STATISTICS IN MEDICINE, VOL. 16, 425–440 (1997)

ESTIMATING THE COMPLETENESS OF PREVALENCE BASED ON CANCER REGISTRY DATA

R. CAPOCACCIA AND R. DE ANGELIS*

Laboratory of Epidemiology and Biostatistics, Istituto Superiore di Sanitá, Viale Regina Elena, 299, 00161, Roma, Italy

SUMMARY

Prevalence data provided by cancer registries are generally biased, since the patients that were diagnosed before the starting of the registry's activity cannot be included in the statistics. The relevance of this incompleteness bias is estimated in this paper. Incidence and relative survival are modelled as parametric functions describing a wide class of cancer diseases. Prevalence estimates are then computed considering different hypotheses on disease reversibility. The ratio between the prevalence observed by the registry and the total estimated prevalence is used as an index of completeness. An analytical evaluation of this ratio, as a function of the parameters characterizing the observational process and the biological behaviour of the disease, is given.

1. INTRODUCTION

The knowledge of cancer prevalence is of major importance to assess the impact of the disease on the population and on the health system. Population-based estimates of prevalence can be obtained in countries where cancer registration is active for a sufficiently long period of time.[1–3] For this purpose, the life status of all incident cases observed during the activity period of the registry is ascertained.

Apart from considerations concerning the completeness of incidence collection and of follow-up, a certain degree of underestimation is intrinsic in this method of measurement. Persons who had a cancer diagnosis before the beginning of the registration period, and possibly are still living at the reference time for prevalence estimation, cannot indeed be included in the statistics. A measure of cancer prevalence is therefore available only from registries with an observational period long enough to give a reasonably negligible underestimation bias. Feldman *et al.*[2] published, for instance, prevalence rates based on the incident cases observed by the Connecticut Cancer Registry during a period of 47 years. The availability of such a long series of data is quite exceptional among the presently active cancer registries, many of which were established during the late-1970s.

The above cited authors reported that the calculated prevalence of all cancers combined would drop by 5 per cent and 16 per cent if based on data for only 27 and 17 years, respectively. These estimates are useful indications, but are not necessarily valid in other cases. The bias depends, in fact, on factors, such as the survival rates, the age increase of incidence, and the time trends, which are all highly variable both between cancer sites and between populations.

* The research of this author was partially supported by the Progetto Finalizzato A.C.R.O. of the Italian National Research Council (CNR), contract 94.01254.PF39.

CCC 0277–6715/97/040425–16

Received October 1994
Revised March 1996

In the present work, the dependence of prevalence incompleteness bias on the length of the registry follow-up, and on the epidemiological characteristics of the disease, is examined in mathematical terms. Quantitative indications are then given about the magnitude of the incompleteness bias and for the correction of the observed prevalence rates.

2. MATHEMATICAL MODELLING OF CANCER PREVALENCE

Consider a single birth cohort, and let $G(x)$ and $I(x)$ be the general mortality and incidence hazards at age x, respectively. Let also $B(t, x)$ be the death hazard at the same age for people who had a cancer diagnosed at age t. The probability, for a person of the birth cohort, to be alive at age x is

$$\exp\left(-\int_0^x G(u)\,\mathrm{d}u\right)$$

while the probability to be alive and to have a previous cancer diagnosis is

$$\int_0^x I(t)\exp\left(-\int_0^t G(u)\,\mathrm{d}u\right)\exp\left(-\int_t^x B(t,u)\,\mathrm{d}u\right)\mathrm{d}t.$$

The proportion in the population of individuals with cancer diagnosis at age x is therefore given by the ratio of the latter expression to the former, that is

$$N(x) = \int_0^x I(t)\exp\left(-\int_t^x (B(t,u) - G(u))\,\mathrm{d}u\right)\mathrm{d}t \tag{1}$$

The difference $D(t, u) = B(t, u) - G(u)$, which appears in (1), is the excess hazard of patients with respect to the general population. The corresponding exponential integrated function:

$$S(t, x - t) = \exp\left(-\int_t^x D(t,u)\,\mathrm{d}u\right) \tag{2}$$

is an approximation to the net survival function, that is, to the survival function of patients in absence of mortality from other causes. This is usually referred to as the cumulative relative survival function.[4,5] Note that the difference D is not constrained to be nonnegative, and then S may be, strictly speaking, also increasing.

In terms of S, the expression (1) takes the form:

$$N(x) = \int_0^x I(t)\,S(t, x - t)\,\mathrm{d}t. \tag{3}$$

If we assume that the disease is irreversible, that is, that each person who had cancer remains a prevalent case during the rest of her/his life, prevalence is directly given by $N(x)$ in the above expression. This is the definition commonly used in estimating and measuring prevalence.

A more general approach considers the possibility of being removed from the prevalence, as a consequence of disease recovery. For this purpose, let $1 - k(d)$ be the proportion of cases, among the survivors at time d since diagnosis, that are cured and are consequently removed from the prevalence. Thus, the prevalence at age x of cancer patients still in care is given by

$$N(x) = \int_0^x I(t)k(x - t)S(t, x - t)\,\mathrm{d}t \tag{4}$$

where $k(x-t)$ specifies the hypotheses made on disease reversibility. The choice $k(x-t)=1$ corresponds to irreversible disease and equation (4) reduces to (3). A suitable criterion to define recovery, and consequently to remove patients from the number of prevalent cases, has not been clearly stated, being still a matter of discussion.

3. A PARAMETRIC MODEL OF PREVALENCE

$N(x)$ can be directly computed, by means of equation (4), when the functions I, k, S are known. For the purposes of this paper and for many partical applications, it is convenient to express these functions in parametric form.

For incidence age trend an exponential shape, whose validity has been confirmed for a quite general class of cancers,[6] has been considered:

$$I(x) = ax^b \tag{5}$$

where x is the current age, a is a scale parameter, characterizing the birth cohort dependence, and the calender period component is neglected.

The relative survival function of the disease is parametrized by means of a 'mixture model'. We suppose that only a fraction of the patients presents an excess of mortality rate, with respect to the general population. Let A indicate the proportion of these fatal cases among the patients and λ their constant excess hazard rate. The remaining fraction is considered as 'cured' and is supposed to be exposed to the same mortality rates of the general population. Then, the cumulative relative survival function for the whole group of patients is given,[7] as a function of current age x and age of diagnosis t by

$$S(t, x-t) = (1-A) + A\exp(-\lambda(x-t)) \tag{6}$$

where $(1-A)$ represents the proportion of sick people not exposed to any excess of death risk.

Consider now the function $k(d)$ giving, as a function of disease duration $d = x-t$, the proportion of actual prevalent cases among cases with past diagnosis. The following three definitions of prevalence, according to different assumptions on disease reversibility, can be proposed:

(i) *Diagnosis prevalence.* The disease, once diagnosed, is irreversible and both 'cured' and fatal cases contribute to prevalence estimate, thus we have

$$k(d) = 1.$$

(ii) *Fatal cases prevalence.* Only fatal cases are considered as prevalent cases, and the 'cured' patients do not contribute at all to prevalence. This definition gives a lower bound for prevalence, and, of course, has not got an operative meaning, since fatal and cured cases are not distinguishable at the moment of diagnosis. Looking at expression (6) for S, $k(d)$ is given by

$$k(d) = \frac{A\exp(-\lambda d)}{S(t, d)}.$$

(iii) *Care prevalence.* An intermediate approach, between hypotheses (i) and (ii), considers as prevalent either fatal cases, until death, and non-fatal cases that were diagnosed

since a given disease dependent time to cure T_c, that is

$$k(d) = 1 \quad \text{for } d \leqslant T_c$$

$$k(d) = \frac{A \exp(-\lambda d)}{S(t, d)} \quad \text{for } d > T_c. \tag{7}$$

Note that care prevalence (7) includes the previous two definitions as limiting cases for $T_c = \infty$ (or actually $T_c = x$) and $T_c = 0$, respectively. Thus, substituting (6) and (7) into equation (4), we obtain a formally unique expression for total prevalence at age x:

$$N(x) = (1 - A) \int_{x - T_c}^{x} I(t)\,\mathrm{d}t + A \int_{0}^{x} I(t) \exp(-\lambda(x - t))\,\mathrm{d}t. \tag{8}$$

To simplify numerical computations and results presentation, without a significative lack of generality, we will substitute (5) into equation (8) assuming an integer log-linear coefficient b for the incidence age trend, thus obtaining the following final expression for $N(x)$:[8]

$$N(x) = \frac{(1 - A)a}{b + 1}(x^{b+1} - (x - T_c)^{b+1}) + \frac{Aa}{(\lambda)^{b+1}}(-1)^b b! \left(\sum_{m=0}^{b} \frac{(-\lambda x)^m}{m!} - \exp(-\lambda x) \right). \tag{9}$$

4. COMPLETENESS INDEX OF THE OBSERVED PREVALENCE

Suppose that the incidence of a given birth cohort is observed by a registry only for a time period of L years. Total prevalence may be separated into a part $N_O(x, L)$, which derives from the incident cases *observed* between the age interval $[x - L, x]$, and a part $N_U(x, L)$ deriving from the *unobserved* cases diagnosed at previous ages and still living at x, that is

$$N(x) = N_O(x, L) + N_U(x, L)$$

where

$$N_O(x, L) = \int_{x-L}^{x} I(t)k(x - t)S(t, x - t)\,\mathrm{d}t \tag{10a}$$

$$N_U(x, L) = \int_{0}^{x-L} I(t)k(x - t)S(t, x - t)\,\mathrm{d}t. \tag{10b}$$

A measure of the closeness of the observed prevalence to the total one is given by their ratio R:

$$R = \frac{N_O(x, L)}{N(x)} = 1 - \frac{N_U(x, L)}{N(x)}. \tag{11}$$

R can be viewed as a 'completeness index' and varies from a maximum of 1, when all the prevalent cases have been diagnosed during the period of the registry activity, to a minimum of 0 in the opposite case.

R can be directly calculated when the functions I, S and k are expressed in the parametric forms (5), (6), (7), and the resulting total prevalence is given by (9). As a consequence, all the results

presented in the next sections, if not explicitly specified, will refer to the following R expression:

$$R = 1 - \frac{P_n + Q[\exp(Y_n)\sum_{m=0}^{b}(-Y_n)^m/m! - 1]}{P_d + Q[\exp(Y_d)\sum_{m=0}^{b}(-Y_d)^m/m! - 1]} \tag{12}$$

where

$$P_n = \frac{(1-A)}{(b+1)}((\max[x - T_c, x - L])^{b+1} - (x - T_c)^{b+1})$$

$$P_d = \frac{(1-A)}{(b+1)}[x^{b+1} - (x - T_c)^{b+1}]$$

$$Q = Ab!(-1)^b \left(\frac{1}{\lambda}\right)^{b+1} \exp(-\lambda x)$$

$$Y_n = \lambda(x - L); \quad Y_d = \lambda x$$

and where $\max[x - T_c, x - L]$ is the maximum value between $x - T_c$ and $x - L$.

Parameter T_c in expression (12) governs the cancer reversibility hypothesis. Diagnosis prevalence, implying a potentially infinite time to cure, is assumed when $T_c = x$. Fatal cases prevalence corresponds to $T_c = 0$, and an intermediate value, $0 < T_c < x$, gives R estimates under care prevalence hypothesis.

We note that the R estimate given by (12) does not depend on parameter a, describing the incidence level of each birth cohort, since it acts like a scale parameter for either total prevalence or unobserved prevalence.

5. RESULTS

The completeness index R has been modelled as a function of six basic parameters: the length L of the registry follow-up; the age x of the considered population group; the slope b of the relationship between incidence hazard rate and age; the fatality A; the excess death hazard λ of the fatal cases, and the time to cure T_c for non-fatal patients. The survival parameter can be equivalently expressed in terms of the reciprocal of λ, that is, $T = 1/\lambda$, which has the more intuitive meaning of mean survival time for fatal cases in absence of competing death causes. Variables L and x arise from the characteristics of the observational process. The remaining variables depend, on the other hand, on the biological behaviour of the considered disease.

Table I reports the values of parameters b, A, λ and T estimated for a list of cancers. The estimates of the slope parameter b were derived from available literature.[9] The parameters of the survival curve were estimated by least square fitting of recently published cancer survival data from European population-based cancer registries.[10] The values of root mean square error (RMSE) for each survival fit, also reported in Table I, show a lower adaptation of the model for oral and respiratory apparatus, for kidney and brain cancer sites.

The slope of incidence age trend varies between a minimum of 1 (for cervix uteri) to a maximum of 7 (for female lung cancer), with a median value of 5. The parameter A presents the maximum value for pancreas and oesophagus cancers (0·96), and the minimum for corpus uteri (0·29), with a median value of 0·68. Finally, the mean survival time of fatal cases falls, in most cases, between 12 and 24 months, and reaches its maximum value for breast cancer (63 months).

Tables II–IV present the expected values of the R index calculated, for various ages and lengths of follow-up periods, by means of the parameter values reported in Table I. The computation of

Table I. Incidence and relative survival parameter values estimated for a list of cancer sites: b is integer slope parameter of log-log linear relationship between incidence and age; A is proportion of fatal cases; T is mean survival time for fatal cases (months), and RMSE is root mean square error of relative survival fit

Site	b	A	T	RMSE
Tongue	4	0·67	17·6	0·021
Oral cavity	5	0·61	21·1	0·024
Nasopharynx	3	0·70	21·1	0·026
Oropharynx	5	0·74	19·4	0·021
Hypopharynx	5	0·85	14·1	0·025
Head and neck (m)	5	0·76	19·4	0·031
Head and neck (f)	5	0·64	25·5	0·047
Oesophagus	6	0·96	7·5	0·009
Stomach	6	0·82	8·0	0·013
Colon	5	0·59	11·7	0·017
Rectum (m)	6	0·68	19·7	0·015
Rectum (f)	5	0·36	17·4	0·018
Pancreas	6	0·96	5·3	0·004
Larynx (m)	4	0·54	35·3	0·011
Lung (m)	6	0·92	8·2	0·008
Lung (f)	7	0·90	7·4	0·012
Breast	3	0·54	63·2	0·003
Cervix uteri	1	0·46	24·0	0·003
Corpus uteri	3	0·29	18·5	0·005
Ovary	2	0·71	14·5	0·013
Penis	4	0·36	18·8	0·015
Kidney	4	0·60	15.2	0·038
Brain	2	0·86	8·3	0·030

(m) = males (f) = females

R has been performed using the 'diagnosis prevalence' definition, that is, a value $T_c = x$, for comparability with the usual cancer prevalence measures.

The completeness of prevalence achieved with a follow-up of 10 years is generally very low (Table II). Only oesophagus cancer presents a value of R greater than 0·8 for all the considered ages. For only three sites (oesophagus, pancreas and lung) similar values are obtained at the age 70, which is higher than the mean age of the prevalent cases for many cancers.

When 20 years of observation are considered (Table III), the same three cancer sites present a value of R higher than 0·9 for all ages. This level is also reached, at 70 year, for oropharynx, and hypopharynx, stomach and male rectum. Finally, at 30 years (Table IV), most cancer sites present R values greater than 0·9 for all ages, and greater than or near to 0·95 for ages under 70. This is not true for nasopharynx, brain, breast, uterus (both cervix and corpus) and ovarian cancers; the estimated prevalence remains heavily biased for these sites also when based on such a long observational period.

5.1. Dependence of R on registry and morbidity parameters

To analyse separately the dependence of R on the model parameters, let us choose for them a set of standard values, as general as possible, and let us compute R varying one parameter at a time, taking the remaining ones as fixed. From Table I we derive for incidence and survival parameters

Table II. Values of the completeness parameter (R) for a list of cancer sites. Follow-up period of 10 years

Cancer site	Age (years)				
	40	50	60	70	80
Tongue	0·821	0·742	0·673	0·614	0·564
Oral cavity	0·867	0·795	0·730	0·672	0·622
Nasopharynx	0·767	0·684	0·614	0·557	0·508
Oropharynx	0·886	0·821	0·760	0·705	0·656
Hypopharynx	0·904	0·846	0·790	0·739	0·692
Head and neck (m)	0·892	0·830	0·771	0·717	0·669
Head and neck (f)	0·876	0·808	0·745	0·689	0·639
Oesophagus	0·963	0·933	0·900	0·866	0·834
Stomach	0·910	0·850	0·791	0·736	0·687
Colon	0·850	0·773	0·704	0·644	0·592
Rectum (m)	0·910	0·851	0·793	0·739	0·691
Rectum (f)	0·866	0·793	0·728	0·669	0·619
Pancreas	0·949	0·910	0·868	0·827	0·787
Larynx (m)	0·819	0·741	0·673	0·615	0·565
Lung (m)	0·941	0·897	0·851	0·807	0·765
Lung (f)	0·950	0·907	0·862	0·817	0·775
Breast	0·761	0·681	0·614	0·559	0·513
Cervix uteri	0·479	0·399	0·341	0·298	0·265
Corpus uteri	0·700	0·608	0·535	0·478	0·430
Ovary	0·650	0.561	0.492	0.438	0.394
Penis	0·783	0·696	0·623	0·562	0·512
Kidney	0·803	0·720	0·648	0·588	0·538
Brain	0·675	0·587	0·518	0·463	0·418

(m) males, (f) females

the following standard values: $b = 5$; $A = 0{\cdot}7$; $T = 2$ years (or equivalently $\lambda = 0{\cdot}5$ year^{-1}). The mean age of prevalent cases for most cancer sites is $x = 65$ years, thus we choose it as standard value for x. Diagnosis prevalence is assumed as standard definition.

Figures 1(a) to (c) present R as a function of the length of follow-up L, with varying x, b and A parameters, respectively. The incompleteness bias decreases, as expected, with the length of follow-up. The values of R are low for very short lengths, thus follow-up periods shorter than 10 years are not represented, and tend to one for periods of 30 years or more.

In Figure 1(a) the effect of age on R estimate is analysed; the error increases with the age of prevalent cases, almost linearly for follow-up lengths smaller than 20 years and for ages higher than 50.

The R–L curve family that is obtained varying the slope parameter b is plotted in Figure 1(b). The bias appears to be very sensitive to the steepness of the incidence age trend. Cancers characterized by a mild age increase present a lower proportion of recently diagnosed patients, and thus, all the other parameters being equal, a more biased prevalence estimate.

The effect of the fatality parameter A (Figure 1(c)) is straightforward; the higher is the fatality, the lower is the underestimation bias. The dependence on the mean survival time T (Figure 2) is more complex. When the survival time of fatal cases is approximatively null, only non-fatal cases contribute to the prevalence, and the underestimation bias is maximum, that is, R is minimum. As parameter T increases, a progressively higher proportion of fatal cases is included in the prevalence. However, if T remains below a given threshold value, corresponding approximately

Table III. Values of the completeness parameter (R) for a list of cancer sites. Follow-up period of 20 years

Cancer site	Age (years)				
	40	50	60	70	80
Tongue	0·977	0·939	0·893	0·845	0·798
Oral cavity	0·988	0·964	0·929	0·890	0·850
Nasopharynx	0·954	0·900	0·842	0·786	0·735
Oropharynx	0·990	0·968	0·937	0·901	0·864
Hypopharynx	0·992	0·973	0·945	0·913	0·878
Head and neck (m)	0·991	0·970	0·940	0·905	0·869
Head and neck (f)	0·989	0·966	0·933	0·896	0·857
Oesophagus	0·998	0·991	0·979	0·963	0·943
Stomach	0·995	0·980	0·956	0·926	0·894
Colon	0·987	0·960	0·922	0·881	0·838
Rectum (m)	0·995	0·980	0·957	0·927	0·895
Rectum (f)	0·988	0·963	0·929	0·889	0·849
Pancreas	0·997	0·988	0·972	0·952	0·928
Larynx (m)	0·976	0·939	0·894	0·846	0·800
Lung (m)	0·997	0·986	0·969	0·946	0·920
Lung (f)	0·998	0·991	0·977	0·957	0·934
Breast	0·956	0·904	0·849	0·795	0·746
Cervix uteri	0·769	0·662	0·579	0·513	0·460
Corpus uteri	0·941	0·876	0·810	0·748	0·693
Ovary	0·896	0.815	0.740	0.675	0.618
Penis	0·971	0·928	0·876	0·824	0·774
Kidney	0·974	0·933	0·885	0·835	0·786
Brain	0·904	0·826	0·753	0·689	0·633

(m) males, (f) females

to a value $T = L/2$, most of these cases are actually observed by the registry, and the bias correspondingly decreases. Further increase of T produces an increasing proportion of fatal cases arising before the start of the registry activity. This increasing contribution to the prevalence due to 'missing' cases makes the completeness parameter R decrease accordingly. This phenomenon is shown in Figure 2, where the parameter R is now represented as a function of T, and the resulting curves are drawn for various lengths of follow-up L.

The results presented above have been derived, in order to generalize our analysis as much as possible, under some simplifying assumptions on the morbidity modelling. Now we will discuss and test the validity limits of these assumptions in some detail.

5.2. Sensitivity of R to incidence model

The age curve for incidence is derived from a model, proposed in the framework of the multistage theory of carcinogenesis,[9] which has fitted adequately the observed incidence data for many, but not all, cancers. In particular, a log-log linear age trend does not properly describe the data of breast, cervix and ovarian cancers.

The sensitivity of the obtained results to the particular shape of the incidence age curve has been tested for breast cancer. Table V presents the incidence rates of breast cancer estimated for the Italian population[11] using the regression on mortality data proposed by Verdecchia *et al.*,[12]

Table IV. Values of the completeness parameter (R) for a list of cancer sites. Follow-up period of 30 years

Cancer site	Age (years)				
	40	50	60	70	80
Tongue	0·999	0·992	0·975	0·949	0·919
Oral cavity	1·000	0·997	0·987	0·971	0·950
Nasopharynx	0·997	0·980	0·950	0·912	0·872
Oropharynx	1·000	0·997	0·989	0·974	0·954
Hypopharynx	1·000	0·998	0·990	0·977	0·959
Head and neck (m)	1·000	0·997	0·989	0·975	0·956
Head and neck (f)	1·000	0·997	0·988	0·973	0·952
Oesophagus	1·000	0·999	0·997	0·992	0·984
Stomach	1·000	0·999	0·994	0·985	0·970
Colon	1·000	0·996	0·986	0·969	0·946
Rectum (m)	1·000	0·999	0·994	0·985	0·971
Rectum (f)	1·000	0·997	0·987	0·971	0·949
Pancreas	1·000	0·999	0·996	0·990	0·980
Larynx (m)	0·999	0·992	0·975	0·950	0·920
Lung (m)	1·000	0·999	0·996	0·989	0·978
Lung (f)	1·000	1·000	0·998	0·993	0·985
Breast	0·997	0·981	0·953	0·917	0·878
Cervix uteri	0·942	0·850	0·763	0·688	0·625
Corpus uteri	0·996	0·976	0·940	0·897	0·852
Ovary	0·987	0.945	0.890	0.833	0.779
Penis	0·999	0·990	0·971	0·942	0·909
Kidney	0·999	0·991	0·973	0·946	0·914
Brain	0·988	0·948	0·896	0·841	0·788

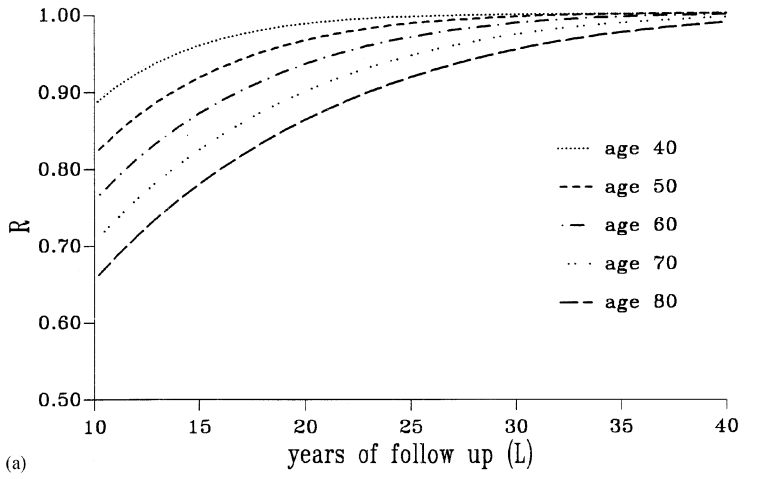
(m) males, (f) females

the corresponding rates expected from the log-log model of incidence, and the estimates of R obtained at different ages from the two sets of values. The incidence age curve used in this comparison is a logistic function having as argument a polynomial function of age, whose degree and coefficients are estimated by means of a regression on mortality data. The predictions derived with this method were suitable for many cancer sites, and in particular for breast cancer.[11] To compute R with this incidence curve a numerical integration of expressions (8) and (10b) has been implemented to derive total and unobserved prevalence.

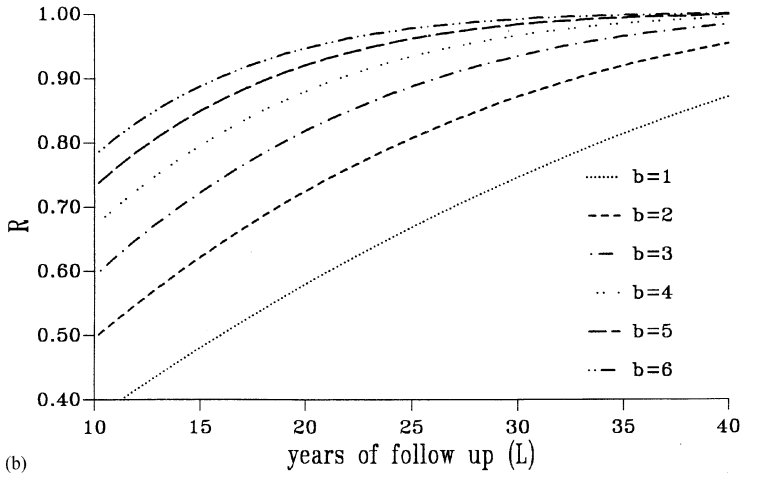
Major differences for the R values are obtained for the ages under 50, due to the different behaviour of the two incidence functions at young ages. Very similar values are, however, estimated for ages equal to or older than 60 years, that give the most important contribution to the overall prevalence rates.

5.3. Sensitivity of R to relative survival model

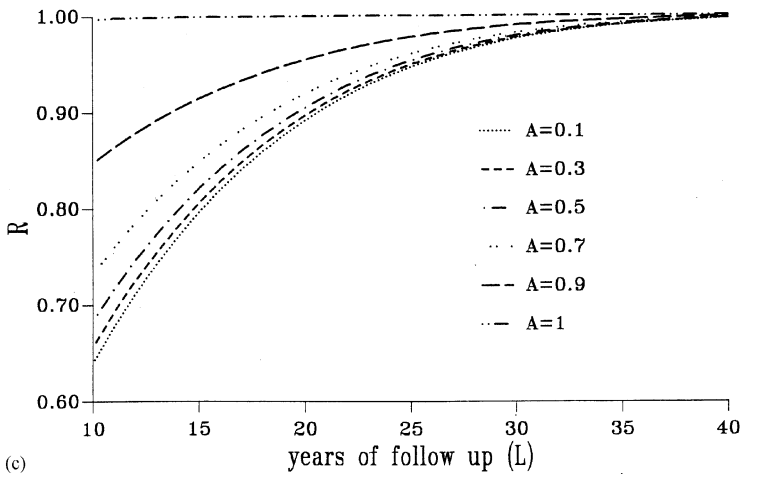
Long-term relative survival rates are a major determinant of prevalence, but they are in principle unobservable when time since diagnosis becomes longer than the observational period of the registry. Thus, assuming a survival parametric model was essential for the estimates carried out in this paper. The mixed survival model given by equation (6) has been already successfully applied in previous studies[13,14] on cancer survival. Moreover, its parameters have a direct meaning in terms of long-term survival. After a sufficient period of time, indeed, when the contribution of the



(a)



(b)



(c)

Figure 1. Prevalence completeness index R as a function of the length L of registry's follow-up, for different values of the relevant variables: (*a*) age; (*b*) slope of the incidence age curve; (*c*) proportion A of fatal cases for the considered cancer site. The other parameters are fixed at the following standard values: age = 65 years; $b = 5$; $A = 0{\cdot}7$; $T = 2$ years, and T_c = age

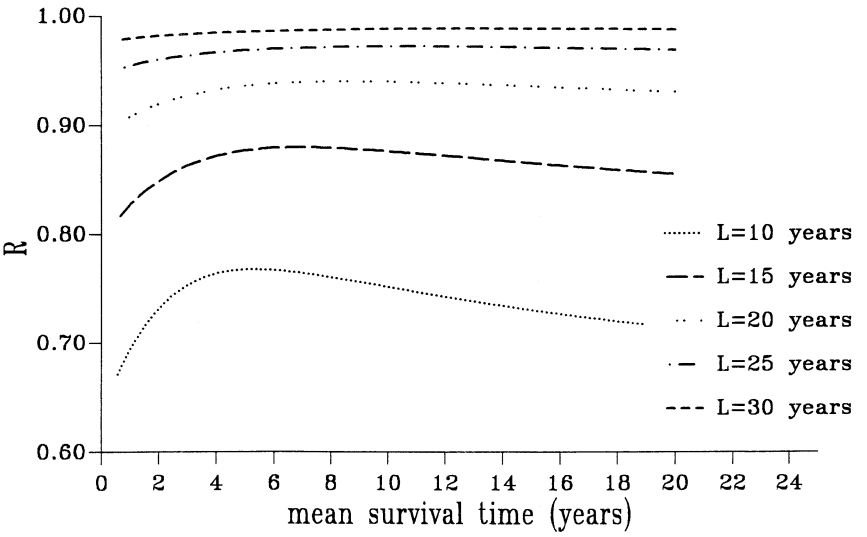


Figure 2. Prevalence completeness index R as a function of mean net survival time T for fatal cases, for various lengths of registry's follow-up (L). The other parameters are fixed at the following standard values: age = 65 years; $b = 5$; $A = 0{\cdot}7$, and T_c = age

Table V. Breast cancer incidence rates $\times 10{,}000$ and corresponding R parameter values by age for two different incidence age models: (A) polynomial logistic function and (B) log-log linear function

Model	Age (years)	Incidence rates	Registry observational period (years)		
			10	20	30
(A)	30	1·5	0·953	0·971	1·000
	40	8·1	0·915	0·994	0·996
	50	17·5	0·795	0·977	0·998
	60	22·5	0·620	0·902	0·988
	70	34·0	0·514	0·780	0·940
	80	55·0	0·508	0·722	0·870
(B)	30	2·8	0·857	0·992	1·000
	40	6·7	0·761	0·956	0·997
	50	13·0	0·681	0·904	0·981
	60	22·4	0·614	0·849	0·953
	70	35·6	0·559	0·795	0·917
	80	51·6	0·513	0·746	0·878

fatal cases fraction becomes negligible, the relative survival nearly approaches the proportion of cured patients expressed by $1 - A$.

Relative survival rates have been assumed as constant with time for each given cohort of patients. For many cancer sites, however, survival decreases with age. On the other hand, prognosis tends to become better with increasing calendar year of diagnosis. The balance between these two effects is generally *a priori* unpredictable, and the assumption of independency of survival on time variables seemed to us the most convenient one. Anyway, the effect of a possible

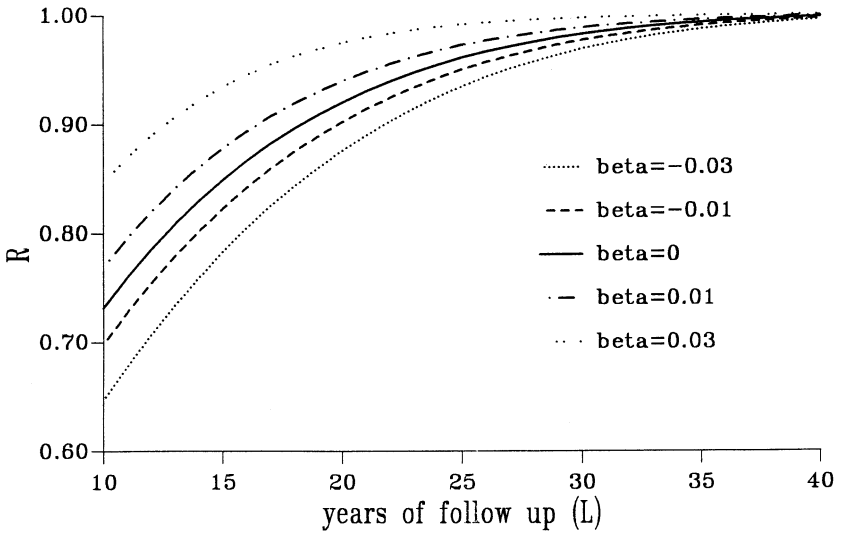


Figure 3. Prevalence completeness index R as a function of the length L of registry's follow-up, for varying exponent β that describes time-dependency of relative survival rates. The other parameters are fixed at: age = 65 years; $b = 5$; $A = 0{\cdot}7$; $T = 2$ years, and T_c = age

overall time trend in relative survival rates on the estimates of R has been evaluated. A relative risk for the excess death hazard of patients, due to the combined effect of age and period, has been introduced as an exponent for the stationary relative survival function considered up to now. For a given cohort at age x we have:

$$S'(t, x - t) = S(t_0, x - t_0)^{\exp - \beta(t - t_0)} \tag{13}$$

where $S(t, x - t)$ is given by equation (6) and t_0 is a reference age that we fixed at 65 years. The values of R are plotted in Figure 3 against the length L of the observational period for different values of β, ranging from $-0{\cdot}03$ to $0{\cdot}03$ (year^{-1}). The same standard values considered in the previous figures have been assumed for the other parameters.

As expected, for increasing values of β the R-curves tend to move toward the top of the figure, indicating better estimates when the global effect of age and period components is an increase of excess hazard with time. In this case, indeed, a lower proportion of prevalent cases were diagnosed before the start of the registry's activity. At 20 years of follow-up, for example, a yearly increase of 1 per cent in excess hazard leads to an increase of R of about 3 per cent.

5.4. Sensitivity of R to time to cure (T_c) definition

The possibility of cure has been introduced in the prevalence estimate by means of a disease dependent time to cure (7). Fatal cases are always included in the prevalence, whereas only non-fatal patients that were diagnosed since a time T_c are considered as prevalent cases. Until now we considered diagnosis prevalence, that is, a potentially infinite time to cure (actually a time equal to patient's age: $T_c = x$), where disease irreversibility is assumed. Now we will adopt the care prevalence assumption ($0 < T_c < x$). This hypothesis will reduce the incompleteness bias, because a significant proportion of unobserved cases will be classified as cured and thus excluded from prevalence. To have a general idea of the effect of T_c definition we plotted in Figures 4(a)

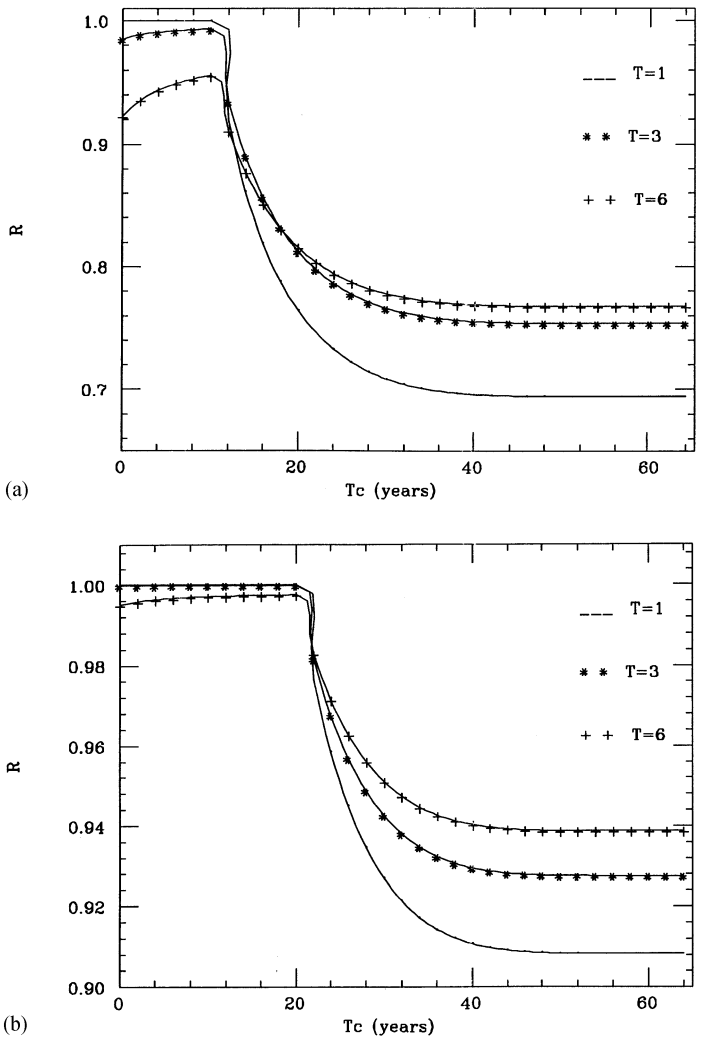


Figure 4. Prevalence completeness index R as a function of time to cure T_c for registry's follow up lengths L equal to: (*a*) 10 years; (*b*) 20 years. The mean survival times of fatal cases are: $T = 1, 3, 6$ years. The other parameter values are fixed at: age = 65 years; $b = 5$; $A = 0{\cdot}7$

and (*b*) the completeness index R, for $L = 10$ and $L = 20$, respectively, as a function of T_c varying in the interval $[0, x]$. The other parameter values are $A = 0{\cdot}7$, $x = 65$ years, $b = 5$ and $T = 1, 3, 6$ years.

When $T_c = 0$ fatal cases only contribute to prevalence, and the proportion of unobserved cases depends on the length of follow up L compared to mean survival time T. This proportion is always negligible for $L = 20$ and equal to a maximum of 8 per cent for $L = 10$ in the case $T = 6$.

When $T_c > 0$ a progressively increasing proportion of non-fatal patients is included in the prevalence and two definite different trends of R can be distinguished in Figure 4:

(i) $0 < T_c \leqslant L$, non-fatal cases are always observed prevalent cases, thus their inclusion increases the value of R with respect to the case $T_c = 0$ (this increase is more marked for $L = 10$ than for $L = 20$);

Table VI. Values of completeness index (R) by age for follow-up period $L = 10$ years. Different time to cure (T_c) since diagnosis, corresponding to different proportions of still alive fatal cases (P), are considered

Site	T_c (years)	P (%)	Completeness index by age 40	50	60	70	80
Breast	∞	0	0·76	0·68	0·61	0·56	0·51
	24	1	0·77	0·72	0·68	0·65	0·63
	16	5	0·84	0·81	0·79	0·78	0·77
	0	100	0·94	0·93	0·92	0·91	0·90
Larynx	∞	0	0·82	0·74	0·67	0·62	0·57
	13	1	0·91	0·88	0·87	0·86	0·85
	9	5	1·00	1·00	0·99	0·99	0·99
	0	100	0·99	0·99	0·98	0·98	0·98

(ii) $L < T_c < x$, the proportion of non-fatal patients unobserved by the registry is increasingly higher with increasing T_c. This effect is responsible for the decrease of R down to an asymptotic minimum level reached even for values smaller than $T_c = x$.

Time to cure is related to the specific cancer prognosis. We define it here as the time within which the relative survival for fatal cases is reduced to a given negligible level (P). Looking at expression (6), this definition takes the following mathematical form:

$$T_c = T \ln(1/P) \tag{14}$$

Proportions of still alive fatal cases equal to 5 per cent and 1 per cent correspond to T_c values equal to $3T$ and $4{\cdot}6T$, respectively, thus giving quite different times to cure according to the specific mean survival time T. We used these two values of T_c to estimate R for the list of cancer sites given in Table I with follow-up periods $L = 10$, 15, 20 years. The resulting completeness index values become very close to 1 for all sites, even for $L = 10$ years, apart from those cancers presenting high mean survival times, that is, larynx and breast. In Table VI the corresponding R values are represented for $L = 10$ with varying time to cure definitions. The prevalence incompleteness of breast cancer is always significant and is not negligible, due to unobserved fatal cases, even for $T_c = 0$. Larynx cancer presents a less marked bias for $P = 1$ per cent, whereas the case $P = 5$ per cent, which corresponds to a time to cure lower than the considered length of follow-up, gives, as expected from Figure 4(*a*), R values equal to 1.

6. DISCUSSION

The main purpose of this work was to study the degree of reliability of prevalence measures based on cancer registries. Publications of cancer registries' data generally consider all people with past diagnosis of cancer as prevalent cases, irrespective of the possibility of cure. This definition certainly overestimates the actual number of persons suffering from the disease. Cancer can be, indeed, in many cases, really cured. Other definitions of prevalence have been discussed[15] to take into account the fact that many cancer patients live in a disease-free status during the rest of their life, dying for causes unrelated to cancer. Therefore, the concept of cure has been included in this paper expressing the probability of being disease-free as a function of time since diagnosis.

Clinical disappearence of the disease does not necessarily mean, however, a true reversibility of the patient's health status and behaviour. Therapies, such as surgery or radiotherapy, may cause

permanent impairment or also serious invalidating effects. Past history of cancer may cause, also in successfully treated patients, a particularly strong concern about their health, and a more frequent than usual recourse to health services. The relevance of these long-term consequences of the disease strongly depends on the cancer site. It is also likely to vary according to socio-economical status and cultural attitudes. Data for their evaluation are not generally collected from cancer registries, and specific epidemiological studies on long-term survivors should therefore be implemented for this purpose. For these reasons, and also for comparability reasons, we feel that the classical concept of diagnosis prevalence still remains an important indicator of the burden of cancer on a population.

Incidence and relative survival functions are expressed in this paper in a parametric form suited to describe a wide class of cancers. We used mathematical models that are based on biological assumption that are not always valid for all the various cancer sites. Furthermore, the parameters involved in incidence and survival models do vary in general between population, genders and age classes. The values of the slope of incidence age trends for the considered cancers were taken from the Cook *et al.* paper.[9] They considered incidence rates from eleven cancer registries established in developed western countries, and whose data were available at that time. While having a longitudinal meaning, their estimates are based, however, on age specific rates observed at a single point in time. Exposure is therefore assumed as constant across cohorts. The effect of a changing exposure on the validity of the model and on its parameter values is extensively discussed in the same paper. The estimates presented in Table I, therefore, give average and approximate parameters values useful to perform example applications of the theoretical expressions presented in the paper.

For 'real world' applications, available data from the considered population should be carefully considered, and the following approach is consequently suggested:

(i) If the population incidence and survival data fulfil the parametric models with parameter values given in Table I, then corrected prevalence estimates can be obtained dividing the observed prevalence by a value of R selected in Tables II–IV.
(ii) If the parameter values are different, but the models can be still assumed as valid, then expression (12) can be used to compute an unbiased estimate of R.
(iii) In all other cases, the observed values of incidence and survival rates for the considered population can be substituted in expressions (8) and (10b) to numerically compute total and unobserved prevalence. The correction parameter R is then directly derived from equation (11).

Finally, the following general conclusions can be definitely derived from this study.

1. The bias in estimating diagnosis prevalence from cancer registries incidence series is in general high. When the average age of prevalent cases is quite high (over 65 years for many cancer sites), an observational basis of 30 years or more is needed to obtain acceptable estimates.
2. Such a long period is not even sufficient to correctly estimate the prevalence of cancers with high survival rates and a relatively early age of occurrence, such as breast or cervical cancer.
3. A considerably shorter period is sufficient for estimating the prevalence of highly fatal cancers such as lung or pancreas. In these cases, however, the usefulness of prevalence measures is uncertain.
4. Exclusion of cured cases from prevalence drastically increases the accuracy of the estimates for follow-up periods equal to or longer than at least 10 years.

REFERENCES

1. Teppo, L., Hakama, M., Hakulinen, T., Lehtonen, M. and Saxen, E. 'Cancer in Finland 1953–70: Incidence, mortality, prevalence', *Acta Pathologica Microbiologica Scandinavica* (*A*), Suppl 252, (1975).
2. Feldman, A. R., Kessler, L., Myers, M. H. and Naughton, M. D. 'The prevalence of cancer: estimates based on Connecticut Tumor Registry', *New England Journal of Medicine*, **315**, 1394–1397 (1986).
3. Adami, H. O., Gunnarsson, T., Sparen, P. and Eklund, G. 'The prevalence of cancer in Sweden 1984', *Acta Oncologica*, **28**, 463–470 (1989).
4. Hakulinen, T. 'On long-term relative survival rates', *Journal of Chronic Diseases*, **30**, 431–443 (1977).
5. Esteve, J., Benamou, 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).
6. Armitage, P. and Doll, R. 'The age distribution of cancer and a multi-stage theory of carcinogenesis', *British Journal of Cancer*, **8**, 1–15 (1954).
7. Capocaccia, R. 'Relationships between incidence and mortality in non-reversible diseases', *Statistics in Medicine*, **12**, 2395–2415 (1993).
8. Gradshteyn, I. S. and Ryzhik, I. M. *Table of Integrals, Series and Products*, Academic Press, NY, 1965.
9. Cook, P. J., Doll, R. and Fellingham, S. A. 'A mathematical model for the age distribution of cancer in man', *International Journal of Cancer*, **4**, 93–112 (1969).
10. Berrino, F., Sant, M., Verdecchia, A., Capocaccia, R., Hakulinen, T. and Estéve, J. (eds) *Survival of Cancer Patients in Europe. The EUROCARE Study*, IARC Scientific Publication No. 132, Lyon, 1995.
11. Capocaccia, R., Verdecchia, A., Micheli, A., Sant, M., Gatta, G. and Berrino, F. 'Breast cancer incidence and prevalence estimated from survival and mortality', *Cancer Causes and Control*, **1**, 23–30 (1990).
12. Verdecchia, A., Capocaccia, R., Egidi, V. and Golini, A. 'A method for the estimation of chronic disease morbidity and trends from mortality data', *Statistics in Medicine*, **8**, 201–216 (1986).
13. Goldman, A. I. 'Survivorship analysis when cure is a possibility: a Monte Carlo study', *Statistics in Medicine*, **3**, 153–163 (1984).
14. Gamel, J. W., McLean, I. W. and Rosenberg, S. H. 'Proportion cured and mean long survival time as functions of tumor size', *Statistics in Medicine*, **9**, 999–1006 (1990).
15. Coldman, A. J., McBride, M. L. and Braun, T. 'Calculating the prevalence of cancer', *Statistics in Medicine*, **11**, 1579–1589 (1992).