Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2009;**38**:1199–1201
© The Author 2009; all rights reserved. Advance Access publication 22 September 2009
doi:10.1093/ije/dyp299

Commentary: Cornfield, Epidemiology and Causality

Joel B Greenhouse

Accepted 3 August 2009

Throughout the 1950s, attacks on the accumulating evidence for a causal link between cigarette smoking and lung cancer frequently centred on the role of confounding variables that might explain the apparent association between the putative agent, cigarette smoke, and lung cancer. Because of the lack of control for omitted variables, there was a strong belief by many in the scientific community that evidence from observational studies was of less value than evidence generated by experiments. In two classic, but now mostly forgotten papers, Jerome Cornfield[1,2] responded to these attacks 'by providing a concise, explicit and lucid philosophic basis for the validity of information obtained from non-experimental studies'.[3] Cornfield wrote:

> We all have a vague feeling that if we can make an event occur, we understand it better than if we simply observe it passively. On analysis, this feeling seems to reduce to two propositions like the following: We are initially skeptical of any relationship based upon simple observation because the effects of other possibly important variables are not controlled and may account for the observed association. We are initially impressed by any relationship established by experiment because we feel that the effects of other important variables are controlled and cannot account for the observed association. The distinction we feel between a relationship based upon a statistical association and one based upon direct experimentation is thus a distinction between relationships that may be explained by other variables and those that cannot[1] (p. 20).

Cornfield then explained that many observational studies can be analysed so as to practically eliminate the possibility that extraneous variables account for the observed association. He cited Snow's demonstration that cholera was transmitted through polluted water as one example and the use of statistical control via cross-classification as another. But he also noted that 'there is no automatic guarantee in any particular instance that extraneous variables have been controlled by direct experimentation'.[1] (p.19). The history of randomized trials is replete with such examples. Having argued that there is no difference in kind between evidence generated by observational and by experimental studies, 'There are merely associations, whether observational or experimental'[1] (p. 20), Cornfield does acknowledge that there are important differences in degree between the possibility of spurious effects in an observational study versus a randomized study.

If there are no differences in kinds of evidence, then the procedure for interpreting an effect or evaluating the validity of an inference, he argued, must be the same whether the evidence is observational or experimental. Cornfield proposed an approach for establishing a cause-and-effect relationship based on the systematic elimination of alternative or competing hypotheses. He wrote,

> If important alternative hypotheses are compatible with available evidence, then the question is unsettled, even if the evidence is experimental. But, if only one hypothesis can explain all the evidence, then the question is settled, even if the evidence is observational. The proposition that some inherent logical incompetence attaches to an inference based on observational, as distinguished from experimental evidence seems to have little to commend it beyond the great positiveness with which it is sometimes asserted[2] (p. 250).

Cornfield understood quite well the scientific climate of the time and appreciated that no one approach was sufficient to establish an aetiologic relationship with high enough probability to convince the most sceptical of the scientific community. He recognized that a demonstration of a causal relationship between an agent and a disease, like a legal case, would be built upon a synthesis of all known facts concerning the agent and the disease. Cornfield, Haenszel, Hammond, Lilienfeld, Simkin, and Wynder, in the paper[4] published 50 years ago that we celebrate in this issue, set out to do exactly that for the specific case of cigarette smoking and lung cancer. They

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA. E-mail: joel@stat.cmu.edu

considered and rebutted each alternative hypothesis point by point: unproved diagnosis, the effect of an ageing population, necroscopy data, recall bias, other aetiologic factors, selection of study groups, confounding variables, the constitutional hypothesis, the argument that a measure of relative risk is inappropriate and many others.

Among the many topics discussed in the article, there is one that is unusually important for epidemiological methodology and is worth reviewing here. This has to do with the concern, noted earlier, that an observed association between a putative agent and disease might actually be due to a confounding variable that is associated with both disease status and the agent. Cornfield *et al.*[4] formally showed that if such a variable existed then its prevalence in the population would have to be at least as great as the observed relative risk of the disease for the putative agent. In the case of cigarette smoking and lung cancer, no such confounder had been found.

Although the proof of this result is given in Appendix A,[4] it might be of interest to present a slightly different derivation.[5] Let D denote the presence of disease, B the presence of the causal agent (B' the absence of the causal agent) and A the unobserved, non-causal agent (A' absence of the unobserved, non-causal agent). We denote the prevalence of the unobserved variable among those with B by $f_1 = P(A|B)$, and among those without B by $f_0 = P(A|B')$. By definition, the observed risk ratio for the putative agent B is $R_o = P(D|B)/P(D|B')$, and the unobserved risk ratio linking the presence of disease with the unobserved variable A is $R_u = P(D|A)/P(D|A')$. The question of interest is whether the unobserved variable A could fully explain the observed relative risk R_o, which appears to be due to B. If this were the case, then D would be independent of B given A. Formally, we have

$$\begin{aligned} p &= P(D|A', B) = P(D|A', B') = P(D|A') \\ pR_u &= P(D|A, B) = P(D|A, B') = P(D|A) \end{aligned} \tag{1}$$

Although perhaps obvious, we rewrite R_o applying the law of total probability, the definition of conditional probability and the assumption of conditional independence given in (1) to yield:

$$\begin{aligned} R_o = \frac{P(D|B)}{P(D|B')} &= \frac{P(D, A|B) + P(D, A'|B)}{P(D, A|B') + P(D, A'|B')} \\ &= \frac{P(D|A, B)P(A|B) + P(D|A', B)P(A'|B)}{P(D|A, B')P(A|B') + P(D|A', B')P(A'|B')} \\ &= \frac{pR_u f_1 + p(1 - f_1)}{pR_u f_0 + p(1 - f_0)} = \frac{R_u f_1 + (1 - f_1)}{R_u f_0 + (1 - f_0)} \end{aligned} \tag{2}$$

For a fixed $R_u \geqslant 1$, expression (2) is maximized when $f_1 = 1$ and $f_0 = 0$, leading to the inequality

$$R_o \leq R_u \tag{3}$$

Similarly, for fixed values of f_0 and f_1, expression (2) is maximized by letting $R_u \to \infty$, yielding the inequality

$$R_o \leq \frac{f_1}{f_0} \tag{4}$$

Equations (3) and (4) provide the conditions under which an unobserved variable A could fully explain the observed relative risk R_o. The unobserved risk ratio R_u must be at least as large as the observed risk ratio R_o, and the prevalence of A among those with B must be at least R_o times the prevalence of A among those without B. To drive the point home, the authors put into words what this result means with respect to the previously intractable constitutional hypothesis that the observed association between cigarette smoking and lung cancer was due to another factor. They explained,

> . . . [I]f cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone-X producers among cigarette smokers must be at least 9 times greater than nonsmokers. If the relative prevalence of hormone-X-producers is considerably less than ninefold, then hormone-X cannot account for the magnitude of the apparent effect[4] (p. 194).

Equations (3) and (4) together are called Cornfield's Inequality. It was the first formal method for sensitivity analysis in observational studies and continues to provide a formal response to one of the most difficult criticisms of observational studies. No longer could one refute an observed causal association by simply asserting that some new factor (such as a genetic factor) might be the true cause. Now one had to argue that the relative prevalence of this potentially confounding factor was greater than the observed relative risk of the putative causal agent.

Cornfield *et al.*[4] is a brilliant case study in the accumulation and synthesis of evidence for the purpose of demonstrating the causal role of cigarette smoke in lung cancer. At the same time, the authors helped advance the practice of epidemiology and scientific discovery through the development of new methods and an approach to causal inference that is still valued today. Their methodological contributions are significant and have clearly stood the test of time, but in establishing a causal link between smoking and lung cancer their collective contributions to public health have been nothing short of heroic.

Funding

National Institutes of Health (grants MH7862 and HL076852/076858 in part).

Conflict of interest: None declared.

References

[1] Cornfield J. Statistical relationships and proof in medicine. *Am Stat* 1954;**8:**19–23.

[2] Cornfield J. Principles of research. *Am J Ment Defic* 1959; **64:**240–52.

[3] Greenhouse SW. Jerome Cornfield's contributions to epidemiology. *Biometrics Supplement: Current Topics in Biostatistics and Epidemiol* 1982;**38:**33–45.

[4] Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. *J Natl Cancer Inst* 1954;**22:**173–203. Reprinted *Int J Epidemiol* 2009;**38**:1175–91.

[5] Gail M. Cornfield's inequality. In: Gail MH, Bénichou J (eds). *Encyclopedia of Epidemiologic Methods*. Chichester: John Wiley & Sons, 2000.