When Can Risk-Factor Epidemiology Provide Reliable Tests?

*Deborah G. Mayo** *and Aris Spanos*†

Can we obtain interesting and valuable knowledge from observed associations of the sort described by Greenland and colleagues[1] in their paper on risk factor epidemiology? Greenland argues "yes," and we agree. However, the really important and difficult questions are when and why. Answering these questions demands a clear understanding of the problems involved when going from observed associations of risk factors to causal hypotheses that account for them. Two main problems are that 1) the observed associations could fail to be genuine; and 2) even if they are genuine, there are many competing causal inferences that can account for them. Although Greenland's focus is on the latter, both are equally important, and progress here hinges on disentangling the two to a much greater extent than is typically recognized.

A THEORY-DOMINATED VERSUS A NEW-EXPERIMENTALIST PHILOSOPHY

In advocating the role of observational studies as the provider of facts from which theory follows, Greenland and his colleagues allude to "one popular philosophy of science"—presumably Popper's, but, in fact, Popper was the exemplar of a "theory-first" philosopher. Nevertheless, Popper's demand is wrongheaded (except in the sense that one begins with an interest or problem—one doesn't just "observe"!); moreover, Popper denied that even the best-tested theory could be regarded as reliable. So the question arises as to which philosophy of science best supports the position of Greenland et al. We think the most congenial would be the "new experimentalism," in which experiment and statistical analysis could "have lives of their own" quite independent of substantive scientific theories.[2,3]

Despite their championing of "purely descriptive (atheoretical) approaches," it is evident by their emphasis on "precision and replication" and "precise null results" that Greenland et al. have in mind, not "mere observations," but rather something more in line with what has been called "experimental or statistical knowledge": knowledge of genuine regularities and of what would be expected to occur were certain experiments carried out.[3] ("Experiment" is understood broadly to cover empirical inquiries in which there is adequate error control.)

WHEN DO WE HAVE A GOOD TEST?

To promise, as do Greenland et al., that "with enough description, ...researchers have on hand immediate tests of predictions from proposed theories" is to gloss over the trials and tribulations of testing (of which Greenland and colleagues are well aware). Such claims weaken their arguments in support of what we take to be their overriding (and

From the Departments of *Philosophy and †Economics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.
Correspondence: Deborah G. Mayo, Department of Philosophy, Major Williams Hall 235, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. E-mail: mayod@pop.vt.edu

Copyright © 2004 by Lippincott Williams & Wilkins
ISSN: 1044-3983/04/1505-0523
DOI: 10.1097/01.ede.0000135912.17991.72

right-headed) thesis, namely that it is possible to obtain useful and reliable experimental knowledge apart from substantive theories.

For observable data to genuinely test a hypothesis, it does not suffice that they be "explained by" the hypothesis. Mere accordance with the data is too easy; we must be able to show the hypothesis has withstood a reliable or severe probe of the ways the hypothesis could be wrong. We would need to argue, for example, that were a causal claim not at least approximately correct, it would not have been able to account as well as it does for numerous and deliberately varied effects; it has withstood what we might call a "severe test." For data to serve in severe tests, they need not be "theory-free"; it is enough that the overall reliability of the test not be weakened by any uncertain aspects of theories.

THE 'NOVELTY' REQUIREMENT

When observational reports constitute experimental knowledge, Greenland et al. are correct that they "can be used regardless of whether they were gathered for the purpose of testing a given theory." Data could severely test a hypothesis even if it is the "same" data used in its construction (remodeled to ask a different question). Admittedly, the best-known philosophies of testing fail to discriminate between unwarranted and warranted cases of using data both to arrive at and test a hypothesis. If one is allowed to search through several factors and report just those that show impressive correlations, there is a high probability of erroneously inferring a real correlation. In other cases, when a genuine effect is already established, we could reliably use the "same" data both to identify and to test the cause (eg, pinpointing the overheated tiles as cause of the Columbia shuttle crash). Requiring tests to be severe explains (apparently) conflicting intuitions about preferring novel predictions.[4]

The example of lipid peroxidation and renal cell carcinoma with which Greenland and colleagues illustrate their thesis is a reasonable success story precisely because of the deliberate manner in which the various studies combined background knowledge and interrelated checks of potential errors, biases, and alternative explanations. The example does not support a looser thesis about the value of mere "descriptive reports," because the evidence here supplies not mere observations, but experimental knowledge of the cluster of risk factors involved. Although the studies were not testing full-blown theories of renal cell carcinoma, each involved testing statistical (or "experimental") hypotheses with a variety of their own assumptions. In particular, several of the studies described by Greenland et al. were specifically designed to examine the risk of renal cell carcinoma in relation to such risk factors as hypertension, diet, and smoking.

STATISTICAL ASSUMPTIONS

For inferences to statistical or experimental knowledge to be reliable, certain assumptions must be met. Although we have no doubt that Greenland and his colleagues are fully aware of this necessity, we think the lack of explicit recognition here obscures the most powerful basis for their intended argument. To obtain reliable experimental knowledge by means of correlation analysis between a response and several potential risk factors requires appropriate statistical analysis, which in turn depends on valid underlying assumptions, eg, normal, independent, and identically distributed random variables. With invalid statistical assumptions, the estimators of the correlation coefficients are not reliable enough to establish the statistical—let alone the substantive—correlation.

The ability to establish the validity statistical assumptions without appeals to substantive theories (by misspecification testing[5]) is key to the experimental knowledge of real effects—the very thing Greenland and colleagues require to make their case. Statistical inferences could give us experimental knowledge, which in turn serves as the basis for the kind of genuine and severe tests that the authors espouse because it remains even through changes in background theories and through reinterpretations of effects.

We endorse the suggestion of creating "a bank of observations" or, more accurately, a repository of experimental knowledge (epidemiologic, laboratory, clinical). Shrewdly combining this knowledge to form severe and informative tests will demand systematic, although not purely formal, methodologic strategies and analogic arguments. To carry out this promise, however, demands addressing, not merely "straw men" skeptics (eg, Skrabanek; see Greenland et al.[1]) but the twin problems with which we began. It will require going beyond impressive success stories, to articulating the experimental knowledge that enables severely discriminating between potential causal factors.

ABOUT THE AUTHORS

DEBORAH MAYO is a Professor of Philosophy. Her work is in the epistemology of science and philosophy of statistical inference. She received the 1998 Lakatos Prize for her book *Error and the Growth of Experimental Knowledge (1996)*.

ARIS SPANOS is a Professor of Economics. His most recent book is *Probability Theory and Statistical Inference* (1999). His current research interests include the methodology of empirical modeling.

REFERENCES

1. Greenland S, Gago-Dominguez M, Castelao JE. The value of risk-factor ("black-box") epidemiology. *Epidemiology*. 2004;15:529–535.
2. Hacking I. *Representing and Intervening*. Cambridge: Cambridge University Press; 1983.
3. Mayo DG. *Error and the Growth of Experimental Knowledge*. Chicago: The University of Chicago Press; 1996.
4. Cox D. The role of significance tests. *Scand J Stat*. 1977;4:49–70.
5. Spanos A. *Probability Theory and Statistical Inference*. Cambridge: Cambridge University Press; 1999.