NOWOTWORY Journal of Oncology • 2001 • volume 51
Number 4 • 339–354

Invited article

Lognormal modelling for the prediction of long-term survival rates from short-term follow-up data

Richard F. Mould

Mathematical modelling is widely used and in terms of health is, for example, associated with predictions of the incidence of leukaemia and solid cancers in the surviving Japanese population after the Hiroshima and Nagasaki atomic bombs and in the exposed populations after the Chernobyl accident. In addition, for Chernobyl, modelling is used to predict the radiation risk of thyroid cancer incidence among emergency accident workers, with BEIR [1] for instance, quoting an excess absolute risk of 1.25 per 10^4 person.year.Gy.

The term *model* is also associated with multivariate analysis, such as the Cox proportional hazards model and the lognormal has also been used within the framework of a regression model for multivariate analysis, to study, for example, prognostic factors in breast cancer [2, 3, 4]. However, in the field of radiation oncology, mathematical modelling seems seldom to be employed apart from multivariate analysis regression models and the obvious example of radiobiological modelling for which oncologists will be familiar with the linear-quadratic model and the various attempts to use biological dose as distinct from physical dose [5].

It is largely forgotten that some 50 years ago, well before any radiobiological modelling was proposed, the lognormal distribution was used as the underlying basis for predicting long-term survival rates from short-term follow-up data [6, 7]. However, this was never used extensively except for three studies. (**a**) In 1975 some 5,000 carcinoma cervix patients from university teaching hospitals in London, the Christie Hospital in Manchester, the M.D. Anderson Hospital in Houston and the Norwegian Radium Hospital in Oslo [8]. (**b**) In 1984 for 14,731 cases of breast cancer in Norway [9]. (**c**) In 1985 for 8,750 cases of breast cancer in Sweden [10]. The reasons for this lack of use were threefold.

Firstly, the necessary computing power was not generally available in the hospital environment until the late 1970s and then the software programmes had to be written by prospective users because unlike actuarial life-table calculations, the so-called Kaplan-Meier method [11], which could be purchased in commercial software packages, lognormal survival rate prediction modelling software has never been available commercially. Currently, commercial software is available for some aspects of lognormal usage but still not for prediction modelling of long-term survival rates.

This review will save prospective users of lognormal survival rate prediction modelling the need to refer back to the original papers such as that of Boag in the *Journal of the Royal Statistical Society* in 1949 [6].

Secondly, for the validation of the lognormal model, or indeed any other model such as the skew exponential [8,12], a large body of data with long-term follow-up has to be available in cancer registries and then retrieved, stored in a study database and analysed into the format required for the modelling. The data retrieval and storage was very labour intensive until the 1960s when computers were first used to any great extent in a hospital environment, and also, not all cancer registries possessed good enough quality data in sufficient numbers with sufficient follow-up. Now in the 21st century computing power is not a problem and neither is software writing and even a few general flow diagrams are available [13].

Thirdly, the generalised lognormal formula has two parameters, mean μ and standard deviation σ, which with the proportion of cured patients C, makes a total of three unknown parameters in the model described by Boag [6]. Three variable parameters often cause the model to be unstable and the solution to this problem, fixing *a priori* the value of σ is not always possible. Indeed when studying cancer patient groups with all disease stages combined, and not separated for example into T stages from the TNM classification, or even simply into early and late stage groups (to ensure larger number of patients per study group), then fixing σ *a priori* is impossible.

This review is written to inform oncologists and medical statisticians about this prediction method which when validated can be of great use in studying the results of cancer treatment, either in a planned prospective study, or using a few years of retrospective records, and then estimating the 15-year and 20-year survival rates. It describes the possibilities of lognormal modelling and the technique of validation which will have to be made for defined cancer site groups, if possible subdivided by disease stage and histology, before the model can be used prospectively as a predictive tool.

Finally, it is emphasised that one should differentiate in the literature between (**a**) the lognormal model as used for long-term survival rate prediction modelling and

determination of a proportion C of cured survivors, and (**b**) the lognormal which is part of a multivariate analysis model. The lognormal is not the only parametric distribution which has been studied for (**a**) and (**b**). These have been summarised [2] after Kalbfleisch and Prentice [14] and are reproduced here. "The choice of model would be dependent on the hazard (risk) pattern of an event (recurrence or death) for a particular cancer in the time period for the study. An *exponential* model would be appropriate if the risk remained constant across the time period; a *Weibull* model if the risk was monotone decreasing or increasing with time; a *lognormal* model if the risk is zero at the beginning of the study, increases to a maximum, and then decreases approaching zero with long follow-up; and a *log-logistic* if the risk either increases monotonely like the Weibull or shows a similar pattern to the lognormal with *heavier tails*".

Normal & lognormal formulae

Full details of the lognormal distribution can be found in a Cambridge University Press monograph [15] and illustrated descriptions of the model and method of determining the predicted proportion of cured patients, C, can be found in several references by Boag and Mould [6-8, 13, 16-18] between 1949 and 1998.

Before the model itself is discussed, relevant formulae for the lognormal are presented, including its relationship to the normal, Gaussian distribution, Eq.6 is for the *standard* normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ and the total area beneath this curve = 1. The *general* normal curve, Eq. 7, is for any values of μ and σ and in it x is replaced from Eq.6 by

$$\xi = (x-\mu)/\sigma \quad \text{..... Eq.1}$$

which is termed the *unit normal deviate*.

Areas beneath the standard normal distribution curve, which has the well known symmetrical bell-shape, are tabulated in almost all introductory statistics textbooks. They are presented in terms of the unit normal deviate ξ usually in the range $0 \le \xi \le 4.0$. Because of the symmetry property of the normal distribution the area beneath the curve is equal to 0.5 for both $-\infty \le \xi \le 0$ and $0 \le \xi \le +\infty$. Such tables are applicable also for the lognormal distribution curve because the lognormal has a unit normal deviate of

$$\xi = (\{\log_e x\} - \mu)/\sigma \quad \text{..... Eq.2}$$

and is the logarithmic transformation of the normal curve when x becomes $\log_e x$, but from here on we will replace x by t since we are interested in the lognormal as a distribution curve for the survival time t of cancer patients who die with their disease present. The equation of the general lognormal is given in Eq.8 where μ is the mean and σ the standard deviation of the lognormal. The properties of the lognormal are such that the value of t at which the mean occurs is

$$t_{Mean} = \mu.\exp(-\tfrac{1}{2}\sigma^2) \quad \text{..... Eq.3}$$

the value of t at which the mode occurs is

$$t_{Mode} = \mu / \exp(\sigma^2) \quad \text{..... Eq.4}$$

and the value of t at which the median occurs is

$$t_{Median} = \mu \quad \text{..... Eq.5}$$

$$y = (1/\sqrt{2\pi}).\exp(-\tfrac{1}{2}x^2) \quad \text{.....Eq.6, Standard Normal}$$

$$y = (1/\{\sigma\sqrt{2\pi}\}).\exp(-\tfrac{1}{2}[\{x-\mu\}^2/\sigma^2]) \quad \text{.....Eq.7, General Normal}$$

$$y = (1/\{t.\sigma\sqrt{2\pi}\}).\exp(-\tfrac{1}{2}[\{\log_e[t/\mu]\}^2/\sigma^2]) \quad \text{.....Eq.8, Lognormal}$$

In the above equations y is usually expressed as a function of the variable on the right-hand side of the equation and thus for Eq.6 and Eq.7 (see Fig. 1) we can write f(x) instead of y and for Eq.8 write f(t) instead of y.

Fig. 1. Equation 7 for the general Normal curve is seen on this 10 Deutschmark banknote, which will be history on 1 January 2002 when the Euro is introduced. The picture is of the German mathematician Carl Friedrich Gauss (1777-1855) whose name is often associated with the Normal curve, so much so, that it is also called the Gaussian curve. In fact, though, it was not discovered by Gauss, but by Abraham de Moivre (1667-1754) in 1773, a refugee French mathematician living in London He was solving problems for wealthy gamblers! The curve was apparently forgotten until later in the 18th century when it was rediscovered by those investigating the theory of probability and the theory of errors [7].

Polynomial approximation for the area beneath a Normal curve

Several polynomial expressions exist to provide the area beneath a Normal curve between defined limits [19]. A suitable polynomial for use with the lognormal model is given in Eq.9 and is the integral of Eq.6 between the limits $-\infty$ and x which in Eq.9 is given the notation P(x).

$$P(x) = 1 - \tfrac{1}{2}[1 + d_1x + d_2x^2 + d_3x^3 + d_4x^4 + d_5x^5 + + d_6x^6]^{-16} + \varepsilon(x) \quad \text{..... Eq.9}$$

where

$$|\varepsilon(x)| < 1.5 \times 10^{-7}$$

$d_1 = 0.04986\ 73470$ $\qquad$ $d_4 = 0.00003\ 80036$

$d_2 = 0.02114\ 10061$ $d_5 = 0.00004\ 88906$

$d_3 = 0.00327\ 76263$ $d_6 = 0.00000\ 53830$

Tables of areas beneath the standard Normal curve, i.e. the curve with $\mu = 0$ and $\sigma = 1$, are to be found in most statistics textbooks and can be used to verify Eq.9. As an example, when $x = 0.5$ the area $P(x)$ beneath the curve from $-\infty$ to $+0.5$ equals 0.69146.

Lognormally distributed variables

Table I lists examples of variables which have been shown to be lognormally distributed. Some are obviously of more practical use than others! Table II lists examples of cancer sites which have been studied in terms of the lognormality of survival times of patients who died with their disease present. Figure 2 is a photograph of an early 20th century analogue machine for generating a positively skewed frequency curve.

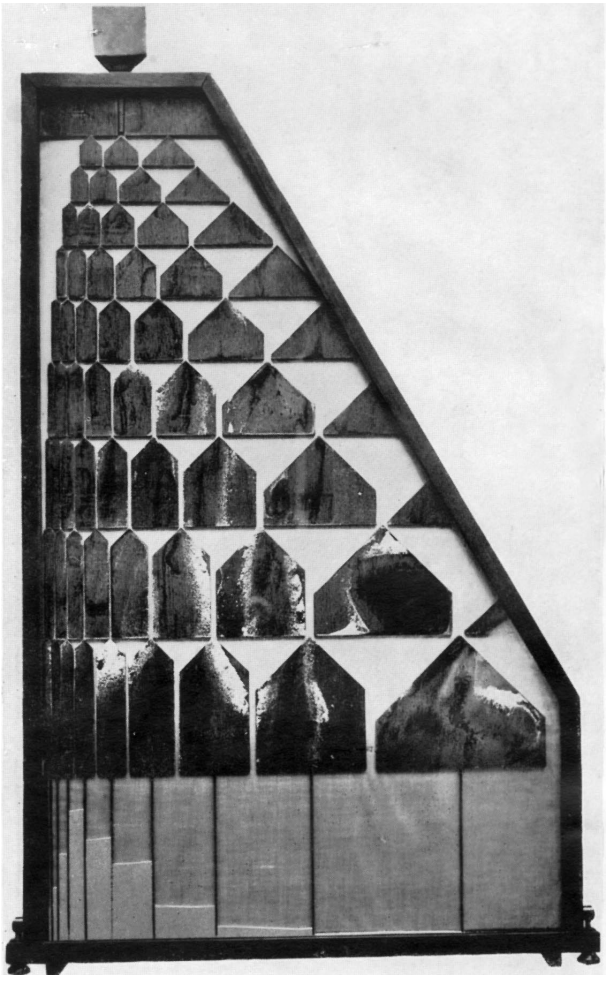

Fig. 2. With the intention of convincing sceptics that skew frequency curves could arise from natural causes, J. C. Kapteyn had built at the end of the 19th century an analogue machine for demonstrating skew frequency curves. It consists of nine rows of wedges shaped like the cross-section of a house in a game of Monopoly, attached to a wood and glass frame 104 cm high. The wedges are of varying width being proportional to the distance of the vertex of the wedge from the left-hand side of the frame. Sand is poured into a funnel at the top of the frame directly above the middle wedge in the top row. The sand arriving at the bottom of the machine forms a two-parameter lognormal distribution. The machine was in the 1950s to be seen in the laboratory Huize de Wolf adjacent to the Genetics Laboratory of the University of Groningen, The Netherlands [15,19].

Tab. I. Selected examples of lognormally distributed variables

Induction time of tumours in mice [20]
Response times for different drugs [21]
Infant mortality rates [22]
Combination of elementary errors [23]
Size of foreheads of crabs [24]
Number of petals on a buttercup [25]
Number of words in a sentence by George Bernard Shaw [21]
Cancer patient symptom duration in the range 0-2 years [26]
Cytokinetics of human solid tumours [27]
Red blood cell volumes [28, 29]
Lung cancer incidence in smokers [30]
Carcinoma-free probability in rats exposed to carcinogens [31]

Tab. II. Cancer patients who die with their disease present and whose survival time distribution has been shown to be lognormally distributed

Cancer site/histology
Cervix uteri [8, 32]
Head & neck [6, 16, 18, 33-35]
Breast [2, 3, 6, 9, 10, 36-38]
Malignant melanoma of skin [39, 40]
Non-Hodgkin's lymphoma [41]
Lung [42]
Bladder [43]

Transformation from t to log(t) and graphical testing for lognormality

Figure 3(a) is a frequency histogram of the survival time of 338 patients treated for cancer of the mouth and throat and who subsequently died with their cancer present [18]. If the number of cases were large enough and the time intervals chosen were small enough, the boundary of the histogram would approach a smooth curve. This would not be symmetrical but would be skewed. If this curve in Figure 3(a) is redrawn, taking the logarithm of the survival time to base 2, i.e. $\log_2 t$, as the variable, the histogram shown in Figure 3(b) is obtained.

The graphical test in Figure 3(c) confirms that the transformed distribution is now sufficiently normal to justify the use of significance testing to verify lognrmality, $P > 0.05$. The logarithmic transformation can be to $\log_e t$ as in Eq. 8, or to $\log_{10} t$ or to any other base. The graphical test for lognormality in Figure 3(c) is one of the first in the literature which is related to radiation oncology. The logarithm base chosen by Boag for this 1950 schematic [18] is $\log_2$ and therefore the logarithmic scale on the horizontal axis in Figures 3(b) and 3(c) is 0.5, 1, 2, 4, 8, etc. However, in later papers Boag, and other authors, used $\log_{10} t$, see Figure 4.

Printed graph paper is commercially available for testing for normality (arithmetic probability graph paper) and for lognormality (logarithmic probability graph paper, Figure 4), using $\log_{10}$. Figure 3(c) has the same

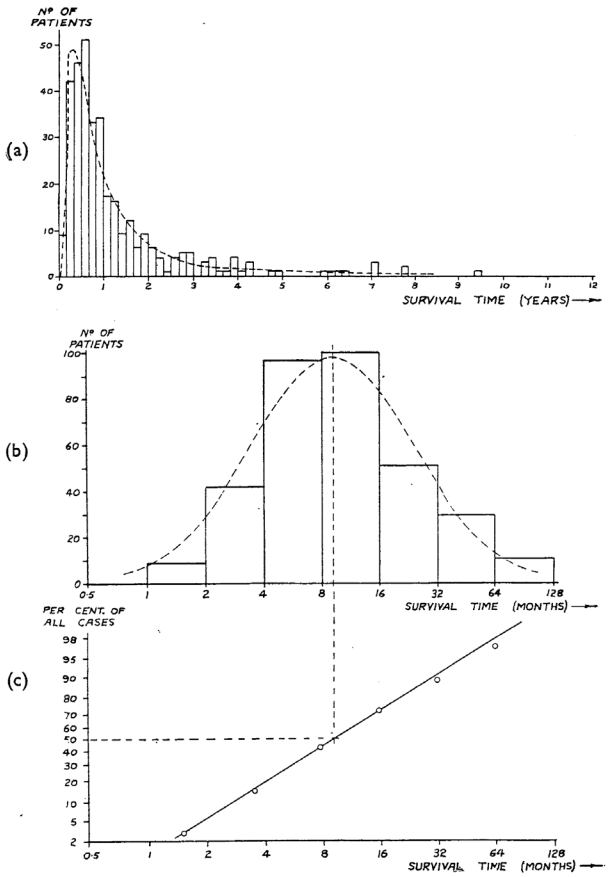

Fig. 3. (a) Frequency distribution of survival times of 338 patients with mouth and throat cancer who died with their disease present. (b) Histogram drawn to a logarithmic time scale, $\log_2 t$, for the same series of 338 cases. The superimposed dotted curve is symmetrical and bell-shaped Normal distribution curve. (c) Graphical test for the Normal distribution of the $\log_2$ of survival time [18].

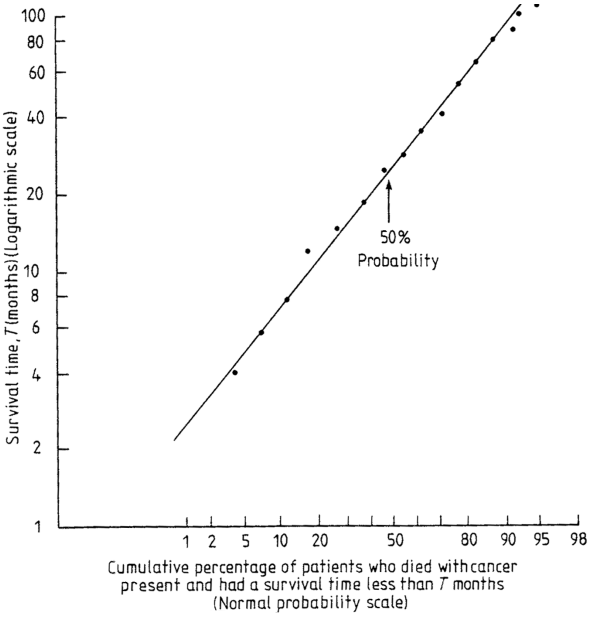

Fig. 4. Logarithmic probability graph plot for testing for lognormality. The data are 449 patients with cancer of the larynx who died with their disease present [13].

graphical format as Figure 4 but convention is now usually to have survival time on the vertical axis and the Normal probability scale on the horizontal axis. Spelt out in full, the Normal probability scale is for '*The cumulative percentage of patients who died with cancer present and had a survival time ≤ T months*' whereas the vertical scale is the survival time T months.

The mean logtime μ (as distinct from the log of the mean time) of the lognormal distribution is given by the value of the survival time which corresponds to a 50% probability. Thus for example if, as in Figure 5, $T_{50\%}$ = 27.4 months then the graphical estimate is μ = 1.44. The graphical estimate of σ is given by

$$\sigma = (T_{95\%} - T_{50\%})/1.645 \text{..... Eq. 10}$$

because a property of the Normal curve is such that 5% of the area beneath the normal curve lies outside the standard deviation limit +1.645σ and $T_{50\%}$ is the mean [7, 13]. It should also be remembered that for Figure 4 we are working with $\log_{10} t$ and therefore μ is the mean $\log_{10}$time.

Examples of graphical demonstrations of lognormality using data for eight sites within the head and neck [13,33] are given in Figure 5 with the $T_{50\%}$ indicated for each straight line. The horizontal axis is termed a scale of probits since a probit is defined as 'a unit for measuring probability in relation to an average frequency distribution' [44]. However, it should be noted that if for a given cancer site lognormality occurs P > 0.05, it would be expected that if a total of 100 series for this site were studied, then a subtotal of 5/100 the lognormality tests would fail, P < 0.05.

Lognormal Prediction Model Description

Figure 6 shows schematically the theory underlying a parametric statistical prediction model [45], not only the lognormal model, but also those models using as an alternative the negative exponential and the skew exponential [8,12] distributions. It is also noted that other forms of

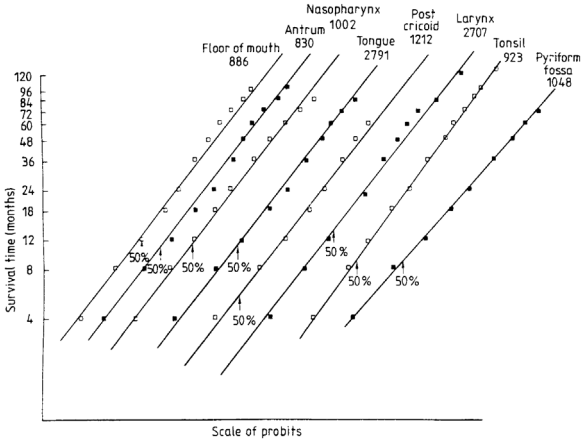

Fig. 5. Log-probability graph plots for eight series of head & neck cancer patients who died with their disease present. In order to illustrate these eight series the horizontal 'Cumulative percentage who died with cancer present and had a survival time ≤ T months' axis, which can also be termed a scale of probits, has been compressed [13].

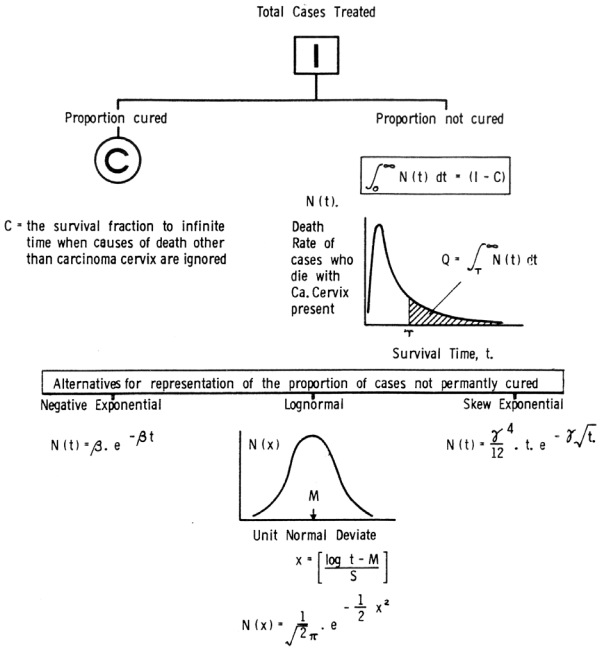

Fig. 6. Schematic diagram of a parametric statistical prediction model in which the cured proportion of cancer patients is denoted by C and three possible distributions are shown for the representation of the (1-C) groups of patients who died with their cancer present: lognormal, negative exponential and skew exponential [45].

prediction model exist, and for breast cancer, the extrapolated actuarial model of Haybittle, see later Figure 13, [46-50] has been shown to be successful.

C is an index of statistical cure but is effectively equal to a long-term τ-year survival rates. Using the model, the τ-year survival rate, see Figure 6, is given in Eq. 11 where Q is the integral of the lognormal distribution between the limits τ and $+\infty$, the shaded area beneath the lognormal curve in Figure 6.

Determination of C by the method of maximum likelihood

The determination of C and μ with the value of σ assumed *a priori* enables the estimation to be made of the τ-year survival rate, Figure 6, which is given in Eq. 10.

τ-year percentage survival rate
= 100 x {C + (1-C).Q}..... Eq. 11

The method of estimation used is the method of maximum likelihood and is taken from Appendix A on pages 138-141 in the *Medical Research Council* publication [18] by Wood and Boag in 1950. It will be seen that the patient data has to be subdivided into four groups and that Group (2) are those patients who die of an intercurrent disease. In practice, it is sometimes very difficult to determine if a death is a true intercurrent death and a good follow-up database is essential. I have retained the Table XXVIII numbering (see page 346) of Wood and Boag for the derivatives of the log likelihood.

Table III gives examples of maximum likelihood estimates of C, the proportion of cured patients, for cancer of the cervix uteri [8, 45] and for cancer of the breast [9, 10]. These examples indicate for cancer of the cervix that squamous cell carcinoma has a better prognosis than adenocarcinoma. The better stage 1 results for surgery will reflect the bias of cases allocated for surgery alone being early stage 1 cases. For breast cancer in Sweden there has been an upward trend in survival between the two periods 1961-63 and 1971-73 [10].

Tab. III. Examples of maximum likelihood estimates of C with associated standard errors given in brackets

Cancer population & treatment period	Estimate of C [±1SE] {%}
Cervix cancer, 1945-59 [8,45]	
Stage 1	61.7 [1.9]
Stage 2	40.3 [1.4]
Stage 3	20.0 [1.5]
Squamous cell ca. Stage 1	64.6 [2.1]
Adenocarcinoma, Stage 1	52.6 [6.8]
Squamous cell ca. Stage 2	42.2 [1.6]
Adenocarcinoma, Stage 2	26.0 [5.9]
Surgery, Stage 1	78.1 [5.0]
Radiotherapy, Stage 1	59.6 [2.5]
Radiotherapy + Surgery, Stage 1	60.2 [3.7]
Breast cancer, Sweden 1961-73, age < 70 years [10]	
1961-63	33 [2]
1971-73	40 [3]
Breast cancer, Norway, 1953-67 [9]	
Stage 1	54 [3]
Stage 2	27 [1]

Lognormal Prediction Model Validation

To validate a mathematical model for predicting survival rates in cancer patient populations, the procedure is divided into two phases when the model is of the type in Figure 6 which assumes an analytical form for the distribution of survival times of the (1-C) patient group.

Phase I
Test of the analytical form of the survival time distribution of those patients who died with their cancer present.

Phase II
Estimation of long-term survival rates when only relatively short-term follow-up is available, and validation of the predicted rates by comparison with the true long-term results as calculated by an actuarial life table method, such as that described by Kaplan & Meier [51].

Phase I can be achieved by using a minimum χ^2 test as a goodness of fit test to the data. For the lognormal model the test will commence with the values of μ and σ which have been estimated graphically, Figures 4 and 5, and these values will then be varied by small amounts δμ and δσ until the minimum value of χ^2 is obtained. It can then be determined whether the observations are not significantly different, $P > 0.05$, from the theoretical lognormal expected distribution. If $P > 0.05$ then Phase II

APPENDIX A

Determination of the Proportion Cured by the Method of Maximum Likelihood

LET it be assumed that a proportion, c, of all patients treated is cured, and that the survival times of those who die with the original cancer present have a frequency distribution given by

$$z(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$$

where $x = \dfrac{\log t - \mu}{\sigma}$

t = survival time reckoned from the commencement of treatment

Further, let π_t' = probability that a patient shall survive to time t when only causes of death other than cancer are considered

π_t = probability that a patient who has not been permanently cured shall not die with cancer present before time t

$$= \int_{x(t)}^{\infty} z(x)\, dx$$

$$= q(x), \text{ say.}$$

At the time when the results of treatment are analysed, a patient, chosen at random from amongst those treated, must fall into one of the following four groups:

Group (1) those who died with cancer present at time t, the growth being either a persistence or recurrence of the original cancer

Group (2) those who died from intercurrent disease at time t with no evidence of a recurrence of the original cancer

Group (3) those who are alive and have been symptom-free for time t

Group (4) those who are alive but with cancer progressing at time t

To represent the distinction between Groups (3) and (4) we should introduce a further factor f_t for the probability that an uncured but still living patient shall show a visible recurrence of cancer at time t.

The probability of finding the patient in each of these groups is then

Group (1) $\pi_t'(1-c) \cdot \dfrac{z}{\sigma} \cdot \dfrac{dt}{t}$

Group (2) $-\dfrac{d}{dt}(\pi_t')\, dt\{c+(1-c)(1-f_t)q\}$

Group (3) $\pi_t'\{c+(1-c)(1-f_t)q\}$

Group (4) $\pi_t'(1-c) f_t q$

Now π_t' can be assumed to be independent of μ, σ and c and hence, in maximizing the likelihood for variations in these parameters, the factors involving only π_t' may be ignored.

The factor f_t cannot be disposed of so easily, since it may depend on μ and σ and will vary with t. If we were to ignore the distinction between Groups (3) and (4) then the likelihood factor for a patient in either group would be simply $\pi_t'\{c+(1-c)q\}$. We would then be discarding the information that patients

in Group (4) were already suffering from a recurrence of cancer and this method should therefore lead to too high an estimate of c. On the other hand, a pessimistic estimate of c will be obtained if we put $f_t = 0$ for patients in Groups (2) and (3) thus assigning to them too small a chance of being cured and at the same time retain the correct expression for Group (4). This we shall do and we shall also ignore the dependence of f_t on μ, σ and c. The number of patients in Group (4) is usually small and the differences between the estimates of c arrived at by the various approximate treatments of f_t will almost always be small compared with their standard error. The log likelihood for a given sample of cases containing patients in all four groups is then

$$L = \sum_1 \log \left\{\frac{(1-c)\,z}{\sigma}\right\} + \sum_{2,3} \log \{c+(1-c)q\} + \sum_4 \log \{(1-c)q\}$$

where each summation is to be extended over all patients in the group whose number is placed below the summation sign. L will be a maximum for the values of μ, σ and c which satisfy the equations

$$\frac{\partial L}{\partial \mu} = \frac{\partial L}{\partial \sigma} = \frac{\partial L}{\partial c} = 0.$$

These equations may be solved by an iterative process. If m_1, s_1, c_1 are rough estimates of μ, σ and c, then second approximations m_2, s_2, c_2 may be obtained by adding corrections δm, δs, δc to m_1, s_1, c_1 respectively, these corrections being obtained from the equations:

$$\frac{\partial L}{\partial m_1} + \frac{\partial}{\partial m_1}\left(\frac{\partial L}{\partial m_1}\right)\delta m + \frac{\partial}{\partial s_1}\left(\frac{\partial L}{\partial m_1}\right)\delta s + \frac{\partial}{\partial c_1}\left(\frac{\partial L}{\partial m_1}\right)\delta c = 0$$

$$\frac{\partial L}{\partial s_1} + \frac{\partial}{\partial m_1}\left(\frac{\partial L}{\partial s_1}\right)\delta m + \frac{\partial}{\partial s_1}\left(\frac{\partial L}{\partial s_1}\right)\delta s + \frac{\partial}{\partial s_1}\left(\frac{\partial L}{\partial s_1}\right)\delta c = 0$$

$$\frac{\partial L}{\partial c_1} + \frac{\partial}{\partial m_1}\left(\frac{\partial L}{\partial c_1}\right)\delta m + \frac{\partial}{\partial s_1}\left(\frac{\partial L}{\partial c_1}\right)\delta s + \frac{\partial}{\partial c_1}\left(\frac{\partial L}{\partial c_1}\right)\delta c = 0$$

The solution of these equations is then

$$\delta m = \frac{s_1}{\Delta}\begin{vmatrix} A & G & H \\ B & E & K \\ C & K & F \end{vmatrix}$$

$$\delta s = \frac{s_1}{\Delta}\begin{vmatrix} D & A & H \\ G & B & K \\ H & C & F \end{vmatrix}$$

$$\delta c = \frac{1}{\Delta}\begin{vmatrix} D & G & A \\ G & E & B \\ H & K & C \end{vmatrix}$$

where $\Delta = \begin{vmatrix} D & G & H \\ G & E & K \\ H & K & F \end{vmatrix}$, while A, B, C ..... K are the expressions defined in

TABLE XXVIII

Derivatives of the Log Likelihood for Patients in Each of the 4 Groups Defined on p. 138

Quantity	*Symbol (see p. 139)*	*Contribution by a patient in Group 1*	*Contribution by a patient in Group 2 or 3*	*Contribution by a patient in Group 4*
$\sigma\frac{\partial L}{\partial \mu}$	A	x	ϕ	$\frac{z}{q}$
$\sigma\frac{\partial L}{\partial \sigma}$	B	x^2-1	$x\phi$	$x\frac{z}{q}$
$\frac{\partial L}{\partial c}$	C	$-\frac{1}{(1-c)}$	ψ	$-\frac{1}{(1-c)}$
$-\sigma^2\frac{\partial^2 L}{\partial \mu^2}$	D	1	$-x\phi+\phi^2$	$-x\frac{z}{q}+\left(\frac{z}{q}\right)^2$
$-\sigma^2\frac{\partial^2 L}{\partial \sigma^2}$	E	$3x^2-1$	$2x\phi-x^3\phi+x^2\phi^2$	$2x\frac{z}{q}-x^3\frac{z}{q}+x^2\left(\frac{z}{q}\right)^2$
$-\frac{\partial^2 L}{\partial c^2}$	F	$\frac{1}{(1-c)^2}$	ψ^2	$\frac{1}{(1-c)^2}$
$-\sigma^2\frac{\partial^2 L}{\partial \sigma \partial \mu}$	G	$2x$	$\phi-x^2\phi+x\phi^2$	$\frac{z}{q}-x^2\frac{z}{q}+x\left(\frac{z}{q}\right)^2$
$-\sigma\frac{\partial^2 L}{\partial c \partial \mu}$	H	0	$\frac{\phi^2}{z(1-c)^2}$	0
$-\sigma\frac{\partial^2 L}{\partial c \partial \sigma}$	K	0	$\frac{x\phi^2}{z(1-c)^2}$	0

In the above Table,

$$\phi = \frac{z}{(q+c/1-c)} \qquad \psi = \frac{1}{(c+q/1-q)}$$

$$z = \frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{1}{2}x^2} \qquad x = \frac{(u-\mu)}{\sigma}$$

$$q = \int_x^\infty z\,dx \qquad u = \log_e t$$

t = survival time

Table XXVIII, summed over all cases in the sample, and are evaluated for $\mu = m_1$, $\sigma = s_1$, $c = c_1$.

If σ can be postulated in advance then the above three equations reduce to two and the corrections δm, δc are given by

$$\delta m = s_1 \frac{\begin{vmatrix} A & H \\ C & F \end{vmatrix}}{\begin{vmatrix} D & H \\ H & F \end{vmatrix}} \qquad \delta c = \frac{\begin{vmatrix} D & A \\ H & C \end{vmatrix}}{\begin{vmatrix} D & H \\ H & F \end{vmatrix}}$$

Similarly, if both σ and μ can be postulated in advance the correction δc for c is given by

$$\delta c = \frac{C}{F}$$

The standard error associated with a statistic derived by the Method of Maximum Likelihood may be calculated directly from the derivatives of the log likelihood. It is the standard error in c which is of particular interest, and if both μ and σ are known then the standard error in c is given by

$$(\text{S.E.})_c = \frac{1}{\sqrt{F}}$$

If σ only is known and both μ and c are estimated by maximum likelihood then

$$(\text{S.E.})_c = \sqrt{\frac{D}{\begin{vmatrix} D & H \\ H & F \end{vmatrix}}}$$

while in the general case

$$(\text{S.E.})_c = \frac{\sqrt{\begin{vmatrix} D & G \\ G & E \end{vmatrix}}}{\sqrt{\begin{vmatrix} D & G & H \\ G & E & K \\ H & K & F \end{vmatrix}}}$$

can proceed. This second phase is described schematically [13] in Figure 7 and subdivided into three parts.

Examples of validation results are shown in Figures 8 and 9 for carcinoma cervix [8] in which a comparison of observed and predicted 10-year and 15-year survival rates are given for minimum follow-up periods of only two, three and four years. These results show that the lognormal model with σ fixed at an appropriate value, Table IV, is of wider validity than any other model tested, including the skew exponential and the extrapolated actuarial, and gives reliable predictions of long-term survival rates for separate disease stage groups of carcinoma cervix.

A simplified flow chart for *Phase I* validation is given in Figure 10 and for *Phase II* validation in Figure 11.

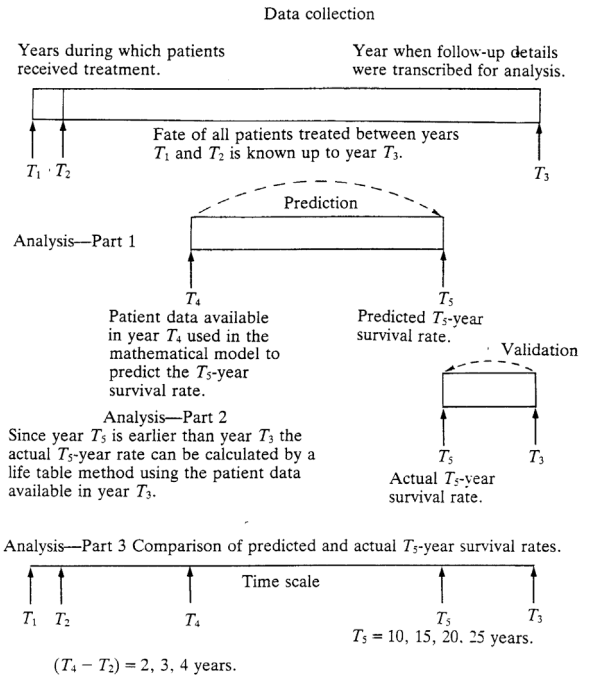

Fig. 7. Schematic diagram illustrating the procedure for validating a parametric statistical model where a specified analytical form such as the lognormal is used for the distribution of survival times for those patients who die with cancer present [13]

Amount of information relative to C for the lognormal model

In planning for a clinical trial it is essential to consider the number of patients required in order, from a statistical point of view, to be able to provide an answer in a reasonable period of time. If the lognormal prediction model is to be used the length of the necessary follow-up observation period can be studied in terms of reducing the standard error in the estimate of C to a desired value. In dealing with

Tab. IV. Summary of conditions for use of the lognormal model to predict long-term survival fractions for carcinoma cervix [8, 45]

Stage	Values which may be assumed for the lognormal parameter σ	Minimum waiting period after a 5-year treatment period closes before use of the lognormal model (n years)	Number in the series of cases tested
1	$0.35 \leq \sigma \leq 0.40$	n=3	101-553
2	$0.35 \leq \sigma \leq 0.40$	n=3	68-152
3	$0.35 \leq \sigma \leq 0.40$	n=2	77-170

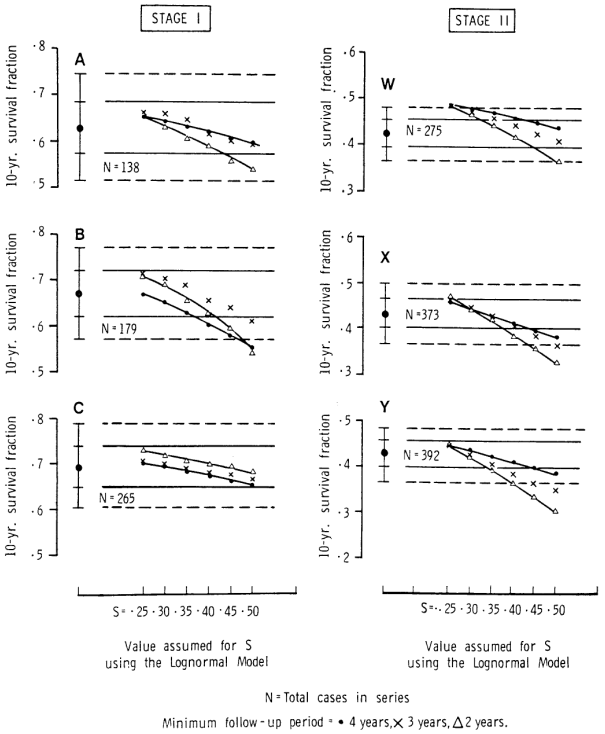

Fig. 8. Comparison of observed and predicted 10-year survival fractions (100 x survival fraction = % survival rate) for stage I and stage II carcinoma cervix [8].

this question the concept of '*amount of information*' introduced in 1922 by Fisher [25] and used by Boag [6] and Mould [45] specifically for the lognormal model. The amount of information contained in an estimate of any parameter is defined by the reciprocal of the sampling variance of that estimate. As the amount of information increases, so the precision of the estimate improves. The total information relative to C varies as the duration of follow-up period increases. However, the ideal amount of information will never be reached because this is only attained after an infinitely long follow-up period. Because of this Boag [6] proposed the use of the ratio i, Eq. 12.

$$i = \frac{\text{[Information utilised]}}{\text{[Information ultimately available]}} \quad \text{..... Eq. 12}$$

When μ and C are estimated simultaneously, as has been described, this ratio i becomes

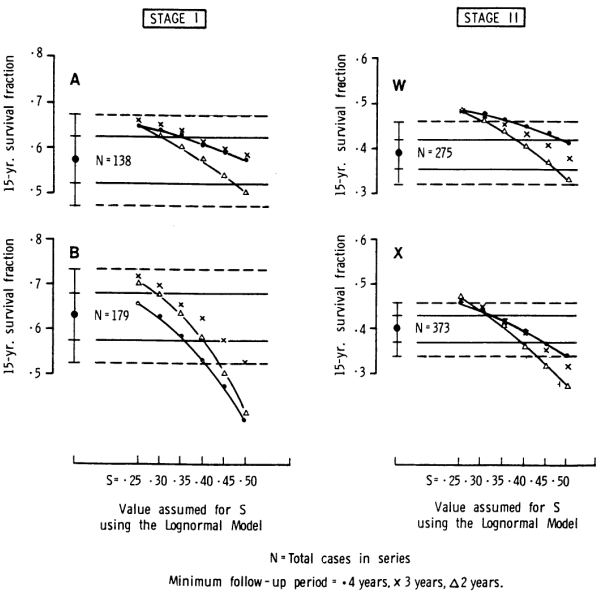

Fig. 9. Comparison of observed and predicted 15-year survival fractions (100 x survival fraction = % survival rate) for stage I and stage II carcinoma cervix [8].

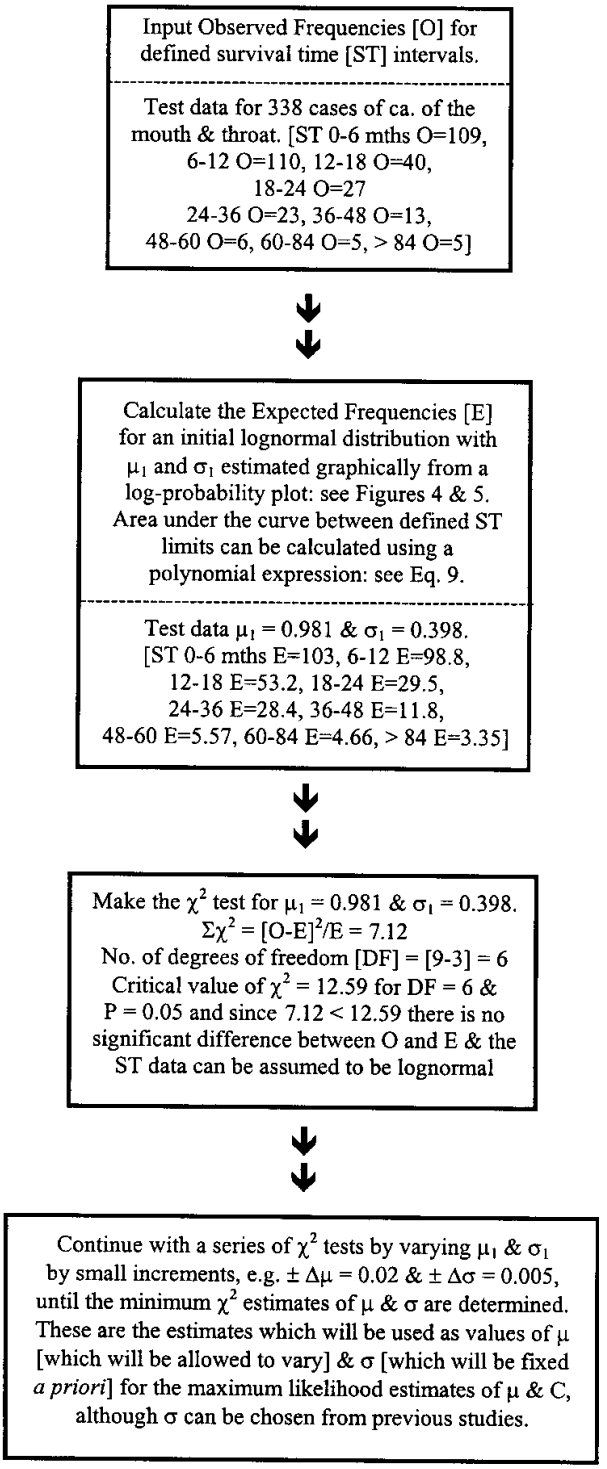

Fig. 10. Flow chart for *Phase I* validation procedure: minimum χ^2 testing. Test data for 338 cases of cancer of the mouth & throat, after Boag [6].

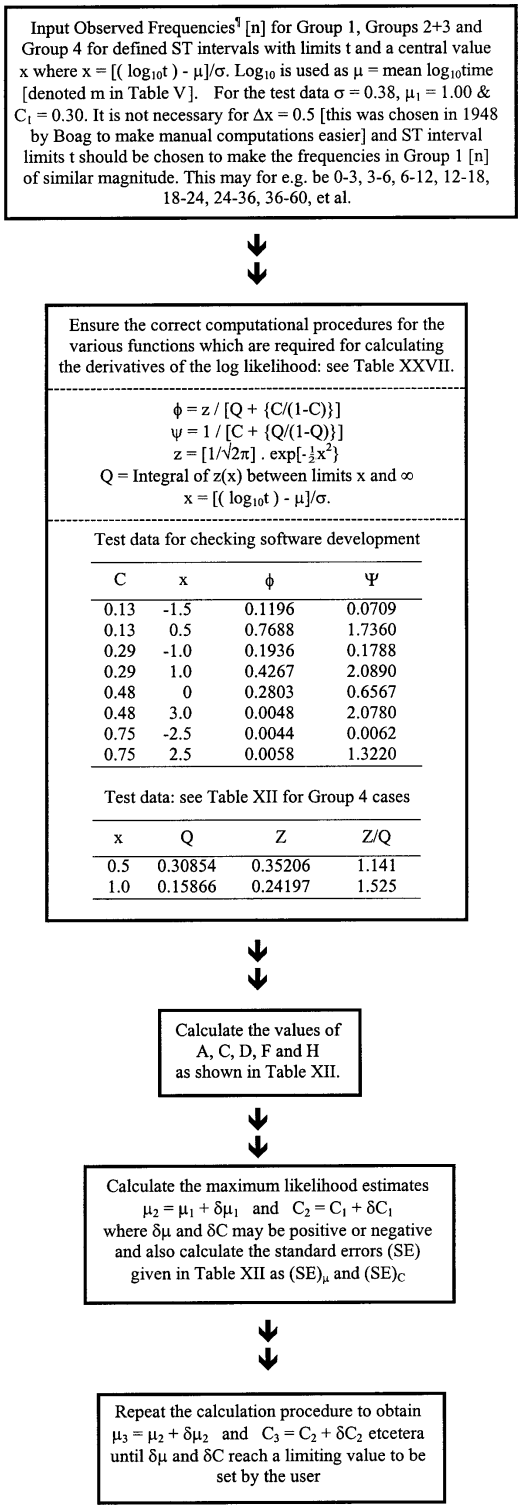

Fig. 11. Flow chart for *Phase II* validation procedure: maximum likelihood estimation of μ and C when there is an assumed value *a priori* for σ. Table V is included from Boag's 1948 paper in the *British Journal of Radiology* [16] and gives a worked example for 58 patients with cancer of the tonsil. Notation used in the flow chart is that used in Table V. The Observed Frequencies ¶ referred to in the first box in this flow chart, will, when the validation procedure is taking place, be for time $t=T_4$ and also for $t=T_2$ (see Figure 7) and the long-term T_5 year survival rate (%) is calculated using the formula $100.[C + \{1\text{-}C\}.Q_{t=T5}]$. Verification is achieved by taking the observed frequencies known at $t=T_3$ and using the life-table (actuarial, Kaplan-Meier) method to calculate the T_5 year survival rate (%), as shown schematically in Figure 7. When the model has been validated and is to be used for further cancer patient series, the T_5 year survival rate (%) is calculated from the latest available follow-up data, i.e. that at T_4.

$$i = C\Psi - ([C\phi^2] / [(1 - C)^2.\{1 - Q + Z(\phi\text{-}x)\}]) \quad \text{Eq. 13}$$

Thus for a series of N patients entering a trial at different points throughout the duration of the trial the information relative to C at the time of estimation of μ and C is given by Eq. 14.

$$I = \sum_{j=1}^{j=N} [\{i_j\} / \{C(1\text{-}C)\}] \quad \text{Eq. 14}$$

and the standard error in C is given by $[1/\sqrt{I}]$. However, neither of these quantities I and $[1/\sqrt{I}]$ provides a satisfactory numerical scale by which the accuracy of any estimate of C can be assessed prospectively for a specified patient series [45]. A more suitable parameter is the *mean information fraction per patient* ξ where ξ is defined in Eq. 15.

$$\xi = i/N \quad \text{Eq. 15}$$

which always lies in the range $0 < \xi < 1$. A theoretical example [45] of the variation of ξ as a function of μ [here termed M] and C for an annual patient intake of 20 cases per year for five years, for a lognormal model with σ= 0.30 [here termed S] for analyses at two and at three years after trial closure, [in terms of the notation of Figure 7, $T_1 \rightarrow T_2$ = 5 years and the two analyses are made with follow-up data available at T_4 such that $T_2 \rightarrow T_4$= 2,3 years].

As stated earlier this is a theoretical exercise and in practice a prospective study would seldom take place with a C = 0.05 and $\log_{10}$meantime μ = 0.90 [equivalent to a $T_{50\%}$ = 8 months] but if it did, then from Table VI it is seen that ξ = 0.953 for a minimum follow-up of three years and 0.875 for a minimum follow-up of two years. A more realistic situation is for a C = 0.50 and μ = 1.60 [equivalent to a $T_{50\%}$ = 40 months] and for this set of parameters. ξ = 0.416 for a minimum follow-up of three years and 0.283 for a minimum follow-up of two years. Such data for ξ can be calculated for any values of μ and C and for any annual pattern of patient intake into the study and can be helpful when deciding when a prospective study can be analysed.

Clinical trial planning decisions

It is emphasised that the use of a lognormal prediction model in planning prospective clinical trials is only a part of the overall spectrum [see 4.1 and 4.3 in Table VII] which has to be considered at the design stage. It can, though, be very useful if it can be shown that such a prediction model will shorten the delay in waiting for the definitive trial result.

The number of patients obviously affects the overall efficiency of a trial, as well as specifically affecting ξ and Figure 12 illustrates [52] the numbers required as a function of the differences in the proportions cured C_1 and C_2 for a P=0.05 level of significance [the α risk] and for three different powers [1-β]. Data such as in Ta-

Tab. V. Calculation schedule, termed method B by Boag [16], to determine, using maximum likelihood, the corrected estimates of the lognormal mean and of the cured proportion of patients when a value of the lognormal standard deviation is fixed *a priori*

CALCULATION SCHEDULE FOR METHOD B

σ = 0·38 — 1ST ESTIMATES m = 1·00 c = 0·30 — DATA ON 58 TONSIL CASES FROM FIG. 2b — DATE

INTERVALS		GROUP ①		GROUPS ②&③										GROUP ④				
LIMITS t	CENTRAL x	n	nx	n	ϕ	$n\phi$	$n\phi x$	$1/z$	$n\phi^2$	$n\phi^2/z$	$n\phi^2x/z$	ψ	$n\psi$	$n\psi^2$	n	z/q	nz/q	nxz/q	$n(z/q)^2$
0·9	-3·0							225·6											
1·4	-2·5	1	2·5		·014			57·05											
2·2	-2·0	2	4	1	·039	·039	·078	18·52	·0015	·028	·056	·024	·024	0					
3·4	-1·5	1	1·5		·095			7·721											
5·2	-1·0	4	4	1	·191	·191	·191	4·133	·0365	·151	·151	·179	·179	·032					
8·1	-0·5	3	1·5	1	·315	·315	·158	2·840	·0991	·282	·141	·394	·394	·155					
12·4	0	7	13·5	1	·430	·430	·427	2·507	·184	·461	·348	·770	·770	·593					
19·3	0·5	5	2·5	6	·479	2·874	1·437	2·840	1·375	3·903	1·952	1·34	8·04	10·77	3	1·140	3·42	1·71	3·90
29·8	1·0	2	2	5	·413	2·065	2·065	4·133	·852	3·52	3·52	2·05	10·25	21·0	1	1·524	1·52	1·52	2·32
46·3	1·5	2	3	2	·262	·524	·786	7·721	·137	1·059	1·589	2·69	5·38	14·47					
71·5	2·0			10	·120	1·200	2·400	18·52	·144	2·663	5·326	3·09	30·9	95·3					
111·0	2·5				·040			57·05				3·27							
	3·0				·010			225·6				3·32							
			7·5				6·688				12·39								
TOTALS		27	-6·0	27		7·638	6·261		2·829	12·07	12·04		55·94	142·32	4		4·94	3·23	6·22
		N_1	ⓐ	N_{2+3}		ⓒ	ⓓ		ⓖ	ⓛ	ⓜ		ⓝ	ⓞ	N_4		ⓒ	ⓓ	ⓙ
						4·94	3·23		6·22										
						12·58	9·49		9·05										

A = ⓐ + ⓒ = 6·58
C = ⓝ − (N₁+N₄)/(1−c)² = 11·65
D = N₁ + ⓖ − ⓓ = 26·56
F = ⓞ + (N₁+N₄)/(1−c)² = 205·6
H = ⓛ/(1−c)² = 24·6

$(S.E.)_m = \sigma \times \sqrt{\frac{F}{\Delta}}$

$(S.E.)_c = \sqrt{\frac{D}{\Delta}}$

D	H	=	26·6	24·6	+ 5470	Δ =	
H	F		24·6	206	− 605	+ 4865	
A	H	=	6·58	24·6	+ 1353		
C	F		11·65	206	− 286	+ 1067	
D	A	=	26·6	6·58	+ 310		
H	C		24·6	11·65	− 162	+ 148	

$\delta_m = \frac{1067}{4865} \times 0{\cdot}38 = +{\cdot}0834$

$\delta_c = \frac{148}{4865} = +{\cdot}0304$

$(S.E.)_m = \sqrt{\frac{206}{4865}} \times 0{\cdot}38 = \pm{\cdot}078$

$(S.E.)_c = \sqrt{\frac{26{\cdot}6}{4865}} = \pm{\cdot}074$

CORRECTED ESTIMATES
m = 1·083 c = 0·330

Tab. VII. Clinical trial design aims and objectives

1. The clinical questions, i.e. what treatment methods are being investigated.
2. The clinical material, i.e. what population is being studied.
3. The design of the study, e.g. phase I, II or III; randomised or non-randomised.
4. Statistical analyses and quality assurance considerations.
   - 4.1 How is the criterion of success to be defined and measured and for what improvement in success is it considered worthwhile organising a clinical trial.
   - 4.2 What level of statistical significance are we prepared to accept when analysing the results.
   - 4.3 Given the number of patients available for entry into the trial, what is likely to be the duration of the trial.
   - 4.4 Can historical controls be used.
5. Endpoints, i.e. what measure(s) of patient welfare.

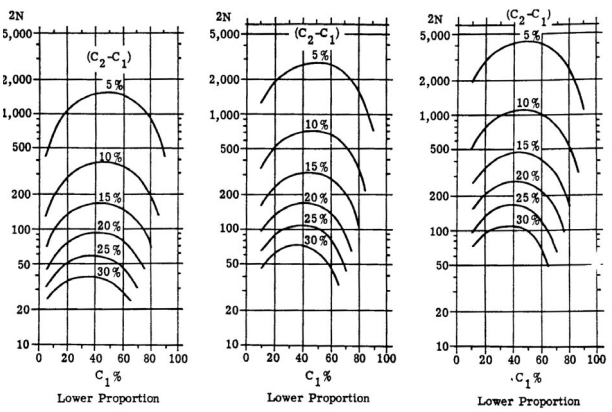

Fig. 12. Charts to determine the number of patients required in a clinical trial for different combinations of variables: α risk [i.e. P = 0.05], power [1-β] and the difference between the proportions cured in the two treatment groups [$C_2 - C_1$]. N = number of cases in each treatment group. (**Left**) Number of cases required in a clinical comparison of two treatments in order that the observed difference [$C_2 - C_1$] should be statistically significant at the P = 0.05 level, [1-β] =0.50. (**Centre**) Number of cases required in a clinical comparison of two treatments in order to stand a 3 in 4 chance [i.e. [1-β] =0.75] of detecting at the P = 0.05 level a difference [$C_2 - C_1$]. (**Right**) Number of cases required in a clinical comparison of two treatments in order to stand a 9 in 10 chance [i.e. [1-β] =0.90] of detecting at the P = 0.05 level a difference [$C_2 - C_1$].

ble VI should be used in conjunction with data such as in Figure 12.

Alternative prediction models to the lognormal

It has already been mentioned that three other models have been studied, as well as the lognormal. These are the

Tab. VI. Values of the mean information fraction per patient $\bar{\xi}$ for the lognormal model with an assumed value of $\sigma = 0.30$. The top half of the table is for an analysis *three years* after the five years of patient intake of 20 per year, and the bottom half of the table is for an analysis *two years* after the intake closed [45]

M =	0.90	0.95	1.00	1.05	1.10	1.15	1.20	1.25	1.30	1.35	1.40	1.45	1.50	1.55	1.60	1.65	1.70	1.75
C = .05	.9530	.9272	.8914	.8444	.7857													
.10	.9752	.9605	.9391	.9090	.8688	.8174	.7550	.6826										
.15	.9831	.9729	.9576	.9355	.9048	.8641	.8125	.7500	.6777	.5977	.5131							
.20			.9674	.9500	.9252	.8915	.8475	.7925	.7270	.6523	.5708	.4857	.4007					
.25			.9736	.9591	.9384	.9096	.8712	.8222	.7624	.6925	.6145	.5310	.4455	.3620	.2842			
.30				.9655	.9476	.9225	.8885	.8442	.7891	.7236	.6490	.5676	.4826	.3977	.3170			
.35						.9321	.9016	.8613	.8103	.7486	.6771	.5980	.5139	.4286	.3458	.2694		
.40							.9119	.8749	.8274	.7691	.7007	.6237	.5409	.4555	.3715	.2926	.2224	
.45								.8860	.8416	.7864	.7207	.6460	.5645	.4795	.3946	.3138	.2407	.1782
.50									.8536	.8011	.7380	.6654	.5854	.5009	.4155	.3332	.2578	
.55										.8139	.7532	.6826	.6040	.5202	.4346	.3512	.2739	.2059
.60										.8251	.7666	.6979	.6208	.5378	.4522	.3679	.2889	.2187
.65										.8349	.7785	.7117	.6360	.5538	.4684	.3835	.3031	.2309
.70										.8437	.7892	.7242	.6499	.5686	.4834	.3981	.3165	.2426
.75										.8516	.7989	.7355	.6626	.5822	.4974	.4117	.3292	.2537
.80										.8587	.8077	.7459	.6743	.5949	.5104	.4246	.3413	.2644
C = .05	.8746	.8279	.7724	.7086	.6376													
.10	.9262	.8939	.8528	.8028	.7443	.6781	.6055	.5284										
.15	.9475	.9228	.8901	.8486	.7984	.7397	.6735	.6011	.5243	.4455	.3676							
.20			.9120	.8766	.8324	.7795	.7184	.6502	.5764	.4990	.4204	.3437	.2720					
.25			.9266	.8956	.8561	.8079	.7511	.6866	.6157	.5401	.4619	.3839	.3092	.2408	.1812			
.30				.9095	.8738	.8293	.7762	.7151	.6469	.5731	.4958	.4174	.3409	.2694	.2056			
.35						.8463	.7964	.7381	.6724	.6005	.5243	.4460	.3685	.2947	.2277	.1700		
.40							.8129	.7572	.6938	.6238	.5488	.4709	.3927	.3174	.2479	.1869	.1361	
.45								.7735	.7122	.6439	.5701	.4928	.4144	.3379	.2664	.2027	.1489	.1057
.50									.7281	.6614	.5889	.5122	.4339	.3565	.2834	.2175	.1610	
.55										.6770	.6057	.5298	.4515	.3737	.2993	.2314	.1725	.1242
.60										.6910	.6208	.5456	.4677	.3894	.3140	.2445	.1835	.1329
.65										.7036	.6345	.5601	.4825	.4040	.3278	.2568	.1940	.1413
.70										.7150	.6470	.5735	.4962	.4176	.3407	.2685	.2041	.1494
.75										.7255	.6585	.5858	.5089	.4303	.3529	.2797	.2137	.1573
.80										.7351	.6692	.5971	.5208	.4423	.3644	.2902	.2229	.1649

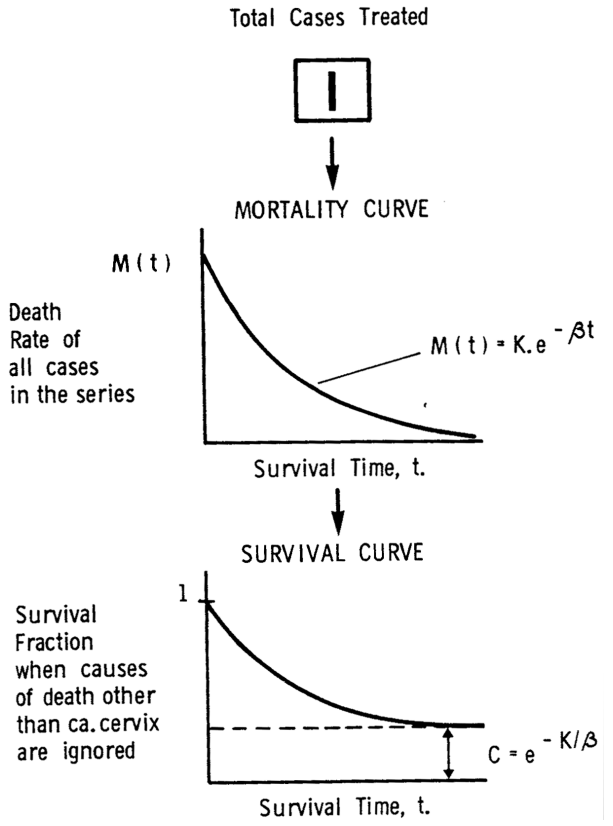

Fig. 13. Schematic diagram of the extrapolated actuarial model proposed by Haybittle in 1959 [46].

negative exponential (Fig. 6), the skew exponential (Fig. 6) and the extrapolated actuarial (Fig. 13) and given below are the derivatives of the log likelihood for these three models using a presentation similar to that in Table XXVIII on page 346 for the lognormal. None of them have been as intensively studied as the lognormal and it is therefore possible, particularly for the skew exponential, that for certain cancer populations [certain tumour sites and subgroups by stage, histology and age group, for example] that the lognormal will not be the optimum prediction model. To determine this, a validation procedure will have to be undertaken, following the procedure in Figure 7. This will include a knowledge of the derivatives of the log likelihood for the other models.

Negative exponential model

The equation for the negative exponential which is analogous to Eq. 8 [which is the *probability density function*] for the lognormal, is Eq. 16, but whereas the lognormal has two variables μ and σ the negative exponential has only one variable α.

$$y = N(t) = \alpha . \exp(-\alpha.t) \quad \text{..... Eq. 16}$$

For the estimation of the parameters of this model, α and C, Haybittle [46] suggested combining Group 3 and Group 4 patients and therefore only two columns of derivatives are given in Table VIII. The corrections $\delta\alpha$ and δC which must be applied to the first estimates α and C are given by Eq. 17 and Eq. 18 where A, C, D, F and H are defined in Table VI.

$$A - D.\,\delta\alpha - H.\,\delta C = 0 \quad \text{..... Eq. 17}$$

$$C - H.\,\delta\alpha - F.\,\delta C = 0 \quad \text{..... Eq. 18}$$

Tab. VIII. Derivatives of the log likelihood for the negative exponential prediction model

Quantity	Symbol	Contribution by a patient in group 1	Contribution by a patient in groups 2 and 3
$\frac{\partial L}{\partial \alpha}$	A	$(1/\alpha) - t$	$-t.\phi$
$\frac{\partial L}{\partial C}$	C	$-\frac{1}{(1-C)}$	ψ
$-\frac{\partial^2 L}{\partial \alpha^2}$	D	$1/\alpha^2$	$-\frac{C.t^2.\phi^2}{(1-C).Q}$
$-\frac{\partial^2 L}{\partial C^2}$	F	$\frac{1}{(1-C)^2}$	ψ^2
$-\frac{\partial^2 L}{\partial C.\partial \alpha}$	H	0	$-\frac{t.\phi^2}{(1-C)^2.Q}$

In the above table:

$$Q = e^{-\alpha t}$$

$$\phi = \frac{1}{1 + \frac{C}{Q(1-C)}}$$

$$\psi = \frac{1}{C + \frac{Q}{(1-Q)}}$$

Skew exponential model

A family of seven skew exponentials of the general form in Eq. 19 were first considered by Mould [45] as alternative analytical forms to the lognormal for the specification of the distribution of survival times of cancer patients who died with their disease present.

$$y = N(t) = N_0.t.\exp(-\gamma.t^n) \quad \text{..... Eq. 19}$$

The seven skew exponentials were defined by m = 1,7, where m is given in Eq. 21 and Eq. 22. where $z = (-\gamma.t^n)$, and (n-2) = -m.n and

$$t.dt = [(dz/[n.\gamma]).(z/\gamma)]^{-(n-2)/n} \quad \text{..... Eq. 20}$$

The integral of Eq. 19 can be written in the form of a gamma function $\Gamma(m+1)$ where m is an integer.

$$\int_0^\infty N(t).dt = [1/[n.\gamma^{m+1}]].\int_0^\infty z.\exp(-z).dz = [1/[n.\gamma^{m+1}]].\Gamma(m+1)$$

..... Eq. 21

For the limits 0 to T the integral in Eq. 21 may be evaluated using the expression in Eq. 22.

$$\int_0^T N(t).dt = \left[-(1/m!).\, z[z^m + m.z^{m-1} + \right.$$
$$\left. + m.(m-1).z^{m-2} + \ldots\ldots + m!] \right]_0^T$$

..... Eq. 22

For cancer of the cervix data the optimum skew exponential was found to be that with m=3, Eq. 23, but it is emphasised that this will not necessarily always be the optimum curve for other cancer sites.

$$N(t) = [\gamma^4/12].t.\exp(-\gamma.t^{\frac{1}{2}}) \ldots\ldots \text{Eq. 23}$$

The corrections $\delta\gamma$ and δC which must be applied to the first estimates γ_1 and C_1 are given by Eq. 24 and Eq. 25 where A, C, D, F and H are defined in Table IX.

$$A - D.\,[\delta\gamma/\delta] - H.\,\delta C = 0 \ldots\ldots \text{Eq. 24}$$

$$C - H.\,[\delta\gamma/\delta] - F.\,\delta C = 0 \ldots\ldots \text{Eq. 25}$$

Tab. IX. Derivatives of the log likelihood for the skew exponential, m=3, Eq. 23, prediction model

Quantity	Symbol	Contribution by a patient in Group 1	Contribution by a patient in Groups 2 and 3	Contribution by a patient in Group 4
$\gamma \frac{\partial L}{\partial \gamma}$	A	$(4-\gamma.t^{\frac{1}{2}})$	$-\phi$	$-\theta$
$\frac{\partial L}{\partial C}$	C	$-\frac{1}{(1-C)}$	ψ	$-\frac{1}{(1-C)}$
$-\frac{\gamma^2 \partial^2 L}{\partial \gamma^2}$	D	4	$\phi(\phi+3-\gamma t^{\frac{1}{2}})$	$\theta.(3-\gamma.t^{\frac{1}{2}})+\theta^2$
$-\frac{\partial^2 L}{\partial C^2}$	F	$\frac{1}{(1-C)^2}$	ψ^2	$\frac{1}{(1-C)^2}$
$-\frac{\gamma.\partial^2 L}{\partial C \partial \gamma}$	H	0	$\left[\frac{-\psi\phi}{(1-C)(1-Q)}\right]$	0

In the above table:

$$\phi = \frac{2Zt}{Q + \frac{C}{(1-C)}} \qquad \theta = 2Z.t/Q$$

$$\psi = \frac{1}{C + \frac{Q}{(1-Q)}} \qquad Z = \frac{\gamma^4.t}{12} e^{-\gamma.t^{\frac{1}{2}}}$$

Extrapolated actuarial model

The extrapolated actuarial model differs from the lognormal, skew exponential and negative exponential models by assuming a certain time variation in the death rate from cancer instead of postulating explicitly a fraction cured C, Figure 13. This assumption, however, leads implicitly to a fraction of long-term sruvivors which can be calculated from the model parameters.

The other models initially assume an analytical distribution of survival times for the unsuccessful group of cases, the fraction (1-C), and solve directly for the parameter C and the location and scale parameters of the assumed distribution of survival times.

The assumption in the extrapolated actuarial model of Haybittle [46] is that the probability of dying from cancer per unit time is given by the expression $K.\exp[-\beta.t]$. The number dying in the interval between t and (t+dt) is then given by Eq. 26

$$-dn = K.\exp[-\beta.t].\,N.\,dt \ldots\ldots \text{Eq. 26}$$

where N = the number of cases at risk at time t. If the total number of patients in a series is N_0 the survival fraction derived from Eq. 26 is given in Eq. 27.

$$[N/N_0] = \exp[\{K/\beta\}.[1 - \exp(-\beta t)]\,] \ldots\ldots \text{Eq. 27}$$

As t tends towards infinity $[N/N_0]$ tends towards $\exp[-K/\beta]$ and this is the fraction which we can identify with C, Eq. 28. The two parameters K and β are estimated simultaneously for this model and the parameter C is obtained from these estimates. The derivatives of the log likelihood for this model are given in Table X.

$$C = \exp[-K/\beta] \ldots\ldots \text{Eq. 28}$$

Tab. X. Derivatives of the log likelihood for the extrapolated actuarial model

Quantity	Symbol	Contribution by a patient in group 1 only	Contribution by any patient
$\frac{\partial L}{\partial \beta}$	A	$-t$	$\frac{K}{\beta^2}(1-e^{-\beta t}) - \frac{K}{\beta}.t.e^{-\beta t}$
$\frac{\partial L}{\partial K}$	C	$\frac{1}{K}$	$-\frac{1}{\beta}(1-e^{-\beta t})$
$-\frac{\partial^2 L}{\partial \beta^2}$	D	0	$\frac{2K}{\beta^3}(1-e^{-\beta t}) - \frac{2K}{\beta^2}.t.e^{-\beta t} - \frac{K}{\beta}.t^2 e^{-\beta t}$
$-\frac{\partial^2 L}{\partial K^2}$	F	$\frac{1}{K^2}$	0
$-\frac{\partial^2 L}{\partial K.\partial \beta}$	H	0	$\frac{t.e^{-\beta t}}{\beta} - \frac{(1-e^{-\beta t})}{\beta^2}$

Acknowledgements

I am most grateful to Professor John Boag and Dr John Haybittle for their guidance when I first commenced studying the Lognormal Model. I would also like to thank Dr John Gamel, Dr Patricia Tai and Mr Joseph Wong for helpful discussions and Dr Edward Towpik for his encouragement with the preparation of this Review for *Nowotwory*.

Richard F. Mould MSc, PhD
41, Ewhurst Avenue
South Croydon
Surrey CR2 0DH
United Kingdom
e-mail address: richardfmould@hotmail.com

References

1. National Academy of Sciences Committee on the Biological Effects of Ionizing Radiation. *Health effects on populations of exposure to low levels of ionizing radiation*. BEIR V Reports. Washington DC: US National Academy of Sciences, 1990.
2. McCready DR, Chapman JW, Hanna WM et al. Factors affecting distant disease-free survival for primary invasive breast cancer: use of a log-normal survival model. *Ann Surg Oncol* 2000; 7: 416-426.
3. Chapman JW, Hanna W, Kahn HJ et al. Alternative multivariate modelling for time to local recurrence for breast cancer patients receiving lumpectomy alone. *Surg Oncol* 1996; 5: 265-271.
4. Gore SM, Pocock SJ, Kerr GR. Regression models and non-proportional hazards in the analysis of breast cancer survival. *Appl Statist* 1984; 33: 176-195.
5. Baier K, Baltas D eds. *Modelling in clinical radiobiology*. Freiburg Oncology Series Monograph No.2. Freiburg: Albert-Ludwigs-University-Freiburg; 1997.
6. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J Roy Stat Soc Series B* 1949; 11: 15-53.
7. Mould RF. *Introductory medical statistics*, 3rd edn. Bristol: Institute of Physics; 1998, 297-301.
8. Mould RF, Boag JW. A test of several parametric statistical models for estimating success rate in the treatment of carcinoma cervix uteri. *Br J Cancer* 1975; 32: 529-550.
9. Rutqvist RE, Wallgren A, Nilsson B. Is breast cancer a curable disease? A study of 14,731 women with breast cancer from the cancer registry of Norway. *Cancer* 1984; 53: 1793-1800.
10. Rutqvist RE. On the utility of the lognormal model for analysis of breast cancer survival in Sweden 1961-1973. *Br J Cancer* 1985; 52: 875-883.
11. Kaplan EL, Meier P. Non-parametric estimation from incomplete observations. *J Amer Stat Assoc* 1958; 53: 457-482.
12. Myles DR. *Skew exponential distributions applied to cancer symptom durations*. Coventry Lanchester Polytechnic BSc Physics Sciences dissertation, 1988.
13. Mould RF. *Cancer statistics*. Bristol: Adam Hilger [Institute of Physics Publishing]; 1983, 214-225, 241-253.
14. Kalbfleisch JD, Prentice RL. *The statistical analysis of failure time data*. New York: Wiley; 1980, p. 21-35.
15. Aitchison J, Brown JAC. *The lognormal distribution*. University of Cambridge Department of Applied Economics Monograph 5. Cambridge: Cambridge University Press; 1957.
16. Boag JW. The presentation and analysis of the results of radiotherapy. Part II. Mathematical theory. *Br J Radiol* 1948; 21: 189-203.
17. Boag JW. Statistical problems which arise in cancer therapy. *Clin Radiol* 1960; 11: 150-155.
18. Wood CAP, Boag JW. *Researches on the radiotherapy of oral cancer*. Medical Research Council Special Report Series No. 267. London: His Majesty's Stationery Office; 1950, 107-122.
19. Abramowitz M, Stegun IA. *Handbook of mathematical functions*. New York: Dover Publications; 1965, p. 932.
20. Lea DEA. The biological assay of carcinogens. *Cancer Res* 1945; 5: 633--640.
21. Gaddum JH. Lognormal distributions. *Nature* 1945; 156: 463-466.
22. Schrek R, Lipson HI. Logarithmic frequency distributions. *Human Biol* 1941; 13: 75-22.
23. McAlister D. The law of the geometric mean. *Proc Roy Soc* 1879; 29: 367.
24. Kapteyn JC. *Skew frequency curves in biology and statistics*. Groningen: Noordhoff; 1903.
25. Fisher RA. *The mathematical theory of probabilities*. London: Macmillan; 1922.
26. Mould RF, Hanham IWF, McSweeney BFD, Myles DR. The lognormal distribution as a fit to symptom duration in the range 0-2 years for 26,000 cases. *Br J Cancer* 1987; 56: 687-689.
27. Spratt JS, Meyer JS, Spratt JA. Rates of growth of human solid neoplasms. *Part 1. J Surg Oncol* 1995; 60: 137-146.
28. McLaren CE, Brittenham GM, Hasselblad V. Analysis of the volume of red blood cells: application of the expectation maximisation algorithm to grouped data from the doubly truncated lognormal distribution. *Biometrics* 1986; 42: 143-158.
29. McLaren CE, Wagstaff M, Brittenham GM et al. A detection of two-component mixtures of lognormal distributions in grouped doubly truncated data: analysis of red blood cell volume distributions. *Biometrics* 1991; 47: 607-622.
30. Whittemore A. Lung cancer incidence in cigarette smokers: further analysis of Doll and Hill's data for British physicians. *Biometrics* 1976; 32: 805--816.
31. Chung SJ. Formula predicting carcinoma-free probability in rats exposed to carcinogen DMBA. Int *J Biomed Comput* 1990; 26: 171-181.
32. Rabbe A. Radiation treatment of cancer of the cervix of the uterus at the Radium Institute in Copenhagen from 1951-54. *Acta Obstet Gynecol Scand Suppl 30 1974 and Mould RF.* Radiation treatment of cancer of the cervix of the uterus at the Radium Institute in Copenhagen from 1951-54. *Acta Obstet Gynecol Scand* 1975; 54: 389-391.
33. Mould RF, Hearnden T, Palmer M, White GC. Distribution of survival times of 12,000 head and neck cancer patients who died with their disease. *Br J Cancer* 1976; 34: 180-190.
34. Gamel JW, Jones AS. Squamous carcinoma of the head and neck: cured fraction and median survival time as functions of age, sex, histologic type, and node status. *Br J Cancer* 1993; 67: 1071-1075.
35. Berg JW. The distribution of cancer deaths in time: a survey test of the lognormal model. *Br J Cancer* 1965; 19: 695-711.
36. Gamel JW, Vogel RL, McLean IW. Assessing the impact of adjuvant therapy on cure rate for stage 2 breast carcinoma. *Br J Cancer* 1993; 68: 115--118.
37. Gamel JW, Vogel RL. A model of long-term survival following adjuvant therapy for stage 2 breast cancer. *Br J Cancer* 1993; 68: 1167-1170.
38. Gamel JW, Vogel RL, Valagussa P, Bonnadonna G. Parametric survival analysis of adjuvant therapy for stage II breast cancer. *Cancer* 1994; 74: 2483-2490.
39. Gamel JW, George SL, Stanley WE, Seigler HF. Skin melanoma. Cured fraction and survival time as functions of thickness, site, histologic type, age, and sex. *Cancer* 1993; 72: 1219-1223.
40. Chung SJ. Formula predicting survival in patients with invasive cutaneous malignant melanoma. *Int J Biomed Comput* 1991; 28: 151-159.
41. Denham JW, Denham E, Dear KB et al The follicular non-Hodgkin's lymphoma: the possibility of cure. *Eur J Cancer* 1996; 32A: 470-479.
42. Yamashita N. Prediction of survival in lung cancer patients with radiation therapy. Maximum likelihood estimation of two parameters of lognormal curves from selected survival patterns. *Nippon Igaku Hoshasen Gakkai Zasshi* 1974; 34: 102-107.
43. Ito S, Suzaki K, Tsujii H et al An analysis of survival data of urinary bladder cancer patients after radiotherapy. *Nippon Igaku Hoshasen Gakkai Zasshi* 1977; 37: 685-690.
44. *Chambers English dictionary*. Cambridge: Chambers; 1988.
45. Mould RF. S*tatistical models for studying long-term survival results following treatment for carcinoma of the cervix*, London University PhD thesis, 1973.
46. Haybittle JL. The estimation of the proportion of patients cured after treatment for cancer of the breast. *Br J Radiol* 1959; 32:725-733.
47. Haybittle JL. The early estimation of the results of treatment for cancer. *Br J Radiol* 1960; 33: 502-507.
48. Haybittle JL. The estimation of T-year survival rate in patients treated for cancer. *J Roy Stat Soc Series A* 1962; 125: 268-283.
49. Haybittle JL. The cured group in series of patients treated for cancer. *Anglo-German Med Rev* 1964; 2: 422-436.
50. Haybittle JL. A two-parameter model for the survival curve of treated cancer patients. *J Amer Stat Assoc* 1965; 60: 16-26.
51. Kaplan EL, Meier P, Non-parametric estimation from incomplete observations. *J Amer Stat Assoc* 1958; 53: 457-482.
52. Boag JW, Haybittle JL, Fowler JF et al The number of patients required in a clinical trial. *Br J Radiol* 1971; 44: 122-125.
53. Berkson J, Gage RP. Survival curve for cancer patients following treatment. *J Amer Stat Soc* 1952; 47: 501-515.

Accepted: 17 July 2001