SKELETAL PLAN OF A COMPREHENSIVE STATISTICAL HEALTH-POLLUTION STUDY

JERZY NEYMAN
STATISTICAL LABORATORY, UNIVERSITY OF CALIFORNIA, BERKELEY

1. *Objective.* The objective of the proposed comprehensive statistical health-pollution (CSHP) study is to estimate the relationship between selected characteristics of health conditions and the proliferating pollutants as they appear in the actual environment.

2. *Selection of health characteristics and of the pollutants to study.* The selection of health characteristics and of pollutants to be studied falls within the field of competence of specialists in biology, in health sciences, in chemistry and in physics.

Two aspects of the problem appear to require separate consideration. First, there are suspected deleterious effects of pollutants on health of "normal" humans now living. Second, it is presumed that certain pollutants are mutagens which affect adversely future generations. The subject of study could have been simplified if these two different aspects would be treated separately. As things stand now, there is a substantial overlap: mutagenic effects seem to parallel carcinogenic effects, which manifest themselves in the now living generations. Also [1], mutagens are being suspected as causes of abnormalities at birth. These points, as well as difficulties in monitoring, will be discussed at the conference during the Thursday morning session, July 22nd.

3. *Necessity of simultaneous treatment of all the suspected deleterious pollutants with reference to a number of localities.* Even though many current studies refer to just one pollutant (frequently radiation), it must be clear that, in order to be able to evaluate the effect of a single pollutant, it is unavoidable to evaluate, perhaps only summarily, the effects of all others. The point is that all the deleterious pollutants "compete" with each other for human health and lives. The number of victims claimed by a particular pollutant A, in a given locality and during a particular year, can be small or large depending on whether another pollutant B kills many or only a few people (respectively), preventing them from succumbing to A. This remark applies to cases where the cause of death (or other condition) is unambiguously defined as, for example, death from cancer. The quantities discussed in some current pollution-health studies are what is technically called "crude rates," for example, of deaths from cancer. The proper

measure is the "net rates" computable taking into account deaths from other causes [2].

In most pollution-health studies the causes of deaths (or other health conditions) are not identified and the purpose of such studies is precisely to find out, tentatively, whether the actual causes might have been the suspected pollutants. An unambiguous answer could be obtained in ideal conditions of having two localities, say L_1 and L_2, inhabited by identical populations and affected by identical pollutants, with one exception: in addition to pollutants affecting L_1, the locality L_2 is affected by a single extra pollutant A. If such two localities could be found, then the differences in health conditions in L_1 and L_2 would be ascribable to A operating with all other pollutants in the background.

Clearly, no such two localities exist in the real world and the results of an actual CSHP study must be only tentative, subject to indirect confirmation by experiments with animals, etc. However, in order to be able to obtain even such tentative evaluations of particular pollutants it is unavoidable to study simultaneously a number of localities with different patterns of pollution. A study of this kind, involving 38 cities in the United States, was recently performed [3] and the authors, Hickey *et al.*, deserve recognition for their effort and for the initiation of what might be described as multivariate-multilocality health-pollution studies.

In a sense, the proposed CSHP study might parallel the study of Hickey *et al.*, with a few modifications. One is that radioactive pollution, omitted by Hickey *et al.*, should be included in the CSHP study. The other modifications refer to statistical methodology and, probably, to selection of localities to be studied. One example is as follows.

The statistical methodology used by Hickey *et al.* is based on multivariate linear regression techniques. Contrary to this, a substantial biological literature indicates that the regression of unsatisfactory health conditions on two or more pollutants might well be far from linear. The literature in question, to be discussed in the Thursday afternoon session, is concerned with the term "cocarcinogens." Apparently, certain agents exist, say A and B, which, by themselves, are only mild carcinogens if at all. A small exposure to either A or B produces only a few cancer cases in experimental animals. Thus, the regression of cancer incidence in A alone may be linear but with only a small regression coefficient, and similarly for B.

On the other hand, if a small dose of A is followed by exposures to increased doses of B, the incidence of cancer grows fast. In these conditions, the regression of cancer incidence on both A and B simultaneously will be far from linear.

The above applies to what might be labeled "positive" co-carcinogens. Professor F. N. David tells me that there are also "negative" co-carcinogens. Another point to consider: if co-carcinogens exist, referring specifically to cancer, is it not plausible that some other ailments, perhaps respiratory diseases, are affected by some co-factors or co-pollutants? The *a priori* adoption of a linear repression scheme may bring out misleading conclusions.

The nonlinearity of regression is just one of the points of divergence between the methodology adopted by Hickey *et al.* and what would be my own preference. There are other such points.

4. *Localities to be included in the proposed study.* Because of the public concern about deleterious effects of radiation, it is proposed that the CSHP study cover (a) all the localities of the operating nuclear power facilities (Messrs. Patterson and Thomas promised to include the list of such facilities in their paper to be given during the afternoon session of July 21st). (b) For comparison, it is proposed to include in the study the sites of a comparable number of large factories using mineral fuels. These factories should be randomly selected from a reasonably complete, appropriately stratified list. Possibly Dr. Waksberg from the Bureau of the Census could help to secure such a list. (c) The same public concern with radioactivity suggests the desirability of inclusion in the study of localities surrounding the nuclear weapons test sites in Nevada. Here again Drs. Patterson and Thomas will be helpful with information about monitoring facilities. (d) In order to provide a comprehensive picture of the health-pollution relationships in the country, the CSHP study should include representative samples of major cities and of countryside localities, particularly those exposed to chemical pollutants. Here the suggestions of Dr. G. Morgan are likely to be very useful. Also I expect important information from Drs. R. Risebrough and T. Sterling.

5. *Difficulties with the data.* It is to be anticipated that the CSHP study will encounter considerable difficulties in securing reliable data. The prospects in this respect may be judged by the already quoted article of Dr. Hook [1], describing a special two-day conference held last year in Albany, N.Y. The specific purpose of this conference was the finding of means to improve the monitoring of human birth defects. It appears that, in a particular state, a change in the method of monitoring malformations resulted in an increase in the records in the ratio of 1 to 3.5! While this applies to malformations of the new-born, it appears plausible that the accuracy of other health records must also vary from one state to another and, probably be a greater extent, from one county to the next. (Inspection of actual data and also conversations with some knowledgeable persons convinced the present writer that this supposition is very plausible.)

If the proposed CSHP study is to bring out reliable information, special care must be taken to see that differences in the health data reflect true differences in health conditions rather than differences in routines of data collection. How to achieve this? What existing organizations can be helpful? It is possible that some agency of the Federal Public Health Services conducts routinely some spot checking of the precision of data collection. Hopefully, information on this point will be forthcoming in the discussions at the conference.

The incompleteness of records is not the only possible source of uncertainties regarding health data. For instance, the incidence of various diseases may depend very much on the stratum of the population. Many diseases rather rare in conditions of relative comfort are likely to be quite frequent in the slums.

Thus, in order to be able to assign deteriorations in health conditions to pollutants it is essential to eliminate biases resulting from equal treatment of data referring to slums and to comfortable suburbs. It follows that the proposed CSHP study will need data on particular racial and economic strata of the population in the various localities. The difficulties in this respect are enhanced by the fact of apparently widespread uneasiness about proximity of a nuclear facility. It is not implausible that, when such a plant is constructed in a given locality, the economically comfortable inhabitants move out, prices of property decrease and the immediate vicinity rapidly turns into a slum.

Clearly, all the above is just speculation which can be easily countered by other speculations. For example, it may be argued that the construction of a nuclear power facility leads to closing down of some smoky mineral fuel power plants, to cleaning of the atmosphere and the consequent increase in the general standard of living. However, if the CSHP study is to yield reliable information, it must be based not on speculations of one kind or another, but on verifiable data. Here questions arise: how and where could one secure reliable data on the stratification of inhabitants of localities selected for the study? Special sample surveys? What agency could and would conduct them? How much would such surveys cost? The participation of Dr. Waksberg in the discussion of these and some other similar problems will be greatly appreciated.

In addition to uncertainties of health data, there are also uncertainties about monitoring of pollutants. The already mentioned paper by Hickey *et al.* deals with an impressive array of pollutants including SO_2, NO_2, Cu, Ti, *etc.*, *etc.* In addition, however, one must also think of residues of pesticides and of defoliants found in milk and other foods. In an optimistic mood one is inclined to take it for granted that all the relevant data are reliable, even though possibly scarce. The disquieting detail in this connection is the paper by Fred S. Goulding concerned with "an improved analytical tool for trace element studies." The title of the paper suggests the possibility that the monitoring of, say, Cu and Ti in the atmosphere is conducted by not just one method in the United States, but by a variety of different methods, using different tools. If this be so, then the data on pollutants must be subject to the same kind of uncertainties as the data on health conditions: the records of trace metals, as well as the records of particulates, and so forth, coming from different localities may reflect not only the real differences in the degree of pollution but also the differences in the method of ascertainment. If the proposed CSHP study is to be reliable, this particular point requires serious attention. How uniform are the techniques of monitoring pollutants? Is it at all feasible to reduce the observations made in different localities by different methods to some kind of a common denominator? Here, Dr. George Morgan may provide very important information.

6. *Recommended structure of the* CSHP *study.* The authoritativeness of the proposed CSHP study will be greatly increased if it is to be performed by not just one but by at least two or three academic statistical groups, perhaps one in the East, one in the Midwest and one on the West Coast. It would be essential

to arrange that the three groups could work in frequent contact with each other and in cooperation with governmental agencies, but with a guaranteed independence from these agencies.

One of the reasons for such multiplicity of effort is the fact that statistical study of a given complex phenomenon allows a number of different approaches (or "models") which, at first sight at any rate, may appear equally plausible. The choice between such possibilities must depend upon earlier experiences of the individuals concerned and, undoubtedly, on the ubiquitous preconceived ideas that affect all of us.

The interactions between cooperating groups, combined with unlimited access to the same data and to all other sources of information are likely to result in an increase in objectivity and reliability of the findings. It may be hoped that the cooperative study will result in a single joint factual report. In addition there will be interpretations and/or conclusions. These are expected to be somewhat different, depending upon personal attitudes of the authors. However, the public and the Government are likely to gain from having at their disposal the presentations from several different points of view.

7. *Practical steps.* The above discussion suggests that, in order that the proposed comprehensive statistical health-pollution study be reliable, it must be preceded by substantial interdisciplinary preparation.

(i) A preliminary list of health parameters of pollutants and of calendar years to be studied must be established.

(ii) A preliminary list of localities must be compiled.

(iii) A group of knowledgeable persons must investigate the reliability of health data available for the chosen localities with possible deletions in lists (i) and (ii).

(iv) Presumably, another group of knowledgeable persons must investigate the availability and the comparability of data on contemplated pollutants. This is likely to result in further deletions in the two tentative lists under (i) and (ii).

(v) The results of efforts under the above four points must be somehow collated, presumably in a substantial conference.

(vi) Ordinarily, desirable things do not just happen. Also, ordinarily, they cost money. If the CSHP study is to be performed, whether conforming with the above skeleton plan or in some other way, a small group of interested persons must pick up and carry the ball, including budgeting, search for funds and all the innumerable organizational details.

In the end, the conclusion may be that the CSHP study is now impossible or that it is too messy to attempt!

◇ ◇ ◇ ◇ ◇

A Postscript. In reply to a question, I wish to make it clear that my generally favorable reference to the paper by Hickey *et al.* does not imply my endorsement of the conclusions they reached. Specifically, I question the statistical methodology leading to the conclusion that, in cases specified, some of the pollutants, like SO_2 and NO_2, tend to increase mortality and some other pollutants, like Cu, tend to decrease it. The conclusions may or may not be true but the statistical analysis which led to them is in my opinion invalid.

REFERENCES

[1] E. B. HOOK, "Monitoring human birth defects and mutations to detect environmental effects," *Science*, Vol. 172 (1971), pp. 1363–1366.

[2] C. L. CHIANG, *Introduction to Stochastic Processes in Biostatistics*, New York, Wiley, 1968.

[3] R. J. HICKEY, D. E. BOYCE, E. B. HARNER, and R. C. CLELLAND, "Ecological statistical studies concerning environmental pollution and chronic diseases," *IEEE Trans. Geosci. Electron.*, Vol. GE-8 (1970), pp. 186–202.