STATISTICS IN MEDICINE, VOL. 11, 1579–1589 (1992)

CALCULATING THE PREVALENCE OF CANCER

ANDREW J. COLDMAN, MARY L. McBRIDE AND TERRY BRAUN

British Columbia Cancer Agency, 600 West 10th Avenue, Vancouver, British Columbia, Canada, V5Z 4E6

SUMMARY

Incidence, prevalence and mortality are commonly used measures to assess the impact of disease on human populations. Prevalence, although regularly assessed for a number of different diseases, has only had recent use to measure the impact of cancer. The calculation of the prevalence of cancer presents several difficulties since there is no reporting mechanism established to measure the proportion of the community that has the disease. In the absence of such a mechanism, mortality data linked to incidence data from cancer registries have been used. The assumption is made that once diagnosed with cancer an individual remains a prevalent case until death. In this paper we present alternatives to this assumption and use them to produce estimates of cancer prevalence. We illustrate the effect of these assumptions on the calculated prevalence of cancer using data from the British Columbia Cancer Registry.

1. INTRODUCTION

Cancer is the second largest cause of death in Western countries and the treatment and care of cancer patients account for a substantial proportion of the health care budgets of those countries. Three indices are commonly used to monitor the impact of the disease: incidence, mortality and survival. Incidence, whether age specific or cumulated over time, is the proportion of individuals diagnosed with the disease in a calendar period and is an indicator of disease risk. Mortality indicates, in the same way as incidence, what proportion die of cancer. The bases for both these measures are populations. Survival, however, is calculated from the mortality experience of a group of individuals with disease, usually measured by the probability that a case will survive a specified period of time from diagnosis. Most estimates of survival are based on select groups (such as attendees at a particular hospital) although there are reports available that calculate survival using a complete series of cases drawn from a geographic area.[1]

Another index of the impact of cancer on the community is prevalence. Prevalence is defined as the 'proportion of a population that is affected by disease at a given point in time'.[2,3] Health planners frequently use prevalence to determine the demand for disease-specific services, since prevalence enumerates the number of individuals with disease and thus those likely to seek care. For example, in the future, projected prevalence numbers will be used to estimate the demand for follow-up visits at cancer facilities in British Columbia.

Only patients diagnosed with cancer at some time are potential prevalent cases and this suggests the use of cancer registry data, where available, to calculate prevalence. Cancer registries usually contain information on demographic variables, date of diagnosis, disease type and extent at diagnosis, whether the individual is alive or dead, and if dead the date and cause of death. Approximately half of the individuals diagnosed with cancer will die from the disease.

Unfortunately, registry data do not usually indicate the current disease status of the individual. Many individuals diagnosed with cancer will survive for a long period and subsequently die (from another cause) without any evidence of persistent disease. Also many patients diagnosed with

0277–6715/92/121579–11$10.50
© 1992 by John Wiley & Sons, Ltd.

Received September 1991
Revised February 1992

Material may be protected by copyright law (Title 17, U.S. Code)

cancer experience a period of no detectable disease but subsequently die from the original cancer. Being clinically disease free is thus not a reliable indicator of disease absence. An example of the latter is breast cancer, where the disease may recur after a period of 10 or more years of no detectable disease. The designation of the disease state among individuals currently disease free clinically presents a difficulty in calculating cancer prevalence.

Previous attempts to calculate prevalence have assumed that a person still alive with a previous diagnosis of cancer represents a prevalent cancer case.[4] There is partial justification for such an assumption in that it is impossible to state with absolute confidence that an individual is free of disease, but many individuals will die long after treatment of their original cancer with no evidence of persistent disease. It would thus seem that the assumption that all incident cases of cancer remain prevalent cases until death is not in accord with the disease process. To calculate prevalence rates from incident data, however, some assumptions about the disease process are necessary. In the absence of specific information on individuals, these assumptions represent generalisations that can provide estimates of prevalence rates but not the status of an individual. In what follows we develop simple models for the estimation of prevalence.

2. METHODS

First we develop some notation. For simplicity, we assume discrete calendar time t. Let $R(t)$ represent all individuals who have had a diagnosis of cancer at some time prior to t, are alive at t and are members of the target population at the time of their diagnosis. We consider $R(t)$ to consist of a set of vectors, $\{x\}$, where each vector consists of data on a single individual. These data would include the usual information obtainable from a cancer registry: name, date of birth, sex, diagnosis etc. For simplicity we refer to an individual as $\{x\}$.

Let t' denote the latest time for which data on incidence and mortality are considered complete (that is, the data include information on all new cases and all deaths that have occurred prior to this time). Then for $t < t'$ and $x \in R(t')$ we define

$$d(x) = \text{time of diagnosis},$$

$$g(x) = \text{time of last ascertainment of vital status},$$

$$f(x) = g(x) - d(x) = \text{follow-up time of } x \text{ at time } t',$$

$$f(x; t) = \begin{cases} f(x) & g(x) \leqslant t, \\ t - d(x) & g(x) > t, \end{cases}$$

$$h(x) = \text{numeric code for cancer type},$$

where we suppress reference to t'. For those with more than one diagnosis, we consider each tumour to have a corresponding set $\{d(x), f(x), h(x), f(x; t)\}$. In what follows we wish to calculate the incidence of a tumour, a tumour subtype or a group (such as leukaemias), so that when reference is made to type h this may represent a variety of situations.

Let $p(h, t; x)$ denote the probability that individual x has cancer type h present at time t. Then an estimate of the number of prevalent cases of type h in the population at time t, $P(h, t)$, is

$$\hat{P}(h, t) = \sum_{x \in R(t)} p(h, t; x). \tag{1}$$

For the simplest case where we assume that every individual once diagnosed with cancer has the disease until death, define the set of functions

$$p_1(h, t, h(x); x) = \begin{cases} 1 & \text{if } h(x) = h, d(x) \leqslant t, \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

Material may be protected by copyright law (Title 17, U.S. Code)

where there is a separate function, $p_1(\cdot)$, for each diagnosis of the individual that lies in the set of interest h, that is $h(x) \in h$. Then set

$$p(h, t; x) = 1 - \prod_{h(x) \in \{h(x)\}} (1 - p_1(h, t, h(x); x)) \tag{3}$$

where $\{h(x)\}$ is the set of diagnoses of $\{x\}$. Equation (3) assumes that the probabilities of tumour presence of each type are independent. The effect of equations (2) and (3) is to set the probability of prevalence equal to unity from the first date of an appropriate diagnosis up to the time of death.

We refer to the system of equations (1), (2) and (3) as model 1. Clearly model 1 does not admit the possibility of cure so that we classify individuals with no residual disease as prevalent cases. Since cure is quite common for some cancers (for example testicular tumours),[5] this indicates a need to construct other models of prevalence.

A common medical model for cancer prognosis is to assume that an individual is cured (that is, no longer has the disease) if he/she survives longer than some minimum specified time, the 'time to cure', f_h, which is a function of cancer type. This suggests

$$p_2(h, t, h(x); x) = \begin{cases} 1 & \text{if } h(x) = h, d(x) \leqslant t, f(x; t) < f_h, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where, as for equation (2), we define a separate function for each diagnosis. We refer to the system of equations (1), (4) and (3) (with $p_2(\cdot)$ replacing $p_1(\cdot)$) as model 2. The effect of model 2 is to define an individual as a prevalent case of type h if the cure time has not been exceeded for all of the subject's cancers.

We can obtain estimates of f_h from the literature. However, quite substantial variation in reported values makes it difficult to obtain appropriate estimates. In most cases there will be several years of registry data available and it is possible to use these data to generate estimates of f_h. One method is to choose f_h as the maximum survival time in the total registry data base (alive and dead) for which the subject's cause of death is h. It is well recognized that the recorded cause of death on a death certificate may be incorrect, with the frequency of misclassification dependent on tumour type and other factors. Although it is possible to correct for cause of death misclassification by using the results of published studies, any such adjustment is likely to be crude.

An alternative approach to defining cure among a group of individuals is to assess when the pattern of mortality equals that expected from a demographically similar sample of the general population.[6] The methodology commonly used in this situation is that of the relative survival rate where one calculates the conditional expected mortality rate using annual survival probabilities obtained from life tables. We recognize that the relative survival curve is subject to survivorship biases[7] but it is possible to use the person-years method[8] to calculate annual expected numbers of deaths and to compare these with those observed with use of a Poisson test or a similar statistic. Implicit in the technique is the assumption that cancer is either cured or causes death (that is, it cannot be controlled at some sublethal burden). This assumption seems reasonable for the vast majority of cancers even though there are instances (for examples, some chronic leukaemias) where disease control for prolonged periods is possible.

The preceding technique assumes that patients 'cured' of their cancer should have the same mortality experience as a similar subgroup of the general population. Exposures that increase the risk of specific cancers, however, may also increase the likelihood of developing other cancers or other potentially fatal diseases, such as smoking and heart disease. Subjects diagnosed with a particular cancer do not represent a random sample of the population and will often have overrepresentation of individuals who have above average expected mortality rates from other causes. Thus, even though the disease may be eliminated, the survivors may suffer excess

Material may be protected by copyright law (Title 17, U.S. Code)

mortality because of their exposure history. In this case an appropriate approach to estimate f_h is to attempt to estimate a change point in the mortality experience (with length of follow-up). Cusum techniques are generally useful in such applications; however, estimation of the 'break-point' may not be efficient since relative mortality excesses due to the cancer may vary with length of follow-up. Such an approach also has the drawback that one might discount excess mortality due to the original tumour or treatment effects. In light of this difficulty it seems reasonable to consider cure achieved only when the mortality experience parallels that of the general population. Having selected f_h, we can then use equations (3) and (4) to calculate $p(h, t; x)$.

Model 2 does not seem compatible with one aspect of the treatment of cancer. In practice, the majority of primary treatment for the disease occurs in a relatively short time after diagnosis, although there are some indolent tumours that one can treat profitably for long periods. Unfortunately, treatment of recurrent disease is seldom curative so that cures achieved occur, in the majority of cases, soon after diagnosis. Thus, as a generalization, 'cured' individuals are those cured early in the course of their disease at some point in time prior to f_h when cure becomes manifest. This suggests specification of some follow-up time, $u(x)$, prior to which disease is present with certainty (the active treatment period), and after which disease is present with some probability that reflects the likelihood of recurrence. An appropriate choice for $u(x)$ is the period of primary treatment for the disease. This period is typically short for treatments that involve surgery or radiotherapy only, somewhat longer for regimens that include chemotherapy and longest for some hormonal therapies. This motivates specification of

$$p_3(h, t, h(x); x) = \begin{cases} 1 & \text{if } f(x; t) \leqslant u(x), d(x) \leqslant t, h(x) = h \\ p_h(x) & \text{if } u(x) < f(x; t) \leqslant f_h, d(x) \leqslant t, h(x) = h \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $p_h(x)$ is the probability that an individual has disease, given that he or she has survived the active treatment period but has not been followed to the time to cure. An estimate of $p_h(x)$ is

$$p_h(x) = 1 - \frac{R_h(f_h)}{R_h(u(x))},$$

where $R_h(\cdot)$ is the relative survival function for tumour type h, and $p_h(x)$ is thus the probability of dying from disease conditional on surviving to $u(x)$. We can improve the estimate of $p_h(x)$ by replacing $R_h(u(x))$ by $R_h f(x; t))$, the relative survival rate at follow-up time $f(x; t)$, that is

$$p_h(x) = 1 - \frac{R_h(f_h)}{R_h(f(x; t))}, \tag{6}$$

so that $p_h(x)$ is the estimated probability that the individual will die from disease given that he/she has survived the observed follow-up period. If no time to cure, f_h, has been found for the disease then $R_h(f_h) = 0$ and equation (5) is equivalent to equation (4). We refer to equations (1), (3) (replacing $p_1(\cdot)$ by $p_3(\cdot)$), (5) and (6) as model 3. We estimate the relative survival function from the registry data in the usual way.[6] If we consider $u(x)$ in equation (5) as a random variable, we can then replace $p_3(\cdot)$ with $E[p_3(\cdot)]$, where we take expectation with respect to the distribution of $u(x)$.

Given the nature of cancer, it is well recognized that a new diagnosis in a person aged 50 implies presence of the tumour at age 49 and probably earlier. In the calculation of age-specific incidence rates, there is no attempt made to estimate when the tumour first came into existence. Thus, the incidence of cancer is the incidence of clinically detectable cancer with no attempt to adjust estimates to reflect the biological history of the disease. Applying the same reasoning to cancer prevalence indicates that we should include as prevalent cases only those with clinically

Material may be protected by copyright law (Title 17, U.S. Code)

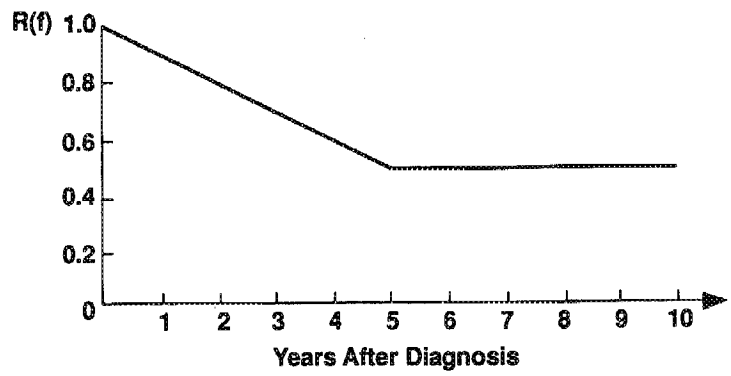

Figure 1. Hypothetical relative survival curve for a tumour in which survival linearly decreases for the first five years and then stabilizes at 0·5

demonstrable disease. When dealing with a registry population the dates of clinical disease disappearance and reappearance are usually unavailable and therefore require estimation. One possible approach is to assume that individuals have disease for some interval after diagnosis $u(x)$ (the period of primary treatment as in equation (5)) and that some proportion then become disease-free. A subgroup of these then recur at some later time and die after a further period, $V(x; t)$, of follow-up. We therefore identify $p_h(x)$ as the probability of disease recurrence given that the individual is alive and has follow-up $f(x; t)$. Assuming competing risks to be independent and estimating the distribution of disease-specific survival by $R_h(f)$, then we can easily show that

$$p_h(x) = \frac{\int_0^{f(x;t)} [1 - F_{V(x;t)}(f(x;t) - r)] \mathrm{d}F_{TR(x;t)}(r)}{R_h(f(x;t))}, \tag{7}$$

where $F_{V(x;t)}$ is the cumulative distribution function of $V(x; t)$ and $F_{TR(x;t)}$ is the cumulative distribution functioin of the time to recurrence $TR(x; t)$. Note that the disease-specific survival time is given by the sum of V and TR. We refer to equations (1), (3) (with $p_1(\cdot)$ replaced by $p_4(\cdot)$), (5) and (7) as model 4.

We notice that for tumour types for which $f_h = \infty$ (that is, excess mortality occurs throughout the follow-up period), models 2 and 3 are equivalent to model 1. Model 4 does not depend directly on f_h; however, by the definition of f_h for $f \geqslant f_h$ we have no deaths from the tumour and hence no recurrences. Thus in general for all x, t we have

$$p_1(h, t, h(x); x) \geqslant p_2(h, t, h(x); x) \geqslant p_3(h, t, h(x); x) \geqslant p_4(h, t, h(x); x) \tag{8}$$

and the estimated prevalence rates will decrease monotonically for models 1 to 4. Also we have that

$$p_i(h, t, h(x); x) \downarrow \quad \text{as } t \uparrow \text{ for } i = 1, 2 \text{ and } 3,$$

so that the probability of being a prevalent case is non-increasing with length of follow-up for these three models. The behaviour of $p_4(h, t, h(x); x)$ as t increases is not monotonic, in general.

We illustrate the basic properties of the models in the following simple example. Consider a tumour with a relative survival curve as given in Figure 1, where the probability of survival decreases linearly in the first 5 years then stabilizes at 0·5. Letting $u(x) = 1$ year and $V(x; t) = 2$ years with probability 1, then one may calculate the estimated probabilities of prevalence for each of the models. An individual diagnosed with this tumour who dies after eight years of follow-up

Material may be protected by copyright law (Title 17, U.S. Code)

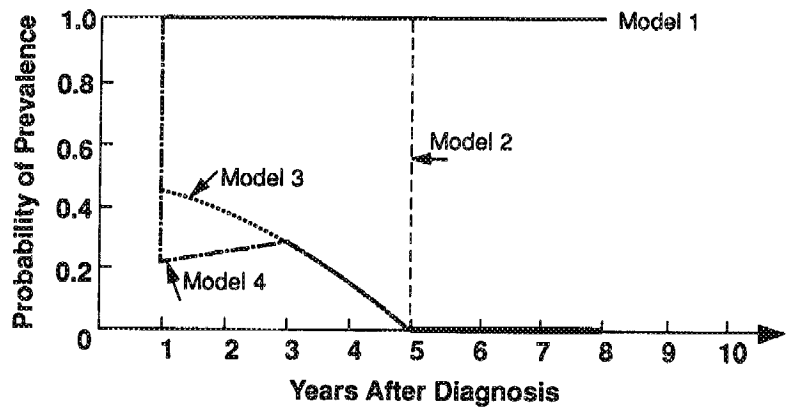

Figure 2. Probability of prevalence of a tumour with the relative survival curve given in Figure 1, as a function of follow-up time for the four models for an individual who dies 8 years after diagnosis of the tumour, where $u(x) = 1$ and $V(x; t) \equiv 1$

will then have estimated probabilities of prevalence as shown in Figure 2. We see that the models have the ordering relationship expressed in (8) and that the probability of prevalence is not monotonic, non-increasing for model 4.

One can perform all the calculations indicated on subsets of the data (for example, age and sex specific) for the purposes of standardization etc., but we have not indicated this explicitly in the notation.

The modifications to equation (2) to calculate cancer prevalence suggested here seek to bring the calculation of prevalence more in line with prevalence calculation for other diseases and with the calculation of cancer incidence. Cancer is a chronic disease, but, unlike some other chronic diseases (such as diabetes, arthritis), it can be cured but not controlled, and this behaviour merits inclusion in the calculation of prevalence. We illustrate the effects of these various modifications with the following example.

3. EXAMPLE: DATA FROM THE BRITISH COLUMBIA CANCER REGISTRY

The British Columbia Cancer Agency (BCCA) has had a population-based registry since 1969. Notification of individuals with a new diagnosis of cancer come to the registry from multiple sources, including pathology reports and treatment clinic admission reports. The registry utilizes a two-tier system for collecting information on the current vital status of patients. There is active follow-up of patients seen at BCCA treatment clinics and death notifications received by this institution are passed on to the registry. Approximately 50 per cent of cases seen in the province are followed in this way. For cases not seen at BCCA clinics, linkage of registry records with provincial death notifications provides the vital status of previously reported cases.

We calculated expected number of deaths for any period using the observed sex- and age-specific follow-up distribution of the group and mortality rates for the Canadian population.[9] We constructed Poisson tests by comparing the observed and expected numbers of deaths occurring k or more years after diagnosis. We selected f_h as the minimum integer k' for which the observed number of deaths did not significantly exceed those expected at $p = 0{\cdot}1$ for all $k \geqslant k'$. This process resulted in estimates of f_h as shown in Table I. In only one case did estimates of f_h differ by more than one year between the sexes. We calculated relative survival curves using the ratio of the

Material may be protected by copyright law (Title 17, U.S. Code)

Table I. Estimated time to cure, f_h, using the British Columbia data, for selected cancers

Site	Time to cure, f_h (years) Males	Females
Lip	6	5
Stomach	8	8
Lung	14	12
Breast	∞	∞
Endometrium	—	11
Testis	6	—

Table II. Sex-specific prevalence rates age standardized to the 1980 U.S. population for selected cancers using models 1, 2, 3 and 4

	Males				Females			
Site	1	2	3	4	1	2	3	4
Lip	41·4	14·3	2·3	2·2	7·2	0·5	0·5	0·6
Stomach	29·4	20·1	10·7	9·5	17·4	10·5	5·7	5·2
Lung	114·5	106·7	62·3	50·6	69·5	64·5	37·2	30·2
Breast	4·2	4·2	4·2	0·6	819·8	819·8	819·8	152·3
Endometrium	—	—	—	—	240·8	136·5	26·3	24·9
Testis	56·5	25·2	4·7	4·7	—	—	—	—
All cancers	1568·6	1272·3	658·3	367·5	2107·8	1791·3	1347·1	385·0

observed overall survival rate divided by the expected survival rate estimated for each year of follow-up with use of only members of the risk set for that year. We required the estimated relative survival rates to be monotonic non-increasing, and in situations where the ratio for one year exceeded one, we set it to unity. In no case did a value in excess of unity appear greater than that reasonably attributable to chance. We then used the resulting estimates of the relative survival curves in model 4. Table II gives estimated prevalences for selected tumour sites for the four models. Models 3 and 4 have $u(x) = 1$ year. $V(x; t)$ can be estimated using data from clinical series; however, in this illustration we have elected to fix it at two years for all tumour types. We calculated the rates for the category 'all cancers', by estimation of the site-specific rates and use of equation (3). To conform to standard practice, we excuded non-melanoma skin cancers. It is clear that the four models produce quite different estimates of the prevalence of various tumours and that this results in different estimates of the overall prevalence of cancer. It is interesting to note that they produce quite different estimates of prevalence for cancers that have poor survival rates (such as lung cancer).

As we used no notifications of disease prior to 1970, the estimates of prevalence contained in Table II must underestimate the true rates. To examine the effect of incomplete data on each of the models we repeated the preceding analyses using only incidence data for later years. Let P_n represent the calculated prevalence rate with use of only incident cases diagnosed in the n years prior to t. Figure 3 plots the relationship between n and the proportional prevalence rate, P_n/P_{18} for each of the models. This figure shows, as expected, that for all models recent years contribute more to the estimated prevalence. The pattern of contribution is, however, not the same and the early years contribute proportionately least to overall prevalence for model 4. Table III shows

Material may be protected by copyright law (Title 17, U.S. Code)

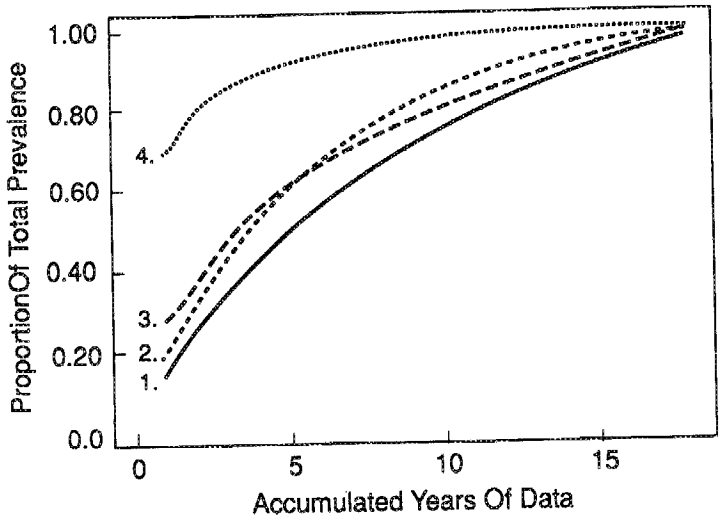

Figure 3. Proportion of prevalence (with year 18 = 100 per cent) as a function of number of years of data for models 1 to 4. All prevalence rate estimates have been standardized to the 1980 U.S. population

computed prevalence rates for each of the models for the last 5 and the last 10 years. The rather regular form of the relationship between n and P_n/P_{18} found in Figure 3 suggests the possibility simply to model this relationship and then extrapolate to estimate the prevalence for $n > 18$. We used the following functional form to model the relationship between n and $r(n) = P_n - P_{n-1}$, the change in the prevalence rates between $n - 1$ and n:

$$\ln r(n) = \alpha + \beta_1 n + \beta_2 \ln n,$$

where α, β_1 and β_2 are constants. We fitted this model separately to the sex-specific rates calculated using each of the four models by means of linear regression for $n = 1$ to $n = 18$. The fit of the model to each data set was good (as judged by the R^2 values) but in no case did the estimate of β_2 differ significantly from zero. Thus we set β_2 to zero and fitted the reduced model to the data. As expected, the estimates to β_1 were negative for each model so that the estimated prevalences for $n = \infty$ were finite. The extrapolated values for $n = \infty$ appear in Table III. As expected, the extrapolated prevalence for model 1 is considerably greater than that calculated at $n = 18$, somewhat different for models 2 and 3 and virtually the same for model 4. We note that the difference between the extrapolated prevalence and that at $n = 18$ was generally greater for females than for males.

Although we cannot, in general, assess the accuracy of the preceding extrapolations, Feldman *et al.*[4] did calculate estimates of prevalence in Connecticut using subsets of their data. They reported that for the overall cancer prevalence rate calculated by model 1 (standardized to the same population and excluding non-melanoma skin cancers), $P_{17}/P_{47} = 0{\cdot}84$. Using the sex-specific estimates of α and β_1 for model 1 obtained from the British Columbia data, we can estimate this quantity. For males the British Columbia data give $P_{17}/P_{47} = 0.903$ and for females $P_{17}/P_{47} = 0{\cdot}784$. In the standard population 51·4 per cent are females so that the resulting overall estimate is 0·84, in close agreement to that observed in the Connecticut study. Using the extrapolation formula for each model we can estimate the number of years of data collection required before the calculated value is a particular proportion of the asymptotic value. Setting this proportion to 0·99, the required numbers of years appear in Table III. Generally the numbers

Material may be protected by copyright law (Title 17, U.S. Code)

Table III. Effect of duration of accumulated incidence data on the estimates of the sex- specific prevalence rates and projected prevalence rates for all cancers standardized to the 1980 U.S. population for models 1, 2, 3 and 4

	Males				Females			
Data available	1	2	3	4	1	2	3	4
Last 5 years	855·8	854·4	492·6	346·9	950·0	948·1	705·5	349·1
Proportion*	0·54	0·67	0·75	0·94	0·45	0·53	0·52	0·91
Last 10 years	1247·5	1182·9	589·9	364·4	1504·4	1457·6	1025·1	374·4
Proportion*	0·80	0·93	0·90	0·99	0·71	0·81	0·76	0·97
Last 18 years (all data)	1568·6	1272·3	658·3	367·5	2107·8	1791·3	1347·1	385·0
Predicted total	1741·3	1286·5	681·6	367·7	2747·0	2114·6	1622·4	386·1
Proportion*	1·11	1·01	1·04	1·00	1·30	1·18	1·20	1·00
Years to 0·99†	37	19	28	13	59	35	50	17

* Proportions are calculated with respect to the relevant all data estimates.
† Number of years of data required to obtain 0·99 of estimated asymptotic value.

of years required is large except for model 4 where we can obtain 99 per cent estimates with as few as 13 years of data. We did not conduct site-specific analyses, but one could try these in an attempt to improve the accuracy of the extrapolations.

Comparison of the four methods shows a wide range of results with the British Columbia data indicating that the different assumptions do result in quite different estimates of prevalence. In models 2 and 3, we estimated the time to cure using the method of excess mortality described earlier. Because of the nature of the method used, this is equivalent to time until site-specific excess mortality is no longer statistically significant. A test that this procedure does not lead to a systematic underestimation of the f_h's is obtained by summing the observed and expected number of deaths in the 'cure year' across sites for those sites that have calculated times to cure. This shows an excess of observed (560) over expected (530·0), indicating that the calculated 'times to cure' may be low. The excess of observed over expected may result from inappropriate designation of some sites as curable when they are not, underestimation of time to cure, the excess mortality associated with previous exposure to carcinogens or some combination of these factors. Accurate cause of death information would help resolve evidence for cure. Other approaches that use more powerful methods to calculate survival rates are likely to be useful.[10]

In the examples, although not in the models, we have assumed that the period from diagnosis to remission and from recurrence to death is independent of tumour type. This is an oversimplification. We feel, however, that the selection of a constant one-year period from diagnosis to remission (if occurring) is reasonable in that we should include a case as a prevalent case in the year of diagnosis. It is unlikely that a patient will have a first remission after one year of active treatment, so that one year also represents a reasonable upper limit for this quantity. For model 4, the duration of the second period, $V(x, t)$, from recurrence to death, is likely to be more variable.

Even if we vary it consistently over sites, however, that is one year or four years, it has comparatively little impact on the calculated overall prevalence rates (for example, $V(x, t) = 1$ year, M:325·1, F:335·8; $V(x; t) = 4$ years, M:419·8, F:456·1). Thus the impact of quite large changes in the assumed periods of time from recurrence to death for model 4 is less than that of other assumptions that lead to the various models presented.

It is also interesting to examine the relationship between incidence and prevalence. Over the data collection period the incidence rate of cancer has not varied substantially and has been

Material may be protected by copyright law (Title 17, U.S. Code)

about 300/100,000 for males and 270/100,000 for females. Thus for model 1 the ratio of prevalence to incidence is approximately 5·2 for males and 7·8 for females, whereas for model 4 the ratios are 1·2 and 1·4. This indicates the relative dependency of each method on historical versus current cases.

4. DISCUSSION

We have presented a number of simple algorithms for the calculation of the prevalence of cancer from registry data. The models attempt to simulate various scenarios for the disease presence. Model 1 assumes that, once diagnosed, disease is always present. Since this model does not recognize the possibility of cure it always overestimates the true prevalence of disease. Model 2 assumes that an individual no longer has disease if their follow-up time exceeds the 'time to cure' for their particular tumour, otherwise they have disease with probability one. This is a common medical model for the cure of cancer and is designed to provide guidance in the management of patients. Since definitive treatment typically occurs in a time period much shorter than the time to cure, most cases are cured before they are followed a sufficient time to satisfy the definition. It therefore seems that model 2 overestimates cancer prevalence for sites where cure is possible. Model 3 also assumes that individuals followed for longer than the 'time to cure' are disease free, but estimates the probability of disease presence for those followed for shorter periods by the probability that they will not subsequently die from disease. Model 4 estimates the probability of disease recurrence using the time to recurrence distribution and time from recurrence to death distribution. Model 4 attempts to estimate the prevalence that one would measure by means of a community survey with use of current medical technology.

Completeness of data is an important consideration in the calculation of prevalence. With the data source based on incidence notifications, it is important that the registry has been in operation for some time. The British Columbia Registry provided data on 18 years' incidence (1970–87) at the time of analysis. The effect of using incident cases over shorter periods (last 5 and 10 years) reduced the prevalence estimates for all of the models, but this decrease was greatest for model 1 and least for model 4. In their Connecticut study. Feldman *et al.*[4] had 47 years of data available and using model 1 they found that use of the last 17 years of data resulted in a prevalence rate of 84 per cent of the total. This implies that given complete information on a whole population, then for at least 16 per cent of the prevalent cases (by model 1), 17 years or more will have elapsed since their original diagnosis. Clinical experience would indicate that the vast majority of such individuals will die without recurrence of their original cancer. It seems appropriate that models of cancer prevalence should not be so influenced by long-surviving individuals. We see from Table III that models 2, 3 and 4 are less influenced than model 1 by length of registry follow-up.

We have not formally considered the effects of migration and the difficulties this poses in the indirect methods used here for the calculation of prevalence. For example, immigrants to the population may have had a previously treated cancer of which the registry is unaware unless recurrence occurs. Similarly, one may include emigrants in the calculation of prevalence even though they are no longer members of the population. One would hope that these two effects might approximately balance, although one cannot verify this directly.

For comparative purposes it may be desirable to standardize results, which is usually done by calculating age- and sex-specific rates and using them to calculate an overall rate in a standard population. We have done this in the tables presented using the 1980 U.S. population as a standard. Although standardization will certainly reduce artificial differences between populations it will not necessarily remove differences in past migration patterns that, as already

Material may be protected by copyright law (Title 17, U.S. Code)

discussed, affect methods for calculating prevalence based on reports of incidence. The removal of such effects will require data on migrants that are rarely available. These effects, however, are likely to be small and should not unduly alter the calculated values.

Commonly, in considerations of the impact of cancer, one omits non-melanoma skin cancer from incidence rates. Feldman *et al.*[4] omitted these tumours in their calculation of prevalence. Basal and squamous cancers of the skin are frequently treated by minor surgery, conducted on an outpatient basis, and not consistently reported to a central registry. Since treatment is effective and skin cancer is also the most common malignancy, omission of these tumours prevents them dominating the estimated prevalence rate of cancer. Although skin cancers should be excluded, because of their poor reporting, there is also a need to incorporate the prognosis of individuals with other tumour types. The effect of models 2, 3 and 4 is to weight individuals in relation to their prognosis. We may view these three models, therefore, as ways to formalize the process of discounting tumours whose prognosis is good.

In summary we have shown that changes to the calculation of cancer prevalence, that are in accord with what is known about the behaviour of the disease and the definition of prevalence, can lead to large differences in the estimated values. We assert that the assumption of 'once a case, always a case' leads to large overestimates of the true prevalence of cancer and that one must take care in the calculation and use of such statistics.

ACKNOWLEDGEMENTS

The authors would like to thank the editor and the reviewers for their suggestions which have improved the manuscript. This research was supported by grants from the British Columbia Health Care Research Foundation and the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

1. Waterhouse, J. A. H. *Cancer Handbook of Epidemiology and Prognosis*, Churchill-Livingstone, London, 1974.
2. Rothman, K. J. *Modern Epidemiology*, Little, Brown, Boston, 1986.
3. Elwood, J. M. *Causal Relationships in Medicine*, Oxford, 1988, p. 26.
4. Feldman, A. R., Kersler, L., Myers, M. H. and Naughton, M. D. 'The prevalence of cancer: estimates based on the Connecticut Tumour Registry', *New England Journal of Medicine*, **315**, 1394–1397 (1986).
5. DeVita, V. T., Hellman, S. and Rosenberg, S. A. (Eds.) *Cancer, Principles and Practice of Oncology*, 3rd edn., Lippincott, Philadelphia, 1989.
6. Ederer, F., Axtell, L. M. and Cutler, S. J. 'The relative survival rate: a statistical methodology', *National Cancer Institute Monographs*, **6**, 101–121 (1961).
7. Hakulinen, T. 'On long-term relative survival rates', *Journal of Chronic Diseases*, **30**, 431–443 (1970).
8. Breslow, N. E. and Day, N. E. *Statistical Methods in Cancer Research. Vol. II: The Design and Analysis of Cohort Studies*, IARC, 1987.
9. Nagnur, D. *Longevity and Historical Life Tables. 1921–1981* (abridged), Statistics Canada, catalogue 89-506, Ottawa, 1986.
10. Est'eve, J., Benhamau, E., Croasdale, M. and Raymond, L. 'Relative survival and the estimation of net survival: elements for further discussion', *Statistics in Medicine*, **9**, 529–538 (1990).

Material may be protected by copyright law (Title 17, U.S. Code)