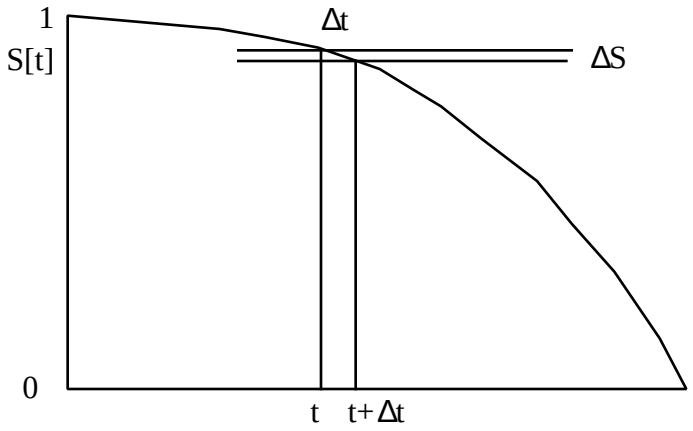
Incidence, cumulative incidence, survival function...

Expression in Rothman page 31 $CI[T] = 1 - \exp[-\int I[t]\, dt]$

He derives this by integrating the equation $I(t) = \frac{-dS[t]}{S[t]dt}$

where S[t] and dS[t] are:



Expression in Miettinen page 249, using the notation ID (Incidence Density) rather than I,

$CI[t_0 \text{ to } t_1] = 1 - \exp[-\int ID[t]\, dt]$ with integration limits t_0 to t_1

Miettinen gives Chiang, Introduction to Stochastic processes in Biostatistics, Wiley New York, 1968 as a reference.

How does the exp[ ] function come into it?

Consider a short period of time (e.g. 1 year) where the Incidence Density, ID, is constant. To make matters concrete, say the ID for the event is 2×10^2 PY^{-1}. What fraction of individuals free of the event at time t = 0, would still be event-free at time t = 1? (to use notation from survival analysis, we might call this quantity S[1], and say that S[0] = 1).

The first instinct is to say that the proportion event-free would be 1 - 0.02 = 0.98; but writing this is mixing metaphors, so to speak: one cannot subtract 0.02, which is a density (with units PY^{-1}), from 1, which is a unitless quantity. Technically, the 0.98 has no meaning.

However, consider the 1 year as 365 days, and do the arithmetic day by day. With such a small unit of time, the use of $1 - \frac{0.02}{365}$ as a conditional probability is less of a transgression.

Then, we can approximate the fraction of individuals free of an event at time t=1 as the product of 365 conditional probabilities

$$S[1] = \left(1 - \frac{0.02}{365}\right) \times \left(1 - \frac{0.02}{365}\right) \ldots \left(1 - \frac{0.02}{365}\right) = \left(1 - \frac{0.02}{365}\right)^{365}$$

$$= 0.980\ 198\ 136$$

We could go even further and calculate hour by hour, multiplying 365×24 conditional probabilities of $1 - \frac{0.02}{365 \times 24}$ each to get the fraction surviving as

$$S[1] = \left(1 - \frac{0.02}{365 \times 24}\right)^{365 \times 24}$$

$$= 0.980\ 198\ 651$$

Dividing the 1 year into an even larger number of infinitely small time increments has very little effect, and one can check that in the limit, as n-> ∞

$$S[1] = \left(1 - \frac{0.02}{n}\right)^{n}$$

$$= 0.980\ 198\ 673$$

In fact, the Limit of $\left(1 - \frac{x}{n}\right)^{n}$ as n goes to infinity is a very special function in mathematics: it's the exp[-x] function

$$e = e^1 = \exp[1] = \text{ limit as n-> of } \left[1 + \frac{1}{n}\right]^{n}$$

$$\exp[x] = \text{ limit as n-> of } \left[1 + \frac{x}{n}\right]^{n} \qquad \exp[-x] = \text{ limit as n-> of } \left[1 - \frac{x}{n}\right]^{n}$$

Thus, of those free at t=0, the proportion event-free at t=1 is $S[1] = \exp[- ID \times 1]$

From S[1] to S[T]

Now that we have an expression for the proportion event-free at the end of 1 time unit, we can write the general formula for the proportion event-free at the end of T unit of time as a product of the time unit by time unit conditional probabilities Prob[> t | > t -1] , i.e.

$$S[1] = \text{Prob}[> 1 \mid > 0]$$

$$S[2] = \text{Prob}[> 1 \mid > 0] \times \text{Prob}[> 2 \mid > 1]$$

$$S[3] = \text{Prob}[> 1 \mid > 0] \times \text{Prob}[> 2 \mid > 1] \times \text{Prob}[> 3 \mid > 2]$$

or

$$S(T) = \prod \text{Prob}[> i \mid > i\text{-}1]$$ ($\prod$ means product, just like $\sum$ means Sum)

Since each Prob[> i | > i-1] can be computed from the ID for that interval as $\exp[-ID_i]$, we can write the S[T] product as

$$S[T] = \prod \exp[-ID_i] = \exp[-\sum ID_i]$$

If we want to, we can replace ID_i by the integral or summation $\int ID[t]dt$ where the integration or summation is over the little time units in interval i, and if we further take the summation over all intervals, then S[T] simplifies to

$$S(T) = \exp[-\int ID[t]dt] = \exp [-H[T]]$$

where the summation or integration is from 0 to T.

The integral $H[T] = \int ID[t]dt$ is called the "integrated hazard" or "cumulative hazard" and denoted H(T) or $\Lambda(T)$.

Note the reverse relation

$$H[T] = -\log[S[T]].$$

It is common to plot the log of a survival curve against T to see if certain assumptions concerning ID are borne out by the data.

Special Cases

- If ID[t] is constant over t, then

  $$H(T) = \int ID[t]dt$$

  simplifies to

  $$ID \times T$$

  so that

  $$S(T) = \exp[- ID \times T]$$

  Thus, the cumulative incidence up to time T, CI(T), simplifies to

  $$CI(T) = 1 - S[T] = \exp[- ID \times T].$$

- Moreover, if ID[t] is constant over t, and $ID \times T$ is small

  $$CI[T] = 1 - \exp[- ID \times T]. \qquad ID \times T$$

  or

  $$S[T] = 1 - ID \times T$$

  That's why the fancy calculations above came so close to the naive 0.98; had the ID been 50×10^2 PY^{-1} , the approximation would have not been very good.

  Plots of functions of log[S[T]] versus t or log[t] can help suggest parametric forms for ID[t].

Another way to link ID[t] and S[t]

Relate, at time t, the number of persons who become positive to the number negative at time t , i.e. as a rate,... if f[t]dt persons become positive in the interval (t,t+dt) then the rate or hazard at time t is

$$ID[t] = \text{limit} \frac{f[t]dt}{S[t]} \text{ as } dt \text{ –> } 0$$

or

$$ID[t] = \text{limit} \frac{-dS[t]}{S[t]} \text{ as } dt \text{ –> } 0.$$

We, as Rothman does, can reverse this equation to give

$$S[t] = \exp[-\int ID[t]dt].$$

Cumulative Rate and Cumulative Risk

[for comparison of cancer incidence]

(author N.E. Day)

Chapter 10 (pp 668-670) from Cancer Incidence in Five Continents VOL IV

Editors J WATERHOUSE, C MUIR, K. SHANMUGARATNAM , J POWELL In collaboration with D PEACHAM, S WHELAN, *Technical Editor for IARC* W. DAVIS, IARC SCIENTIFIC PUBLICATIONS No. 42, Lyon, 1982.

In Volume lll of this series, the cumulative rate was advanced as a new age standardized incidence rate. Some of the advantages that would follow from its widespread adoption were described. In this volume rates standardized to the so-called European and African populations have been dropped and replaced by rates cumulated over the age ranges 0-64 and 0-74. This change must be regarded as an encouraqing step towards the objective of removing arbitrary rates of weight completely from descriptive epidemiology. The use of different standard populations, not infrequently without specifying which one has been used, has been a definite hindrance to communication in the past, and interfered with the primary aim of age standardization, comparability. Since the cumulative rate is still not widely used, and since each volume in this series is intended to be self-contained, the reappearance of a virtually identical chapter to the one in Volume III was considered warranted.

The purpose of this chapter, as in the corresponding chapter of Volume III, is to encourage the greater use of the cumulative rate, which is both a directly standardized incidence rate and a good approximation to the actuarial or cumulative risk.

Cumulative rate: Definition

Before defining the *cumulative rate,* the concept of the *cumulative risk* will be introduced. The cumulative risk is the risk an individual would have of developing the disease in question during a certain age period if no other causes of death were in operation. The age period over which the risk is accumulated has to be specified, and would depend on the comparison being made. Thus for childhood tumours one would take age 0-14, for example. However, in general the whole life span risk would be the appropriate measure, which can be taken as 0-74. If the instantaneous incidence rate at an age t is given by I(t), then the cumulative risk between apes t_1 and t2 is given by

(1) $1 - \exp[- \int I[t]dt]$ with limits t_1 and t_2

The expression inside the exponential is closely approximated by the sum of the age specific incidence rates for each year of age between the two limits t_1 and t_2.

Now, the risk for any specific tumour up to age 75 is very rarely over 0.1, and for a value of 0.1, one can approximate expression (1) by

(2) $\int I[t]dt$

The accuracy of the approximation is shown in Table 10.2 (for example, if expression (2) equals 0.1, expression (1) has a value of 0.095).

Expression (2) is dimensionless, and so not strictly a rate which should have dimension $(\text{time})^{-1}$. However, if the integration is looked upon as a weighted sum of incidence rates, with dimensionless weights, which would be the same numerically, then expression (2) can be regarded as an incidence rate.

The proposed measure is, *then, the sum* over *each* year of age of the *age-specific incidence* rates, *taken* from *birth* to age 74+ (as age-specific incidence rates are usually computed for 5-year age intervals, the cumulative rate is five times the sum of the age-specific incidence rates calculated over five year age-groups, or during the first five years of life, where the group 0-1 is often given separately, by the rate in the 0-1 age-group plus four times the rate in the 1-4 age-group. It can be interpreted either as a directly age-standardized rate with the same population size in each age-group, or as an approximation to the cumulated risk. It has *been proposed* to *call* the measure the Cumulative Rate. It is more conveniently expressed per hundred (per cent) than per hundred thousand.

A corresponding measure for, childhood cancer would sum the age-specific rates over each year of age, 0-14.

Standard error of the cumulative rate

The variance and standard error of the cumulative rate can be derived directly from the general expression for the standard error of a directly standardized rate given in Chapter 11. If we have k age groups and the age-specific rate in age group i is based on r cases and n person years, then the age-standardized rate given by

$$\sum w \frac{r}{n}$$

has a variance (based on the Poisson distribution) of

$$\sum r \left\{ \frac{w}{n} \right\}^2$$

and a standard error equal to the square root of the variance.

Advantages of the proposed measure

1 As a form of direct age-standardization, the arbitrariness in choosing a standard population is removed, and the calculations are simpler.

2 As an approximation to the cumulative risk, it

(a) has greater intuitive appeal than that of an incidence rate standardized to some arbitrary population, and is more directly interpretable as the carcinogenic load of the specific environment for the specific site;
(b) is a natural way of expressing the tumour experience of cohorts defined by year;
(c) can be combined with the relative risk outlined from analytical i.e.. case control) studies to obtain measures of risk for particular groups see below) This definition of risk for subgroups would seem of particular importance both for the individuals concerned and those treating them, and
(d) is directly comparable with the risks observed in animal experiments, when the latter are analysed by the correct life table approach. This point may not seem of much importance, but nevertheless consideration of Table 10.1 shows the error in the often repeated belief that tumour incidence in animal experiments is almost always an order of magnitude greater than for human populations. Thus the risk of stomach cancer in males in Japan up to age 75 is 12%, which is within the range observed in experimental situations. However, it should be remembered that the effect of other diseases such as those occurring in middle age in developing countries is ignored.

Some examples

To demonstrate the range of values one obtains using the cumulative rate, and to compare these values with the truncated or age-standardized rate for the same neoplasm, examples are given in Table 10.1.

Table 10.2 demonstrates the correction needed to convert the proposed measure into a mathematically precise measure of risk (other causes of death excluded). For values under 10% the change is small.

Combination with relative risks

Suppose a factor X occurs in p% of the population and is associated with a relative risk of r for a particular tumour. If the cumulative risk for the whole population is R, then the cumulative risk, R', for those with factor X is given by

$$pR' + (1—p)\frac{R'}{r} = R$$

or

$$R' = rR/(1 + rp—p)$$

Thus consider a woman in Birmingham, UK, whose mother has a cancer of the breast. The associated relative risk is of the order of 3. If one assumes that 5% of women are thus affected, the risk is given by:

$$\frac{5.58x3}{1 +0.15 -0\ 05} = 15.2$$

indicating a very substantial risk (close to the figure for a second primary among long term breast cancer survivors).

Taking as a second example lung cancer among Birmingham males, the 10% of the population with the highest cigarette consumption have a risk for lung cancer 6 times higher than the rest of the population. The risk for such men is then given by:

$$\frac{9.73x5}{1+0.5-0.1} = 34.75$$

a very high value indeed.

The absolute risks thus measured appear to be precisely the quantity one would want to know for different groups under consideration, and considerably more meaningful than some annual incidence rate.

Cancer epidemiology often appears divided into two disjointed activities. the establishing of incidence rates on the one hand and the identification of risk factors and quantification of relative risks on the other. Both, however, should have as their aim the measurement of the absolute risk, as much for populations as a whole as for individuals with given characteristics. The aim of this note is to propose a measure of incidence which should assist in unifying the diverse measures in present use.

N . E. Day

Note from JH: See also Section 2.3 (Cumulative Incidence Rates) in Volume I of Breslow and Day Statistical Methods in Cancer Research, IARC Scientific Publications No. 32, Lyon 1980