Some Statistical Considerations in the Study of Cancer in Industry

SIDNEY J. CUTLER; MARVIN A. SCHNEIDERMAN; and SAMUEL W. GREENHOUSE

Although this statistical discussion deals with cancer in industry, the general principles enunciated are applicable equally to other diseases and in any population group.

With the general agreement that there are such things as occupational cancers and that occupational cancer, in some sense, can be controlled, studies of industrial and occupational hazards are not only warranted, but are essential. However, if such studies are to be conducted in a manner to yield usable and useful results, two statistical questions must be answered at the very outset. First, how big a study must we conduct; how many people should be under inquiry; and for how long? Second, how do we get these people for study; where do we locate them; where, and how do we get information about them? This paper will attempt to answer these two questions, plus an obvious third (and nonstatistical) question: After observing a large enough group of people for a sufficiently long period to yield valid results, what recommendations for action can one make?

Determination of Study Size

The question of "how big a study" is important for two reasons. There are scientific requirements to be met, if the results are to be statistically valid, and there is thé budgetary problem to be solved. If the scientific requirements call for a large study, with many people followed for a long time, this must be known at the outset, so that sufficient money can be made available to carry it out. If sufficient money is not available to carry out a proper study, then perhaps no study should be undertaken.

From the arithmetic point of view, there are two things that determine the size of a study: How big a difference is it important to find, and how certain do we want to be that we will be correct when we conclude our study and say that a difference does (or does not) exist? By "how big a difference" we mean a difference between a selected "normal" rate and the rate for the particular industry, plant, or operation under suspicion. The age-sex-color specific rates in the general population provide a basis for computing a "normal" rate against which the rate for the plant (let us say) can be compared.

The second part of this question, "How much of a risk of arriving at an incorrect conclusion are we willing to take?", is the key to further work, once we have decided how big a difference (or really small a difference) it is essen-

The authors are statisticians associated with the National Cancer Institute, Public Health Service, U. S. Department of Health, Education and Welfare, Bethesda, Md.

This paper was presented before a Joint Session of the American Association of Registration Executives and the Epidemiology, Food and Nutrition, Industrial Hygiene, and Statistics Sections of the American Public Health Association at the Eighty-first Annual Meeting in New York, N. Y., November 12, 1953.

tial to find. It is possible to arrive at an incorrect conclusion for either of two reasons: we may conclude that the rate in the plant exceeds that of the general population, when it does not, or the converse.* In conducting our study we want protection against both types of error. If we make the first error, we may lead management into an extensive search for a nonexistent hazard. If we make the second, we will lend an air of assurance and complacency to a dangerous and hazardous situation.

The decision of "how big a difference" and "what degree of protection" against both sorts of error is a substantive decision and must be made by persons who are able to evaluate the costs in money and in lives of the alternative course of action.

Consider a specific example. It is suspected that a respiratory cancer hazard exists in an industrial plant that employs 2,500 persons. One thousand of these are production workers, all males, who are in operations in which they are exposed to the suspected cancerigenic agent. We know that in the general population of males from 15 to 65 years of age the annual incidence of respiratory cancer is about 40 per 100,000.† Economic and medical considerations have led us to decide that if the rate among the 1,000 males is 200 per 100,000, or more, a definite hazard exists.

When we have concluded our study, and if we say that the plant rate is "normal," we want to be (let us say) 95 per cent certain that it really is; i.e., in repeated trials, we would incorrectly conclude that the plant rate is above "normal" 5 per cent of the time. Similarly, if the plant rate actually exceeds the general rate by five times or more, we want 90 per cent assurance that the plant rate is higher than "normal." (The five times the general rate, and the 95 per cent and 90 per cent assurance have been selected here for illustration only.) If other conditions suggest other factors, these must be modified, of course. These are the substantive decisions that must be made at the outset of the study.

We now have sufficient data to compute a "sample size." The results of just such computations are summarized in Table 1.‡ The rate for the general population is 40 per 100,000. Column 3 of Table 1 is concerned with this rate. We want a 90 per cent assurance that we will say "above normal" when the true rate in the plant is five times the general rate. Five times the general rate data are given in column 9. We then read down column 9 until we find a value of 0.90, or more (the second line). Reading back along the second line to column 6 we find the number 6; in column 5 we find the number 2; and in column 3, we find the number 5,000. This means that we will need 5,000 person-years of observation; i.e., we will have to observe the 1,000 production workers for 5 years. At the rate in the general population, 40 per 100,000 per year, we would expect 2 cases of respiratory cancer in 5,000 person-years of observation. If we find 6 or more cases during the 5-year interval of observation, we will conclude that the plant rate is above "normal," and

* These two errors correspond to the errors of Type I and II introduced by Neyman and Pearson in the theory of testing statistical hypotheses.

† This is an estimate of age-specific incidence, based on the National Cancer Institute series "Cancer Illness in Ten Urban Areas of the United States." In our discussion, this rate serves as a norm, the basis for entering Table 1. However, in a specific investigation, it may be desirable to use a different norm, e.g., rate in the state, rate in the local area, rate in a selected industrial population, or perhaps even a rate using the "unexposed" plant population as a control. This last procedure imposes additional technical problems which require the development of tables not included here.

‡ Table 1 is constructed to provide information on sample size when the general population rates are 10, 20, 40, or 80 per 100,000, and when the degree of assurance desired against falsely saying "above normal" is 95 per cent. Tables, of course, can be constructed for other rates and for other degrees of assurance.

Table 1—Determination of Sample Size

Number of Person-Years Yielding the Expected Number of Cases Given in Column 5, for a Selected Set of Rates:				Expected Number of Cases at General Population Rate	Minimum Number of Cases Required to Conclude that Plant Rate Exceeds General Population Rate*	Probability of Concluding that Plant Rate Exceeds General Population Rate When in Fact It Is:		
Rate per 100,000 Population								
10	20	40	80			Twice as High	Three Times as High	Five Times as High
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
10,000	5,000	2,500	1,250	1	4	0.14	0.35	0.73
20,000	10,000	5,000	2,500	2	6	0.21	0.55	0.93
30,000	15,000	7,500	3,750	3	7	0.39	0.79	0.99
40,000	20,000	10,000	5,000	4	9	0.41	0.84	†
50,000	25,000	12,500	6,250	5	10	0.54	0.93	†
60,000	30,000	15,000	7,500	6	11	0.65	0.97	†
80,000	40,000	20,000	10,000	8	14	0.73	0.99	†
100,000	50,000	25,000	12,500	10	16	0.84	†	†
120,000	60,000	30,000	15,000	12	19	0.87	†	†
140,000	70,000	35.000	17,500	14	21	0.93	†	†
160,000	80,000	40,000	20,000	16	24	0.94	†	†
180,000	90,000	45,000	22,500	18	26	0.97	†	†
200,000	100,000	50,000	25,000	20	29	0.97	†	†

* If the minimum number of cases is observed, the probability of incorrectly concluding that the plant rate exceeds the population rate, when in fact it does not, is less than 0.05.

† Greater than 0.995.

we will make this conclusion with the required degree of assurance (95 per cent).

Consider another example: Say that a difference of three times the general rate is important to find; we want 95 per cent assurance against falsely crying "wolf," and we want 90 per cent assurance that we will find an "above normal" rate if so large a difference does exist. Column 8 in Table 1 is the "3 times" column, and we look down this column until we find a value of 0.90 or more. This occurs on the fifth line. In column 3, the 40 per 100,000 column, the number is 12,500, and in column 6, the number is 10. Thus, we would need 12,500 person-years of observation. If we found 10 or more cases during this period of observation, we would again conclude that the plant rate exceeds the general population rate with the required degree of assurance.

Having decided how large a study must be conducted, in terms of the number of person-years of observation required, it is now important to consider how we will get this population for study.

Selection of Employee Population

In studying the incidence of cancer among members of an industrial population, it is essential that the study group be made up of individuals employed for a period long enough to constitute effective exposure to the suspected cancer hazard and that these individuals be followed for a long enough period so that the effects of exposure may manifest themselves in a diagnosable condition. Changes in methods of manufacture present a complicating factor, because an individual with a particular job title may have experienced different types of exposure as manufacturing methods changed. An individual's work experience prior to employment in the plant under investigation also presents a complicating factor.

It is desirable to restrict the study group to categories of employees coming into intimate contact with the suspected cancer hazard. Unless the suspected cancerigenic agent affects the general atmosphere of the plant, the inclusion of administrative, warehouse, and other categories of employees who are not directly exposed to the suspected hazard will materially dilute the apparent effects of the cancerigenic agent and the existence of a hazardous operation may be missed in a plant which is generally not hazardous. For example, let us assume that out of a total of 5,000 male employees, 500 are directly exposed to a cancer hazard. The incidence rate of respiratory cancer for these 500 employees is 5 times normal (200 per 100,000 per year), while the incidence rate for the remaining 4,500 employees is normal (40 per 100,000). With no special cancer hazard in this plant, an average of 2 respiratory cancer cases would be diagnosed annually among the 5,000 employees. In this particular plant, however, we would expect an average of 3 cases per year—2 cases drawn from the population of 4,500 employees not coming in contact with the cancer hazard, plus one case drawn from the population of 500 exposed employees. However, an observation of 3 cases compared to an expectation of 2 cases would not be statistically significant. Thus, data based on the experience of the total population of male employees would not point to the existence of a hazard though one does exist.

If we had knowledge which could lead us to restrict the study to the employees directly exposed to the suspected cancerigenic hazard, the size of the study group would be reduced to 500. Under the conditions stated above, in one year, we can expect one case of respiratory cancer. However, the observation of just one case would not permit us to reach a statistically sound conclusion concerning the existence of a cancer hazard. It would be necessary to keep the study group under observation for a number of years, as the data in Table 1 show.

Collection of Data

Having defined the employee population to be studied, one must establish a procedure for collecting the necessary information. A number of alternative methods are possible.

Company Medical Records—Medical records maintained in some industrial plants may be sufficiently complete to serve as the primary source of information concerning the incidence of specific disease categories among employee groups. In some instances, plant medical records may be supplemented by information available from medical care programs and insurance carriers. Under this procedure, the records for some period of time are reviewed and the number of diagnosed cancer cases are enumerated. An incidence rate is computed by relating the number of cancer cases to the average number of persons employed during the study period. This rate may then be compared to the rate for a "normal" population, with appropriate adjustments for the composition of the employee group with respect to race, sex, and age.

This procedure is valid only if the medical records contain information on every case of cancer diagnosed among all members of the employee group under study. As a minimum, the medical records would have to be sufficiently complete so that all episodes of illness which may have led to a diagnosis of cancer can be selected for follow-up. However, even if the company's medical records provide a complete and reliable source of information, relatively rapid employee turnover, coupled with an extended latent period, may result in a gross understatement of the cancer incidence rate.

Company Personnel Records—Since many plants do not have comprehensive medical care and health insurance programs, an alternative approach would be to obtain a list of all persons dropped from the payroll due to illness. These persons are then followed in order to determine the nature of the illness. This procedure is predicated on the assumption that the company records can be used to identify every employee who stopped working due to illness. In practice, even if the company records reasons for termination of employment, the plant records may appreciably understate the number of employees that stopped work due to illness. Persons discontinuing work due to ill health may not inform the company unless they benefit by having this information on record.

This approach requires that every employee dropped from the payroll due to illness be followed so that a definitive diagnosis may be determined. Consequently, if a company's records can serve as a reliable source of information concerning employees stopping work due to illness, an extensive follow-up program would have to be carried out. This involves locating and interviewing the employees who stopped work, visiting individual physicians, examining hospital and clinic records, and reviewing death certificate files. Employees who left the community subsequent to termination of employment may be lost to follow-up.

Follow-Up of Persons Employed as of a Given Date—The most direct method for doing a plant survey is to obtain a list of employees on the payroll as of a given date. Individuals who were employed for some minimum length of time in the suspected operation can then be selected for follow-up in order to determine the number that developed cancer during a specified period. The group to be followed may be selected in one of two ways: (1) employees on the payroll as of a current date may be followed forward, or (2) employees on the payroll as of a prior date may be followed through a current or future date. In either case, it is necessary to follow up every (or nearly every) employee in the study group in order to identify the ones who developed a serious illness during the study period and to ascertain the nature of the illness.

It must be emphasized that the employee population to be studied should be made up of individuals employed for a period long enough to constitute effective exposure to the suspected cancer hazard and that these individuals must be followed for a period long enough for the effects of exposure to manifest themselves. From this point of view, starting with a payroll list of 5, 10, 15, or 20 years ago would seem advantageous. On the other hand, a large proportion of individuals employed on this prior date may have left the plant and the community and be lost to follow-up. It is undoubtedly easier to follow members of the study group if the study is carried forward from a current date. The decision will depend on information concerning the rate of employee turnover, the average period of exposure to the suspected hazard, and estimates of the latent period for the form of cancer under study.

Evaluation of Results

Assume that funds were available for conducting a study of the proper size,

Table 2—Interpretation of Observed Number of Cases

Observed Number of Cases	Minimum Number with Probability of:			Maximum Number with Probability of:		
	0.05	0.10	0.20	0.05	0.10	0.20
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	*	*	*	4.8	3.9	3.0
2	*	*	*	6.3	5.4	4.3
3	*	1.1	1.5	7.8	6.7	5.6
4	1.3	1.7	2.2	9.2	8.0	6.8
5	1.9	2.4	3.0	10.6	9.3	8.0
6	2.6	3.1	3.9	11.9	10.6	9.1
7	3.2	3.8	4.7	13.2	11.8	10.3
8	3.9	4.6	5.5	14.5	13.0	11.4
9	4.6	5.4	6.4	16.0	14.3	12.6
10	5.4	6.2	7.2	17.0	16.0	13.7
11	6.1	7.0	8.1	19.0	17.0	14.8
12	6.9	7.8	9.0	20.0	18.0	16.0
13	7.6	8.6	9.9	21.0	19.0	18.0
14	8.4	9.4	10.7	22.0	21.0	19.0
15	9.2	10.2	11.6	24.0	22.0	20.0
16	10.0	11.1	12.5	25.0	23.0	21.0
17	10.8	11.9	13.4	26.0	24.0	22.0
18	11.6	12.8	14.3	27.0	25.0	23.0
19	12.4	13.6	16.0	28.0	26.0	24.0
20	13.2	14.5	16.0	30.0	28.0	25.0

* Less than 1

that the employee population selected for study was successfully followed, and that a given number of cases of (let us say) respiratory cancer have been observed. For this observed number of cases we can compute probable upper and lower limits on the actual rate in the plant. These are analogous to confidence limits on a simple average. Table 2 presents 90, 80, and 60 per cent confidence limits for observations of from one to 20 cases.* In the first example given above assume that 8 cases were found in 5,000 person-years of observation. The best estimate of the plant rate is, of course, 160 per 100,000. To obtain 90 per cent confidence limits on the rate read down column 1 in Table 2 to the number 8, the number of cases found. Column 2 gives minimum number of cases and column 5 maximum number of cases, each with probabilities of 0.05. The numbers in these columns opposite 8 are 3.9 (minimum) and 14.5 (maximum). These numbers are associated with rates of 78 and 290 per 100,000. If 4 cases were observed in the same plant population, the best estimate of the rate in the plant is 80 per 100,000, with 90 per cent confidence limits of from 26 to 184 per 100,000.

From a knowledge of these limits, the evaluation of the results of the investigation is greatly enhanced. The confidence limit concept can be used to arrive at meaningful decisions about

* Confidence limits have traditionally been quoted with the probability that the interval includes the true population value. However, Table 2 gives the probability that the true population value will be either below the lower limit or above the upper limit of the interval.

further work. If the number of observed cases is smaller than the critical number given in Table 1, we will conclude that the plant rate appears to be compatible with the general population rate. However, it does not necessarily follow that the plant is perfectly safe and that the investigation should be dropped, for the question might be asked: "Given the number of cases found, what is the maximum possible rate, on a probability basis, that might exist in this plant and still produce as few cases as we found?"

Conversely, if the observed number is greater than the critical number, we will conclude that the plant rate is above "normal." However, this does not mean that the situation is necessarily serious enough to warrant the expenditure of large sums on safety measures. We would like to know whether the plant rate is substantially greater or only slightly greater than normal. This leads to the question: "Given the number of cases found, what is the smallest possible rate, on a probability basis, that might exist in the plant and produce as many cases as we found?"

As an illustration, we refer to the previous example cited. Say eight cases were found in 5,000 person-years of observation. This is more than 6, the critical number given in Table 1, column 6, and we conclude that the plant rate is "above normal." Our best estimate of the plant rate, 160 per 100,000 is 4 times the general population rate of 40 per 100,000. However, 8 cases in 5,000 person-years of observation could have arisen out of a smaller basic rate than 160 per 100,000. We now wish to use Table 2 to compute a possible minimum rate on a probability basis. Suppose that we wish to be very cautious before undertaking corrective action; we therefore would want to give our plant every opportunity to be called nonhazardous. We would then look under the 0.05 column (column 2) to find the lowest rate that could, with a 5 per cent probability, yield as many as 8 cases. The number in this column, opposite 8, is 3.9. This implies a rate of 78 per 100,000. We can now ask: "Is corrective action warranted, if the plant rate is really as low as 78 per 100,000, which is less than twice the rate in the general population (40 per 100,000)?"

It must be remembered that we have stacked the cards against a decision favoring corrective action by taking a minimum with a probability of only 5 per cent. If we wanted the smallest rate that could produce as many as 8 cases, 20 chances in 100, we would look under column 4 of Table 2. This would yield a minimum rate of 110 per 100,000 to compare with the general population rate of 40.

The opposite situation must be considered, too. Suppose the study has turned up fewer than the critical number of cases given in Table 1, column 6. Assume that in the preceding example only 4 cases of respiratory cancer were observed in 5,000 person-years of observation. Although the observed number, 4, is twice the number expected at the general population rate, 2, it is less than the minimum number of cases, 6, required to conclude that the plant rate is "above normal." Are we safe in concluding that no hazard exists, or should the investigation be continued? This question can be answered by determining, in some probability sense, the maximum plant rate that could have produced 4 cases. This determination can be made by referring to columns 5, 6, and 7 of Table 2. If a hazard exists, we want to protect ourselves against understating the possible maximum, so we select the 0.05 level (column 5) on which to work. The number in column 5 opposite the number 4 (the actual number of cases observed) in column 1, is 9.2. Dividing this by the expected number of cases, 2, and multiplying by

40 per 100,000, yields a rate of 184. Is a plant with a respiratory cancer rate that is possibly (with 0.05 probability) 4.6 times as high as the general rate worth further investigation? This is, of course, a substantive question, and each separate investigation will have to be considered on its own merits.

Technical Note

It is assumed that the chance distribution of the observed numbers of cases for a given incidence rate follows the Poisson probability law. Thus, if "p" is the incidence rate to which "n" person years are subject to risk, the average number of cases is np = a. Then according to the Poisson law, the probability of observing exactly X_0 cases is

$$\frac{e^{-a} a^{x_0}}{x_0!}$$

and the probability of observing X_0 cases or more is

$$\sum_{X=X_0}^{\infty} \frac{e^{-a} a^{x}}{x!}.$$

Tables 1 and 2 depend upon the evaluation of these sums for particular values of "a," or determining "a" for a given value of the sum. For these purposes, Molina's tables on the Poisson law were used (Molina, E. C., Poisson's Exponential Binomial Limit. New York: Van Nostrand, 1947).

Medical Research Fellowships

Awards designed to offer research experience for promising individuals planning investigative careers are announced by the Division of Medical Sciences of the National Academy of Sciences-National Research Council. Applications for postdoctoral research fellowships for 1955–1956 are now being accepted and must be postmarked before December 10, 1954. Ordinarily, only persons 35 years of age or under are eligible.

The programs for which the division selects candidates include fellowships in cancer research and British-American exchange fellowships in cancer research, awarded by the American Cancer Society. The former, available to United States citizens, are for study in the biological, chemical, and physical sciences and of clinical investigation related to either typical or malignant growth. The latter are for advanced study in Great Britain. The British Empire Cancer Campaign in turn provides similar fellowships for British citizens to study in the United States.

The Medical Fellowship Board of the division administers fellowships for the Rockefeller Foundation, the Lilly Research Laboratories, and the National Tuberculosis Association. The first two are for research in the basic medical sciences. The first is open to citizens of both the United States and Canada, the second to United States citizens only. The fellowships for tuberculosis investigators are open to United States citizens who are graduates of American schools.

Further details and application blanks from Fellowship Office, National Academy of Sciences-National Research Council, 2101 Constitution Ave., N. W., Washington 25, D. C.