MAXIMUM UTILIZATION OF THE LIFE TABLE METHOD IN ANALYZING SURVIVAL

SIDNEY J. CUTLER, M.A., AND
FRED EDERER, B.S.
BETHESDA, MD.

(Received for publication Aug. 18, 1958)

MEASUREMENT of patient survival is necessary for the evaluation of treatment of usually fatal chronic diseases. This is particularly true for cancer. The American College of Surgeons, recognizing this, requires the maintenance of a cancer case registration and follow-up program for approval of a hospital cancer program.[1] Acceptance of survival as a criterion for measuring the effectiveness of cancer therapy is also attested to by the very large number of papers published every year reporting on the survival experience of cancer patients.

Although the proportion of patients alive 5 years after diagnosis (5-year survival rate) is the most frequently used index for measuring the efficacy of therapy in cancer, an increasing number of investigators are reporting on the manner in which patient populations are depleted during a period of time, e.g., survival curves. A popular and relatively simple technique for describing survival experience over time is known as the actuarial or life table method. Whereas the method and its uses have been admirably described by a number of authors,[2-6] one important aspect has received relatively little attention. A principal advantage of the life table method is that it makes possible the use of all survival information accumulated up to the closing date of the study. Thus, in computing a 5-year survival rate one need not restrict the material to only those patients who entered observation 5 or more years prior to the closing date. We will show that patients who entered observation 4, 3, 2, and even one year prior to the closing date contribute much useful information to the evaluation of 5-year survival.

Let us consider a group of patients entering observation continuously beginning with Jan. 1, 1946. Sometime early in 1952, we decide to analyze the survival experience of these patients to obtain a 5-year survival rate. We choose Dec. 31, 1951, as the closing date, i.e., the follow-up status and survival time of each patient is recorded as of that date.

Of the patients entering the study during the 6 years ended on Dec. 31 1951, only those diagnosed in 1946 were exposed to the risk of dying for at least 5 years.* The exposure time for patients entering in each of the calendar years is shown in Table I.

TABLE I

CALENDAR YEAR OF DIAGNOSIS	YEARS OF EXPOSURE TO RISK OF DYING
1946	5 to 6
1947	4 to 5
1948	3 to 4
1949	2 to 3
1950	1 to 2
1951	Less than 1

It might be supposed, intuitively, that the patients who entered observation from 1947 to 1951 are of no value in computing a 5-year survival rate as of Dec. 31, 1951, since each of these patients was under observation for less than 5 years. This, however, is not true. Merrell and Shulman[5] have pointed out that patients for whom less than the required number of years' survival information is available should not be discarded from the analysis. Wilder[7] has demonstrated that, through maximum utilization of the life table method, it is possible to compute reliable 5-year survival rates for a large series even when the longest possible exposure time is just short of 5 years.† The primary objective of this paper is to show how partial survival information can be included in the life table and to show how much is gained by doing so. Data from the Connecticut Cancer Register are used for illustrative purposes.‡

THE ANATOMY OF THE LIFE TABLE§

Table II provides the basic facts, as of Dec. 31, 1951, concerning 126 male patients with localized cancer of the kidney, diagnosed during the period 1946 through 1951. The cases are divided into 6 cohorts, one for each year of diagnosis. The columns of Table II are described here.

Column 1. Years After Diagnosis (x *to* $x+l$).—This column gives the time elapsed from the date of diagnosis in intervals of one year, i.e., 0-1, 1-2, etc. For example, a patient who was diagnosed Jan. 20, 1946, and died on Oct. 5, 1948, died during the third year after diagnosis, i.e., during interval 2-3. The number of patients that left observation during each interval is entered in the appropriate column (3, 4, or 5), according to the reason for removal from observation.

Column 2. Alive at Beginning of Interval (l_x).—The entry on the first line of this column indicates the number of cases alive at diagnosis, i.e., the initial number of patients in the cohort.

Column 3. Died During Interval (d_x).—

*In this example, date of entry into the study is defined as date of diagnosis. In practice, other reference dates, such as date of initiation of a particular course of therapy, may be used.

†In the series reported by Wilder, the range of exposure time was from one day to, but not including, 5 years.

‡We wish to thank Dr. Matthew H. Griswold, Director, Division of Cancer and Other Chronic Diseases, Connecticut State Department of Health, for his courtesy in making these data available.

§We borrowed the phrase "anatomy of the life table" from Pearl's[2] excellent textbook *Biometry and Medical Statistics.*

TABLE II. SURVIVAL DATA FOR SINGLE YEAR COHORTS
(126 Male Connecticut Residents With Localized Kidney Cancer; Diagnosed 1946-1951 and Followed Through Dec. 31, 1951)

YEARS AFTER DIAGNOSIS (1) x TO x + 1	ALIVE AT BEGINNING OF INTERVAL (2) l_x	DIED DURING INTERVAL (3) d_x	LOST TO FOLLOW-UP DURING INTERVAL (4) u_x	WITHDRAWN ALIVE DURING INTERVAL* (5) w_x
Patients diagnosed in 1946 (1946 cohort)				
0–1	9	4	1	—
1–2	4	—	—	—
2–3	4	—	—	—
3–4	4	—	—	—
4–5	4	—	—	—
5–6	4	—	—	4
Patients diagnosed in 1947 (1947 cohort)				
0–1	18	7	—	—
1–2	11	—	—	—
2–3	11	1	—	—
3–4	10	2	2	—
4–5	6	—	—	6
Patients diagnosed in 1948 (1948 cohort)				
0–1	21	11	—	—
1–2	10	1	2	—
2–3	7	—	—	—
3–4	7	—	—	7
Patients diagnosed in 1949 (1949 cohort)				
0–1	34	12	—	—
1–2	22	3	3	—
2–3	16	1	—	15
Patients diagnosed in 1950 (1950 cohort)				
0–1	19	5	1	—
1–2	13	1	1	11
Patients diagnosed in 1951 (1951 cohort)				
0–1	25	8	2	15

*Alive at closing date of study.

TABLE III. COMBINED LIFE TABLE AND COMPUTATION OF 5-YEAR SURVIVAL RATE
(126 Male Connecticut Residents With Localized Kidney Cancer Diagnosed 1946-1951 and Followed Through Dec. 31, 1951)

YEARS AFTER DIAGNOSIS (1) x TO x + 1	ALIVE AT BEGINNING OF INTERVAL (2)* l_x	DIED DURING INTERVAL (3)* d_x	LOST TO FOLLOW-UP DURING INTERVAL (4)* u_x	WITHDRAWN ALIVE DURING INTERVAL (5)* w_x	EFFECTIVE NUMBER EXPOSED TO THE RISK OF DYING (COL. 2 − ½ COL. 4 − ½ COL. 5) (6) l'_x	PROPORTION DYING (COL. 3 ÷ COL. 6) (7) q_x	PROPORTION SURVIVING (1 − COL. 7) (8) p_x	CUMULATIVE PROPORTION SURVIVING FROM DIAGNOSIS THROUGH END OF INTERVAL ($p_1 \times p_2 \times \ldots \times p_x$) (9) P_x
0–1	126	47	4	15	116.5	0.40	0.60	0.60
1–2	60	5	6	11	51.5	0.10	0.90	0.54
2–3	38	2	—	15	30.5	0.07	0.93	0.50
3–4	21	2	2	7	16.5	0.12	0.88	0.44
4–5	10	—	—	6	7.0	0.00	1.00	0.44‡
5–6†	4	—	—	4				

*Columns 2 through 5 of this table were obtained by summing, cell by cell, the survival data of the 6 yearly cohorts of Table II.

†This line is not needed for computing the 5-year survival rate; it is included here merely to complete the account of the initial 126 patients. Four were alive at the closing date of the study.

‡Five-year survival rate.

Column 4. Lost to Follow-up During Interval (u_x)*.—In this column we enter the number of patients whose survival status as of the closing date, Dec. 31, 1951, was unknown. The length of observation for each patient lost to follow-up is the time elapsed from date of diagnosis to date last known to be alive. Thus, a patient observed for 3 years and 4 months is entered on the fourth line, i.e., interval 3-4.

In applying the life table method it is usually assumed that *subsequent to date of last contact*, the survival experience of lost cases was similar to that of cases remaining under follow-up. In contrast, complete omission of lost cases from the analysis is equivalent to assuming that from *date of diagnosis* the survival experience of lost cases was similar to that for cases with complete follow-up information.

Column 5. Withdrawn Alive During Interval (w_x).—In this column we enter the number of patients known to have been alive on the closing date, Dec. 31, 1951. The interval during which these patients withdrew from observation depends on their date of diagnosis. For example, all patients diagnosed in 1949 and alive on Dec. 31, 1951, are recorded as withdrawals from observation during the third year after diagnosis, interval 2-3. Note that, for each cohort in Table II, zeros (symbolized by dashes) are entered in this column for all intervals but the last.

Although only one of the cohorts (1946) provided survival information for a full 5 years, we used the available information on all 6 cohorts. Table III was obtained by pooling all the information in Table II, summing cell by cell. For example, by summing the entries on the first line of Column 3 for each yearly cohort in Table II, we obtained the total of 47 cases who died within one year of diagnosis, shown in Table III.

In practice, the data for the pooled cohort of 126 cases would be tabulated directly, as in Table III, rather than by summing tabulations for 6 individual cohorts. We used the latter procedure to show how much information each of the cohorts contributed to the pooled data. For example, by comparing Tables II and III, we find that of the 5 patients known to have died in the second year after diagnosis (Table III, Line 2, Column 3), one was diagnosed in 1948, three in 1949, and one in 1950. Similarly, of the 18 patients diagnosed in 1947, 6 were alive 4 years after diagnosis; of the 21 patients diagnosed in 1948, 7 were alive 3 years after diagnosis. Thus, each cohort contributes some information to our knowledge of patient survival during a period of 5 years after diagnosis. A statistical measure of the gain in precision resulting from this procedure will be discussed later. First, however, we will explain how the basic data summarized in Columns 1 through 5 of Table III are used to compute survival rates.

COMPUTATION OF SURVIVAL RATES

The first step in preparing a life table is to distribute the deaths, losses, and withdrawals with respect to the interval in which they left observation.† This

*We are using the letter "u" to represent "untraced" cases, rather than the letter "l" which comes to mind as a symbol for "lost" cases, because "l" is a standard life table notation for "alive at beginning of interval."

†For a detailed account of the mechanics of recording and tabulating survival data, see Berkson and Gage,[4] pp. 4-5.

information is summarized in Columns 3, 4, and 5 of Table III. The sum of the entries in Columns 3, 4, and 5 equals the total number of cases in the study, which is entered on the first line of Column 2 (126 cases). Successive entries in this column are obtained according to the formula:

$$l_{x+1} = l_x - (d_x + u_x + w_x).$$

For example, the number alive at the beginning of the second year (60) was obtained by subtracting from the number alive at the beginning of the first year (126), the sum of the deaths, losses, and withdrawals during the first year (47 + 4 + 15).

The life table is completed by a series of four computations for each follow-up year (Columns 6 through 9).

Column 6. Effective Number Exposed to Risk of Dying (l_x').—It is assumed that patients lost or withdrawn from observation during an interval were exposed to the risk of dying, on the average, for one-half the interval.* For example, of the 25 patients diagnosed in 1951, 15 were alive on Dec. 31, 1951 (withdrawn alive). It is reasonable to assume that the date of diagnosis for these 15 patients was roughly equally distributed during the calendar year 1951 and that, on the average, each patient was observed for one-half year.

The effective number exposed to risk is obtained by subtracting from the number alive at the beginning of the year, one-half the sum of the number lost and withdrawn during the year. Thus,

$$l_x' = l_x - (u_x + w_x)/2.$$

Column 7. Proportion Dying During Interval (q_x).—This is also referred to as the probability of dying during the interval. It is obtained by dividing the

---

*The computing procedure given here is based on the assumption that, for cases withdrawn alive and cases lost to follow-up, survival subsequent to date of last contact is similar to that for cases with complete follow-up information. For cases withdrawn alive, this assumption introduces no bias, because there is no reason to believe that patients alive on the closing date are different from patients observed for a longer period. However, for cases lost to follow-up, this assumption may introduce a bias.

Patients lost to follow-up were alive when last observed, and whether their survival experience is better than, worse than, or equal to the survival of patients remaining under follow-up is highly speculative. For example, cancer patients may be lost to follow-up for a variety of reasons. Far-advanced cases may leave their usual place of residence to enter the household of a relative; successfully treated patients may stop reporting to the tumor clinic, because they feel that no further medical care is required. It is therefore important to keep the proportion of cases lost to follow-up at a minimum. Survival rates based on a series in which a substantial proportion of patients have been lost to follow-up are of highly questionable value, because it is impossible to determine the extent to which they are biased.

Some investigators, such as Paterson and Tod[8] recommend that lost cases be counted as dead "to avoid undesirable uncertainty. . . although (it) may result in a slight bias against the efficacy of treatment." Other investigators, such as Ryan and his colleagues[9] omit lost cases from the analysis of survival. The latter procedure involves the assumption that *from date of diagnosis* the survival experience of lost cases is similar to that of cases with complete follow-up.

We prefer the first of the several possible assumptions regarding lost cases, namely that subsequent to date of last contact their survival is similar to that for cases with complete follow-up. The complete omission of lost cases from the computation of survival rates discards available information. The assumption that lost cases died immediately after the date of last contact is contrary to fact. Registry experience with intensive field investigation of lost cases, which resulted in recovery of some, indicates that such patients often live for several years beyond the initial date of last contact.[10]

Although cases withdrawn alive and cases lost to follow-up are treated alike in the computations described here, we distinguish between the two in the life table for reasons mentioned: (1) it is important to be aware of the number of cases lost to follow-up because of their potential bias, and (2) other computational methods may treat the two groups differently.

number of deaths by the effective number exposed to risk:

$$q_x = \frac{d_x}{l_{x'}}.$$

To express as a percentage, multiply by 100.

Column 8. Proportion Surviving the Interval (p_x).—This is referred to alternately as the probability of surviving the interval, or the survival rate. It is obtained by subtracting the proportion dying during the interval from unity:

$$p_x = 1 - q_x.$$

To express as a percentage, multiply by 100.

Column 9. Cumulative Proportion Surviving From Diagnosis to End of Interval (P_x).—This is generally referred to as the cumulative survival rate. It is obtained by cumulatively multiplying the proportion surviving each interval:

$$P_x = p_1 \times p_2 \times p_3 \times \ldots p_x.*$$

Note that successive entries in this column give the 1-year, 2-year, 3-year, 4-year, and 5-year cumulative survival rates (Table III). The successive cumulative survival rates are plotted in drawing a survival curve.

Although the computations illustrated in Table III were carried out in intervals of one year after the date of diagnosis, the life table may be set up in terms of days, weeks, months, years, etc. In fact, the life table may be organized in intervals of varying length. For example, one might record experience during the first year in monthly intervals, and the experience thereafter in annual intervals. This type of presentation may be desirable when a large proportion of deaths occur during the first year. The method of computing survival rates described here may be used whatever the size of the intervals.

GAIN IN UTILIZING EXPERIENCE OF COHORTS WITH PARTIAL FOLLOW-UP

The standard error provides a measure of the confidence with which one may interpret a statistical result. Thus, the standard error of the survival rate indicates the extent to which the computed rate may have been influenced by sampling variation.† For example, by adding and subtracting twice the standard error to and from the computed survival rate, one obtains an approximate 95 per cent confidence interval. This means that in repeated observations under the same conditions the true survival rate will lie within a range of two standard errors on either side of the computed rate, an average of 95 times in 100.

Thus, the computed 5-year survival rate for male patients with localized cancer of the kidney is 44 per cent. The standard error, computed according to the method explained in the Appendix, is 6 per cent. It is therefore likely that

*This formula is based on the assumption that the various interval survival probabilities are statistically independent.

†The 126 cases of localized cancer of the kidney are in effect a sample from a population of male patients with localized kidney cancer.

An illustration of sampling variation may be drawn from baseball. A 0.250 hitter may, in four times "at bat," get one hit. Frequently, though, he will get no hits or two hits. And not too infrequently he will get three hits. If we watch a game and see a player get two hits in four times "at bat," it is difficult for us to judge how good a hitter this player really is. We have to watch this player for many games before we can get a reliable estimate of his batting average.

Survival rates are similar to batting averages in the sense that they are relative frequencies, i.e., the numerator is part of the denominator. For each hit there must be at least one time "at bat," and for each death there must be at least one case exposed to the risk of dying.

the true 5-year survival rate is not smaller than 32 per cent and not larger than 56 per cent.

Admittedly, the computed rate does not yield a very firm estimate of the true survival rate, but we must bear in mind that it was based on a series of only 126 patients and only 9 of these patients were diagnosed a full 5 years prior to the date of study. Furthermore, whereas the survival rate based on all information available on these 126 patients provides at least a rough idea of the true rate (one-third to one-half), discarding the information on the cohorts with partial follow-up information would result in an extremely unreliable estimate. This is explained in the discussion that follows.

The computing method applied to the total series of 126 cases, illustrated in Table III, can be applied to any selected portion of the group. We have therefore used it to compute a series of 5-year survival rates based on successively larger patient cohorts. A 5-year survival rate was computed for the 9 patients diagnosed in 1946, all of whom had a 5-year exposure time. We then added the 18 patients diagnosed in 1947, who had a 4-year exposure time, and computed a 5-year survival rate based on the available information for these 27 patients. This procedure was continued until the known experience of all 126 patients was utilized in estimating the 5-year survival rate. The successive rates and their corresponding standard errors are shown in the uppermost section of Table IV.

The 1946 cohort of 9 cases yielded a 5-year survival rate of 53 per cent, with a standard error of 17 per cent. The large standard error tells us that this is a very unreliable estimate; the true rate is probably between 19 and 87 per cent,* a very wide range. The combined experience of the 1946 and 1947 cohorts yielded a survival rate of 46 per cent, with a standard error of 10 per cent. Thus, the addition of information on cases with 4 full years of exposure reduced the standard error from 17 per cent to 10 per cent, a relative decrease of 43 per cent. The addition of the available information on cohort 1948 (3 full years of exposure) reduces the standard error to 7.5 per cent, etc. The utilization of all available information on all the cohorts results in a standard error of 6.0 per cent. Thus, the standard error of the survival rate based on all available information is 65 per cent less than the standard error based on cases with a full 5 years of exposure.

We then computed survival rates and corresponding standard errors for series of successively enlarged cohorts of patients for each of four additional groups of patients: kidney cancer with regional involvement, in men; localized breast cancer, in women; breast cancer with regional involvement, in women; and cancer of the lip, both sexes combined (Table IV). We did this in order to illustrate the advantage of utilizing all available experience for patient groups of varying size and with varying mortality experience. The results are shown graphically in Fig. 1. In every instance, the standard error of the 5-year survival rate based on the combined experience of cohorts 1946 through 1951 is smaller than the corresponding standard error for the 1946 cohort by at least one-third

*These are the 95 per cent confidence limits: $53 \pm 2(17)$.

TABLE IV. FIVE-YEAR SURVIVAL RATES AND THEIR STANDARD ERRORS FOR FIVE GROUPS OF CANCER PATIENTS, SHOWING THE REDUCTION IN STANDARD ERROR WITH INCREASE IN COHORT SIZE

COHORT	NUMBER OF CASES DIAGNOSED	5-YEAR SURVIVAL RATE	STANDARD ERROR OF 5-YEAR SURVIVAL RATE	PER CENT REDUCTION IN STANDARD ERROR OF 5-YEAR SURVIVAL RATE
Kidney, localized				
1946	9	0.53	0.171	—
1946–1947	27	0.46	0.098	43
1946–1948	48	0.43	0.075	56
1946–1949	82	0.43	0.064	63
1946–1950	101	0.45	0.063	63
1946–1951	126	0.44	0.060	65
Kidney, regional				
1946	11	0.18	0.116	—
1946–1947	23	0.33	0.101	13
1946–1948	30	0.28	0.091	22
1946–1949	39	0.25	0.074	36
1946–1950	43	0.23	0.069	41
1946–1951	47	0.24	0.070	40
Breast, localized				
1946	225	0.64	0.033	—
1946–1947	454	0.64	0.025	24
1946–1948	695	0.64	0.023	30
1946–1949	963	0.64	0.022	33
1946–1950	1,227	0.65	0.022	33
1946–1951	1,490	0.65	0.021	36
Breast, regional				
1946	208	0.42	0.035	—
1946–1947	443	0.38	0.025	29
1946–1948	708	0.39	0.021	40
1946–1949	967	0.39	0.020	43
1946–1950	1,239	0.39	0.020	43
1946–1951	1,531	0.39	0.020	43
Lip				
1946	61	0.71	0.060	—
1946–1947	109	0.65	0.048	20
1946–1948	169	0.68	0.042	30
1946–1949	224	0.68	0.040	33
1946–1950	283	0.68	0.040	33
1946–1951	332	0.67	0.039	35

The advantage of utilizing information on patient cohorts with less than 5 years of exposure was greater for localized kidney cancer than for the other groups. This is because: (1) particularly few cases (9) of localized kidney cancer were diagnosed in the first year (1946), compared with an average of 23 cases in each of the subsequent years; and (2) the mortality rate during the first year of follow-up (0.40) was much larger than in succeeding years (annual average of 0.07). Thus, because of this mortality pattern, the information on survival during the first year after diagnosis contributed very substantially to the information on survival over a 5-year period. Five of the 6 annual cohorts (1946-1950) contributed complete information regarding survival during the first year after diagnosis. In general, the relative gain in utilizing survival information on patient cohorts with partial follow-up information will vary directly with: (1) the relative increase in the initial size of the cohort*; (2) the relative completeness of the added survival information; and (3) the relative magnitude of the mortality rates during the first few follow-up intervals.

In cancer, and in other diseases, mortality is often relatively high shortly after diagnosis and tends to taper off thereafter. For some diseases, such as lung or stomach cancer, the patient group may be so depleted within one year that little is gained by waiting more than one year before evaluating therapeutic results. Therefore, it may frequently be unnecessary to wait until a 5-year survival rate can be computed to evaluate the effects of therapy. A 1-year, 2-year, or 3-year rate may provide important information. In other instances, survival data for only 5 years may be inadequate, because of significant changes in the mortality pattern at a later time.[11]

DISCUSSION

A category of patients with relatively few new cases per year was intentionally chosen as the principal example to illustrate the advantage of utilizing all available information for the computation of survival rates—only 126 cases of localized kidney cancer in men were diagnosed in 6 years. This was done because it is frequently desirable to describe the survival experience of relatively small groups of patients. For example, if we were interested in evaluating the survival experience of patients with localized breast cancer treated by surgery in combination with radiation, we would find that in any one year the number of patients receiving the combined therapy is small. As an illustration, only 25 of the 225 cases diagnosed in Connecticut in 1946 were treated by the combined therapy. Similarly, if survival is to be evaluated for a specific subgroup with respect to age, the number of cases per year would usually be small. Therefore, in order to increase the reliability of survival rates computed for various patient groups of clinical interest, it is important to utilize all available information.

It is of paramount importance to use all available survival information in computing survival rates if the rates are going to be used as criteria in a clinical trial. For example, a 3-year survival rate may have been selected as a criterion in a clinical trial. It may be possible to determine which of the several treatments

*See the Appendix for a discussion of effective sample size.

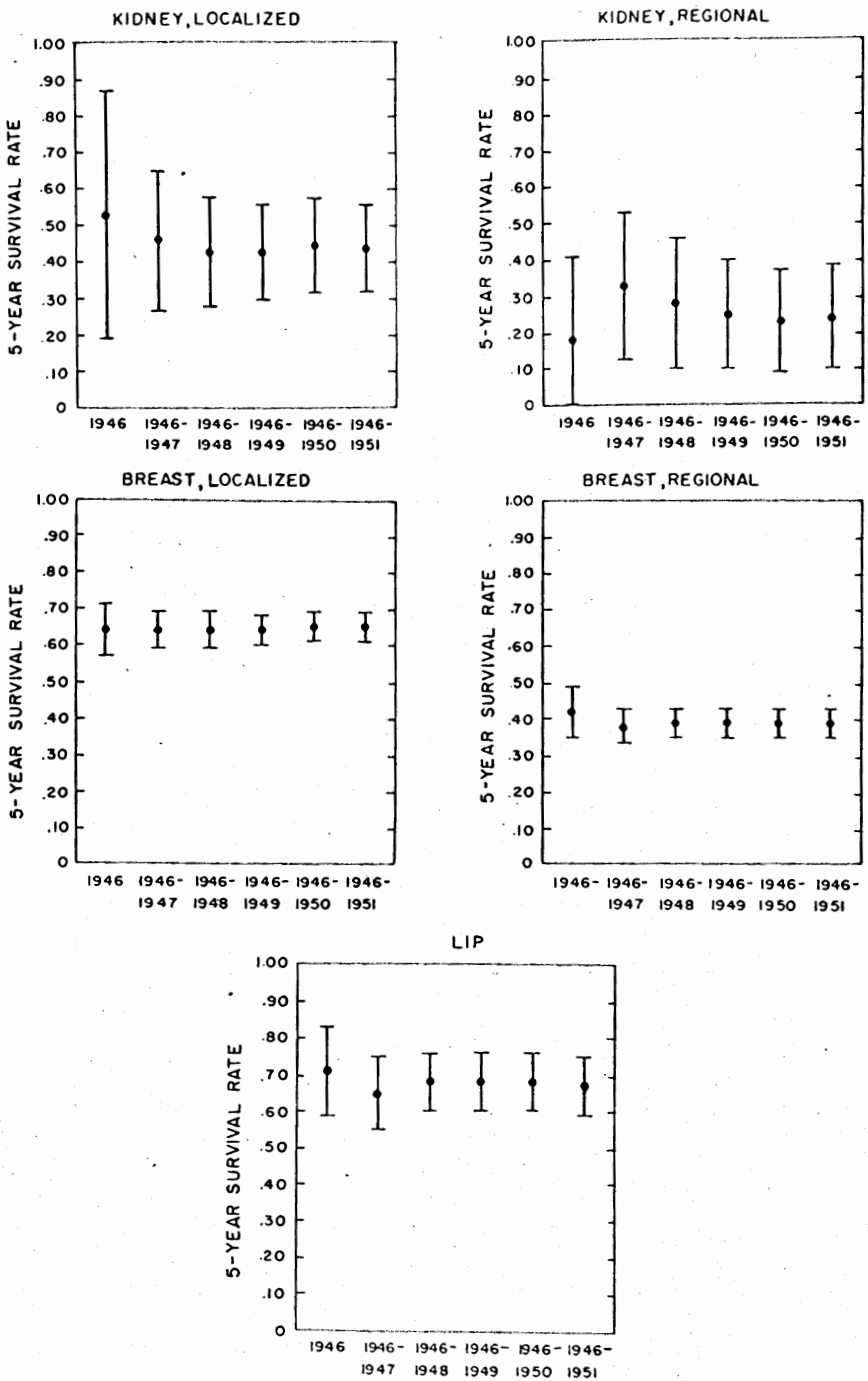

Fig. 1.—Decrease in the 95 per cent confidence interval for the 5-year survival rate as cases with less than 5 years' exposure to the risk of dying are added. (The 95 per cent confidence interval is obtained by adding ±2 standard errors to the survival rate.)

Source: Table IV.

being tested yields the best survival before all patients have been followed to death or for a full 3 years. Thus, inferior treatments would be discontinued at the earliest possible point.

SUMMARY

We have illustrated the life table method for computing survival rates with 5-year survival data for cancer patients, emphasizing the advantage gained by including survival information on cases which entered the series too late to have had the opportunity to survive a full 5 years. The advantage is measured in terms of reduction in standard error of the survival rate. For the five series of patients in this paper, the reduction in standard error ranged from one-third to two-thirds.

REFERENCES

1. Manual for Cancer Programs, Bull. Am. Coll. Surgeons **38**:149, 1953.
2. Pearl, R.: Introduction to Medical Biometry and Statistics, Philadelphia and London, 1923, W. B. Saunders Company.
3. Hill, A. B.: Principles of Medical Statistics, New York, 1956, Oxford University Press.
4. Berkson, J., and Gage, R. P.: Calculation of Survival Rates for Cancer, Proc. Staff Meet., Mayo Clin. **25**:270, 1950.
5. Merrell, M., and Shulman, L. E.: Determination of Prognosis in Chronic Disease, Illustrated by Systemic Lupus Erythematosus, J. CHRON. DIS. **1**:12, 1955.
6. Griswold, M. H., Wilder, C. S., Cutler, S. J., and Pollack, E. S.: Cancer in Connecticut, 1935-1951, Hartford, 1955, Connecticut State Department of Health, pp. 112-113.
7. Wilder, C. S.: Estimated Cancer Survival Rates Confirmed, Connecticut Health Bulletin **70**:217, 1956.
8. Paterson, R., and Tod, M. C.: Presentation of the Results of Cancer Treatment, Brit. J. Radiol. **23**:146, 1950.
9. Ryan, A. J., et al.: Breast Cancer in Connecticut, 1935-1953, J.A.M.A. **167**:298, 1958.
10. Griswold, M. H.: Personal communication.
11. Cutler, S. J., Griswold, M. H., and Eisenberg, H.: An Interpretation of Survival Rates: Cancer of the Breast, J. Nat. Cancer Inst. **19**:1107, 1957.
12. Greenwood, M.: Reports on Public Health and Medical Subjects, No. 33, Appendix 1, The "Errors of Sampling" of the Survivorship Tables, London, 1926, H. M. Stationery Office.

APPENDIX

Computing the Standard Error of the 5-Year Survival Rate.—The method for computing the standard error of the 5-year survival rate was developed by Greenwood (see ref. 12) and is also described by Merrell and Shulman (see ref. 5). The formula is

$$s_5 = P_5 \sqrt{\sum_{x=1}^{5} \frac{q_x}{l'_x - d_x}} = P_5 \sqrt{\frac{q_1}{l'_1 - d_1} + \frac{q_2}{l'_2 - d_2} + \ldots + \frac{q_5}{l'_5 - d_5}},$$

where s_5 is the standard error of the 5-year survival rate. In general, the standard error of the k-year survival rate is

$$s_k = P_k \sqrt{\sum_{x=1}^{k} \frac{q_x}{l'_x - d_x}}.$$

Columns 10 and 11 of Appendix Table I show how the calculation of the standard error of the 5-year survival rate is carried out as a continuation of the computation of the survival rate.* The first 9 columns are a replica of Table III. (1) Subtract d_x from l'_x for each line (Column 10); (2) divide q_x by $l'_x - d_x$ for each line (Column 11); (3) total the entries in the first 5 lines of Col-

*The standard error computed in this illustration is, itself, only an estimate of the true standard error. And, since it is based on relatively small numbers of cases, it is not a very reliable estimate. For example, had there been, due to sampling variation, one death in the last interval, rather than none, the computed standard error would be 0.0216 rather than 0.0187.

APPENDIX TABLE I. COMPUTATION OF THE 5-YEAR SURVIVAL RATE AND ITS STANDARD ERROR.
(Data from Table III)

YEARS AFTER DIAGNOSIS (1) X TO X + 1	ALIVE AT BEGINNING OF INTERVAL (2) l_x	DIED DURING INTERVAL (3) d_x	LOST TO FOLLOW-UP DURING INTERVAL (4) u_x	WITHDRAWN ALIVE DURING INTERVAL (5) w_x	EFFECTIVE NUMBER EXPOSED TO THE RISK OF DYING (COL. 2 − ½ COL. 4 − ½ COL. 5) (6) l'_x	PROPORTION DYING (COL. 3 ÷ COL. 6) (7) q_x	PROPORTION SURVIVING (1 − COL. 7) (8) p_x	CUMULATIVE PROPORTION SURVIVING FROM DIAGNOSIS THROUGH END OF INTERVAL ($p_1 \times p_2 \times \ldots \times p_x$) (9) P_x	(COL. 6 − COL. 3) (10) $l'_x - d_x$	(COL. 7 ÷ COL. 10) (11) $q_x/l_x' - d_x$
0–1	126	47	4	15	116.5	0.40	0.60	0.60	69.5	0.0058
1–2	60	5	6	11	51.5	0.10	0.90	0.54	46.5	0.0022
2–3	38	2	—	15	30.5	0.07	0.93	0.50	28.5	0.0024
3–4	21	2	2	7	16.5	0.12	0.88	0.44	14.5	0.0083
4–5	10	—	—	6	7.0	0.00	1.00	0.44*	7.0	0.0000
5–6‡	4	—	—	4						
										0.0187†

*Five-year survival rate.

†This is the sum of the five entries in Column 11. The square root of this number, when multiplied by the 5-year survival rate, yields the standard error of the 5-year survival rate: $s = (0.44)\ 0.0187 = (0.44)\ (0.37) = 0.060$.

‡See footnote *, Table III.

umn 11: 0.0187; (4) take the square root of this number: $\sqrt{0.0187} = 0.137$; (5) multiply the result by P_5: $0.137 \times 0.44 = 0.060$. This is the standard error of the 5-year survival rate.

The standard error of survival rates for end-points other than 5 years is computed similarly. For example, to compute the standard error of the 3-year survival rate, the first three entries in Column 11 must be totaled, the square root taken, and multiplied by P_3.

Effective Sample Size.—The concept, *effective sample size*, provides another way of assessing the benefit of including in the life table cases with partial survival information. The concept relates to the fact that the reliability of a statistical result depends on the size of the sample, i.e., the number of cases observed. For example, the standard error of a survival rate, P, when all cases have been followed until death or for the required time interval (i.e., no losses from observation or withdrawals alive prior to the cut-off date) is given by the binomial formula

$$s = \sqrt{\frac{P(1-P)}{l_1}}, \quad (1)$$

where l_1 is the sample size, i.e., the initial number of cases. In formula (1), the standard error is inversely proportional to the square root of the sample size.

APPENDIX TABLE II

	SAMPLE SIZE		
	1946–1951 COHORT*	EFFECTIVE SAMPLE SIZE	1946 COHORT†
Kidney, localized	126	68	9
Kidney, regional	47	37	11
Breast, localized	1,490	516	225
Breast, regional	1,531	595	208
Lip	332	145	61

*Since the cut-off date was Dec. 31, 1951, cases diagnosed in 1947 or later were eligible for less than 5 years of observation.

†Actual number of cases eligible for 5 years of observation.

Let us consider the 1946-1951 localized kidney cancer cohort (Appendix Table I), for which the survival rate is 0.44, and its standard error, 0.060. Of the initial 126 cases in this cohort, a substantial number were withdrawn alive less than 5 years after diagnosis. We now ask how large a cohort, with a 5-year survival rate of 0.44 and with all cases followed to death or for a full 5 years, would have a standard error equal to 0.060. To answer this question, we solve equation (1) for l_1, placing a circumflex over the l_1 to indicate that this is a hypothetical value:

$$\hat{l}_1 = \frac{P(1-P)}{s^2}. \quad (2)$$

Substituting $P = 0.44$ and $s = 0.060$, we obtain

$$\hat{l}_1 = \frac{(0.44)(0.56)}{0.0036} \doteq 68.$$

The result, 68, is the *effective sample size*, which we interpret as follows. Had we started with about 68 cases (instead of 126) and followed them all until death or survival for 5 years and found that 44 per cent survived 5 years, then the standard error would have been equal to that we actually obtained in our cohort of 126 cases. Thus, the survival rate we obtained is as reliable as one based on 68 cases. This is in sharp contrast to 9 cases which were eligible for 5 years of observation. These three values are compared for the five cancer groups discussed in the text. In each instance, the effective sample size based on the 1946-1951 cohort is substantially larger than the number of cases eligible for 5 years of observation (1946 cohort).