The Use of Hospital Data In Studying the Association Between a Characteristic And a Disease

By ARTHUR S. KRAUS, M.S.

WHEN STUDYING the possible association between a characteristic and a disease, it is common practice to compare a group of hospital patients suffering from the disease with a control group in the same hospital suffering from other diseases. The objective is to find the proportion of each group possessing the characteristic in question. The control group is usually limited to diseases that can reasonably be assumed to have no relationship with the characteristic. When the characteristic occurs significantly more often in the hospital group with the disease in question, it is concluded that the characteristic is associated with the disease.

The current conclusion that cigarette smoking is associated with lung cancer rests largely on this type of evidence from hospital data.

Berkson (*1*) has discussed the hazards in using such hospital data to reach general conclusions. White (*2*) has more recently discussed a wide range of sampling problems in medical research, including those involved in hospital samples. While these authors have correctly indicated the need for caution when drawing conclusions from hospital samples, there may exist at present an oversuspiciousness regarding all conclusions based primarily on hospital data. The study of hospital groups is usually the most convenient and often the only way to obtain information pertinent to the problem under consideration. This paper indicates some of the circumstances under which valid conclusions can be drawn from hospital data.

Mr. Kraus is a biostatistician with the New York State Department of Health.

If the Characteristic Is a Disease

Berkson illustrated the problem with a hypothetical example dealing with the possible role of cholecystic disease as a causative or aggravating agent in diabetes. Hospital diabetes cases were compared with hospital cases with refractive errors (the control group), regarding the proportion with active cholecystic disease. He showed that an apparent association between diabetes and active cholecystic disease could be found when no such association existed in the population as a whole, because of the ordinary compounding of the independent probabilities that individuals with each of the three diseases would come to a hospital. Spurious correlations would be especially likely if the probability of a diabetes case coming to a hospital differed considerably from the probability of a refractive error case coming to a hospital. However, the risk of spurious correlations was great in this problem primarily because the characteristic under study, cholecystic disease, was itself an active disease with a fairly high probability of bringing those who have it to a hospital. I agree with Berkson and with White that in a study of the possible relationship between one active disease and another active disease, each with a fairly high probability of bringing cases to a hospital, there is a considerable risk of finding a spurious correlation among hospital series of cases.

If the Characteristic Is Not a Disease

Let us consider the more frequent class of problems in which the characteristic under study is not an active disease with a symptomatology that is likely to bring those who have

it to a hospital. The possible association between the characteristic, smoking, and lung cancer or between the characteristic, physician, and heart disease are examples.

In the lung cancer problem, we want to know if the prevalence rate of lung cancer is significantly higher among smokers than among nonsmokers of comparable age and residence. Owing to the difficulties of following healthy smokers and nonsmokers until enough get lung cancer to permit a conclusion to be drawn, the problem is first approached by asking what proportion of lung cancer cases smoke in comparison to the rest of the population of comparable age. Cornfield (*3*) has shown that if the proportion of smokers among lung cancer cases is higher than among the rest of the population, it follows that the lung cancer prevalence rate among smokers is higher than among nonsmokers.

Because of the difficulties in obtaining for interview nonhospital lung cancer cases, or nondiseased controls of comparable age and residence, a hospital sample of lung cancer cases and a hospital control group is usually involved in the study. A number of such studies have shown a significantly higher proportion of smokers among lung cancer cases than among controls, no matter what disease groups were included in the control. The question arises as to whether the probabilities of lung cancer cases or of smokers coming to a hospital are such as to cause a spurious correlation among hospital groups, when no such correlation exists in the population.

Let us assume that for a population of 1 million table 1 represents the true proportion of smokers and of lung cancer cases, with no association between the two. If smoking does not influence the chance of a person being in a hospital, then hospital data will show the same relationship between lung cancer and smoking that exists in the population, as Berkson and White have both pointed out. Table 2 illustrates this situation, under the assumptions that 50 percent of the lung cancer population is hospitalized at a given time and that 0.5 percent of both the smoking and nonsmoking population with no lung cancer is hospitalized at a given time for a set of diseases obviously unrelated to smoking and thus suitable for a control group.

Table 1. Lung cancer and smoking in population X

	Smokers	Nonsmokers	Total	Percent smokers
Lung cancer cases	600	400	1, 000	60
Rest of population	599, 400	399, 600	999, 000	60
Total	600, 000	400, 000	1, 000, 000	60

Table 2. Lung cancer and smoking in hospitals serving population X

	Smokers	Nonsmokers	Total	Percent smokers
Lung cancer cases	300	200	500	60
Control cases	2, 997	1, 998	4, 995	60
Total	3, 297	2, 198	5, 495	60

In table 2, the relationship between lung cancer and smoking in a hospital group is the same as in the population of table 1, namely, there is no correlation. The important point is that our assumption that smoking per se does not change the probability of hospitalization is most probably correct. In Berkson's illustration, cholecystic disease itself was likely to bring a person to a hospital. He assumed that even if diabetes were unrelated etiologically to cholecystic disease a person who had both cholecystic disease and diabetes would have a greater chance of hospitalization than one who had diabetes alone and similarly for a person with both cholecystic disease and refractive errors. This assumption, which resulted in the possibility of spurious associations in hospital data, is untested and may or may not be true. However, the probability of hospitalization for a smoker with lung cancer is based on the nature of his disease and not on his habits of life. Thus, he should have the same chance of hospitalization as a nonsmoker, of comparable age and economic status, with the same kind of lung

cancer. Similarly, a smoker with a control disease has the same chance of coming to the hospital as a nonsmoker with the same disease.

Smoking may be etiological in various diseases. However, whether it is etiological for a certain disease or not, smoking adds nothing to the probability of hospitalization for a person with such a disease, and it certainly doesn't cause a person without disease to come to a hospital. Smoking thus belongs in the category of characteristics which have no probability by themselves of hospitalizing a person. Thus hospital data will not show any spurious correlations involving smoking, where none actually exist. Similarly, it would appear that any other characteristic which is not itself a disease condition, such as occupation, diet, physical activity, and habits of life can be studied for etiological significance in particular diseases, using hospital data, without the risk of spurious correlations of the type illustrated by Berkson. This is the type of characteristic under study in a large part of chronic disease research.

If Characteristic Is Related to Control

If we mistakenly include in the control group of diseases in the hospital sample a disease which has an unsuspected etiological relationship to the characteristic under study, we will tend to decrease the chance of demonstrating a positive correlation between the characteristic and the disease we are studying, when such an association actually exists. This is illustrated in tables 3 and 4. Table 3 represents a population with a true association between smoking

Table 3. Lung cancer and smoking in population Y

	Smokers	Nonsmokers	Total	Percent of smokers
Lung cancer cases	800	200	1, 000	80. 00
Rest of population	599, 400	399, 600	999, 000	60. 00
Total	600, 200	399, 800	1,000,000	60. 02
Prevalence of lung cancer (per 100,000 population	133. 3	50. 0	100. 0	

Table 4. Lung cancer and smoking in hospitals serving population Y

	Smokers	Nonsmokers	Total	Percent of smokers
Lung cancer cases	400	100	500	80. 0
Control cases	8, 392	1, 998	10, 390	80. 8
Total	8, 792	2, 098	10, 890	80. 7

and lung cancer. In table 4, it has been assumed that 1.4 percent of non-lung-cancer smokers in population Y are hospitalized for the diseases included in the control group, while only 0.5 percent of the non-lung-cancer nonsmokers are hospitalized with these diseases. This difference is due only to a higher prevalence rate of the control diseases among smokers, and not to smoking increasing the chance that an individual with one of these diseases will enter a hospital. It is still assumed that 50 percent of all lung cancer cases are hospitalized. By assuming that the ratio of hospital prevalence of the control diseases among smokers compared to nonsmokers was as high as the ratio of hospital prevalence of lung cancer among smokers compared to nonsmokers, we were able to completely obscure in the hospital sample the true association of lung cancer with smoking that pertained in the population. Inclusion in the control of one or two important diseases with some true but unsuspected association with smoking might thus make a true association between lung cancer and smoking appear statistically insignificant in a hospital study.

However, it would seem that in studying the association of smoking and lung cancer in hospital samples, the inclusion of diseases in the control which had an unsuspected relationship to smoking could cause error in one direction only. A true association between lung cancer and smoking might be obscured, but a spurious positive correlation should not appear if no true correlation exists. Smoking should have either no etiological significance or a positive etiological significance for any disease in the hospital control group, but it could hardly be expected to provide specific protection against any of the important diseases in that group. Therefore,

the finding of repeated significant associations between smoking and lung cancer in hospital studies indicates a true association in the population, regardless of the composition of the control groups used in the hospital studies. When studying the possible etiological relationship between some other characteristic and some other disease, by the use of hospital samples, one should be sure that the characteristic is not suspected of being a protective agent against any of the diseases included in the control. If it is, a spurious association between the characteristic and the disease under study might appear in the hospital sample.

REFERENCES

(*1*) Berkson, J.: Limitations of the application of fourfold table analysis to hospital data. Biometrics Bull. 2:47–53 (1946).

(*2*) White, C.: Sampling in medical research, Brit. M. J. No. 4849: 1284–1288 (Dec. 12, 1953).

(*3*) Cornfield, J.: A method of estimating comparative rates from clinical data. J. Nat. Cancer Inst. 11:1269–1276 (1951).

Immunization Information for International Travel

Persons planning cruises, including world cruises, touching the yellow fever infected countries, Trinidad, Venezuela, Colombia, and Honduras, are reminded by the Division of Foreign Quarantine of the Public Health Service of the necessity of possessing a valid yellow fever certificate. A valid certificate will be required at the next port of call from anyone who has touched any of these countries. In addition, anyone expecting to stay in any part of Central America should have a yellow fever vaccination for his own protection.

The yellow fever vaccination requirements are presently being strictly enforced in all Caribbean area ports on account of the appearance of yellow fever in many places where it had not been heard of for over 20 years. They are also strictly enforced by the Union of South Africa, Egypt, India, and Pakistan.

Travelers coming directly from the United States can enter some, but not all, yellow fever infected areas without presenting a yellow fever vaccination certificate, but usually they cannot leave again without receiving yellow fever vaccination. Presentation of a valid certificate is compulsory for departure from Trinidad and Colombia, and probably also from certain ports of Venezuela.

A yellow fever vaccination certificate does not become valid until 10 days after vaccination (in India and Pakistan, after 12 days). The certificate is valid for 6 years, except in Curacao, Aruba, and other Dutch possessions, where it is valid for only 4 years.