Five-year age-specific incidence rates

I Their nature and limitations

I. K. CROMBIE

From the Cancer Epidemiology Research Unit, University of Birmingham

N. CRAMER

From the School of Economics and Politics, Kingston Polytechnic

SUMMARY The effect of population structure on five-year age-specific incidence rates was investigated using the one-year population data from life tables and a theoretical age incidence curve of the form:

$$I = bt^k$$

where I is the incidence at age t , and b and k are constants.

The five-year incidence rates differed systematically from the one-year rates of the central year of the five-year period. This difference depended on the change with age of both the population size and the incidence rate. Thus at ages 20–24 the five-year rate overestimates the mid-period one-year rate by about 4%, but the overestimate progressively decreases to become an underestimate of 0.5% at ages 75–79. In consequence the one-year and five-year rates produce fitted age incidence curves with different slopes; the value of k in the incidence equation is about 0.7% greater for the one-year rates.

The population structures of developed and underdeveloped countries are markedly different and these were found to affect the five-year incidence rates, but never by more than 0.5%.

The effect of the irregularities in one-year age structure of real populations on the observed five-year rates is also small, of the order of 0.5%. However, when incidence rates are calculated by recording tumours over several calendar years, these irregularities can create difficulties for the estimation of the appropriate denominator population. The use of the census population, even that of the central year of the observation period, can be in error by over 2%. A good method is to calculate the mean annual population of the observation period, estimating the intercensal year populations by interpolation between flanking censuses.

The aetiology of human cancers is frequently investigated by comparing the differences of cancer frequency observed between races, countries, sexes, and occupations.¹ In addition, studies of the age distribution of cancer have led to hypotheses about fundamental carcinogenic processes.² These comparisons are usually made using five-year age-specific incidence rates, which are assumed to reflect accurately the cancer experience of the population and to be sufficiently invariant for such purposes. Many cancers increase rapidly with age,³ so that there will be a considerable difference in the cancer incidence between individuals in the first and last years of each five-year period. This raises the question of whether the investigations and analyses

of cancer frequencies would be more accurate or informative if performed using one-year age-specific incidence rates instead.

The occurrence of tumours in a five-year age group will depend on the underlying one-year age structure of the population. At younger ages this structure will be influenced largely by fluctuations in the birth rate, but at older ages the additional effects of mortality on the one-year age structure will increasingly be felt. The possibility exists that the five-year rate will be affected differently at different points on the age scale.

The differences in one-year age structure could also be important in comparisons between different countries. The age-specific mortality rates in

developed and developing countries are markedly different and give rise to populations whose one-year age structures are distinct.⁴ It is therefore possible that the five-year rates from two such countries could be affected by their differing one-year population structure.

In the present study we investigate the effects of the one-year age structure of the population on the five-year age-specific incidence rates and examine the contributions of fluctuations in the birth rates and differences in mortality rates to the comparability of five-year rates between age groups and populations.

THE BASIC DATA

The study of the effects of the one-year age structure of a population on age-specific incidence rates is complicated by the fact that all real populations have somewhat irregular age structures due to marked fluctuations in the birth rate, changes in age-specific mortality rates, and wars. Such marked irregularities greatly complicate the interactions between population structure and age-specific incidence rates. To clarify these interactions we have taken advantage of the very regular age structure provided by the theoretical population of a life table, which has the additional advantage that the population is tabulated by one-year age groups.

To study the effects of the different age structures of developed and developing countries, the life tables for populations with life expectancies of 68.0 and 43.7 years (the 1961 and 1891 England and Wales Life Tables) were chosen.^{5,6} These populations exhibit a smooth decrease in number with age (Fig 1) enabling the factors affecting incidence rates to emerge clearly.

Fig. 1 The age structure of the England and Wales life table populations (male) for the years 1891 and 1961; data from the Registrar General.^{5,6}

Five-year age-specific incidence rates are obtained by allocating all the tumours which have occurred during a calendar year to the appropriate five-year age groups by the age of the patient at his last birthday. The five-year rate, expressed per 100 000 population, is then given by the ratio of the number of tumours to the number of persons at risk of a tumour, and refers to one complete calendar year. The method of estimation of the number of persons at risk is discussed in the Appendix. Thus the five-year rate can be defined as the probability of a tumour occurring per calendar year among the group of people who are in that age group during that year. A mathematical expression for the incidence rate is given in the Appendix.

To preclude any circularity in the argument we can define an age-specific incidence curve such that the one-year incidence rates I are given by the equation:

$$I = bt^k \quad \dots (i)$$

where I is the incidence at age t , and b and k are constants. It is known that many cancers show such a distribution by age.³ The data for stomach cancer among women from the Birmingham Cancer Registry⁷ in the age range 35 to 75 were selected and the incidence curve was fitted to them. The resulting values of $b = 4.0345 \times 10^{-8}$ and $k = 5.0888$ produced a regressions line with a correlation coefficient of 0.995. The one-year rates obtained in this way are represented in Table 1 (for brevity only the one-year rates of the mid-year of each five-year group are shown).

Table 1 Theoretical one-year age-specific incidence rates^c and the five-year rates which they would give rise to in 1891 and 1961 life table populations

Age group (years)	One-year incidence rates I_3	'Observed' five-year incidence rate		% difference between the five-year incidence rates $\frac{I_{vB} - I_{vA}}{I_{vB}} \times 100$
		1891 I_{vA}	1961 I_{vB}	
20 —	0.31 ^a	0.32	0.32	0.18 ^b
25 —	0.85	0.87	0.87	0.21
30 —	1.99	2.03	2.03	0.23
35 —	4.13	4.18	4.19	0.23
40 —	7.80	7.87	7.89	0.23
45 —	13.74	13.83	13.86	0.22
50 —	22.87	22.96	23.00	0.20
55 —	36.34	36.40	36.46	0.18
60 —	55.54	55.48	55.59	0.21
65 —	82.17	81.89	82.02	0.15
70 —	118.20	117.38	117.59	0.18
75 —	165.96	164.02	164.36	0.21

a. Only the rates of the mid-year of each five-year age group, that is, 22, 27, 32 ... are shown.

b. The calculations were performed using incidence rates specified with more precision than those shown in the Table.

c. Derived from the age incidence curve of cancer of the stomach among women.

Five-year age-specific incidence rates: I Their nature and limitations

DEPENDENCE OF THE FIVE-YEAR INCIDENCE RATES ON THE ONE-YEAR POPULATION STRUCTURE

The five-year incidence rates I_v which would occur in a population whose one-year incidence rates and one-year population sizes are known can be calculated using the equation:

$$I_v = \frac{\sum_{j=1}^5 I_j N_j}{\sum_{j=1}^5 N_j} \quad \dots (ii)$$

where I_j and N_j are the one-year incidence rates and the one-year population sizes respectively. This equation illustrates that the five-year incidence rate I_v is the average of the one-year rates weighted by the one-year population sizes. The set of one-year incidence rates described above was applied to the two life table populations to generate the 'observed' five-year incidence rates. The five-year rates of the 1961 population (I_{vB}) are consistently greater than those of the 1891 population (I_{vA}), but the differences are small and do not exceed 0.23% (Table 1). These differences can be due only to differences in the one-year age structures of the two populations. The proportionate decrease in population size across each five-year age group for the 1891 and 1961 populations is shown in Fig 2 and it is clear that the

1891 population experiences the greater decrease at all ages. The consequence of this is that in each five-year age group for the 1891 population there are relatively more people in N_1 and N_2 . Thus, in the equation for I_v , more weight is given to the incidence rates I_1 and I_2 and less to the higher rates I_4 and I_5 than is the case with the 1961 population. In consequence the 1961 population always exhibits the higher five-year incidence rate.

DIFFERENCE BETWEEN THE FIVE-YEAR AND ONE-YEAR AGE INCIDENCE CURVES

The five-year incidence rate (I_v) also differs from the one-year rate of the mid-year (I_3) of the five-year age group (Table 1). This difference arises because I_v is the (weighted) average of five I_j s, and since these increase in a rapid non-linear fashion, the average is larger than the central values (I_3) at the younger ages. However, at older ages the rate of increase of I_j is slower (from equation (i)), and the weighting effect of the one-year population becomes marked ($N_1 > N_5$), giving more weight to I_1 than to I_5 , so that I_v is smaller than I_3 at ages above 60 (Table 1). Expressed as a percentage, the difference between I_v and I_3 decreases smoothly from +4.0% at ages 20-24 to -0.5% at ages 75-79 (data not shown).

The existence of the difference between I_v and I_3 implies that the slope of the age incidence curves produced by five-year and one-year incidence rates could be different. On the assumption that both sets of rates are of the form of equation (i), an indication of the size of this effect can be obtained by calculating the values of k which would be obtained from the five-year rate data. The 1961 population gave $k = 5.0542$ and the 1891 population 5.0545. These values differ by 0.68% and 0.67% from the one-year rate value of 5.0888, so that the five-year age-specific incidence curve can be used as a good approximation to the true age curve (where the incidence is at least approximately of this form).

EFFECT OF REAL POPULATIONS

Population census data are published by five-year groups, but an estimate of single-year structure of the 1961 census population of England and Wales was obtained using the interpolation method of Keyfitz and Flieger.⁸ Although approximations only, the values obtained will give a good indication of the size of the effect of irregular population structures on five-year incidence rates. The set of one-year rates defined above was applied to the one-year population data derived in this way, and the five-year rates which would be observed were calculated using equation (ii). The per cent difference between these five-year rates and those obtained from the life table population (Table 2) gives a measure of the effect of

Fig. 2 The rate of population decrease exhibited by the England and Wales life table populations (male) for the years 1891 and 1961; the rate was calculated as the difference between the first and the fifth single years of the five-year age group as a percentage of the total in the age group.

population irregularities on five-year rates. The two sets of rates are very similar in size, differences being less than 0.5%, but the population irregularities can cause the rates to be either larger or smaller than those from the life table population.

Table 2 Comparison of the five-year incidence rates which would be observed in a real population with those from a life table population

Age group (years)	Five-year incidence rate 'observed' in 1961 census population	% deviation of census ^{a, b} population rates from life table population rates
20 —	0.32	-0.31
25 —	0.88	+0.11
30 —	2.04	+0.44
35 —	4.19	+0.02
40 —	7.89	-0.03
45 —	13.89	+0.21
50 —	23.00	0
55 —	36.34	-0.32
60 —	55.32	-0.50
65 —	81.78	-0.28
70 —	117.50	-0.08
75 —	164.26	-0.06

a. The 1961 life table population five-year incidence rates are shown in Table 1.
b. The calculations were performed using incidence rates specified with more precision than those recorded in the Table.

EFFECT OF RECORDING INCIDENCE RATES OVER SEVERAL CALENDAR YEARS

To overcome some of the variability due to small numbers, age-specific incidence rates are commonly obtained by recording the occurrence of tumours over several years. This raises the problem of what type of population data should be used to calculate incidence rates. The practices of 57 population-based cancer registries contained in *Cancer Incidence in Five Continents*⁷ are shown in Table 3. Most commonly the population used is that enumerated by a census conducted during the period of observation,

Table 3 Nature of population used in the calculation of incidence rates in 57 cancer registries^a

Location of census	POPULATION ESTIMATE BASED ON:		
	Single-year population		Mean annual population of observation period
	Census population	Mid-period estimate	
Mid-period	16	—	1
Other within period	21	—	3
Outwith period	1	3	2
Other including not clearly specified	1	2	7
Total	39	5	13

a. These data are taken from Waterhouse *et al.*⁷ Although 59 registries are quoted by them, the population estimates used by two registries were too unusual for inclusion in this Table.

although the census year is often not that of the mid-year of the observation period. Some registries estimate the mid-observation year population from census data, and others calculate the mean annual population of the observation period, estimating the population sizes of the intercensal years. The problems associated with each of these methods are discussed below.

(1) MEAN ANNUAL POPULATION OF OBSERVATION PERIOD

The estimation of population sizes for intercensal years is clearly prone to error, but this is likely to be small. The estimate of the population of England and Wales for 1971, carried forward from the 1961 census, was found to overestimate the results obtained from the 1971 census by 0.7%. The difference would have resulted from 10 successive estimates, so that the error in each would have been much smaller. None the less, the Registrar General regarded the difference as unusually large and produced a set of revised estimates of the population for the intercensal years by interpolating between the two flanking censuses.⁹ This method is likely to produce very small errors in the estimates intercensal populations.

(2) SINGLE-YEAR ESTIMATES OF THE POPULATION OF A PERIOD

Using a single calendar year population as the estimate of the persons at risk of a tumour during several years is subject to errors arising from the one-year age structure of the population. This can be seen by inspection of the one-year age groups for the years 10–19 for males in England and Wales in the 1961 census (Fig. 3).

It is clear that the total in the age group 15–19 will increase with successive calendar years as the smaller

Fig. 3 The one-year age structure of the age group 10–19 years of the 1961 male census population (home) of England and Wales; data from the Registrar General.⁹

Five-year age-specific incidence rates: I Their nature and limitations

numbers at ages 19 and 18 are replaced by the larger numbers at ages 14 and 13 (mortality will have little effect on the population sizes at these ages). The Registrar General takes cognisance of this effect in his estimates of population sizes for intercensal years; the population recorded by single years of age is aged one year for each calendar year that passes. In fact the size of the group 15–19 increases from 1 647 200 in 1961 to a maximum of 1 879 800 in 1965, a difference of 14%. It is clear that the use of a single year to estimate a period population is prone to sizeable error. The central year can be seen to give the smallest errors by considering the one-year age structure of the mean annual population and that of each single year. Using the nomenclature illustrated in Fig. 3 for the ages 10–19, it is clear (assuming no mortality) that the structure of the mean annual population for the age group 15–19 over the period 1961–65 will be $1/5 (5f + 4e + 4g + 3d + 3h + 2c + 2i + b + j)$ and that the single year which has the closest structure to this is 1963 with $1/5 (5f + 5e + 5g + 5d + 5h)$. It is also clear that the further from the mid-period is the single year selected, the greater is the error in the estimate of the mean annual population.

Discussion

Several authors have considered the sources of error in age-specific incidence rates, but have limited their considerations to problems concerning the accuracy of diagnosis, the completeness of registration, and the accuracy of the population census.^{10 11 12} In this paper we have considered the nature of the variation imposed on the incidence rates as a result of the underlying population structure.

The effect of the differences in the age structure of populations on overall cancer incidence rates has long been recognised, and methods of standardisation have been developed to overcome these problems.¹³ The calculations use incidence rates from the population subdivided by five-year age groups and tacitly assume that these rates accurately represent the age-specific incidence curve. In this study we have shown that five-year incidence rates themselves depend not only on the underlying tumour incidence, but also on the underlying population structure. This population structure is affected by past fluctuations in the birth rate and the increase with age of mortality rates. By themselves these factors are unlikely to cause the five-year incidence rates of a real population to differ by more than 0.5% from those of a life table population. Thus comparisons of the rates of different age groups or populations will involve a composite error of at most twice this value.

A more serious consequence of the one-year population structure results from the practice of recording the occurrence of tumours over several calendar years. The problem which arises is how best to estimate the period population with which to calculate the incidence rates. The use of the single-year population enumerated by a census conducted during the observation period can produce serious errors. These will be minimised if the census occurs at the mid-point of the observation period. Because the errors in the estimates of the populations for intercensal years are much lower than the foregoing, it would be preferable to use the estimated mid-point population if the census does not occur in the central year. Similarly, if there are an even number of years of observation the average of the estimates of the two central years would give the best population estimate.

Better than these would be to use the mean of the annual estimates of the populations of each year of the period. There will be errors in the population estimates, but these are usually small, and can be minimised if the estimates are based on interpolation between flanking population censuses. Comparison of the population estimated for a census year with that obtained by the census would serve as a check on the accuracy of the estimations.

All the foregoing considerations serve to give an indication of the magnitude of the variation inherent in the calculation of incidence rates. As a general rule it could be considered that the lower limit of the variability of an incidence rate will be of the order of one per cent. In practice, however, because of errors in population estimates, in particular from the sources discussed by other authors,^{10 11 12} the variability will be considerably larger.

The fact that the five-year rate is not equivalent to the one-year rate of the mid-year of the five-year period gives rise to some interesting consequences. This is a result of the non-linear rate of increase of the incidence rate and the way in which it interacts with rapid increase in mortality at older ages. The difference between the two rates changes in sign across the age range, so that the one-year and five-year incidence rates produce age curves with differing slopes. However, the difference in slope is small (less than one per cent) so that the five-year rates give a good approximation to the underlying age incidence curve. The equivalence of the five-year rate and that of the mid-year of the five-year period is sometimes assumed when one-year rates are calculated by interpolation between five-year rates.¹⁴ This error can have important consequences when the interpolated one-year rates are used to calculate the number of tumours expected to occur in a study population.¹⁵ It can be shown that the use of such

interpolated rates can in some circumstances result in larger errors than those from the use of the original five-year rates.

We thank Dr. A. Minawa and Mr. T. Sorahan for helpful advice in the preparation of this manuscript. One of us (I.K.C.) received a grant from the Cancer Research Campaign.

Reprints from Dr. I. K. Crombie, Department of Social Medicine, University of Birmingham, Queen Elizabeth Hospital, Edgbaston, Birmingham B15 2TH.

References

- ¹ Wynder EL. Contribution of the environment to cancer incidence: an epidemiologic exercise. *J Natl Cancer Inst* 1977; **58**: 825–32.
- ² Cairns J. The cancer problem. *Sci Am* 1975; **233**: 64–79.
- ³ Cook PJ, Doll R, Fellingham SA. A mathematical model for the age distribution of cancer in man. *Int J Cancer* 1969; **4**: 93–112.
- ⁴ Preston SH. *Mortality Patterns in National Populations*. New York: Academic Press, 1976.
- ⁵ Registrar General. *Decennial Supplement England and Wales 1961. Life Tables*. London: HMSO, 1968.
- ⁶ Registrar General. *Supplement to the 65th Annual Report 1891–1900, Part I*. Cmdnd No. 2618. London: HMSO, 1907.
- ⁷ Waterhouse J, Muir C, Correa P, Powell J. *Cancer Incidence in Five Continents. Vol. III*. IARC Scientific Publications No. 15. Lyon: IARC, 1976.
- ⁸ Keyfitz N, Flieger W. *Population: facts and methods of demography*. San Francisco: W. H. Freeman, 1971.
- ⁹ Registrar General. *Revised Estimates of the Population of England and Wales, Regions and Local Authority Areas 1961 to 1971*. London: HMSO, 1975.
- ¹⁰ Linden G. The comparability of cancer incidence rates. *Recent Results Cancer Res* 1975; **50**: 141–7.
- ¹¹ Tuyns AJ. The collections of information available. Possibilities and limitations of international standardisation. *Recent Results Cancer Res* 1975; **50**: 14–19.
- ¹² Muir CS, Waterhouse JAH. Reliability of registration. In: Waterhouse J, Muir C, Correa P, Powell J, ed. *Cancer Incidence in Five Continents. Vol. III*. IARC Scientific Publications No. 15. Lyon: IARC, 1976: 45–51.
- ¹³ Doll R, Cook P. Summarising incidences for comparison of cancer incidence data. *Int J Cancer* 1967; **2**: 269–79.
- ¹⁴ Bell R, Case RAM. *Serial Abridged Life Tables: England and Wales 1841–1970. Part II: Interpolated Cohort Life Tables*. London: Chester Beatty Research Institute, 1971.
- ¹⁵ Cramer N, Crombie IK. Five-year age-specific incidence rates. II: Systematic errors produced in the calculations of expected numbers of tumours. *J Epidemiol Community Health* (In press).

Appendix

The estimation of the number of persons in a particular age group during a calendar year is difficult. In principle one would count the number of people alive on each day of the year (or in the extreme, each millisecond), and convert the number of person-days (or person-milliseconds) to person-years. In practice the Registrar General provides an estimate of the numbers in each age group on 30 June (the mid-point of the year). These point estimates of the numbers in the age groups will be the same as the averaged numbers in each age group for a whole year if the events which cause people to move in and out of age groups (birthdays and deaths) occur evenly throughout the year.

This method of estimating the size of each age group leads to a simple mathematical model for the age-specific incidence rate.

Let: $n(x)$ be the number of people of exact age x on 30 June, and assume that an identical $n(x)$ would be observed on all other days of that calendar year.

$I(x)$ be the probability of developing a tumour during one calendar year among people of exact age x .

Thus the incidence rate for the five-year group u to $u + 5$ is:

$$\int_u^{u+5} n(x) \cdot I(x) dx \bigg/ \int_u^{u+5} n(x) \cdot dx$$

and that for one-year rates would be of similar form but with limits u to $u + 1$. It should be noted that equation (ii) in the text, which was used to calculate the five-year incidence rate I_v , is an approximation to the integrated form in which the population term and the incidence rate refer to one complete year of age.