Commentary: Smoking and lung cancer: reflections on a pioneering paper

David R Cox

This fine paper\(^1\) is surely of direct value to all interested in the history of a major issue, possibly the major issue, in non-infectious disease epidemiology. It also sends a strong message to epidemiologists, statisticians and those from the machine learning world who are concerned with potential causal interpretation of their data. It may indeed be helpful to introduce statistical models to represent causal processes, even to call them causal models and to fit them successfully to empirical data, but this is far from demonstrating causality itself.

At a more personal level, the paper is a reminder of the one author whom I knew personally, J. Cornfield. He was a fine statistician, conversations with whom were as stimulating as they were enjoyable.

One broad theme in statistical thinking, no doubt with a long history, is that of the individually secure investigation, whether a randomized experiment or a survey of a population. The emphasis is typically on the use of randomization and various forms of error control to obtain conclusions that are, so far as is feasible, self-contained and do not rely on extraneous assumptions. Yet, Cornfield \etal\(^1\) remind us forcibly that deep conclusions often require synthesis of evidence of different kinds. Indeed R. A. Fisher, who was surely a prime designer of principles for achieving individually secure investigations, when asked by Cochran\(^2\) how to make observational studies more like randomized experiments in the security of their interpretation, replied: make your theories elaborate. Cornfield \etal\(^1\) beautifully illustrate Fisher’s aphorism, in that the strength of their paper comes in substantial part from the wide variety of the evidence that they discuss, population rates, prospective and retrospective studies and laboratory work.

Cochran himself played a major role in the authorship of the subsequent US Surgeon General’s Report, which was surely much influenced by the paper under current discussion.

It is interesting that at the time Cornfield \etal\(^1\) wrote their paper, three leading statisticians, from very different backgrounds, R. A. Fisher, J. Neyman and J. Berkson, all had reservations about whether causality had been established. J. Berkson, from the Mayo Clinic, came from a medical background and was particularly forceful in expressing his doubts. It is an occupational hazard of being a statistician that one may sometimes be the unwelcome, but necessary voice of caution. It is, I think, a delicate issue for individual statisticians to expose genuine difficulties of interpretation without becoming a source of negative thinking and discouragement.

Cornfield \etal\(^1\)’s paper is an overview in an older, and in a way more searching, sense than is nowadays often meant by that term. The numerical comparison, and if possible synthesis, of the conclusions from broadly similar studies is important in reducing the statistical uncertainty inherent in individual studies and providing some assurance of reproducibility. Cornfield \etal\(^1\) are concerned not so much with the statistical reliability of the studies they review, but rather with deeper issues of interpretation.

At the time Cornfield \etal\(^1\) wrote their paper, there was strong renewed interest in statistical circles in the notion of probability as representing uncertainty of knowledge in quite general contexts; Cornfield himself was certainly interested in these developments. These ideas might have led to the calculation of a number to represent the probability of causality given the data currently available. It seems, however, extremely unlikely that any such number would have been as helpful and insightful as the careful qualitative discussion in the paper.

Causality is a word used in those days much less freely in such discussions than it is now, although some such notion has surely always underpinned interpretation of most experimental and much observational data. The authors saw no need to define what they meant by causality. Implicit is more than a descriptive notion that smokers as a group have poorer prognosis than an otherwise comparable group of non-smokers. Rather, the hoped-for interpretation is that the group of smokers have a poorer prognosis than that same group of individuals would have experienced had they been non-smokers, other things being equal. This too is an essentially statistical notion. Understanding of an underlying biological process involves a different and in a sense deeper notion of causality and, while of course

---

Nuffield College, Oxford OX1 1NF, UK.
E-mail: david.cox@nuffield.ox.ac.uk
The reading of paper by Cornfield et al.1 on ‘Smoking and Lung Cancer’ is a real treat to anyone who is interested in the roots of causal reasoning in today’s epidemiology. It is surprising how much was already known: basic notions about confounding, selection and other biases, genetic influences, misclassification, the nature of observational evidence and the threshold for action when evidence is not perfect. All these ideas are present in the paper—not with their present-day names, but with a clear exposition about the concepts in crisp language, as they pertained to the 1950s debate on smoking and lung cancer. These ideas still figure as essential topics for discussion in today’s textbooks. A reading of the paper makes clear how much modern epidemiology was formed by the discussions about smoking and lung cancer. In this commentary, I have tried to elucidate key aspects of the paper, to indicate how they still lead to debate in the 21st century—50 years after the original publication.

A sensitivity analysis that still reverberates

The magnitude of the excess lung-cancer risk among cigarette smokers is so great that the results cannot be interpreted as arising from an indirect association of cigarette smoking with some other agent or characteristic, since this hypothetical agent would have to be at least as strongly associated with lung cancer as cigarette use; no such agent has been found or suggested.1 This conclusion of the paper rests on an algebraic derivation in an appendix and is what the paper is often remembered for nowadays. Although conceptually simple, it represented a gigantic leap forward, and might be seen as the starting point of all sensitivity analyses.2 The notion that large relative risks can be convincing by themselves is still very much alive.3 However, over the past few decades, the concept has often been reversed, to shed doubt on ‘small relative risks’, which led to statements that only relative risks >3 would be credible.4 That is not what the original said. The paper proposed that it is difficult to think of potential confounders to explain a 9-fold relative risk of smoking on lung cancer incidence because a potential confounder should be even more strongly associated with lung cancer as cigarette use; no such agent has been found or suggested.1

References