The Estimation of the Probability of Developing a Disease in the Presence of Competing Risks

JEROME CORNFIELD

For the reader who is—shall we say—not comfortably at home in mathematics, the appearance of the illustrative computation in this discussion may tempt him to turn to the next paper. No matter what his interests, one cannot help but gain from this presentation a wholesome caution about accepting seemingly obvious assumptions.

The life table provides a systematic record of the rate at which members of a cohort withdraw from that cohort as it ages. The cohort can consist of any collection of elements having one or more characteristics in common, with withdrawal taking place whenever an element loses a key characteristic. In public health applications the elements composing the cohort are live human beings, but the common characteristic and reason for withdrawal may vary from problem to problem. Thus, the characteristic might consist simply of being alive and having been born at a certain time, as in many demographic problems, or it might consist of being free of a specified disease after having been treated for it. Withdrawal in the first case would then occur at death, in the second case either after recurrence of the disease or at death.

In some problems it is desirable to subclassify the elements by reason for withdrawal. Thus, when withdrawal takes place because of death, we may classify the deaths by cause; when, as in the second example, withdrawal takes place because an element is no longer alive and disease-free the withdrawn elements may be classified as either (a) alive and disease present, or (b) dead with disease present at time of death, or (c) dead with disease absent at time of death. In any problem in which the key characteristic, whose loss leads to withdrawal, can be subdivided into two or more mutually exclusive categories, we say, speaking purely formally, that these categories compete with each other. Thus, diabetes and coronary heart disease are competing causes of death. The existence of competing risks raises certain problems in the calculation and interpretation of life table probabilities that do not arise when we are concerned only with the fact of withdrawal and are not concerned with subclassification by reason for withdrawal.

Before indicating what these problems are, let me make clear one thing that they are not. The problems arising from the existence of competing risks are in no way related to the important, but independent, set of problems arising in coding causes of death when two or more causes are simultaneously present. In the usual joint cause problem we should like to assign all deaths either to disease category A, or to B, or to C, and so forth, but in some cases death occurs in the presence of two or more diseases, say, A and B, and a decision must be

Mr. Cornfield is mathematical statistician, Biometrics Branch, Division of Research Services, National Institutes of Health, Bethesda, Md.

This paper was presented before a Joint Session of the American Association of Registration Executives and the Statistics Section of the American Public Health Association at the Eighty-Fourth Annual Meeting in Atlantic City, N. J., November 13, 1956.

made as to which category to assign the death. The way in which such decisions are made can have important effects on the results, but no matter how they are made a problem of competing risks remains. This is true whether all deaths for which both A and B are present are assigned entirely to A, or to B, or to a new category called "both A and B." A problem of competing risks exists whenever withdrawals can be subdivided into a set of mutually exclusive classes, and its existence is no way dependent on how these classes are defined.

The nature of this problem is perhaps most easily seen by considering an oversimplified version of a common laboratory experiment. We have a toxic agent whose administration to some laboratory animal will result in the development of a particular disease during some specified time interval. We wish to estimate the probability that an animal exposed to this agent will develop the disease during this time. It is a simple enough matter to design an experiment in which a cohort of animals is exposed to the agent and the number developing the particular disease during the time interval observed. But, suppose, as often happens, that the time interval is a lengthy one. In experiments on tumor induction, for example, six or more months may elapse between the initial application of the carcinogen and the appearance of a tumor. In an interval of this length a number of untoward events may occur; in particular, animals may die of other causes. Now these other causes of death compete with the development of the disease of interest in two quite different senses: (a) In a purely formal sense if an animal dies from some extraneous cause it has no chance to develop the disease in which we are interested. In this formal sense the presence of other causes of death must decrease the observed proportion developing the disease of interest; (b) in an empirical (as contrasted with formal) sense the animals dying of extraneous causes may have a probability of developing the disease which differs from that of other animals.

Now if we wish to eliminate the empirical effects of competing causes of death we have no alternative but to conduct the experiment in such a way that these other competing causes are not present. This is not always easy to do, and in many cases is completely impossible. In such cases the probability of developing the disease of interest is defined only in the presence of these other competing causes. But there is a formal effect still present and one we must recognize in interpreting the results.

Consider for example, a hypothetical experiment designed to compare the probabilities that an agent will cause the development of a disease in young and old experimental animals. We may, by taking sufficient care, reduce the probability that young animals will develop any disease but the one of interest close to zero; in general it will be impossible to do this for older animals, some of which will die during any protracted experiment simply because they are older. We may be unable to do anything about this experimentally, but we must recognize that for this reason alone the older animals will appear to have a lower probability of developing the disease of interest. Thus, if the weekly incidence rate for a surviving young animal for the disease of interest was 0.1, while the equivalent rate for an old animal was twice as high, 0.2, it is an easy calculation to show that nevertheless the older animals would have a smaller proportion developing the disease over the course of seven or more weeks, if their weekly mortality rate from other causes was also 0.2. Thus,

$$\text{Proportion of younger animals developing the disease} = \int_0^7 0.1 \times e^{-0.1t}\, dt = 0.503$$

$$\text{Proportion of older animals developing the disease} = \int_0^7 0.2 \times e^{-0.1t}\, dt = 0.470$$

In the face of difficulties such as this experimenters have sought a numerical way of "correcting" their results for the disturbing effects of competing causes. In the study of dose-response curves we have Abbott's correction for natural mortality,[1] i.e., mortality even in the absence of any dose of the toxic agent. In the study of experimental carcinogenesis several methods have been developed [2, 3] including a direct application of life table methods by J. O. Irwin.[4] In all cases the purpose, although not always made explicit, has been to estimate the probabilities that would have been observed if it had been possible to eliminate the other causes of death without affecting the short-term probabilities for the disease of interest.

At this point it is convenient to introduce some terminology. We shall call the probability of developing the disease of interest during a time interval for a cohort subject to competing risks a mixed probability. We shall call a pure probability that idealized probability that would have been observed had it been possible to eliminate all competing risks, and if these competing risks had no empirical effects.*

Now, when we turn to human populations we find the same problems as in animal experimentation but, since our objectives are usually more complicated than the experimenters, things are not as clear cut. In experimental studies one is usually interested in isolating effects, i.e., in analysis, whereas in studying human populations we are interested both in isolating effects, and in studying them in combination, i.e., in both analysis and synthesis. The experimenter wants a pure probability; the epidemiologist may want both a pure and a mixed probability. In the usual problem of estimating the probability that a human being will develop a specific disease between two specified ages, we proceed by (a) computing the life table implied by current mortality rates, (b) applying incidence rates for the disease to the surviving population at each age, (c) summing the new cases developing between the stated-ages, and (d) dividing that total by the number of survivors at the initial age. The estimate obtained is of a mixed probability. It provides an answer to the question: In a cohort subject for some future number of years to both the pure risk of developing the disease and to the pure risk of dying from some other cause what proportion will develop the disease in question?

If we are interested in isolating effects, however, and wish to study, say, changes in the pure risk of developing a disease, without regard to changes in other causes of death, such a proportionate frequency may be misleading. Thus, if such a calculation tells us that the probability of developing cancer is higher now than it was in the past, this may be either because the pure risk of developing cancer has increased, or because the chance of dying of other causes has decreased. Nor are these difficulties entirely hypothetical. In a recent study of the changing probability of developing cancer in Upstate New York, for example, it was found that the probability of a female developing cancer at some time during her life had increased by 25 per cent over a seven-year period, but that the largest part of this increase was accounted for by decreases in other forms of mortality.[7] From a control point of view the mere decrease in other forms of mortality

* The mixed and pure probabilities are identical with what Fix and Neyman [5, 6] have called crude and net probabilities in a problem in long-term survival involving competing risks. Because net and crude have different meanings in demography than those assigned by Fix and Neyman as well as the fact that it is mnemonically awkward to have the net quantity the larger one, I have used pure and mixed.

makes cancer control a more pressing problem, even in the absence of any change in the pure probability of its development. The synthetic measure is thus a useful one, but we cannot use it without some modification if our purpose is the purely analytic one of measuring the effects of changing risks, in the absence of changes in other causes.

There is no reason, of course, why one could not compare two different periods or two different places by standardizing for the competing causes of death and this is, in fact, what was done in the New York State study. For comparative purposes this provides a useful analytic measure, although the—at least theoretical—difficulty of selecting a standard set of mortality rates for competing causes of death remains. But it does not provide an absolute measure which is independent of competing causes of death and this is the question to which we now turn.

The only thing new about a pure probability is its name. The basic analytic apparatus necessary for making the estimates was developed by Bernoulli and D'Alembert as part of a consideration of the effect of eliminating smallpox as a cause of death on the probability of surviving to a given age. A good summary of their method is given by Todhunter [8] (who finds nothing of "special interest" in D'Alembert's contribution and regards it as "not of practical use") and Karns [9] (who bases her methods on D'Alembert's modification of Bernoulli's procedure). When, 100 years later, the British actuary, W. M. Makeham, was considering theoretical problems involved in preparing life tables for populations subject to multiple sources of decrement he drew upon and simplified the earlier work [10] so that the Bernoulli-D'Alembert-Makeham method provides the basis for current actuarial methods of handling this problem. Good contemporary accounts are given by Jordan [11] and Bailey and Haycocks.[12]

We shall not require much of this theory for our purposes. We start by defining the survivorship curve for a cohort, $l(X)$, the number of survivors at age X. Withdrawal from the cohort can occur either because of death, or the development of a particular disease. X is considered a continuous variable, and the interval of time from X to $X+1$ is referred to as unit time. The age at which the cohort starts is denoted by X_0. Corresponding to any increment of time, $\triangle X$, we shall have some corresponding change in the survivorship curve, say $\triangle l(X)$. Since $l(X)$ is a decreasing function of age, the negative of $\triangle l(x)$ gives the number of withdrawals during the interval. Specify N mutually exclusive causes of withdrawal and denote the number attributable to the ith cause by $-\triangle_i l(X)$. Then

$$(1) \qquad -\triangle l\,(X) = -\sum_{i=1}^{N} \triangle_i\, l(X).$$

The total probability of withdrawing during the interval is $-\triangle l(X)/l(X)$, while the mixed probability of withdrawing because of the ith cause is $-\triangle_i l(X)/l(X)$. We denote the former by $q(X,X+\triangle X)$ and the latter by $q_i(X,X+\triangle X)$. The complements, i.e., the total and mixed probability of surviving, are denoted by $p(X,X+\triangle X)$ and $p_i\ (X,X+\triangle X)$. The total withdrawal rate per unit time for the interval X is then given by

$$(2) \qquad -\frac{\triangle l(X)}{\triangle X l(X)}.$$

Now as the interval of time, $\triangle X$, is shortened the total rate so defined will change slowly and as $\triangle X$ approaches zero the rate itself may approach a limit. This limit is referred to by the actuaries as the total force of decrement per unit time at age X. We denote it by $\mu\ (X)$. Thus, when this limit exists

$$(3)\qquad \lim_{\Delta X \to 0} \frac{-\Delta l\,(X)}{\Delta X l\,(X)} = \frac{-dl\,(X)}{dX l\,(X)} = \mu(X).$$

Integrating both sides of (3) we obtain the fundamental relationship between the survivorship curve and the force of decrement,

$$(4)\qquad l(X) = l(X_o)\ e^{-\int_{X_o}^{X} \mu(t)\,dt}.$$

We may define a partial force of decrement for the ith cause of withdrawal just as we did for all causes, and if we denote it by $\mu_i(x)$, we have

$$(5)\qquad \lim_{\Delta X \to 0} \frac{-\Delta_i l(X)}{\Delta X l(X)} = \mu_i(X).$$

In view of (1) the sum of the partial forces is equal to the total force, i.e.,

$$(6)\qquad \mu(x) = \sum_{i=1}^{N} \mu_i(x).$$

With this machinery we may now make precise the notion of the pure probability of developing a disease. We consider a hypothetical cohort for which all causes of withdrawal but the one of interest, say the ith, have been eliminated without affecting the partial force of decrement for that cause. Then, from (4) the survivorship curve for that cohort, which we denote by $l_i(x)$ is

$$(7)\qquad l_i(x) = l_i(X_o)\ e^{-\int_{X_o}^{X} \mu_i(t)\,dt}.$$

The pure probability that an element of a cohort subject only to the ith cause of withdrawal will survive from ages X_1 to X_2 we denote by $p^*_i\ (X_1, X_2)$. Clearly,

$$(8)\qquad p^*_i(X_1,X_2) = \frac{l_i(X_2)}{l_i(X_1)} = e^{-\int_{X_1}^{X_2} \mu_i(t)\,dt}.$$

The pure probability of withdrawing for the ith cause is then $1-p^*_i\ (X_1, X_2)$.

The pure probabilities as defined have a very interesting property. The actual probability of surviving from age X_1 to age X_2 is the product of the individual pure probabilities. Thus, the individual pure probabilities of surviving combine as if they were independent of each other. This is more than a computational convenience. It provides additional insight into the effect of the individual causes on the total death rate. The proof of this multiplicative relation is straightforward.

$$\prod_{i=1}^{N} p^*_i(X_1,X_2) = e^{-\sum_{i=1}^{N} \int_{X_1}^{X_2} \mu_i(t)\,dt}$$

$$= e^{-\int_{X_1}^{X_2} \sum_{i=1}^{N} \mu_i(t)\,dt}$$

$$= e^{-\int_{X_1}^{X_2} \mu(t)\,dt}$$

$$= p(X_1, X_2)$$

For the actual computation of a pure probability we must find a numerical approximation to the required integral. The usual raw material will be a central withdrawal rate for an interval for a particular cause. If we apply the usual methods and treat these rates as if they were the only rates in a single decrement table, an $l_i(X)$ curve will result.

By how much will a pure probability differ numerically from a mixed one? There is of course no general answer to such a question but the following calculations may be of interest. Several years ago Cutler and Loveland undertook a comparison of the lifetime probabilities of developing lung cancer for different classes of smokers.[13] They did this by: (a) projecting the lung cancer incidence rate for cohort 1910 for each age up to age 80; (b) breaking this rate down among the different smoking classes on the basis of several retrospective studies relating the incidence of lung cancer to amount smoked; and (c) assuming the continued applicability of the 1950 life table for all other causes. These calcu-

lations supply estimates of the mixed probability of developing lung cancer among different classes of smokers.

Such estimates appropriately weight the chance of developing lung cancer and the chance of dying from other causes before lung cancer has had a chance to occur. They supply the appropriate factual basis for decision for those hypothetical individuals whose smoking habits are determined by consideration of consequences. They do not supply an estimate of the pure risk of developing lung cancer for persons exposed to different amounts of tobacco in the absence of competing risk. Cutler and Loveland have kindly made their data available to me. The comparison between their mixed probabilities and the pure ones calculated by me is shown in Table 1 for different smoking classes and for different ages. It will be observed that the pure and mixed probabilities of developing lung cancer by age 60 for this cohort are not very different. The age 70 probabilities differ by more, approximately 25 per cent, while the age 80 probabilities are quite different. Thus, those who smoke 25–49 cigarettes per day have a mixed probability of 0.082 of developing lung cancer by age 80, but a pure probability of 0.134. We may conclude on the basis of these estimates that the pure risk of developing lung cancer by age 80 for the smoker of a pack or more cigarettes per day is about one in eight, while the mixed risk is about one in 13.

Having worked one's way through the theory of multiple decrement tables it is easy to fall into the semantic trap of believing that a calculation which eliminates the formal effects of competing risks also eliminates their empirical effects. Thus, even as careful a writer as Jordan says that the partial force of decrement "being an instantaneous rate of decrement, is not based upon any time interval, and is not affected by the operation of competing causes."[11] Taken literally this statement means that the actual physical elimination of one competing cause would of necessity leave the force of mortality for all others unaltered. It is hard to see why this must be so, and in particular how its truth could be established by anything short of an experimental demonstration. Makeham, it is interesting to note, concluded his article[10] with a caution against this confusion.

Although this paper is concerned primarily with the formal effects of competing causes, an additional word on

Table 1—Comparison, Pure and Mixed Probability that White Males Born in 1910 and Alive at Age 40 Will Subsequently Develop Lung Cancer, by Smoking Class and Age at Which Developed

Smoking Class	Age					
	80		70		60	
	Pure	Mixed	Pure	Mixed	Pure	Mixed
	Probability x 100					
Nonsmoker	0.8	0.5	0.4	0.3	0.1	0.1
Smoker:						
Cigarettes per day						
1–14	5.6	3.4	2.8	2.1	1.0	0.9
15–24	7.9	4.8	3.9	3.0	1.3	1.2
25–49	13.4	8.2	6.7	5.2	2.3	2.1
50 or more	21.6	13.7	11.0	8.7	3.8	3.5

Note: The estimates of the mixed probabilities are due to Cutler and Loveland[12] and are based on observations of Doll and Hill.[14]

the empirical effects may not be out of order. Many applications of mortality and morbidity data involve the tacit assumption that the empirical effects can be disregarded. Thus, when we observe an increase in mortality from lung cancer we are likely to conclude that some new environmental influences have been introduced. But as Berkson points out this need not necessarily be the case. "It is entirely possible and even likely that at least part of the increase in death rate from lung cancer which has been recently noted is attributable to deaths in adulthood from this disease of individuals who have not been eliminated, as in former years they would have been, by death in early life from tuberculosis or some other pulmonary malady." [15] The same problem arises for comparison among different places. Thus, urban-rural differences or differences among different countries in the incidence of a particular disease may or may not be related to differences in intensity of exposure to some environmental agent of interest, but the possible empirical effects of competing risks cannot be overlooked.

With respect to actual knowledge of the magnitude of possible empirical effects of competing risks we seem to have made no advance beyond Bernoulli. One way of expanding our knowledge and insight in this area is by means of the kind of experimental epidemiology that we have not seen since the work of Greenwood, Hill, Topley, and Wilson.[16] Thus it is possible to produce both tuberculosis and lung cancer experimentally. If they could be produced in the same laboratory animal, we could obtain a direct comparison of the pure probabilities of developing lung cancer in animals that were and were not exposed to the risk of mortality from tuberculosis. Needless to say the bearing of these results on human experience would still remain to be established.

Finally, we can ask with John Graunt, "to what purpose tends all this laborious buzzling and groping?" I would say the purpose is to remind us that in the presence of competing risks the question, "What is the probability of developing a particular disease?" is not an unambiguous one, and depending on what was meant, may have several different answers. In particular, it is suggested that the elimination of the formal effects of competing causes by computation of a pure probability may often serve a useful purpose.

REFERENCES

1. Finney, D. J. Probit Analysis: A Statistical Treatment of the Sigmoid Response Curve (2nd ed.). London: Cambridge University Press, 1952.
2. Twort, C. C., and Twort, J. M. Suggested Methods for the Standardization of the Carcinogenic Activity of Different Agents for the Skin of Mice. Am. J. Cancer XVII:293–320 (Feb.), 1933.
3. Bryan, W. R., and Shimkin, M. B. Quantitative Analysis of Dose-Response Data Obtained with Carcinogenic Hydrocarbons. J. Nat. Cancer Inst. 1:807–833 (June), 1941.
4. Irwin, J. O., with the assistance of Nancy Goodman. The Statistical Treatment of Measurements of Carcinogenic Properties of Tars (Part I) and Mineral Oils (Part II). (Brit.) J. Hyg. 44:362–420 (May), 1946.
5. Fix, E., and Neyman, J. A Simple Stochastic Model of Recovery, Relapse, Death and Loss of Patients. Human Biol. 23: 205-241 (Sept.), 1951.
6. Neyman, J. First Course in Probability and Statistics. New York: Henry Holt, 1950.
7. Goldberg, I. D.; Levin, M. L.; Gerhardt, P. R.; Handy, V. H.; and Cashman, R. E. The Probability of Developing Cancer. J. Nat. Cancer Inst. 17:155–173 (Aug.), 1956.
8. Todhunter, I. A History of the Mathematical Theory of Probability. New York: Chelsea Publishing Co., 1949.
9. Karn, M. N. An Inquiry into Various Death Rates and the Comparative Influence of Certain Diseases on the Duration of Life. Ann. Eugenics 4:279–326 (Apr.), 1930.
10. Makeham, W. M. On an Application of the Theory of the Composition of Decremental Forces. J. Inst. Actuaries XVIII:317–322 (Oct.), 1874.
11. Jordan, C. W. Life Contingencies. Chicago: Society of Actuaries, 1952.
12. Bailey, W. G., and Haycocks, H. W. Some Theoretical Aspects of Multiple Decrement Tables. Edinburgh, Scotland: T. and A. Constable, Ltd., 1946.
13. Cutler, S. J., and Loveland, D. B. The Risk of Developing Lung Cancer and Its Relationship to Smoking. J. Nat. Cancer Inst. 15:201–211 (Aug.), 1954.
14. Doll, R., and Hill, A. B. A Study of the Aetiology of Carcinoma of the Lung. Brit. M. J. 2:1271–1286 (Dec. 13), 1952.
15. Berkson, J. The Statistical Study of Association Between Smoking and Lung Cancer. Proc. Staff Meetings Mayo Clinic 30:319–348 (July 27), 1955.
16. Greenwood, M.; Hill, A. B.; Topley, W. W. C.; and Wilson, J. Experimental Epidemiology. Medical Research Council (Gr. Br.), Special Report No. 209, 1936.