AGE OF ONSET DISTRIBUTION OF A DISEASE, I. ESTIMATION FROM INCIDENCE DATA

by

Regina C. Elandt-Johnson

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1119

May 1977

AGE OF ONSET DISTRIBUTION OF A DISEASE. I. ESTIMATION FROM INCIDENCE DATA*

Regina C. Elandt-Johnson
Department of Biostatistics
University of North Carolina
Chapel Hill, N.C. 27514, U.S.A.

Abstract

The age distribution of the *first episode* of a certain event among individuals who have experienced this episode is called the *age at onset distribution*. Using an appropriate life table and model assuming a stationary population, the incidence data can be adjusted by allowing for mortality (by multiplying incidence rates by the L_x column of the life table). From adjusted incidence data, the probability of onset in a given age group, among those who experience onset sometime, and so the cumulative distribution function of age at onset is estimated. An example comparing age at onset distributions for breast cancer in White and Black Females is presented.

Key Words:
Age at onset; Incidence rates; Life table; Breast cancer

*This investigation was supported by NIH research grant number 1 R01 CA17107 from the National Cancer Institute.

1. AGE OF INCIDENCE DISTRIBUTION: BASIC CONCEPTS

1.1. Imagine a cohort of ℓ_0 individuals born, who experience a certain mortality pattern from birth until the last member of the cohort dies. A *life table* (which is usually constructed from population mortality data) can be regarded as a survival model for such a hypothetical cohort. The value ℓ_0 can be chosen arbitrarily; a common choice is $\ell_0 = 100{,}000$. Life tables include the following functions:

q_x - the conditional probability of dying between age x to x+1 given alive at exact age x;

ℓ_x - the expected number of survivors at exact age x out of ℓ_0 initially at age x = 0;

d_x - the expected number of deaths between age x to x+1;

L_x - the expected number of years to be lived by ℓ_0 persons at age x = 0 during the year of age x to x+1 from the date of birth. It is also the number of persons aged x last birthday in the corresponding stationary population.

1.2. Suppose that in such a cohort the *occurrence*, also called the *incidence* of a certain event (e.g. a specified disease) is observed.

Let c_x denote the number of individuals who experience the *first episode* of the event between age x and x+1. In epidemiology, these are called *new cases*.

The (central) *incidence rate* of the first episode of the event per year per person aged x to x+1 is

$$i_x = \frac{c_x}{L_x} \quad . \tag{1.1}$$

Note that the denominator in (1.1) represents a 'reference population' rather than a 'population at risk', since previous new cases (at earlier ages) are included in the denominator.

The probability of experiencing the first episode or *onset* of the event in the age interval x to x+1 is

$$\Pr\{\text{Onset between } x \text{ to } x+1\} = \frac{c_x}{\ell_0} = \frac{i_x L_x}{\ell_0} \quad . \tag{1.2}$$

The (conditional) probability of the onset occurring in the age interval x to x+1 among all those who sooner or later experience the event is

$$\Pr\{\text{Onset between } x \text{ to } x+1 \,|\, \text{Onset sometime}\}$$

$$= \frac{i_x L_x / \ell_0}{(\sum_{t=0}^{\infty} i_t (L_t)/\ell_0} = \frac{i_x L_x}{\sum_{t=0}^{\infty} i_t L_t} = Q_x \ , \quad \text{say.} \tag{1.3}$$

This defines a proper distribution *of age of onset* (or *age of incidence)* with the cumulative distribution function

$$\Pr\{\text{Onset before or at } x \,|\, \text{Onset sometime}\}$$

$$= \Pr\{X \leq x\} = \sum_{y=0}^{x-1} Q_y = F_X(x) \ , \tag{1.4}$$

where the random variable X denotes the age at onset, and $F_X(0) = 0$.

The complement of $F_X(x)$

$$S_X(x) = 1 - F_X(x) \tag{1.5}$$

(the *'survival' distribution function*) is the expected proportion of those who are free of the event at age x, but will experience it sometime in the future.

Of course, if the event is repetitive, one can also speak about age of onset of the second episode given the age at the first episode; about

age of onset of the third episode given the ages at the first and second episodes, etc. In this paper, however, we will always understand by *age of onset* the age of *incidence of the first episode.*

2. ESTIMATION OF AGE AT ONSET DISTRIBUTION FROM POPULATION INCIDENCE DATA

2.1. Incidence data are usually obtained from the *current population* over a relatively short period of time.

Let K_x denote the midyear population age x last birthday, and C_x denote the number of new cases in the same age interval over a given calendar year. Then the *observed* age specific (for ages x to $x+1$) incidence rate *per year per person* is

$$I_x = \frac{C_x}{K_x} . \tag{2.1}$$

Of course, the current population is not a cohort and we need to evaluate the appropriate '*expected*' number of new cases which would occur in a (hypothetical) cohort subject to a certain mortality pattern. The 'expected' number of new cases, $\hat{c}_x$, say, is *age adjusted* (according to the appropriate life table) and is calculated from the formula

$$\hat{c}_x = I_x \cdot L_x . \tag{2.2}$$

The probability Q_x defined in (1.3) is then estimated from the formula

$$\hat{Q}_x = I_x L_x / \left(\sum_{t=0}^{\infty} I_t L_t \right) = \hat{c}_x / \sum_{t=0}^{\infty} \hat{c}_t , \tag{2.3}$$

and the estimated age of onset distribution is

$$\hat{F_X}(x) = \sum_{y=0}^{x-1} \hat{Q}_y . \tag{2.4}$$

($\hat{F_X}(0) = 0$, of course).

Note that if we replace I_t by $I_t \cdot 10^k$, the resulting $\hat{Q}_x$ (and so $\hat{F}_X(x)$) are the same. Therefore, in the calculation of $\hat{Q}_x$ we may use annual incidence rates per 1,000, 10,000, 100,000 etc. persons (consistently) as desired, without affecting the estimates of Q_x and $F_X(x)$.

2.2. Incidence data from the total population are rather seldom available; more often we have *survey data*. Of course, our method is still applicable if the survey data are *representative* of the corresponding total population. It should therefore always be of serious concern for the investigator who designs the survey to determine conditions under which the survey sample is representative.

2.3. Incidence data are usually recorded for n-year (n = 5 is commonly used in practice) age groups rather than for single years of age (as is also often the case for mortality data). This is almost necessary when the event (disease) is rare and the incidence is low. In such situations, we calculate ${}_nI_x$ (i.e. the annual incidence rate in age group x to x+n), and use ${}_nL_x$ either from an abridged life table, or calculated from a complete life table by the formula

$$ {}_nL_x = L_x + L_{x+1} + \cdots + L_{x+n-1} . \tag{2.5} $$

The probability $\hat{Q}_x$ is replaced by ${}_n\hat{Q}_x$, and is estimated from the formula

$$ {}_nQ_x = {}_nI_x \cdot {}_nL_x / \left(\sum_{t=0}^{\infty} {}_nI_t \cdot {}_nL_{tn} \right) . \tag{2.6} $$

The age of onset distribution will then be tabulated at ages: 0, n, 2n etc.

3. CANCER INCIDENCE DATA THE THIRD U.S. SURVEY, 1969-71

The Third National Survey in U.S. on cancer incidence was carried

out over the period 1969-71. A detailed description of the survey can be found in [1]. The census population, 1970 for the total survey area was 21,003,451 and 181,027 cancers were diagnosed during the period of three years 1969-71. Among these were 6,933 persons who experienced more than one type of cancer. There were, of course, individuals, who already had a cancer at the date of beginning of the study; the ages of onset for these persons were not recorded.

Clearly, from such data we cannot estimate the age of onset distribution for "all cancers". Nevertheless, it is very unlikely that a person who has just suffered incidence of a *specific* cancer, has already had the *same* cancer once before. We are then justified in assuming that, for example, cases recorded as incidence of lung cancer represent *new cases* of *lung cancer*. Similar arguments apply, of course, to other specific cancers.

To illustrate the method, we use the data on breast cancer in females.

3.1. *Age of onset distributions of breast cancer in females*

Table 1 illustrates the calculation of the age of onset distribution of breast cancer of White Females.

Column 2, headed ${}_nI_x \cdot 10^5$, gives the observed annual incidence rates per 100,000 persons in age groups x to $x+5$. The ${}_5L_x$ values given in column 3 are calculated from formula (2.5) using the U.S. Life Table, 1969-71, White Females [2]. The 'expected' number of new cases in age group x to $x+5$ is equal to ${}_5I_x \cdot {}_5L_x$ (see column 4). The values of ${}_5\hat{Q}_x$ (in column 5) and of $\widehat{F_X}(x)$ (in column 6) are calculated from formulae analogous to (2.3) and (2.4), respectively, but using five-year instead of one-year age groups. The last column gives the estimated 'survival' function, $\widehat{S_X}(x) = 1 - \widehat{F_X}(x)$.

TABLE 1

Breast cancer in White Females: Incidence distribution
(Third National Survey. Nat. Cancer Inst. Monograph No. 41 (1975)

Age group x to x+n	$_nI_x \cdot 10^5$	$_nL_x$	$_nI_x \cdot {}_nL_x$	$\widehat{{}_nQ_x}$	$\widehat{F_X(x)}$	$\widehat{S_X(x)}$
0-5	0.0	481233	0.0	0.0	0.0	1.00000
5-10	0.0	478232	0.0	0.0	0.0	1.00000
10-15	0.0	476800	0.0	0.0	0.0	1.00000
15-20	0.0	473425	0.0	0.0	0.0	1.00000
20-25	1.1	465200	5.117	0.00107	0.0	1.00000
25-30	8.7	453764	39.477	0.00828	0.00107	0.99893
30-35	22.5	441392	99.313	0.02084	0.00936	0.99064
35-40	52.5	426181	223.745	0.04694	0.03019	0.96981
40-45	103.7	406490	421.530	0.08844	0.07714	0.92286
45-50	159.2	381185	606.847	0.12732	0.16557	0.83443
50-55	171.7	349082	599.374	0.12575	0.29289	0.70711
55-60	191.8	309462	593.548	0.12453	0.41864	0.58136
60-65	226.2	262779	594.406	0.12471	0.54317	0.45683
65-70	234.2	211217	494.670	0.10378	0.66788	0.33212
70-75	259.8	156931	407.707	0.08554	0.77167	0.22833
75-80	294.9	105305	310.544	0.06515	0.85721	0.14279
80-85	301.3	64556	194.507	0.04081	0.92236	0.07764
85+	307.9	57017	175.555	0.03683	0.96317	0.03683
∞					1.00000	0.00000

3.2. *Comparison of age of onset distributions of breast cancer in White and Black Females*

Calculations similar to those in Table 1 were carried out for Black Females. A graphical presentation is used for comparisons of the age of onset distributions for White Females (WF) and Black Females (BF).

(i) Fig. 1 represents the *observed* age specific *incidence rates* of breast cancer for White and Black Females, respectively. Although they do not reflect the age compositions in the two populations, a comparison of these two curves might be of some interest. The incidence in White Females is higher than in Black Females, especially at older ages.

(ii) In Fig. 2 the *cumulative age of onset distributions* are compared. Both are smooth and increasing functions and it is not easy to extract from them all the information on the differences between these two distributions. The two remaining Figures seem to be more informative.

(iii) Fig. 3. represents *'onset curves'*. These are, in fact, the estimated probability density functions. Let $x^* = x + \frac{n}{2}$ be the midpoint of the interval $(x, x+n)$. The density function at x^* is estimated from the formula

$$\hat{f_X}(x^*) \doteq \frac{1}{n} [F_X(x+n) - F_X(x)] \ . \qquad (3.1)$$

The highest incidence is in a rather broad age interval: from age 45 up to age 65. This is observed in both groups, but it is higher in White than in Black Females.

(iv) Finally, Fig. 4 represents estimated *hazard rates* (forces of mortality) functions for these two populations.

The hazard rate of a distribution $F_X(x)$ is defined as

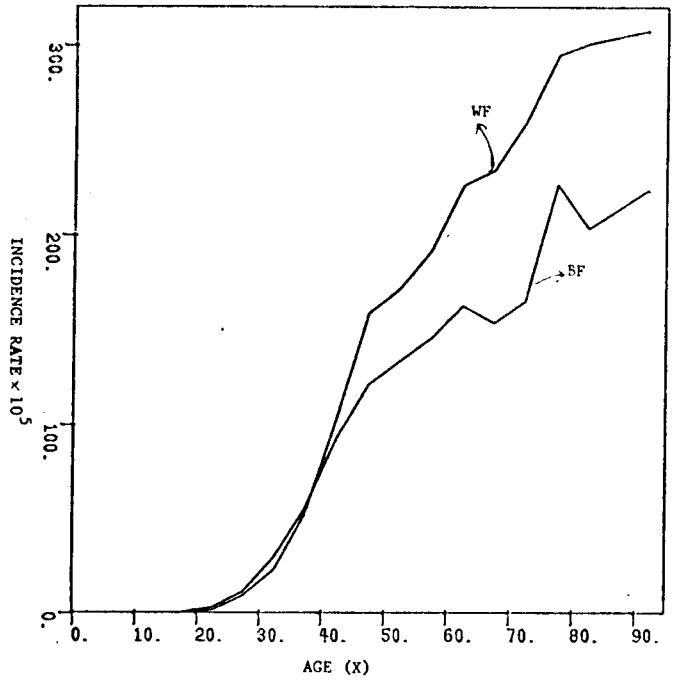

Fig. 1. Breast cancer in females: Observed incidence rates

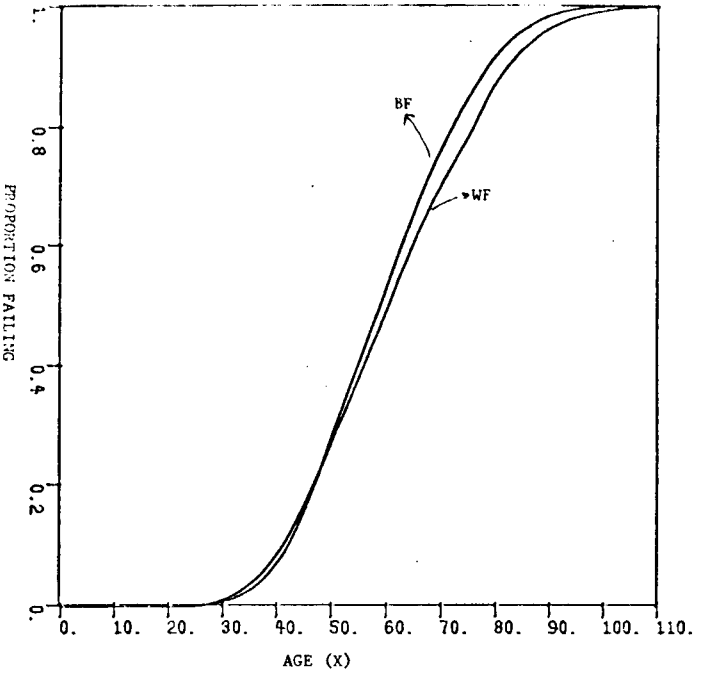

Fig. 2. Breast cancer in females: Age of onset distributions

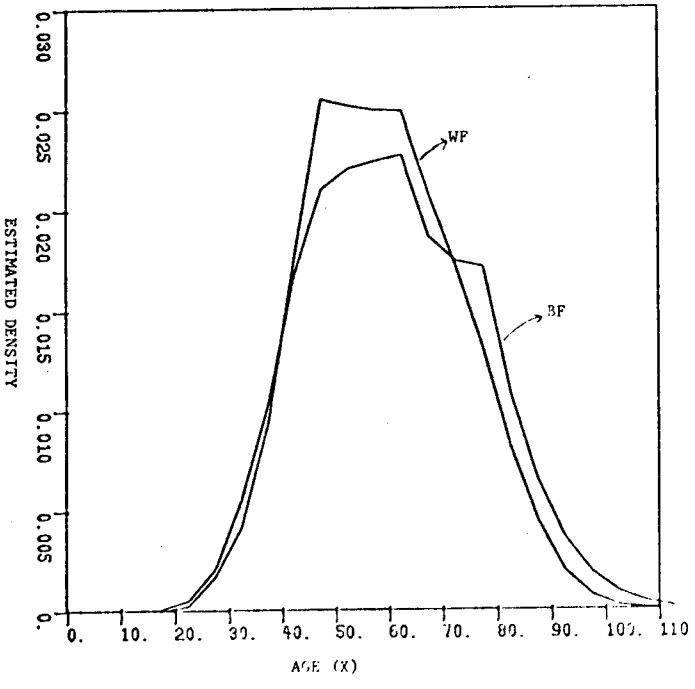

Fig. 3. Breast cancer in females: Onset curves

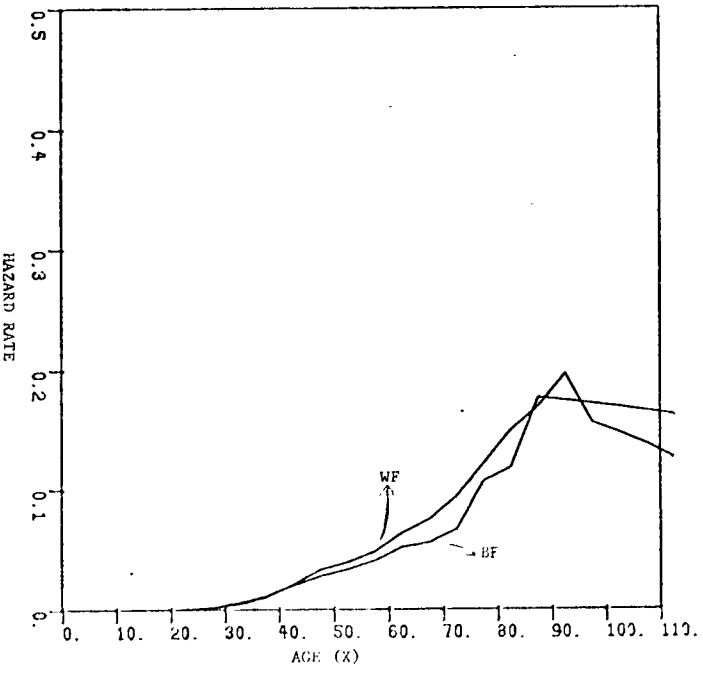

Fig. 4. Breast cancer in females: Hazard rate functions

$$\mu_X(x) = -d\log[1 - F_X(x)]/dx. \qquad (3.2)$$

Assuming that $\mu_X(x)$ is approximately constant over the interval $(x, x+n)$, it is often estimated from the approximate formula

$$\hat{\mu}_X(x) \doteq -\frac{1}{n}\log_n \hat{p}_x \ , \qquad (3.3)$$

where

$$\hat{{}_n p_x} = \frac{1 - \hat{F}_X(x+n)}{1 - \hat{F}_X(x)} \ .$$

The two incidence hazard rates are adjusted for age distribution and mortality in White and Black Females, respectively. Notice that they are fairly close (similar) as compared to those in Fig. 1.

REFERENCES

1. *Third National Cancer Survey: Incidence Data.* National Cancer Institute Monograph No. 41, DHEW, Publication No. (NIH) 75-787, 1975.

2. *United States Life Tables:* 1969-71. Vol. I, Number 1. DHEW Publication No. (HRA) 75-1150, 1975.

Acknowledgement

I would like to thank Mrs. Anna Colosi for the calculation of Table 1 and for plotting the graphs using an IBM 370 Model 155 computer.