Journal of Clinical Epidemiology 62 (2009) 270−277

Journal of Clinical Epidemiology

A weight of evidence approach to causal inference

Gerard Swaen[a,*], Ludovic van Amelsvoort[b]

[a]*Epidemiology Department, the Dow Chemical Company, The Netherlands*

[b]*Epidemiology Department, School for Public health and Primary Care (CAPHRI), Maastricht University, The Netherlands*

Accepted 24 June 2008

Abstract

Objective: The Bradford Hill criteria are the best available criteria for causal inference. However, there is no information on how the criteria should be weighed and they cannot be combined into one probability estimate for causality. Our objective is to provide an empirical basis for weighing the Bradford Hill criteria and to develop a transparent method to estimate the probability for causality.

Study Design and Setting: All 159 agents classified by International Agency for Research of Cancer as category 1 or 2A carcinogens were evaluated by applying the nine Bradford Hill criteria. Discriminant analysis was used to estimate the weights for each of the nine Bradford Hill criteria.

Results: The discriminant analysis yielded weights for the nine causality criteria. These weights were used to combine the nine criteria into one overall assessment of the probability that an association is causal. The criteria strength, consistency of the association and experimental evidence were the three criteria with the largest impact. The model correctly predicted 130 of the 159 (81.8%) agents.

Conclusion: The proposed approach enables using the Bradford Hill criteria in a quantitative manner resulting in a probability estimate of the probability that an association is causal.

Keywords: Causal inference; Bradford Hill criteria; Epidemiology; Risk assessment; Methodology; Epistemology

1. Introduction

One of the main objectives of epidemiological research is to identify causes of disease. A cause of a disease can be defined as a factor that affects its incidence. Elimination of the causal factor would result in a change in disease incidence. There is no universal consensus on the definition of a cause and several types of definitions have been described [1,2].

Causal inference from epidemiological studies is complex and has been the subject of extensive debate. For many associations between risk factors and health effects there is still some room for doubt about their causality. Perhaps the most realistic conclusion on causal inference was drawn by Weed who stated that the purpose of epidemiology is *not* to prove cause−effect relationships but to acquire knowledge about the determinants and distribution of disease and to apply that knowledge to improve public health [3].

Bradford Hill has formulated a set of criteria to assess causality [4]. Hill's criteria were an expansion of a set of criteria formulated in a landmark surgeon general's report on smoking and health. These causality criteria are widely accepted and only marginal changes have been proposed since they were first published. They are not only used in epidemiology. Guzelian et al. proposed a framework for evidence-based toxicology which heavily relies on the Bradford Hill criteria for causal inference [5].

The rigid use of strict causality criteria has been challenged. Kundi for instance pointed out that causality criteria can be used to falsely postpone public health action under the pretext that the available evidence does not fulfill the criteria [1]. Kundi proposed to use a pragmatic approach for assessing causal inference based on prior knowledge, epidemiology, animal studies, and in vitro studies. Kundi proposed that temporal relation, association, and environmental and population equivalence would suffice for a verdict of potential causation.

From the papers published on this topic, it is clear that the Bradford Hill criteria still remain key components to causal inference. However, there is no agreement on what weight should be given to each individual criterion and perhaps trying to draw absolute conclusions on causality is too far fetched. A more fruitful and practical approach would be to use the Bradford Hill criteria to estimate the likelihood that an association is causal. The outcome of such a probabilistic approach could be an estimate of the

* Corresponding author. Tel.: 31 (0)433626042; Fax: 31 (0)115674164. Epidemiology Department, The Dow Chemical Company, P.O. Box 444, Terneuzen, Zeeland, The Netherlands.

E-mail address: gswaen@dow.com (G. Swaen).

0895-4356/09/$ − see front matter © 2008 Published by Elsevier Inc.
doi: 10.1016/j.jclinepi.2008.06.013

What is new?

- The Bradford Hill criteria offer the best guidance for causal inference.
- However, there is no empirical information on how to weigh the criteria against each other.
- This paper provides an original and empirically based approach to causal inference.
- The outcome of the approach is a probability estimate that the association is causal.
- The approach is tested on examples and seems to work well and should be further tested in future causal inference studies.

probability that the association is causal. The estimate of the likelihood that an association is causal should be based on all nine Bradford Hill criteria and some criteria might be more useful than others. However, the weight of evidence that each single criterion contributes to the overall probability is not specified and a matter of subjective interpretation, or "expert judgment." Hill did not believe that any hard-and-fast rules of evidence could be laid down and he emphasized that his nine "viewpoints" were neither necessary nor sufficient for causation [6]. Weed evaluated two accounts of the use of the Bradford Hill criteria and concluded that the criteria of consistency, strength, dose–response, and biological plausibility but not often temporality were used when judging weak associations [7].

In a weight of evidence approach to causal inference, two aspects need to be quantified. First, the probability that a certain criterion is met needs to be estimated based on the available (epidemiological and other) evidence. In most cases, the available evidence in support of a particular causality criterion will not be evaluated to be 100%. For instance, the available evidence for the consistency may be assessed to be 75% in a situation that three out of four studies report an association and the fourth does not. Or the evidence for the plausibility criterion can be estimated to be met by a probability of 60% if only one animal study has found similar results, which essentially is an arbitrary estimate but must be seen in the perspective of a most complete data set of animal in vivo, in vitro experiments, and mechanistic supportive information which would receive a 1,005 probability if all evidence were in support of the association. The analogy criterion can be estimated to be met for 0% if well-designed studies on a similar compound are negative. Such estimates of the probability of the criterion being met need to be made for each of the nine criteria.

Second, a weight must be given to each of the nine criteria expressing the relative importance of each criterion for the overall causality assessment. It is obvious that not all criteria equally contribute to the overall evidence for causality and some may argue that the strength of evidence contributes more than plausibility for instance. Swaen et al. for instance reported that a dose–response relationship or biological gradient decreased the chance of a false positive finding [8]. Again, this weight of evidence can be regarded as being a matter of expert judgment, but preferably it would be based on empirical data.

The Bradford Hill criteria usually are not applied in a very systematic manner. On occasion the criteria that are thought to be met are listed and a conclusion on causality is drawn. A more systematic and predetermined weight of evidence approach would have the advantage that the overall assessment of causality becomes a more transparent and reproducible approach. A more systematic approach could also improve validity and reproducibility compared to the nonsystematic approach commonly used.

The purpose of our analysis is to develop a weight of evidence approach to causal inference resulting in an estimate of the overall probability that an association is causal. We have applied a weight of evidence approach to 159 chemicals or agents evaluated by International Agency for Research of Cancer (IARC) as category 1 or 2A carcinogens. Taking the IARC classification as the gold standard, we estimated the optimal weights for each criterion.

2. The causality criteria

The best and most recent description of Hill's causality criteria is given by Rothman and Greenland [9]. In short, Rothman and Greenland list the following nine causality criteria:

1 *Strength* of the association, the stronger the association the more likely that the association is causal.
2 *Consistency*, if more studies find similar results, the more likely it is that the association is causal.
3 *Specificity*, a specific exposure should exert a specific effect. There are causal associations that are not specific, for example, cigarette smoking, asbestos exposure, which are associated with multiple carcinogenic effects.
4 *Temporality*, the causal exposure should precede the caused disease in time.
5 *Biological gradient* or dose–response. If a dose–response is seen, it is more likely that the association is causal.
6 *Plausibility* depends on the current knowledge of the etiology of the disease. For instance, is it known that the agent or metabolite reaches the target organ, are studies in animal models positive?
7 *Coherence*, refers to other observed biological effects possibly relevant in the etiologic pathway that make a causal association more likely, for example, histological changes in the target organ.

8 *Experimental evidence*, if the disease rates go down after the causal agent has been eliminated, it is support for a causal association.
9 *Analogy*, if a similar agent exerts similar effects, it is more likely that the association is causal.

From Hill's and later Rothman's publications it is clear that these causality criteria should not be applied rigorously and that there always needs to be room for other interpretations. Hill did not intend the criteria to be used as a "tick" list, only as guidelines to interpretation of results. So far these nine criteria have not been rigorously challenged since they were proposed by Hill in 1965 and no claim has been made to add additional criteria.

3. The IARC database

The IARC has evaluated the carcinogenicity of a substantial number of chemicals, mixtures, and exposure circumstances. These evaluations have been carried out by expert interdisciplinary panels of scientists and have resulted in classification of these agents or exposure conditions into human carcinogens (category 1) probable human carcinogens (category 2A), possible human carcinogens (category 2B), not classifiable agents (category 3), and chemicals that are probably not carcinogenic to humans (category 4) (IARC, 2006). Although the IARC Working Groups do not formally use the Bradford Hill criteria to draw causal inferences many of the criteria are mentioned in the individual reports. For instance, the preamble specifically mentions that the presence of a dose–response is an important consideration for causal inference. In this analysis, the IARC database serves as the reference database although we recognize that it may contain some disputable classifications. However, to our knowledge there is no other database containing causal inferences that were compiled by such a systematic process involving leading experts in the areas of toxicology and epidemiology. The database describes in great detail the toxicological and epidemiological evidence that was evaluated and which formed the basis of the evaluation.

4. Methods and statistical analysis

As mentioned earlier, a weight of evidence approach to causal inference based on the Bradford Hill criteria requires two parameters. First, the probability that each criterion out of the nine criteria is true under the given epidemiological (and other toxicological) database, and second, an estimate of the relative weight for the contribution of each criterion to the overall causal inference. Next, these two parameters can be combined to obtain an estimate of the overall probability that the association is causal. The first parameter is a matter of expert judgment. It can be argued that the latter is more of a constant value and that the weights do not change across associations. We have estimated the weights of the individual criteria by means of a post hoc empirical approach. To assess the weights by means of an empirical approach, a series of assessments with respect to causality should be available.

We selected all agents in the IARC categories 1 and 2A. Some of these were excluded because they were similar. For instance Radium 224, Radium 226, and Radium 228 were treated as one agent, Radium. This resulted in a list of 159 agents, 96 in category 1 and 63 in category 2A. We did not include agents classified as 2B, 3, or 4 because for most agents in these categories no epidemiologic studies are available.

All agents (chemicals, mixtures, or exposure situations) in categories 1 and 2A were re-evaluated. The available evidence at the time of the IARC evaluation was used to assess the probability of each Bradford Hill criterion being fulfilled. Assessment of the probability is a matter of expert judgment and the outcomes may vary between individual assessors. The following principles were used as guidelines in the evaluation of the underlying evidence and in assessing the probability that a certain causality criterion was met.

The strength of the association was given a probability of 60% if the relative risk was between 1 and 2 (exceptional outliers in either direction were discarded if multiple studies were available), 80% if the reported relative risks were between 2 and 5 and 95% if the relative risks were higher than 5. If the relative risks from several papers fell into different scoring categories, outliers were disregarded and the category was selected that presented the most reliable and most consistent relative risk estimate. However no formal meta-analysis was conducted to establish the most likely relative risk estimate.

Consistency could only be evaluated if more studies were available. In the case of only one study consistency was scored as 0%. If three out of four studies were consistent (showing the same type of cancer effect), the probability that this criterion was met was scored as 75%. If all four or more studies were consistent in showing the same cancer effect the probability was scored as 95%. If there were conflicting studies but the overall effect was positive, the probability was scored as 60%. If there were several studies, some finding a cancer increase but other studies being negative, the probability that the criterion consistency was met was scored as 60%. Specificity was given a high score in case of one clear effect (e.g., bischloromethyl ether and small cell lung cancer as only reported cancer was scored as 95% met probability). In the case of cadmium, where some studies report an elevation of prostate cancer, and other studies report excesses of lung cancer and breast cancer, the probability that the specificity criterion was met was scored as 40%. For some agents, the epidemiological evidence only consisted of case reports, but they all indicated a similar cancer type. For these agents, the probability that the specificity criterion was met was scored as 60%. The temporality criterion was nearly always met if epidemiological data were available and was scored as

100%. Biological gradient (dose–response) was given a 90% probability if one or more studies clearly reported a positive dose–response gradient. If no clear dose–response was seen, but investigated it was scored as a 30% probability. If one study showed a clear dose–response but another did not the probability was estimated to be 50%. If no dose–response was ever reported for an agent or if the data showed no dose–response relationship a score of 0% was given. Most important component of the plausibility criterion was the availability of positive animal data or mechanistic data. In cases where clear long-term positive animal studies were available the plausibility criterion was scored with a 90% probability. If only one such study was available plausibility was scored as 80% probability. If there were mechanistic considerations why carcinogenic effect would not occur in man the probability was assessed to be 60%. If the only available evidence consisted of decreased DNA repair potential a probability of 50% was scored. If an immunosuppressive agent was shown to affect the immune system, the plausibility criterion was scored as a 40% probability. Early effects or other abnormalities in the target organ contributed to the probability of the coherence criterion being true. For instance, evidence showing that benzene causes bone marrow damage was scored as indicating that the coherence criterion was met with 80% probability. For only a few agents experimental evidence was available in the form of a reduction in the cancer excess after termination of exposure. Smoking cessation is a good example and because in this case the evidence is quite abundant, the probability that the experimental evidence criterion was met was scored as 95% probability. The probability score of the analogy criterion was based on the carcinogenic potential of other similar agents for instance other alkylating agents. The probability that the analogy criterion was met for an alkylating agent was scored as 60%. For an immunosuppressive agent, this probability was scored as 20%.

The probability of a criterion being fulfilled was assessed for each of the nine criteria, for 159 agents. Discriminant analysis was applied to assess the optimal weight for each criterion in such a way that the Bradford Hill criteria, together with the probability assessments, would optimally discriminate the agents between category 1 and category 2A carcinogens in a 100% correct manner. The analysis takes into account that 96 agents were classified as category 1 and the remaining 63 as category 2A and the weights are calculated in such a manner that this empirical distribution would be the ideal result. The statistical analysis was performed in SAS programming using the discriminant analysis module. Logistic regression was used to assess the univariate relationship between the criteria and the IARC classification. Stepwise logistic regression was used to assess the most relevant criteria in a combined model. Linear discriminant analysis was applied to construct the model and assess the individual parameters. This is a more robust approach than logistic regression for the available number of observation [10].

5. Results

All nine Bradford Hill criteria were entered into the discriminant model. The linear discriminant values for the Bradford Hill criteria are given in Table 1. The Appendix describes how the probability for a causal association should be calculated.

In general, the predictive model worked well. The percentage concordant predictions from the model is 130/159 = 81.8%. The model correctly classified 88.5% (85 out of 96) of the IARC category 1 carcinogens and 71.4% (45 out of 63) of the category 2A carcinogens. An example where the model made "wrong" predictions is Chimney sweeping, which was predicted to be a 2A agent, but has been classified by IARC as a category 1 carcinogen. This can be explained by the lack of systematic epidemiological data. UV light was predicted as a category 1 agent by the discriminant model, but was classified by IARC as a 2A agent. Recently IARC re-evaluated UV light into category 1.

Three Bradford Hill criteria contributed most to the predictive model. These were strength, consistency, and experimental evidence. The three criteria combined explain 41% of the total variance in the database.

6. Examples

As an example, we applied the weight of evidence approach to the data on the association between cigarette smoking and cancer (see Table 2). The application of the calculations described in the Appendix, resulted in a probability for classification as a true causal association of 98.9%. A second example is the sometimes-reported increased cancer risk among gasoline station attendants (see Table 3).

Several cohort studies have been conducted on gasoline station attendants. Only one of these has shown an association with leukemia [11]. However, other studies in Sweden and Italy have not reported similar results. The application of the weight of evidence approach to the available data in the literature resulted in a probability of 15.4% for this association to be causal.

As a third example, we have applied the weight of evidence causal inference approach to an agricultural chemical. There has been some concern in the literature about the potential carcinogenic effects of 2,4-D. This concern was mostly triggered by a case–control study by Hardell in Sweden that reported an association between 2,4-D and Hodgkin's disease, non-Hodgkin's lymphoma, and soft tissue sarcoma [12]. Later other epidemiological studies generally were not able to confirm these findings [13]. Chronic bioassays also did not report carcinogenic effects [14]. Table 4 provides an overview of the data on 2,4-D and the assessment of the Bradford Hill criteria (see Table 4). The weight of evidence approach results in a probability of 4.8% that the association is causal.

Table 1
Discriminant functions for the Bradford Hill criteria based on the discriminant analysis of 91 category 1 and 69 2A agents as evaluated by IARC

Hill's criterion	Key features	Discriminant function for category 1	Disciminant function for category 2A
Constant		−14.7799	−10.08346
1. Strength	The strength of the association in terms of relative risk, should be interpreted in the light of the degree that confounders may explain the association	0.06223	0.01923
2. Consistency	Is the association the same across studies?	0.04061	0.01803
3. Specificity	Applies to effect. Is the exposure associated with a specific outcome?	−0.02787	−0.03877
4. Temporality	This condition nearly always is met and does not help distinguish false positives from true positives	0.07657	0.08281
5. Dose–response	Does the relative risk increase with increasing exposure intensity?	−0.03528	−0.03534
6. Plausibility	Does the chemical or metabolite reach the target organ? Is there toxicological evidence for such an effect? Do in vitro and in vivo studies provide supporting evidence?	0.23025	0.21689
7. Coherence	Have effects relevant for the etiologic pathway been observed?	0.0009621	−0.00334
8. Experimental evidence	Does the observed excess decrease after termination of exposure?	0.00843	−0.00659
9. Analogy	Have similar effects in related chemicals been observed?	−0.01294	−0.01011

7. Discussion

Causal inference is one of the principal objectives of epidemiological research. It is generally accepted that the Bradford Hill criteria, as formulated in 1965, are still the most relevant criteria to be used in causal inference. However, there is no consensus on how these criteria should be applied to the epidemiological evidence and how each criterion should be weighed against the others. We have used the IARC category 1 and 2A database as a gold standard to assess the weights for each individual criterion. We have applied these weights to several other examples to assess the probability that these associations are truly causal. Our approach has three advantages.

First, it provides a more systematic and transparent approach. People can disagree on our assessment of the likelihood that a certain criterion is met. The advantage is that it now is clear about what aspect there is disagreement and to what extent the overall assessment of the causality is influenced by this disagreement.

The second major advantage of our approach is that the result of the weight of evidence is an estimate of the probability that the association is causal. This in our perspective is a more realistic outcome, which takes into account all available evidence. This probability can, for example, be taken into consideration in setting public health priorities. Additionally, it can be a tool to set research priorities. For instance, if the probability of an association being causal is already estimated to be above 80% further research can be regarded as a low priority, because a further increase in the probability of a particular criterion will not affect the overall probability of a causal association. On the other hand, if the probability for a causal association is estimated to be 50%, for example, this can be seen as an indicator that further research is needed and given priority. An analysis on the level of the probabilities of the individual criteria could be even a better indication for data gaps. For instance, if the plausibility, coherence, and analogy criteria have a high probability then new research targeted at information on dose–response could be considered instead of new research

Table 2
Weight of evidence approach using the Bradford Hill criteria for the association between cigarette smoking and lung cancer

Hill's criterion	Evidence	Probability (%) of criterion being true	Probability × weight for category 1	Probability × weight for category 2A
Constant			−14.7799	−10.0835
1. Strength	Strong associations reported not likely to be explained by confounders	95	5.91185	1.82685
2. Consistency	Nearly all studies are positive	95	3.85795	1.71285
3. Specificity	More effects have been noted	80	−2.2296	−3.1016
4. Temporality	Clearly smoking precedes lung cancer	100	7.657	8.281
5. Dose–response	Strong dose-response noted in many studies	95	−3.3516	−3.3573
6. Plausibility	Cigarette smoke is genotoxic and mutagenic and animal models for the association exist	90	20.7225	19.5201
7. Coherence	Effects such as epithelial changes have been seen	80	0.076968	−0.2672
8. Experimental evidence	After smoking cessation lung cancer rates go down	95	0.80085	−0.62605
9. analogy	Coal tar derivatives have been shown to cause lung cancer	80	−1.0352	−0.8088
Sum			17.630818	13.09635
Final probability	$e^{17.630818}/(e^{17.630818}+e^{13.09635})$=98.94%			

Table 3
Weight of evidence approach using the Bradford Hill criteria for the association between being a gasoline station attendant and cancer

Hill's criterion	Evidence	Probability (%)	Product of discriminant function and probability, C1	Product of discriminant function and probability, C2A
constant			−14.7799	−10.0835
1. Strength	Jakobson reported a RR of 3.6, but two other studies did not observe leukemia increases. An estimate for the RR would be between 1 and 2	60	3.7338	1.1538
2. Consistency	An association has been reported in one out of three studies	20	0.8122	0.3606
3. Specificity	Other increases in cancer have been noted	30	−0.8361	−1.1631
4. Temporality	Being a gasoline station attendant preceded leukemia, but this criterion does not distinguish	100	7.657	8.281
5. Dose−response	No dose−response noted or analyzed	30	−1.0584	−1.0602
6. Plausibility	I vitro data indicate that benzene is mutagenic. Animal bioassays provide some evidence for carcinogenicity	30	6.9075	6.5067
7. Coherence	None studied, none reported	0	0	0
8. Experimental evidence	Not available, no information for this criterion	0	0	0
9. Analogy	Leukemia found in higher exposed cohorts Gasoline contains small amounts of benzene. However the exposure levels are much lower than for instance in the Pliofilm cohort and exposure is much lower than cohorts with increases of leukemia	50	−0.647	−0.5055
Sum			1.7891	3.4898
Calculation	$e^{1.7891}/(e^{1.7891}+e^{3.4898})=15.4\%$			

that would contribute to the plausibility. If the probability of the association being causal is low, and all other things such as the potential health benefit and necessary costs being equal, a higher public health priority could be given to those risk factors that are more likely to be causal.

Thirdly, the weights assigned to each of the Hill criteria are based on empirical data. So far combining the nine criteria into one overall assessment of causality was a purely subjective exercise. The weight of evidence approach proposed here replaces this subjective assignment by a structured empirically based quantitative estimate of the weights. However, it must be clear that the assessment of the probability that a criterion is true still remains a matter of expert judgment. Our approach still contains the arbitrary element of assessing to what extent the criteria have been met and as such still requires an extensive degree of expert judgment. But even with respect to this arbitrary element it offers a more systematic approach because all the available evidence needs to be separated into the nine criteria and needs to be systematically reviewed. The still arbitrary process of assessing the probability that each of the nine criteria is met could be further strengthened by conducting an interrater and intrarater study with a number of described data sets. For these data sets, the probability that each of the nine criteria is met could be independently assessed and the intrarater variability could be evaluated.

Table 4
Weight of evidence approach using the Bradford Hill criteria applied to the carcinogenicity data for 2,4-D

Hill's criterion	Evidence	Probability (%)	Product of discriminant function and probability, C1	Product of discriminant function and probability, C2A
Constant			−14.7799	−10.0835
1. Strength	Hardell reported elevated risks, most later studies did not report any increased risks	30	1.8669	0.5769
2. Consistency	Most studies did not report excesses	20	0.8122	0.3606
3. Specificity	Soft tissue sarcomas, Hodgkin's Disease, and non-Hodgkin's lymphoma	20	−0.5574	−0.7754
4. Temporality	The cases occurred after the exposure occurred	100	7.657	8.281
5. Dose−response	No dose−response noted or analyzed	30	−1.0584	−1.0602
6. Plausibility	Long-term animal studies are negative [14]. Not mutagenic in *Salmonella typhimurium* [13]	30	6.9075	6.5067
7. Coherence	No early premalignant effects observed.	0	0	0
8. Experimental evidence	Not available, no information for this criterion	0	0	0
9. Analogy	Parent compounds are not carcinogenic in animal experiments	10	−0.1294	−0.1011
Sum			0.7185	3.705
Calculation	$e^{0.7185}/(e^{0.7185}+e^{3.705})=4.8\%$			

This then could be used to construct a number of reference data sets on individual agents that could be used as reference by experts assessing causality.

The approach taken by us has some limitations also. First, we have chosen to compare category 1 carcinogens with category 2A carcinogens and not to include category 2B, 3, or 4 agents. The reason for this is that category 2B, 3, and 4 agents in most cases lack epidemiological data and a discriminant analysis between category 1 and, for example, category 2b carcinogens would be based on many empty cells for the category 2B carcinogens. It must, therefore, be kept in mind that the model is based on a relatively small difference between the two groups of agents. Had the category 2B and category 3 agents been addressed in the model, the more epidemiological weights might have received lower weights. It is, therefore, recommended to use our approach only in those instances that resemble the data sets available for category 1 and 2A agents. The IARC classification of carcinogens was the best available data set known to us that could serve as a gold standard.

Second, the IARC database only contains carcinogenic endpoints and no noncancer endpoints. The weights for the Bradford Hill criteria derived in this article, therefore, may only apply to associations in the field of cancer endpoints and we cannot make inference about the applicability of the weights in case of noncancer endpoints. However, there is no a priori reason to assume that the Bradford Hill criteria should be weighted differently in cancer outcomes than in noncancer outcomes. We would have preferred that the database applied in our analysis would include data on associations between agents and noncancer endpoints. However, no database comparable to the IARC database exists for noncancer endpoints. The only way to include a reference database on noncancer endpoints would have been to carry out the classification ourselves which probably would have led to circular reasoning. A third limitation of our approach is the possibility that other assessors may score differently the probability that a Bradford Hill criterion is met. Different scores for the probabilities that a criterion is met could affect the weights calculated by the discriminant analysis. An additional step to the approach proposed by us would be to standardize the scoring of these probabilities. This could be achieved by having a panel of experts score a number of reference data sets and reaching consensus on the probabilities assigned to evidence packages for each of the criteria.

Using the results from randomized controlled clinical trials as the gold standard instead of the IARC database could have been an alternative approach for our analysis. However, this alternative approach has several disadvantages. First, only a selection of risk factors reported in the literature have been investigated by means of trials, certainly not the occupational and environmental chemicals. Second, there are instances in which randomized trials have yielded contradictory results, for instance, in case of several vitamin supplements and cancer outcomes.

The proposed weight of evidence approach will yield an estimate of the probability of the association being causal. Together with an appropriate cost-benefit analysis this estimate can be used as a basis for public health authorities to set priorities and plan intervention strategies.

By means of the analysis presented here, we have attempted to provide a more systematic and empirically based approach to the assessment of causal inference that can be used to assess the probability of a causal association.

Appendix

Instructions on how to apply the weights and calculate the probability of a causal association (smoking as example)

The probability of an agent being a human carcinogen is equal to $(e^{C1})/(e^{C1}+e^{C2A})$C1 is a constant plus the sum of the products of weight1 × probability per criterion.

C2A is a constant plus the sum of the products of weight2A × probability.

$$\begin{aligned} C1 = & -14.7799 + 95x.0.06223 + 95x0.04061 \\ & -80x0.02787 + 100x.0.07657 - 95x0.03528 \\ & +90x0.23025 - 80x0.0009621 + 95x0.00843 \\ & -80x0.01294 = 17.630818. \end{aligned}$$

$$\begin{aligned} C2A = & -10.08346 + 95x0.01923 + 95x0.01803 \\ & -80x0.03877 + 100x0.08281 - 95x0.03534 \\ & +90x0.21689 - 80x0.00334 - 95x0.00659 \\ & -80x0.01011 = 13.09636. \end{aligned}$$

$$\text{Probability} = \frac{e^{17.630818}}{(e^{17.630818} + e^{13.09635})} = 98.9\%.$$

References

[1] Kundi M. Causality and the interpretation of epidemiologic evidence. Environ Health Perspect 2006;114:969–74.

[2] Parascandola M, Weed DL. Causation in epidemiology. J Epidemiol Community Health 2001;55:905–12.

[3] Weed DL. Environmental epidemiology: basics and proof of cause-effect. Toxicology 2002;181–182:399–403.

[4] Hill AB. The Environment and disease: association or causation? Proc R Soc Med 1965;58:295–300.

[5] Guzelian PS, Victoroff MS, Halmes NC, James RC, Guzelian CP, et al. Evidence-based toxicology: a comprehensive framework for causation. Hum Exp Toxicol 2005;24:161–201.

[6] Phillips CV, Goodman KJ. The missed lessons of Sir Austin Bradford Hill. Epidemiol Perspect Innov 2004;1:3.

[7] Weed DL. On the use of causal criteria. Int J Epidemiol 1997;26:1137–41.

[8] Swaen GG, Teggeler O, van Amelsvoort LGPM. False positive outcomes and design characteristics in occupational cancer epidemiology studies. Int J Epidemiol 2001;30:948–54.

[9] Rothman KJ, Greenland S. Causation and causal inference in epidemiology. Am J Public Health 2005;95(Suppl 1):S144–50.

[10] Grimm La YP. Reading and understanding multivariate statistics. Washington, DC: American Psychological Association; 1995.

[11] Jakobsson R, Ahlbom A, Bellander T, Lundberg I. Acute myeloid leukemia among petrol station attendants. Arch Environ Health 1993;48:255–9.

[12] Hardell L, Axelson O. Soft-tissue sarcoma, malignant lymphoma, and exposure to phenoxyacids or chlorophenols. Lancet 1982;1(8286):1408–9.

[13] Garabrant DH, Philbert MA. Reiew of 2,4-dichlorophenoxyacetic acid (2,4-D) epidemiology and toxicology. Crit Rev Toxicol 2002;32:233–57.

[14] Charles JM, Dalgard DW, Cunny HC, Wilson RD, Bus JS. Comparative subchronic studies on 2,4-dichlorophenoxyacetic acid, amine, and ester in rats. Fundam Appl Toxicol 1996;33:161–5.