New York: Oxford University Press.

Ye, C.J. et al. (2009). Genome based cell population heterogeneity promotes tumorigenicity: The evolutionary mechanism of cancer. *Journal of Cellular Physiology* 219, 288–300.