Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 3466–3468, May 1989
Statistics

Biostatistical analysis of mortality data for cohorts of cancer patients

(Hardin Jones principle/Kaplan–Meier renormalization)

LINUS PAULING

Linus Pauling Institute of Science and Medicine, 440 Page Mill Road, Palo Alto, CA 94306

Contributed by Linus Pauling, February 13, 1989

ABSTRACT **The Hardin Jones principle states that for a homogeneous cohort of cancer patients the logarithm of the fraction surviving at time *t* has a constant slope. With use of this principle, the survival times of the members of a heterogeneous cohort can be analyzed to divide the cohort into subcohorts with different mortality rate constants. Probable values of the additional survival time can be estimated for members surviving at the closing date of a clinical trial, permitting them to be included in the biostatistical analysis of the results of the trial in a more significant way than through Kaplan–Meier renormalization.**

Cancer continues to constitute a great problem. Although there has been significant progress in the prevention and treatment of some kinds of cancer during recent decades, most kinds still cannot be prevented or successfully treated.

I have developed a powerful method of biostatistical analysis of the observed survival times of cancer patients on the basis of the Hardin Jones principle, as described in the following sections of this paper.

The Hardin Jones Principle

In 1956 Hardin Jones made a penetrating analysis of the demography of the cancer problem (1). An important conclusion that he formulated is that with a reasonably homogeneous cohort of cancer patients, such as those with the same kind of cancer who have reached the terminal or untreatable stage (for example, breast cancer patients with metastases who have not responded favorably to high-energy radiation or chemotherapy), the rate of death is given by the equation

$$\frac{dN}{N} = -\alpha t, \qquad [1]$$

where N is the number of survivors at time t and α is a constant, the probability of death in unit time for a member of the homogeneous cohort. Integration of this equation leads to

$$S = \frac{N}{N_0} = e^{-\alpha t}, \qquad [2]$$

in which N_0 is the number of patients in the cohort at $t = 0$ (the beginning of the study or the time of entrance of the patient into the study). This equation describes a first-order reaction; that is, the number of persons dying in unit time is a constant fraction of the number of survivors in the cohort, independent of the time.

If for some reason patients leave the study, renormalization is usually made by the Kaplan–Meier method (2).

Jones (1) reported the results of his analysis of about 50 sets of mortality data for cohorts of cancer patients on the basis of his principle, and Burch (3) reported similar results for 9 sets. An example is given in figure 11.4 of ref. 3, which shows as a good straight line the logarithm of the percentage survival for 9159 women of all ages with localized breast cancer (data were obtained from the 1963 California Tumor Registry). My associate Zelek S. Herman and I have made many similar analyses of the survival data for presumably homogeneous cohorts of cancer patients, verifying the general validity of the Hardin Jones principle (unpublished studies). This principle accordingly provides a sound basis for the formulation of a biostatistical theory of cancer mortality.

The Analysis of Trials Made on Cohorts Consisting of Two or More Significantly Different Subcohorts

Jones (1) found that the logarithm of the fraction of the cohort surviving at time t sometimes could be expressed as the sum of two or three exponential terms, rather than one,

$$\frac{N}{N_0} = \sum f_i e^{-\alpha_i t}. \qquad [3]$$

For example, for women with metastatic breast cancer who were treated by the Halsted operation, 67% had 50% survival time $t_{1/2}$ ($= 0.693/\alpha$) of 0.69 years and 33% had $t_{1/2}$ of 4.25 years (1). For women seen initially without evidence of metastasis, there were three subcohorts: 36% with $t_{1/2} = 1.20$ years, 54% with $t_{1/2} = 5.37$ years, and 10% with $t_{1/2} = 35$ years (1). Burch (3) resolved the logarithm of the percent survival of 13,392 women in California with breast cancer at all stages into the sum of two terms: 30% with 50% survival time $t_{1/2} =$ 1.2 years and 70% with $t_{1/2} = 9.1$ years.

A Theory of Modest Heterogeneity of a Cohort of Cancer Patients

The Hardin Jones plot for some presumably homogeneous cohorts of cancer patients shows some curvature, such as to suggest a moderate amount of heterogeneity. A reasonable assumption is that there is an error function distribution of the activation energy of the rate constant, α, about a mean value of the activation energy corresponding to an intermediate value α_0 of the rate constant. This assumption leads on expansion and integration to the introduction of a quadratic term:

$$\ln S = -\alpha_0 t + \beta t^2. \qquad [4]$$

The parameter β is related to the standard deviation σ in the activation energy error function by the equation

$$\beta = \frac{3}{4}\left(\frac{\sigma}{RT}\right)^2, \qquad [5]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

in which R is the molar gas constant and T is the absolute temperature.

Mean Values of Powers of Survival Times for a Homogeneous Cohort

The mean value $\langle t^n \rangle$ of the nth power of the survival time t for a homogeneous cohort is obtained by integrating the product of t^n and the fraction $-dS/dt$ dying between t and $t + dt$, which is $\alpha e^{-\alpha t}$:

$$\langle t^n \rangle = \int_0^\infty \alpha t^n e^{-\alpha t}\, dt. \quad [6]$$

The known value of the definite integral leads to the equation

$$\langle t^n \rangle = \frac{\Gamma(n+1)}{\alpha^n}. \quad [7]$$

Here the Γ function has values 0.9999422883, 0.886227, 1, and 2 for $n = 0.0001, 0.5, 1$, and 2, respectively. It is seen that the mean $\langle t \rangle$ is equal to $1/\alpha$. It is convenient to use the symbol τ for $1/\alpha$, the reciprocal of the rate constant.

From Eq. 7 we derive the following result:

$$\{\langle t^n \rangle / \Gamma(n+1)\}^{1/n} = \frac{1}{\alpha} = \tau. \quad [8]$$

This equation is valid for every positive value of n for a homogeneous cohort with the values of t distributed in accordance with Eq. 2. Adherence to this equation accordingly provides a test for the homogeneity of a cohort.

Another convenient method for evaluating τ is to make use of the definite integral

$$\langle \ln t \rangle = \int_0^\infty \ln t e^{-t}\, dt = -\gamma, \quad [9]$$

in which γ is Euler's constant, with value 0.5772156649 . . . . In this equation τ has been taken to have the value 1, so that τ is equal to $e^\gamma \exp\langle \ln t \rangle$.

$$\tau = e^\gamma \exp\langle \ln t \rangle = 1.7810 \exp\langle \ln t \rangle. \quad [10]$$

Moreover, $\langle \ln t \rangle^{1/N_o}$, with N_o the number of terms, is equal to $\{\Pi(t_i)\}^{1/N_o}$, the N_oth root of the product of the values of t, giving the following equation (equivalent to Eq. 10):

$$\tau = 1.7810\, \{\Pi(t_i)\}^{1/N_o}. \quad [11]$$

The surviving fraction corresponding to $\exp\langle \ln t \rangle = 1$ is $e^{-\gamma} = 0.56146$ . . . .

Still another method is to calculate the slope of the line connecting the points on the Hardin Jones plot with the origin; the reciprocal of this slope for each point is a value of τ:

$$\tau_i = -t_i / \ln S_i. \quad [12]$$

S_i is N_i, the number of survivors (half integral) on the day t_i when this number decreased by 1, divided by N_o, the number of members in the original cohort. The mean of τ_i is

$$\tau = \langle -t / \ln S_i \rangle. \quad [13]$$

The values of τ for t small may be in significant error because a change in t by 1 day changes τ by t^{-1}, and the values for t large may be in error because truncation of the cohort can make a large change in S.

Table 1 gives results for a representative example of a small cohort (10 patients). The mean of the four values of τ is 43.2 days, and the mean deviation from the mean is 2.7 days, which is close to the 10% error expected for τ determined by any one of the four methods for a cohort of 10. Since the causes of error have different effects for the four methods, the mean of the values provides a better approximation than any one value. Accordingly, I recommend that this mean be presented as the value of τ for a cohort. The value of $\langle t \rangle$ is usually close to this mean.

Table 1. Mean values of survival times for a cohort of 10 stomach cancer patients

Method	τ, days
$1.7810 \exp\langle \ln t \rangle$	48.9
$\{\langle t^{1/2} \rangle / \Gamma(3/2)\}^2$	43.7
$\langle t \rangle$	41.5
$\{\langle t^2 \rangle / 2\}^{1/2}$	38.7
Mean	43.2

The patients had reached the untreatable stage at time $t = 0$ and received no further treatment (group no. 1 in table 1 of ref. 4, all female, ages 56–66; t = 5, 8, 12, 21, 29, 36, 41, 54, 85, and 124 days).

For the 10-member cohort of Table 1, omitting one value of t leads to a mean deviation of ±9% for the value of τ determined by any one of the four methods. For N_o members, the mean deviation is somewhat less than N_o^{-1}.

An Alternative to the Kaplan–Meier Renormalization Procedure in the Biostatistical Analysis of Survival Times of a Cohort of Cancer Patients Some of Whom Are Alive at the Termination Time of the Study

In the Kaplan–Meier renormalization procedure (2), a member of the cohort who changes treatment, becomes unavailable, is alive at the termination of a mortality study, or for some other reason can no longer be considered to be a member of the cohort is removed from the study, decreasing the number at risk by 1. Valuable information may be lost by this procedure if the fraction of dropouts is large, especially if the dropouts tend to occur at larger values of t. In the case of mortality studies of cancer patients, there is an alternative procedure, which is to use the Hardin Jones principle to predict the probable survival time t for the member of the cohort surviving at time t^+ at the termination of the study. If the survivor is a member of a homogeneous cohort, the value of t for this survivor is given by the following equation, in which τ is the mean survival time of the cohort:

$$t = t^+ + \tau. \quad [14]$$

For a cohort consisting of N_o members with N_o^+ alive at the termination date of the trial or on withdrawal from the trial, a first approximation value of τ is τ_o, the mean value of t_i and t_i^+. The self-consistent value of τ is then given by the following equation:

$$\tau = \tau_o / (1 - N_o^+ / N_o). \quad [15]$$

This value of τ is then to be added to each t_i^+ to obtain the estimated value of t_i.

For a cohort consisting of two subcohorts with fractions f_1 and f_2 and mean survival times τ_1 and τ_2, the mean expected additional survival time of a survivor is

$$\tau = \frac{\{f_1\tau_1 \exp(-t^+/\tau_1) + f_2\tau_2 \exp(-t^+/\tau_2)\}}{f_1 \exp(-t^+/\tau_1) + f_2 \exp(-t^+/\tau)}. \quad [16]$$

Values of f_1, f_2, τ_1, and τ_2 are to be obtained in a self-consistent manner by consideration of all values of t, including the predicted values for the survivors. For example, with a cohort of 15 patients with "untreatable" bronchial cancer who received daily doses of ascorbate (4) and who had values of t_i from 17 to 460 days, including one survivor with t^+ =

200^+ days, the value of τ increased from 137 to 146 days when t^+ was replaced by $t^+ + \tau$.

A Subcohort with a Single Member

With a single member, with survival time t, in a subcohort, the probability of survival at time t is $e^{-t/\tau}$. When t/τ is equal to 0.693, this probability is 1/2; this is accordingly the median of the values of τ leading to the value t. The mean survival time is equal to τ. It corresponds to the relation

$$\tau = t. \quad [17]$$

If the sole member of the subcohort is still surviving at time t^+ and the probable corresponding median value of the expected lifetime t is $t^+ + 0.6931\ \tau$, we obtain the equation

$$\tau = 3.2589\ t^+ \quad [18]$$

for τ and the expected lifetime, after entry into the study, for the survivor

$$t = 3.2589\ t^+. \quad [19]$$

Dividing a Cohort into Two Subcohorts

If a cohort can be represented as the sum of two homogeneous subcohorts, with mean survival times τ_1 and τ_2 and coefficients f_1 and f_2 ($f_1 + f_2 = 1$), the three parameters may be evaluated from three independent properties of the set of values of t. For example, the three properties $\langle t^{1/2}\rangle$, $\langle t\rangle$, and $\langle t^2\rangle$ provide a set of equations that can be solved for the unknowns. A large and truly representative cohort is needed for this analysis to be successful.

Another method makes use of three points, $S_1(t_1)$, $S_2(t_2)$, and $S_3(t_3)$, on a smoothed Hardin Jones plot of ln S vs. t. For one set of 130 breast cancer patients (4), this treatment with $S = 0.5$, 0.25, and 0.0625 gave $f_1 = 0.45$, $f_2 = 0.55$, $\tau_1 = 29$ days, and $\tau_2 = 83$ days. This method is unreliable for small cohorts. An alternative is to make a least-squares fit to all the points on a Hardin Jones plot with a two-term function or, if the cohort is very large, to carry out a Laplace transformation.

Treatment of a Cohort with Several Survivors

With the assumption that the cohort, with N_o members, consists of two subcohorts, the following treatment can be used. There are N_o^+ survivors, with survival times greater than t_i^+, and $N_o - N_o^+$ others, with known survival times t_i. The value of τ_1 for the latter is taken to be $\tau_1 = \langle t_i\rangle$. The quantity $\exp(-t_i^+/\tau_1)$ is then calculated for each t_i^+ and assumed to be the probability that this member is a member of the first subcohort. The value of τ_2 is evaluated by Eq. **15** with N_o^+ decreased by subtracting $\Sigma \exp(-t_i^+/\tau_1)$. The values of t_i for the surviving patients are found from the equation

$$t_i = \tau_2 - (\tau_2 - \tau_1)\exp(-t_i^+/\tau_1). \quad [20]$$

Outliers

An outlier is a member of a cohort with such a large value of t that it is likely that the member belongs to a separate subcohort. Many examples could be quoted. In one cohort the survivor had $t^+ = 857^+$ days (alive at the termination date of the study), with the other 16 members of the cohort of 17 untreatable patients with bronchial cancer having values of t from 16 days to 450 days. The value of τ_1 for these 16 members is 152 days (5). With this value of τ_1, the value of $N_o\, e^{-t^+/\tau}$ for $t^+ = 857$ days is 5.7%, and for $t = 2793$ days, the probable survival time of the outlier, it is about 10^{-7}. Accordingly this cohort consists of a 16-member subcohort with $\tau_1 = 152$ days and a one-member subcohort with $\tau_2 = 2793$ days.

Discussion

The Hardin Jones principle that the death rate of members of a homogeneous cohort of cancer patients is constant is supported by much empirical evidence. The use of this principle permits powerful methods of biostatistical analysis of cancer mortality data to be formulated. Heterogeneous cohorts can be resolved into two or more homogeneous subcohorts. Probable values of additional survival times of members of a cohort who have not yet died at the end of a study can be estimated, permitting a method of analysis to be carried out that provides more information than that given by Kaplan–Meier renormalization. Outliers, members of a subcohort with very large survival time, can be identified. These methods are especially useful in interpreting survival times for small cohorts.

I thank Ewan Cameron, Zelek S. Herman, and Dorothy Munro for their help. This investigation was supported by grants from the Japan Shipbuilding Industry Foundation and other donors to the Linus Pauling Institute of Science and Medicine.

1. Jones, H. B. (1956) *Trans. N.Y. Acad. Sci.* **18**, 298–333.
2. Kaplan, E. L. & Meier, P. (1958) *J. Am. Stat. Assoc.* **53**, 457–481.
3. Burch, P. R. J. (1976) *The Biology of Cancer, A New Approach* (University Park Press, Baltimore).
4. Cameron, E. & Pauling, L. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 3685–3689.
5. Cameron, E. & Pauling, L. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 4538–4542.