Epidemiologic Causation: Jerome Cornfield's Argument for a Causal Connection between Smoking and Lung Cancer*

Roger Stanev[**]
rstanev@interchange.ubc.ca

ABSTRACT

A central issue confronting both philosophers and practitioners in formulating an analysis of causation is the question of what constitutes evidence for a causal association. From the 1950s onward, the biostatistician Jerome Cornfield put himself at the center of a controversial debate over whether cigarette smoking was a causative factor in the incidence of lung cancer. Despite criticisms from distinguished statisticians such as Fisher, Berkson and Neyman, Cornfield argued that a review of the scientific evidence supported the conclusion of a causal association. Cornfield's *odds ratio* in case-control studies—as a good estimate of relative risk—together with his argument of "explanatory common cause" became important tools to use in confronting the skeptics. In this paper, I revisit this important historical episode as recorded in the *Journal of National Cancer Institute* and the *Journal of the American Statistical Association*. More specifically, I examine Cornfield's necessary condition on the minimum magnitudes of relative risk in light of confounders. This episode yields important insight into the nature of causal inference by showing the sorts of evidence appealed to by practitioners in supporting claims of causal association. I discuss this event in light of the manipulationist account of causation.

1. INTRODUCTION

As a number of historians, sociologists and epidemiologists have observed, the cigarette-smoking controversy exemplifies how economic, political and social factors can influence disputes over standards of scientific evidence. For instance, it is today understood that "[the tobacco industry] was able to ward off its public health enemies, harnessing its commercial interest to the social, psychological, and physiological dependence on cigarette smoking of people at large" (Susser 1973, 142). For Brandt, "the historical application of innovative methods of causal inference is inextricably tied to proving the harms of smoking … all the while, the tobacco industry worked diligently to disrupt the course of this scientific investigation." (2007, 4)

We have learned that during the 1950s and 1960s, despite several observational studies supporting the causal link between smoking and lung cancer, vocal critics such as Sir Fisher, J. Berkson and J. Neyman rejected the validity of such results in establishing a causal link. Although most today would be quick in pointing out how the ill-judgments of such critics were due to conflict of interest[1], not enough credit has been given to the important methodological

* I would like to thank Paul Bartha for helpful comments on earlier versions.
** University of British Columbia, Vancouver.
[1] Both Fisher and Berkson were hired and paid consultants for the tobacco industry.

questions raised by these leading skeptics. Questions concerning the sorts of evidence that were produced by the "new" epidemiologic studies, and the difficulties involved in establishing *causal association*, have been less well scrutinized by scholars. In part, this reflects the complexities involved in elucidating notions of causation and causal evidence that are implicitly appealed to by practitioners. This short paper is an attempt to improve on this predicament by examining Cornfield's response to critics of smoking as a cause of lung cancer from a manipulationist view of causation. I argue that Cornfield's *odds ratio* in case-control studies, together with his necessary condition on the magnitudes of relative risk in light of confounders, were essential tools in confronting the skeptics. The paper will also serve to explore the scope and certain difficulties with the manipulationist account of causation.

Among the several philosophical accounts of causation—e.g. counterfactual (Lewis 1982), physical connection (Salmon 1981), probabilistic (Cartwright 1979, Hitchcock 1993), contrastive (Schaffer 2005)—I choose Woodward's manipulationist account of causation as my candidate account because it values and focuses on the goals behind causal practices in science, such as the role of experiments (controlled and observational) in causal explanation and causal inference. This aspect of the account makes it an attractive one because it makes the project of describing certain scientific practices we have come to value, and the project of making recommendations about causal claims, intertwined projects that can and should be pursued jointly. This contrasts with competing philosophical accounts of causation that are essentially revisionary, or unificationist in their purposes.

The paper proceeds as follows. Section 2 sketches the arguments from the skeptics, i.e., Berkson's specificity of association argument, and Fisher's common cause argument. Section 3 presents Cornfield's response to the skeptics. Section 4 then raises some difficulties with the manipulationist account of causation in explaining Cornfield's response to critics. Section 5 concludes.

2. THE SKEPTICS

2.1 BERKSON'S SPECIFICITY OF ASSOCIATION ARGUMENT

Among the arguments raised by Berkson against the claim that smoking causes lung cancer—possibly the most forceful one—was the specificity of association argument. In his 1958 article Berkson raised the objection that the fact that smoking, as an exposure, was found to be linked not only with an increase in lung cancer incidence, but also with increases in incidence of seemingly unrelated conditions, such as heart disease, emphysema, bladder and pancreatic cancers, violated the specificity of association postulate for causal identification. The multiplicity of conditions according to him suggested that the causal inference from exposure to effect was no longer trustworthy. As Berkson wrote:

> Are these associations "statistically significant"? We are not concerned, at the moment with whether there is association with some specified category of disease, but with the validity of the evidence that there is association with disease in general. (…) For myself, I find it quite incredible that smoking should cause all these diseases. (…) When an investigation set up to test a theory, suggested by evidence previously obtained, that smoking causes lung cancer, turns out to indicate that smoking causes or provokes a whole gamut of diseases, inevitably it raises the suspicion that something is amiss. (1958, 32-34)

Practitioners like Berkson, took the view that the identification of a specific association demands that a certain exposure is associated with one, and only one, disease. The intuition behind this specificity of effect was that, the fewer the number of diseases which were associated with an exposure, the greater the weight which could be attributed to the relation between exposure and disease.

Although this form of intuition might have had its role in reductionist approaches to causal explanation, the methodological commitment to specificity per se, i.e., one exposure, one disease, was characteristic of a quite successful germ paradigm—Pasteur, Koch—experienced by the previous century. The locus of causal explanations there was on the identification of specific organisms that would provide necessary conditions to specific diseases. With the discovery of *tubercle bacillus* in 1882, Koch's postulates for causation—e.g., an "alien structure" must *always* be found with the disease—became the disease paradigm for the identification of causal associations. As the attention for causal associations began to shift from cases of infectious to chronic diseases, the need for a new form of causal evidence and inference could no longer be ignored. As we will see shortly, Cornfield's response to Berkson was an important step in allowing that shift.

2.2 FISHER'S COMMON CAUSE ARGUMENT [2]

Fisher was another leading critic of the causal link between smoking and lung cancer. Four considerations led Fisher to question the evidence produced by the new epidemiological studies: (i) ethical constraints requiring non-randomized experiments; (ii) study results ascertaining a link that was—according to Fisher—yet unknown (Cornfield et al 1959); (iii) conflicting results showing a negative correlation between inhaling and lung cancer (cf. Doll and Hill 1950), and (iv) a commitment to genetic notions of cancer causality (Box 1978). According to him, of the three logical possibilities, i.e., A causes B, B causes A, or something other causing both A and B, Fisher found the third possibility the most plausible in explaining the association between smoking and lung cancer. He proposed the view that a constitutional, common genetic factor, was leading individuals both to smoke and develop lung cancer, creating a confounding factor. According to him, if the observed association was not an artifact of the statistical data, i.e., not a spurious correlation between smoking and lung cancer, then genetic make-up could be the common cause that explained the association between these two variables.

The force of Fisher's constitutional hypothesis rested on the fact that it became a viable alternative explanation that was almost impossible to refute by observational studies. According to Fisher, since smokers had chosen to smoke, there could be personality traits driving the choice for smoking that could also be influencing other predispositions, such as the risk of developing cancer. Moreover, Fisher's subsequent studies concerning the genetics of smokers substantiated his argument by observing a greater concordance of smoking habits among monozygous than among dizygous twins. In a *Nature* article, Fisher noted that of the 82 recorded pairs of males investigated—51 monozygotic and genetically identical, and 31 dizygotic regular siblings—23% (12) of the identical twins, showed distinct differences of smoking behavior, whereas 52% (16) of the dizygotic brothers were dissimilar to the same extent; an approximate 2:1 ratio findings. As Fisher concluded:

[2] Fisher's common cause argument is also known as the "constitutional hypothesis" argument in epidemiologic literature.

> There can therefore be little doubt that the genotype exercises a considerable influence on smoking, and on the particular habit of smoking adopted. (…) Such genotypically different groups would be expected to differ in cancer incidence … (1958, 108)

Fisher's argument and identical twin study findings would prove unsuccessful in the face of Cornfield's response.

3. CORNFIELD'S RESPONSES

The persuasiveness of Cornfield's response to the skeptics resonates with Woodward's manipulationist account of causation. Let's begin with Fisher's objection. While genetic predisposition can never be completely excluded as a possible causal factor, its presence has not been demonstrated. If we accept the claim that observational study findings corroborate a causal association between personality traits and smoking, an essential counterfactual link in the logical structure of Fisher's explanation seems to be missing. That is, genotypes that supposedly predispose people to smoke have not been found invariant with respect to lung cancer.

According to the manipulationist account, "if a relationship is to qualify as causal, it must be invariant under some interventions." (Woodward 2003, 69) First, if a causal relationship between genotype and lung cancer holds, then it must be true that for *some* interventions on genotype—and holding fix background conditions—the manifestation of lung cancer would have to continue to hold.[3] The claim that a causal relation holds between genetic predisposition and lung cancer in some populations (e.g., smokers) ought to have some implications, even if only weak ones, for what we should expect to observe in other populations or circumstances. This can hardly be said of Fisher's genetic causal factors, since no study results, other than the observational ones pertaining to smokers, had been shown as potential interventions on genetic traits with respect to lung cancer.

Second, generalizations that are invariant under a larger and important number of changes provide better explanations than generalizations that do not. (Woodward 2003, 257) That is, even if a genetic causal link were suggested by further studies, it could hardly help to explain (i) the soaring increase in the incidence rate of lung cancer and (ii) the degrees of invariance. Genetic predisposition could be maintained as a common cause of smoking and possibly some cases of lung cancer, but as a plausible explanation for the vast increase in the incidence of lung cancer, Fisher's common cause would hardly do. After all, why haven't we experienced an increase in lung cancer before the introduction of cigarettes? By providing generalizations that are invariant under a larger and more important set of changes—i.e., larger set of studies and vast increase in the frequency of lung cancer—smoking becomes the crucial intervening variable between a possible predisposition to lung cancer and its manifestation. Cornfield's response capitalizes on this difference vis-à-vis relative risks in light of confounders.

Cornfield's response begins from the premise that observational study results show that cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer. Rather than

[3] A relationship may be invariant under some interventions but not invariant under others. That is, it is not a requirement that the invariance holds under all possible interventions.

giving in to the possibility that the observed association may not imply causation, or that hidden confounders might equally explain the observed association, Cornfield shifts the burden of proof onto the skeptics arguing against causation. Cornfield et al wrote:

> "If an agent, A [smoking], with no causal effect upon the risk of a disease, nevertheless, because of a positive correlation with some other causal agent, B [genetic factor], shows an apparent risk, r, for those exposed to A, relative to those not so exposed, then the prevalence of B, among those exposed to A, relative to the prevalence among those not so exposed, must be greater than r.
>
> Thus, if cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a casual agent, but only because cigarette smokers produce hormone X, then the proportion of hormone-X-producers must be at least 9 times greater than that of nonsmokers. If the relative prevalence of hormone-X-produces is considerably less than ninefold, then hormone X cannot account for the magnitude of the apparent effect." (Cornfield et al 1959, 194)

That is, if one would want to posit a common cause agent as a potential confounder, not any common cause agent would qualify as the explanation. One would need a common cause factor of a particular effect magnitude. If the observed association is strong (i.e., 9 times the relative risk) the candidate common cause factor would need to explain that the association between exposure and effect is *that* large.[4]

Therefore, with the introduction of Cornfield's necessary condition on the minimum magnitudes of relative risk, Fisher's common cause argument and his identical twin results regarding smoking behavior, will not pass muster. The burden shifts onto Fisher to show that his common genetic factor—whichever the factor turns out to be—is nine times as great among smokers as among nonsmokers. Fisher never replied to Cornfield et al 1959.

With respect to Berkson's specificity of association argument, Cornfield et al saw "nothing inherently contradictory nor inconsistent" in the suggestion that one agent could be responsible for more than one disease. (1959, 196) As a matter of fact, it would be even more "incredible" to expect that tobacco smoke—something known to be composed of hundreds of different chemical substances—would always have the same harmful effect. For this reason, and the fact that other historical precedents were available,[5] the evidence that tobacco smoke was a causal agent in the development of diseases other than lung cancer should had been perfectly expected, with each of its harmful effects requiring independent studies.

Berkson's skepticism however was deeper. It reflected a high skepticism of case-control study results. Rather than assessing the relative risk[6] of developing lung cancer between smokers and nonsmokers (i.e., the ratio between the probability of developing lung cancer among smokers and the probability of developing lung cancer among nonsmokers) case-control studies would, at best, measure the proportion of cases (those with lung cancer) that

[4] Cornfield et al (1959) included a relatively simple proof for such requirement in Appendix A.
[5] Cornfield et al noted that in 1952 the Great Fog of London increased death rate for a number of causes, including respiratory and coronary diseases. (ibid)
[6] Relative risk = [# of (smokers with cancer) / # of (smokers with cancer + smokers w/o cancer)] / [# of (nonsmokers with cancer) / # of (nonsmokers with cancer + nonsmokers w/o cancer)]

were smokers and the proportion that were not, compared with the proportion of controls (those without lung cancer) that were smokers and the proportion that were not.

It was here that Cornfield's *odds ratio* would prove an important tool against skeptics like Berkson, by gaining new advocates of case-control studies. Despite the fact that in case-control studies relative risks could not (and cannot) be computed directly, because there is no information about the incidence of lung cancer in smokers versus nonsmokers, case-control can still assess the measure of causal association between smoking and lung cancer vis-à-vis relative odds. In an earlier work published in 1951 in the *Journal of the National Cancer Institute*, Cornfield had shown how the ratio of the odds that the cases were exposed to the odds that the controls were exposed can function as a good estimate of whether a certain exposure (smoking) is associated with a specific disease (lung cancer).[7] For the estimate to hold, three conditions were needed: (i) the controls must be representative of all individuals without the disease in the population from which the cases were drawn; (ii) cases must be representative of all individuals with the disease in the population from which the cases were drawn; and (iii) the disease studied needs to be infrequent.[8] All conditions argued as having been satisfied by the case-control studies summarized in Cornfield et al 1959 review report.

However, despite the introduction of new conceptual tools and "numerous answers to Berkson's critiques, Berkson never relented in his skepticism." (Brandt 2007, 143) Moreover, "by 1963, Berkson's critiques had been repeatedly rebutted (…) and while no one questioned his sincerity, it had become clear that his doubt was impervious to evidence." (Brandt 2007, 224)

4. CAUSATION AS COUNTERFACTUALS IN CASE-CONTROL STUDIES

According to Woodward's manipulationist account of causation, causal associations are counterfactual relations that are potentially exploitable for purposes of manipulation. (Woodward 2003) Since for the manpulationist, the manipulation that has the right sort of structure *is* an intervention, to what extent do case-control studies, most specifically the selection of controls, fit the manipulationist clause of an intervention? That there must be some intervention on smoking such that if it were to occur, then the probability of lung cancer would change, does not strike this author as something underlying the selection of controls.

For Woodward, it is useful to think of an intervention as an idealized experimental manipulation, and if the reader wishes to have a concrete idea of his notion of intervention, "the obvious candidate is randomized experiments." (2003, 95) Woodward treats randomized control trials as indirect methods of intervention, i.e., as providing indirect evidence of what would happen under an intervention. By dividing subjects with the disease into two groups, one that receives the drug and the other that does not, and then observing the incidence of recovery in the two groups, the experimenter's intervention consists in the assignment of treatment to individual subjects (e.g., represented as a binary value, depending on whether or not the subject receives the treatment).

[7] Odds ratio are also known as cross products ratio in epidemiologic studies.

[8] This last condition occurs when very few individuals develop the disease (a) compared to the number of individuals who never develop the disease (b), then (a) + (b) can be approximated to (b).

However, randomized control studies are methodologically distinct from case-control studies, i.e., they produce different types of counterfactual relations. Because in case-control studies we begin with the diseased individuals (cases) the counterfactual relation sought by such design requires the selection of non-diseased individuals (controls). What exactly constitutes the intervention in these situations? Is the manipulation the complete selection of controls? If so, is the intervention the selection of controls that are similar to the cases in *all* respects, or just those relevant epistemic aspects other than having lung cancer? Or is it that in order to qualify as an intervention, the selection has to accord with one of Cornfield's condition, i.e., the selection of controls should be such that is representative of all individuals without lung cancer *in the population from which the cases were selected* ? It is unclear.

Moreover, what about situations in which the characteristics of the non-diseased individuals in the population from which the cases were selected are not well understood, because the reference population might have not been well defined. Will the selection of controls still constitute an intervention for the purposes of causal inference? Perhaps a friendly amendment to Woodward exists, of which I am not aware of.

For another manipulationist such as Freedman (1997) causal inferences differ from other sorts of inference based on associations, because in causal inferences "a change in the system is contemplated; there will be an intervention." (62) But the answer implicit in Freedman's causation with respect to the validity of case-control findings, although arguably more straightforward than Woodward's, is also puzzling. According to him, there exists a fundamental difference between two ideas of conditional probabilities: (A) selecting individuals with $X=x$ (e.g., lung cancer) and looking at the average of their Y's (e.g., smoking habits), and (B) intervening to set $X=x$ and looking at the average of their Y's, license different inferences. The latter licenses causal inference, whereas the former does not, suggesting that for Freedman, the validity of case-control studies in making causal inference is questionable. An even more puzzling manipulationist answer given the importance of observational studies in establishing causal association.

5. CONCLUSION

This brief revisionist episode of Cornfield's argument supporting the claim of a causal association between smoking and lung cancer has pointed out how tools such as case-control studies, odds ratio as estimator of relative risk, and the demand for a minimum effect size, were important in confronting the skeptics. It has also allowed us to begin to see these tools in light of a manipulationist account of causation and a certain difficulty in making the selection of controls as a means for interventions.

It is not so much as these tools were able to meet all the necessary and sufficient conditions in establishing causal associations, but rather working as important contributions to an assessment of the total amount of relevant findings, i.e., epidemiologic, experimental and clinical, while observing a consistency among diverse studies. The conclusion of smoking as causal factor of lung cancer was greatly strengthened when different types of evidence, including those from case-control studies, could support such association. A true account of causation would have to do justice to the role these tools can play in establishing causal associations.

BIBLIOGRAPHY

Berkson, J. (1958), Smoking and Lung Cancer: Some Observations on Two Recent Reports, *Journal of the American Statistical Association* Vol.53, 281:28-38.

Brandt, A. (2007), *The Cigarette Century*, Basic Books Press, New York.

Box, J. F. (1978), *R. A. Fisher The Life of a Scientist*, John Wiley & Sons, New York.

Cartwright, N. (1979), Causal Laws and Effective Strategies, *Noûs* 13, 419-436.

Cornfield, J. (1951), A Method of Estimating Comparative Rates from Clinical Data: Applications to Cancer of the Lung, Breast and Cervix, *Journal of the National Cancer Institute*, 11:1269-75.

Cornfield, J., et al (1959), Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions, *Journal of the National Cancer Institute,* 22:173-203.

Doll, R. and Hill, B. (1950), Smoking and Carcinoma of the Lung: Preliminary Report, *British Medical Journal* 224:742-747.

Freedman, D. (1997) From Association to Causation via Regression, *Advances in Applied Mathematics* 18:59-110.

Hitchcock, C. (1993), A Generalized Probabilistic Theory of Causal Relevance, *Synthese* Vol. 97, 3:335-364.

Lewis, D. (1982), Causation, *Philosophical Papers* Vol. II, Oxford: Oxford University Press.

Salmon, W. (1980), Causality: Production and Propagation*, Proceedings of the Biennial Meeting of the Philosophy of Science 1980* Vol.2, xx-xx.

Schaffer, J. (2005), Contrastive Causation*, The Philosophical Review* Vol.114, No.3 297-328.

Susser, M. (1973), *Causal Thinking in the Health Sciences*, Oxford University Press.

Woodward, J. (2003), *Making Things Happen*, Oxford University Press.