STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:3511–3526 (DOI: 10.1002/sim.1304)

Estimation and projections of cancer prevalence from cancer registry data

Arduino Verdecchia*,†, Giovanni De Angelis and Riccardo Capocaccia

Istituto Superiore di Sanita', *Rome*, *Italy*

SUMMARY

A method, PIAMOD (Prevalence, Incidence, Analysis MODel), which allows the estimation and projection of cancer prevalence patterns by using cancer registry incidence and survival data is presented. As a first step the method involves the fit of incidence data by an age, period and cohort model to derive incidence projections. Prevalence is then estimated from modelled incidence and survival estimates. Cancer mortality is derived as a third step from modelled incidence, prevalence and survival. An application to female breast cancer is given for the Connecticut State by using data from the Connecticut Tumor Registry (CTR), 1973–1993. The age, period and cohort model fitted incidence quite well and allowed us to derive long-term projections up to 2030. Patients' survival was also projected to future years according to a scenario approach based on two extreme hypotheses: steady, that is, no more improvements after 1993 (conservative), and continuously improving at the same rate as during the observation period. Age-standardized estimated incidence shows a changing trend around the year 2005, when it starts decreasing. Age-standardized prevalence is expected to increase and change trend at a later date. Breast cancer mortality is projected as decreasing, as the combined result of no further increase in incidence and improving cancer patients' survival. An easy-to-use PIAMOD software package, on which work is in progress, will be made available to individual cancer registries and/or health planning institutions or authorities once it is developed. The use of the PIAMOD method for cancer registries will allow them to provide results of paramount importance for the whole community involved in the assessment of future disease burden scenarios in an evolving society. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: prevalence; incidence; survival; breast cancer; projections; regression analysis

INTRODUCTION

Incidence, mortality and survival are the basic measures used to describe disease patterns. When related to cancer, they allow the burden to be monitored and evidence the progress achieved in the struggle against cancer. Prevalence, defined as the number and/or the

* Correspondence to: Arduino Verdecchia, Laboratorio di Epidemiologia e Biostatistica, Istituto Superiore di Sanita', Viale Regina Elena, 299 00161 Roma, Italy.
† E-mail: verdeck@iss.it

Received January 2001
Accepted March 2002

proportion of people with past or present diagnosis of a certain disease (for example, cancer), within a well-defined population at a fixed point in time, provides an additional direction in this context. Prevalence is intrinsically related to the other three measures and gives a comprehensive view of the simultaneous effect of incidence, survival and mortality patterns on the cancer load for a population. In addition, prevalence provides relevant information for practical use, that is, for: (i) planning health services; (ii) allocating health resources; (iii) administering medical care facilities; (iv) designating appropriate research expenditures; (v) assessing the relative burden of cancer with respect to mortality and life quality deprivation.

Whereas incidence and mortality are directly derived from data collected, respectively, by cancer registries and national health statistics, survival and prevalence can only be derived from incidence and mortality data. The task is particularly cumbersome for prevalence, which requires a very long time series of cancer registry data to compute complete prevalence measures [1, 2], that is, including all patients in the population, irrespective of time since diagnosis. Prevalence estimates derived from cancer registry data are always partial, as they do not include cases occurring before the start of the registration activities. The degree of this bias is variously relevant depending on the length of observation period. Correction factors can be identified and used to obtain estimates of total prevalence also with limited observation period [3].

The existing approaches to estimate cancer prevalence may be clustered into two categories, according to data used: direct and indirect methods. Any of these two categories can include either numerical or statistical methods.

Direct methods [1, 2, 4–6] exploit individual incidence and life status follow-up data to count patients living in the population at a defined time. Lost-to-follow-up cases cannot be properly taken into account, since their own life status is actually not known, and their own contributions to prevalence have to be estimated statistically [6]. Depending on the length of the cancer registry observation period, direct prevalence estimates of this kind are partial to a variable degree. A theoretical correction factor obtained by modelling the observed and unobserved part of actual prevalence [3] can be computed and used to obtain total prevalence estimates from partial measures. Although existing direct methods are essentially numerical, they necessarily involve some statistical correction factors. The major limits of numerical direct methods are that they can provide prevalence results for areas covered by a cancer registry only and invariably referred to the past.

Indirect approaches include statistical methods [7, 9] which make use of cancer mortality data and an estimate of patient survival to back-calculate cancer incidence and prevalence, and to provide short-term and middle-term projections of both incidence and prevalence. The MIAMOD method [7, 8] was extensively applied to provide national cancer incidence and prevalence estimates and projection in Italy [10–13]. The MIAMOD method is currently being widely applied in European countries, within the EC BIOMED-II EUROPREVAL project, and in the U.S.A. [14].

The aim of this work is to present a new statistical direct method, PIAMOD, developed to provide estimates of complete prevalence from cancer registry data as well as medium-term projections to the future. An application is given to female breast cancer for Connecticut to show the use and the performance of the proposed method. Female breast cancer was chosen as one of the most important cancer sites in the U.S. giving great concern for the expected cancer load.

METHOD

The Prevalence and Incidence Analysis MODel (PIAMOD) is presented here as a method to obtain statistical estimates and projections of prevalence by basically using individual incidence and patient follow-up data, as usually used for population-based survival analyses. The method takes advantage of the MIAMOD formulation [7, 8]. Presentation of the method is given in separate sections, each one devoted to a specific internal task.

Estimating cancer prevalence

Prevalence of an irreversible disease, such as cancer, is defined as the probability v_i of being found in the population at age i, having had present or past diagnosis for the disease.

For a birth cohort, prevalence can be expressed as a convolution of incidence and patients' survival time distribution [7], as follows:

$$v_i = \sum_{j=0}^{i-1} (1 - v_j)\mu_j \sigma_{ij} \tag{1}$$

where the cohort-specific prevalence v_i at exact age i is expressed as the summation over all ages up to i of the probability μ_j of becoming ill between age j and $j+1$ (with j less than i) times the probability of surviving the risk of dying up to exact age i, say σ_{ij}. The term $(1-v_j)$ represents the proportion of healthy people at age j within the cohort, that is, the appropriate denominator where new cancer cases can come from. Since the effect of general mortality pattern on the birth cohort is implied in the age profile of the cohort itself, the appropriate σ_{ij} in equation (1) should express the probability of surviving the extra death hazard specifically due to cancer disease, that is, cumulative relative survival rate. Relative survival is defined as the ratio of the observed survival rate in the group of patients to the survival rate expected in a group of people in the general population, who are similar to the patients, at the beginning of follow-up period, with respect to all possible factors affecting survival except for the disease under observation. In principle, relative survival should represent the survival rate if the disease of interest were the only cause of death, that is, accounting for the extra risk of dying in sick people compared with the general healthy population.

Equation (1) gives the estimated age-specific prevalence probability for a birth cohort, provided that disease incidence and patient survival are known. A system of equations (1), including one equation for each birth cohort involved in the observation period, allows us to reconstruct cross-sectional prevalence series for the entire observation period. Average prevalence proportions over one year $N_i(t)$ is computed from birth cohort specific exact age prevalence $v_i^{(z)}$ as follows:

$$N_i(t) = (v_i^{(z)} + v_{i+1}^{(z-1)})/2$$

where z is the birth cohort and $t = z + i$ the index calendar year.

Starting from complete cohort incidence, either observed or estimated, these prevalence estimates include all cancer patients with a past history of cancer, irrespective of date of diagnosis, that is, total prevalence.

When incidence can be projected into the near future, projected prevalence can be obtained accordingly.

Estimating cancer mortality

By using the relationships between incidence, prevalence, survival and mortality, cancer mortality can be derived from incidence, prevalence and survival [7]. For a birth cohort, expected age specific cancer mortality, M_i, can be expressed as following:

$$M_i = \sum_{j=0}^{i} (1 - v_j)\mu_j \sigma_{ij} \delta_{ij} \tag{2}$$

where δ_{ij} represents the crude probability of death from cancer at age i for cancer patients diagnosed between age j and $j+1$, and surviving up to exact age i. The probability δ_{ij} is derived from the cumulative relative survival curve as follows:

$$\delta_{ij} = \left(1 - \frac{\sigma_{i,j}}{\sigma_{i,j+1}}\right)(1 - q_i^*)$$

where q_i^* is the probability of death from competing causes at age i for people belonging to the birth cohort surviving at i, and $\sigma_{i,j}$ the relative survival rate at age i for patients diagnosed at age j, that is, $i-j$ years before (see equation (6)). Owing to a lack of specific information, mortality from competing causes for cancer patients, q_i^*, can be assumed to be the same as the general population, q_i, in practical applications.

Equation (2) gives the estimated age-specific mortality probabilities for a birth cohort, provided that disease incidence (either observed or estimated), μ_j, prevalence, v_j, and patients' survival σ_{ij} are known. A system of equations (2), including one equation for each birth cohort involved in the observation period, allows us to reconstruct cross-sectional mortality series for calendar years involved in the observation period. When incidence and prevalence can be projected to the near future, projected mortality can be obtained accordingly.

Regression analysis of incidence data

Let m_{it} be the number of incident cases at age i ($i = 0, 1, \ldots, I$) and time t in years ($t = 1, 2, \ldots, T$), as recorded on a defined population by a cancer registry that has been operating for T years at least.

Age, period and cohort trends of incidence are modelled by conventional log or logistic regression in order to identify independent age, period and cohort effects with only a small number of parameters. Incidence probability at age i and time t, as needed to be plugged into equations (1) and (2) is assumed as a polynomial function of age, period of diagnosis and birth cohort, throughout a link function ϕ:

$$\phi(\mu_{it}) = \text{const} + \sum_{k=1}^{k_1} a_k i^k + \sum_{k=1}^{k_2} b_k t^k + \sum_{k=1}^{k_3} c_k (t-i)^k \tag{3}$$

where $\theta = (\text{const}, a_1, \ldots, a_{k_1}, b_1, \ldots, b_{k_2}, c_1, \ldots, c_{k_3})$ is the vector of the parameters to be estimated. According to the generalized linear model theory [15] the link function can be either the natural logarithm or logit. Both logarithm and logit functions give the same results for small argument, such as annual cancer incidence rates usually expressed as per 100000. A log link function was actually used in the application.

The degree of the polynomials, k_1, k_2 and k_3, have to be chosen to give the best fit of the $I \times T$ matrix of observed cancer incidence counts m_{it}. Since the cohort term is a linear combination of age and year (cohort = year − age) the linear term of period of diagnosis will be excluded in order to avoid colinearity problems when estimating the parameters.

Observed cancer incidence counts m_{it} is assumed to be a random variable following the Poisson distribution with expected value

$$E(m_{it}) = \mu_{it} P_{it} \tag{4}$$

where P_{it} represents age and time specific population size. The likelihood function is expressed as

$$\log L(\theta) = \sum [m_{it} \log [P_{it} \hat{\mu}_{it}(\theta)] - P_{it} \hat{\mu}_{it}(\theta)] \tag{5}$$

The maximum-likelihood estimate of the parameter vector θ can be obtained by maximizing equation (5), by using standard techniques. The degree of the polynomials, k_1, k_2 and k_3, are identified by a stepwise procedure based on likelihood ratio statistics (LRS) [7]. An asymptotic standard errors estimate of the maximum likelihood parameters may be obtained from the second-order derivatives of log-likelihood function (5), in the same way as in reference [7].

Estimation of survival

Relative survival estimates are needed for prevalence estimation (see equation (1)) and they can be directly derived from incidence and follow-up data and population life-tables, using standard methods [16, 17]. Tabulated relative survival estimates along with their standard errors can be modelled by mixture models [18] to extrapolate quite far back and forward in time as needed by the PIAMOD method and to facilitate making clear assumptions when projecting into the future. The essential features of the mixture modelling approach are briefly recalled here.

Let k represent the kth stratum for the data, either sex or age group or period of diagnosis, the relative survival model is given by

$$\sigma_k(d) = \alpha_k + (1 - \alpha_k) \exp(-(\lambda_k d)^{\gamma_k}) \tag{6}$$

where α_k represents the proportion of patients potentially cured from cancer, λ_k the hazard and γ_k the scale parameters of the Weibull distribution that describes the time to death for the proportion $(1 - \alpha_k)$ of fatal cases. Relative survival parameters α, γ and λ in the relative survival model were estimated by fitting the relative survival values $S_k(d_l)$ at times since diagnosis d_l $(l = 1, \ldots, L_k)$ by sex, age group and period, together with the corresponding standard errors. The SAS NLIN procedure [19] was used, with the inverse of the variances of the observations used as weights.

Owing to right truncation of survival data, a complete stratified analysis by period of diagnosis is likely to give biased results, particularly for breast cancer with a high and slightly declining survival curve. Including period of diagnosis as covariate in model (6) is a more robust way to proceed. Formally, for each stratum k, for example age class, we express

$$\alpha_k(t) = 1/(1 + \pi_0 \exp(\pi t)), \quad \lambda_k(t) = \beta_0 \exp(\beta t)$$

where π_0, π, β_0 and β are unknown parameters to be estimated and t the time of diagnosis.

Two indicators will be considered to summarize the results obtained from these models. The first one is the proportion of cured patients, α. The second one is the mean survival time for fatal cases, T, which is given by

$$T_k = \lambda_k^{-1}\Gamma(1+\gamma_k^{-1}) \tag{7}$$

where Γ is the gamma function.

Once the two parameters α and T are estimated, a given survival curve is represented by these parameters and easily made available to the PIAMOD computer programme. When projecting to the future, with respect to available data, this formulation allows us to make very clear-cut assumptions on each of the parameters, for example, increasing proportion of cured patients and decreasing, or steady survival time for fatal cases, as a potential effect for screening practices.

Deriving projections

A number of hypotheses need to be adopted to derive and project morbidity rates into the future: major hypotheses involve incidence and survival, minor ones the population evolution patterns.

Projection of modelled incidence to calendar years following the observation period can be derived by assuming the persistence of both age and cohort effect identified into the future. For cancer disease this hypothesis is quite reasonable because cancer risk in adulthood is generally thought of as determined by past exposure to several known and unknown risk factors. Conversely, hypotheses on period effects, that is, changes in incidence affecting all the age groups simultaneously, although present for past years, cannot be kept for subsequent years, and are not then considered for projections. Only the drift already expressed by the cohort linear term, that is indistinguishable from linear period trend, is considered. When the linear extrapolation of the drift is based on a logarithmic link function then implausible exponential asymptotic growth can even occur. When this is the case the use of logit link function is preferable.

Hypotheses on cancer patient survival are also needed for projecting purposes. As in any scenario approach, the most appealing choice is not to try to adopt the best hypothesis that can be taken, but to provide a plausible range for projected rates. To do this, one conservative and one optimistic hypothesis are proposed. A reasonable, but pessimistic, hypothesis consists in assuming cancer patients' survival as remaining stable for projected years, that is, survival improvements will no longer be observed in the projected years as they have been observed in the past years. Conversely, in a rather optimistic scenario, we assume cancer patients' survival as continuing to improve at the same rate as observed in recent past years.

Minor *ad hoc* hypotheses on the population evolution patterns have to be adopted. The number of new born and the age specific general non-breast-cancer mortality are kept constant all along the projection period, being equal to the respective values for the last calendar year for which actual data are available, that is, the last period of estimation. Population at older age classes is estimated by accounting for members of cohort incrementing age and expected number of deaths.

Projected prevalence and specific mortality patterns are then reconstructed from projected incidence and survival according to the technique shown above.

DATA SOURCES

Incidence and follow-up data from the Connecticut Tumor Registry (CTR), 1973–1993, were extracted from the SEER Public Use CD-ROM [20]. Population data, mortality for breast cancer and all-causes mortality for the state of Connecticut were kindly provided by the National Cancer Institute (NCI, NIH, Bethesda, MD, U.S.A.), by single year of age, 0–84 years, for 21 calendar years, 1973–1993.

APPLICATION TO BREAST CANCER PREVALENCE IN CONNECTICUT

An application of the PIAMOD method to breast cancer data for Connecticut is given to show its potential and performance in providing a complete scenario of projected morbidity and mortality indicators that can be useful for cancer control and health planning purposes. Although the CTR is famous for having one of the longest time series of cancer incidence in the world, we chose to use the limited 1973–1993 series in our application just to show the potential of the method to be widely used with most European and U.S. cancer registry data.

As a first step we provide mixture model estimation and analysis of breast cancer relative survival for Connecticut, 1973–1993. The age trend of breast cancer relative survival by time since diagnosis and period of diagnosis is reported in Figure 1. Survival markedly improved during the two-decade period from 1973 to 1993 during the expansion of screening activities in the U.S. [21, 22], both at short-term and long-term. The age profile, although rather flat,

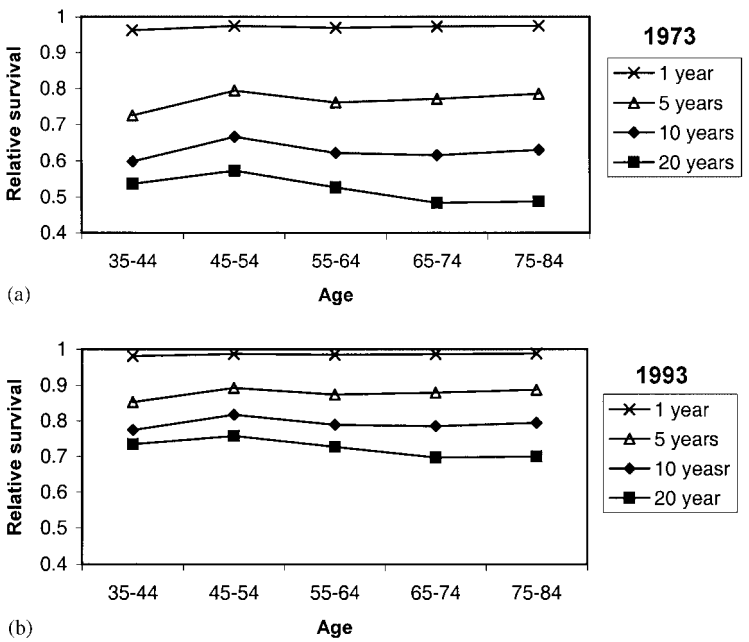

Figure 1. Modelled female breast cancer relative survival by age and time since diagnosis, Connecticut: (*a*) 1973; (*b*) 1993.

Table I. Proportion per cent of breast cancer women cured, P, and mean survival time for uncured patients, T (years), by age and time of diagnosis, Connecticut, 1973–1993.

Age	15–year relative survival	Average SE of P	Time of diagnosis							
			1973–1977		1978–1982		1993–1987		1988–1993	
			P	T	P	T	P	T	P	T
35–44	53	0.61	53	6.0	57	6.5	59	7.1	61	7.8
45–54	58	0.55	57	6.7	60	7.5	62	8.6	65	9.9
55–64	51	0.81	47	8.3	51	9.8	54	11.8	57	14.3
65–74	50	4.28	40	13.2	49	14.0	58	14.8	67	15.6
75–84	44	7.65	34	17.8	45	19.4	57	21.2	70	23.3

*Period of diagnosis 1973–1977.

presents a peak with better survival for women aged 45–54 as frequently found in several breast cancer survival analyses [23, 24]. This effect is not confined to short-term, instead it looks more pronounced as time from diagnosis increases.

Table I reports the estimated proportion of cured breast cancer women (P) (that is, not bound to die from breast cancer) and the mean survival time (T) for uncured patients by age class, as obtained by mixture model application. Average standard error (SE) of P and 15-year relative survival for patients diagnosed in 1973–1977 are reported in addition. The cured proportion decreases with age, although long-term relative survival is not so sensitive to age (see Figure 1). The proportion of women cured of breast cancer markedly increased during the period 1973–1993, as also clearly reflected by the improvement of relative survival figures. Overall, more than 60 per cent of women diagnosed with breast cancer in 1988–1993 are expected to be cured, whereas this proportion was less than 50 per cent in average for those diagnosed during the 1970s. Proportion of cured women is lower than 15-year survival, particularly for elderly women, thus indicating that a residual excess death hazard still persists after 15 years from diagnosis. This apparent inconsistency between the age pattern of 15-year relative survival and the proportion of breast cancer female patients that are cured is due to changes in the shape of the age-specific survival curve. Also improvements in survival attained with time affect the shape of the survival curves, thus leading to different asymptotic expectations.

An incidence model including an eighth-order polynomial in age and a second-order polynomial in cohort was selected in a stepwise procedure as based on LRS test. As an example, to compare two nested models with a and $a+b$ parameters, respectively, we compute the difference of the corresponding LRS and interpret it as a χ^2 statistic with b degrees of freedom. Table II shows the model selection procedure. The minimum of the LRS was found for model 5. Further increasing of LRS, although not expected, might occur in practice for numerical reasons. The eighth-order polynomial is required to model the complex age pattern of breast cancer incidence. This result is consistent with a previous application in Italy [10] in which the same polynomial order was identified as the best fit. The second-order cohort polynomial is required to model the change in incidence age profile with time. Figure 2 reports estimated and observed incidence age profile in 1973 and 1993 as an indication of the goodness of the fit operated by the selected model on CTR incidence data. Although observed incidence present large variability in the rates, the model was able to properly catch the so-called 'Clemmesen's

Table II. APC model identification procedure.

Step	Model	Number of parameters	LRS
1	A6P0C0	7	6192
2	A6P0C2	9	5848
3	A8P0C0	9	2653
4	A8P0C1	10	2541
5	A8P0C2	11	2405
6	A8P0C3	12	2408
7	A8P0C2	14	2440
8	A9P0C2	12	3345

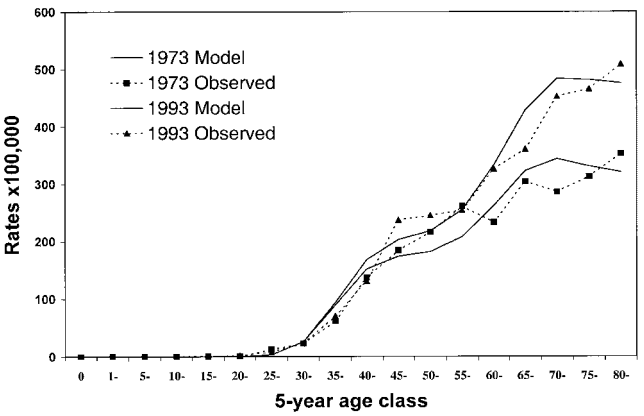

Figure 2. Comparison of estimated and observed female breast cancer incidence for the years, 1973 and 1993. Rates per 100000.

hook' at ages near 45–49 [25]. Incidence increased during the study period, particularly for women aged 60 and over.

Table III reports estimated and projected breast cancer mortality, incidence and prevalence by age class for women in Connecticut. Projections refer to the more conservative (pessimistic) hypothesis on survival, that is, with the relative survival remaining stable in the projection period. Projected population is also reported in addition. For younger women aged less than 50 there is a systematic decline from 1990 onwards. Prevalence is estimated as still increasing during the 1990–2000 period, and declining thereafter. Conversely, for women over 65 all the indicators show an increase. For intermediate age classes between young and older women, changing trends were estimated as for a transition area between the two regimens. Overall ages, crude mortality, incidence and prevalence rates increases from 1990 to 2030. When considering age-adjusted rates, mortality decreases from 41 per 100000 in 1990, to 33 per 100000 in 2030, and incidence increases with a decreasing speed from 1990 to 2000 and it is expected to decrease by near 10 per cent from 2000 to 2030. Also for prevalence a slight decline is expected from 2000 to 2030. The so called 'baby boom' phenomenon, that is, persons born between 1946 and 1964, clearly shown as a moving pinch in the population

Table III. Mortality, incidence and prevalence of female breast cancer, Connecticut, 1990, 2000 and 2030. Rates per 100000 women

Age	Population	1990			Population	2000			Population	2030		
		Mortality	Incidence	Prevalence		Mortality	Incidence	Prevalence		Mortality	Incidence	Prevalence
0	18336	0	0	0	19649	0	0	0	19649	0	0	0
1–4	72494	0	0	0	78926	0	0	0	78926	0	0	0
5–9	83809	0	0	0	98069	0	0	0	98564	0	0	0
10–14	81562	0	0	0	91183	0	0	0	98511	0	0	0
15–19	104933	0	0	0	82413	0	0	0	98411	0	0	0
20–24	120726	0	0	1	83782	0	0	0	98225	0	0	0
25–29	125113	0	3	6	98633	0	3	6	98021	0	2	0
30–34	126896	2	27	60	117746	1	25	61	97752	1	18	0
35–39	115500	8	93	303	130684	6	91	312	97098	4	64	146
40–44	108109	17	167	819	125254	13	166	877	91843	9	122	654
45–49	81415	29	203	1524	116076	22	200	1654	81810	16	158	1326
50–54	71264	41	213	2159	100159	33	223	2402	79619	26	183	2123
55–59	77653	54	245	2717	77273	47	269	3154	89605	38	234	3065
60–64	78445	75	325	3482	66212	66	352	4085	104266	58	335	4300
65–69	73686	112	414	4398	66200	104	455	5282	111185	100	469	5979
70–74	61818	175	466	5438	65128	166	519	6599	100929	171	579	7991
75–79	48420	265	460	6202	56408	261	527	7737	79768	280	619	10110
80–84	32670	410	458	6294	41489	425	534	7837	55921	503	679	11040
85+	29762	648	422	6565	37431	832	540	9416	40610	1019	691	13456
All ages	1512611	56	145	1542	1552648	64	165	1955	1623899	77	183	2550
Age adjusted*		41	124	1290		36	127	1445		33	113	1438
Number of cases		842	2198	23324		987	2567	30349		1244	2974	41414

*1973 Connecticut female population was used as standard.

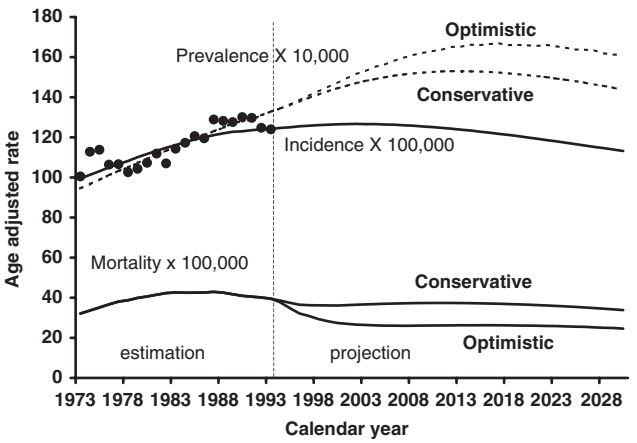

Figure 3. Estimated and projected age-adjusted incidence, mortality and prevalence rates for female breast cancer in Connecticut, according to two hypotheses on projected patients' survival. Observed incidence rates 1973–1993 are reported (dots) for goodness-of-fit evaluation. Female Connecticut population in 1973 was used as standard.

structure all along 1990 to 2030 of Table III, is expected to affect the elderly (65 and over) population in 2030, thus placing a burden on the medical care system. The reduction of breast cancer incidence in young generations completely involves the baby boomers. Owing to this decline in breast cancer incidence, the burden of breast cancer in 2030 is limited with respect to what can be expected by simple linear projection of recently observed incidence and prevalence trends to projected population of 2030.

Figure 3 shows trends of age-adjusted breast cancer mortality, incidence and prevalence rates, 1973–2030. The Connecticut population in 1973 was used as standard in order to eliminate the effect of population ageing. Within the estimation period, 1973–1993, incidence is the resulting curve of model fitting on CTR data, while both prevalence and mortality are the expected rates on the basis of incidence and patients' survival. Within the projection period, 1994–2030, prevalence and mortality are considered in both the conservative and the optimistic hypotheses as discussed above, generating an interval with extremes within which the results for mortality and prevalence are most likely to be found. Of course no effect is given on incidence by differences in relative survival. By the year 2030, breast cancer prevalence is estimated as ranging between 1438 and 1610 per 100000 women, with expected breast cancer mortality ranging from 25 to 34 per 100000 women, and incidence equal to 113. Comparatively, in 1993 we had incidence and mortality almost at the same levels, 124 and 32, respectively, prevalence at lower level, 1260 per 100000 women.

Table IV reports estimates of female breast cancer prevalence from different sources for comparison. PIAMOD estimates are fully consistent with estimates derived from CTR data, that is, by using the counting method PREVAL (PREVALence) (Micheli *et al.*, [26]), and the method by Feldman [2], as reported in references [27, 28]. Conversely, prevalence estimates derived by the U.S. National Health Interview Survey (NHIS) [27] for 1997 appear to be grossly underestimated with respect to any other method.

Table IV. Comparison of female breast cancer prevalence estimates by different sources proportion per 100000 women.

Year	PIAMOD	PREVAL*	NHIS†	CTR†	Merril‡
1987§	1422		1332	1485	
1992§	1616	1612			
1993¶	1637				1581

*Micheli *et al.* [26].
†Byme, 1992 [27].
‡Merril *et al.* 2000 [28].
§Crude proportions, all ages.
¶Crude proportions, 0–89 years of age.

DISCUSSION

In this paper a statistical method is presented that may provide estimates and projections of cancer incidence, prevalence and mortality by fitting incidence data from population-based cancer registries. This method can greatly expand the potential of cancer registries to provide updated and useful information for health planning, resource allocation and cancer control activities.

The PIAMOD method was formulated in discrete time just because practical applications usually deal with discrete data, although it has its more basic representation in continuous time [7]. Some approximation is necessarily involved in the model even when 1 year age and calendar year classes are in use. We assumed that events (that is, diagnosis, death) can only occur at the midpoint between two consecutive birthdays, except for those who become sick and die within the same calendar year and to whom an average disease duration of six months is assigned. This leads to discrete survival times as well. Estimated incidence and mortality account for annual probabilities. Prevalence at exact age is basically estimated by the model, while it is usually needed as a proportion at a defined point in time. Therefore, estimated prevalence at exact age is averaged between contiguous age classes to obtain an average prevalence proportion during the year, as based on the idea that population birthdays are almost uniformly distributed within each calendar year.

To obtain meaningful results for projection, the method involves assumptions on the data that need to be discussed. For the incidence data only effects on generations, that is, age and cohort effects, depending on past exposures, can be considered in the projection process whereas this is not the case for effects involving all ages at the same time, that is, period effects, apart from the first linear term. Patient survival is modelled as mixture model of fatal and cured patients, thus providing an efficient way to summarize information for use in the PIAMOD software and to make specific assumptions on cured and not cured patients for projection. Two extreme hypotheses are adopted, respectively, no more improvements in the projection period (conservative) and steady improvement all along the projected period at the same rate as observed in previous years (optimistic). In this way it is possible to test the two extreme situations, and all the possible results for survival-related measures, that is, mortality and prevalence, are likely to be within the interval defined by the results. The number of

newborns and the non-breast cancer mortality in the population are kept constant all along the projection period as those of the last calendar year for which estimations are possible. General all-causes mortality is then variable within the projected years only as the result of expected breast cancer mortality. Population size is then projected by incrementing each age class with members of each cohort incrementing age and with decrement by the expected number of deaths. Intermediate and older age classes are not affected by the assumption on the number of newborns. This is not the case for younger ages, born at a constant rate after 1993 and diminished at a constant mortality rate, which will be always identical in different projected years after a finite number of steps. Although strongly affected by the assumptions, this younger component of the projected population is not relevant for breast cancer in that it is a very rare disease in the under 30 age groups.

As an application to show all the potential of PIAMOD, a study on breast cancer data from the Connecticut Tumor Registry (CTR), diagnosed in the period 1973–1993, has been performed.

The mixture model technique for relative survival calculation allows the survival trends for the disease to be determined and clearly evidenced, with marked improvements all along the two-decade period. The proportion of women cured from breast cancer is always estimated as lower than 15-year relative survival (see Table I and Figure 1), thus revealing the presence of a residual excess death hazard still persisting after 10 years from diagnosis, particularly for the elderly. Table I reports also 15-year relative survival for cases diagnosed during 1973–1977 to show that a residual excess death hazard still persists even after 15 years from the diagnosis. Completeness of follow-up might be an important issue when looking at very long-term survival figures. In particular, if deaths are missing the effect might be an artificial proportion of cured. CTR is part of the SEER and this implies good quality of the data, including both incidence and follow-up. We checked long-term relative survival for colon cancer where a proportion cured is known to exist. Colon relative survival for the CTR cases diagnosed 1973–1977 was 55 per cent at 5 years, 48 per cent at 10 years and 47 per cent at 20 years, thus remaining almost stable since 6 years since diagnosis. The effect of potential missing deaths would be increasing in long-term relative survival in this case.

For the incidence model the best fit has been obtained with an eighth-order-in-age and second-order-in-cohort polynomial, a model already obtained for the same cancer site in Italy [10], an indication perhaps of an endogenous aetiology related to the physiology of women and unrelated to external exposures. Such a model is anyhow needed to cope with the complicated age pattern shown by breast cancer patients. Period effects did not appear significant and were not included in the model. Increases in mammography utilization, whether or not due to organized screening programmes, are expected to result in an increase in both incidence and patients' survival. For Connecticut, as for most areas in the U.S.A., the utilization of mammography examination spread widely since 1982 [29]. A smoothed rather than sharp increase of breast cancer incidence was observed and modelled [30] as due to diffusion of mammography screening practice in Connecticut and the U.S.A. Improvements in survival were also attained progressively, as also results from our analysis (see Table I). Although dissemination of breast cancer screening spread in the Connecticut population during the 1980's and was responsible for most of the increase in breast cancer incidence, it did not cause any major problem in our modelling application.

Goodness-of-fit of the model to breast cancer incidence data was rather good, as the model was able to well capture both age and secular trends, as can be seen from Figures 2 and 3.

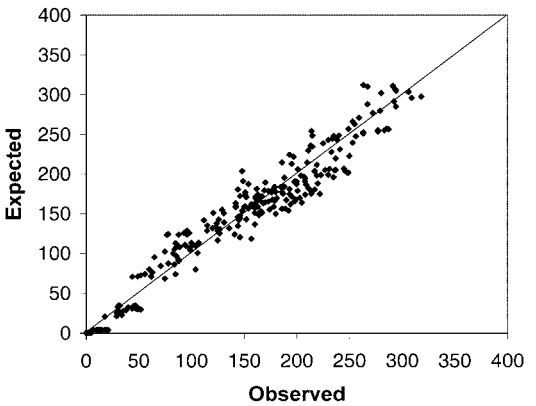

Figure 4. Age class and calendar time-specific expected versus observed incidence cases, 1973–1993.

Random variability strongly affects observed incidence rates since Connecticut is not a large state (1.5 million women). We used a rather parsimonious model to catch most of the overall trends having the objective to derive projections. Although the fit for the first calendar years 1973–1975 was rather poor, the overall trend is modelled satisfactorily. In Figure 2, we reported a comparison of observed and estimated age incidence profile for the two extreme calendar years of the data. Random variability strongly affects age specific observed incidence rates and limits the consistency with estimates. We used a single age-cohort model to fit the 84×21 age-period data cells. Our estimates represent the average age profile among the 21 observed, with modulation made by the cohort effects. Distribution of age and calendar year-specific residuals of estimated and observed incidence counts is reported in Figure 4. Residuals are scattered across the line representing null residual. Although there are clusters in the lower left corner, not very important for projections and corresponding to young ages and first calendar years for which the fit was poor, the overall pattern for the adult and elderly population looks as randomly distributed with mean zero and standard deviation 16.5, that is, less than 10 per cent mean relative residual based on average 200 counts in each age-period cell. The analysis of residuals showed the existence of a moderate extra-Poisson variation ($\phi = 1.57$) indicating that residuals are larger than expected simply on Poisson distribution. The effect of this overdispersion does not influence the estimates and only informs us that standard errors of the estimates might be underestimated to some extent.

In model formulation we assumed relative survival to be an independently known quantity. asymptotic confidence intervals for incidence estimates and projections we can derive from the Poisson regression are partial, as they do not include any uncertainty deriving from relative survival estimation and modelling. Developing a bootstrap procedure to evaluate the total uncertainty for estimates and projections is planned for future work. Actually we validated prevalence estimates with independent estimates from literature and provided a region for plausible prevalence projections according to two hypotheses on patients' survival.

When results for estimation and projection are obtained for both survival scenarios, and corrected for population age structure, incidence is found to increase strongly in the period 1973–1993, slowly until the year 2000, and decreasing in 2000–2030. Towards 2030 an overall stable to decreasing pattern is observed for mortality, 2008 versus 1990, and stable thereafter,

according to the two hypotheses on survival trend. In the same time framework, prevalence shows a changing trend from a moderate increase to decrease starting from 2018 in the stable survival scenario, while faster increasing with delayed start to decrease in the improving survival scenario. Overall we have very different trends among incidence, prevalence and mortality, making it very hard to guess each one from the others, and only a joint analysis of the three indicators can give a consistent picture of how things are.

The 'baby boom' phenomenon, persons born between 1946 and 1964, is well taken into account in the projection. Prevalence of breast cancer in the elderly is then estimated to increase dramatically and almost double between 1990 and 2030, while being almost stable for younger ages (see Table III). Breast cancer prevalence in the elderly is then the phenomenon of major concern in the near future. Age-adjusted prevalence is expected to be no longer increasing after 2020 as the effect of the expected decreasing trend of age-adjusted incidence from 2010.

We compared the breast cancer prevalence estimates we derived by means of the Piamod method applied to SEER CTR data, 1973–1993, with other estimates available from the literature. Consistency between all the methods involved was impressive, with the only exception of the NIH survey estimate that actually refers to the whole of the U.S.A. not to the Connecticut state. Two types of comparison with results of counting processes were involved: (i) partial prevalence derived from CTR data 1973–1993 and corrected for limited observation period; (ii) almost total prevalence obtained from the historical CTR data 1940–1993, with more than 50 years of observation. These results are an important validation step both for the statistical PIAMOD method and the correction factor [3] used to correct partial prevalence for a limited observation period.

The detail of the results obtained with PIAMOD highlights the need to develop an easy-to-use software package (work in progress), to be made available to individual cancer registries and/or health planning institutions or authorities as a fundamental instrument to provide results of paramount importance for the whole community involved in the assessment of the future disease burden scenarios in an evolving society.

REFERENCES

1. Adami HO, Gunnarsson T, Sparen P, Eklund G. The prevalence of cancer in Sweden. *Acta Oncologica* 1989; **28**:463–469.
2. Feldman AR, Kessler L, Myers MH, Naughton MD. The prevalence of cancer. *New England Journal of Medicine* 1986; **315**:1397–1397.
3. Capocaccia R, De Angelis R. Estimating the completeness of prevalence based on cancer registry data. *Statistics in Medicine* 1997; **16**:425–440.
4. Krogh V, Micheli A. Measure of cancer prevalence with a computerized program: an example on larynx cancer. *Tumori* 1996; **82**:1–4.
5. Coldman AJ, McBride ML, Braun T. Calculating the prevalence of cancer. *Statistics in Medicine* 1992; **11**: 1579–1589.
6. Micheli A, Francisci S, Krogh V, Giorgi Rossi A, Crosignani P and the ITAPREVAL Working Group. Cancer prevalence in Italian Cancer Registry areas: the ITAPREVAL Study. *Tumori* 1999; **85**:309–369.
7. Verdecchia A, Capocaccia R, Egidi V, Golini A. A method for the estimation of chronic disease morbidity and trend from mortality data. *Statistics in Medicine* 1989; **8**:201–216.
8. De Angelis G, De Angelis R, Frova L, Verdecchia A. MIAMOD: a computer program to estimate chronic disease morbidity using mortality and survival data. *Computer Methods and Programs in Biomedicine* 1994; **44**:99–107.
9. Mariotto A, Verdecchia A. Using AIDS mortality data to reconstruct HIV/AIDS epidemics. *Statistics in Medicine* 2000; **19**:164–174.
10. Capocaccia R, Verdecchia A, Micheli A, Sant M, Gatta G, Berrino F. Breast cancer incidence and prevalence estimated from survival and mortality. *Cancer Causes and Control* 1990; **1**:23–30.

11. Capocaccia R, Micheli A, Berrino F, Gatta G, Sant M, Ruzza MR, Valente F, Verdecchia A. Time trends of lung and larynx cancers in Italy. *International Journal of Cancer* 1994; **57**:1–8.
12. Capocaccia R, De Angelis R, Frova L, Gatta G, Sant M, Micheli A, Berrino F, Conti E, Gafa L, Roneucci L, Verdecchia A. Estimation and projections of stomach cancer trends in Italy. *Cancer Causes and Control* 1995; **6**:339–346.
13. Capocaccia R, De Angelis R, Frova L, Sant M, Buiatti E, Gatta G, Micheli A, Berrino F, Barchielli A, Conti E, Gafa L, Verdecchia A. Estimation and projections of colorectal cancer trends in Italy. *International Journal of Epidemiology* 1997; **26**:924–932.
14. Mariotto A, Capocaccia R, Verdecchia A, Sant M, Feuer EJ, Clegg LX. Estimating prevalence of cancer in the U.S. through the use of mortality and survival data. *Cancer* (submitted).
15. Dobson A. *An Introduction To Generalized Linear Models*. Chapman and Hall: London, 1990.
16. Ederer F, Axtell, LM, Cutler SJ. The relative survival rate: a statistical methodology. In *End Results and Mortality Trends in Cancer.* National Cancer Institute Monograph No. 6. US Government Printing Office: Washington DC, 1961; 101–121.
17. Hakulinen T, Abeywickrama KH. A computer program package for relative survival analysis. *Computer Methods and Programs in Biomedicine* 1985; **19**:197–207.
18. Verdecchia A, De Angelis R, Capocaccia R, Sant M, Micheli A, Gatta G, Berrino F. The cure of colon cancer: results from the Eurocare Study. *International Journal of Cancer* 1998; **77**:322–329.
19. SAS Institute Inc. *SAS/STAT User's Guide*, *Version 8*. SAS Institute Inc: Cary, NC, 1999.
20. SEER 1973-93 Public-Use CD-ROM. U.S. Department of Health and Human Services, PHS/NIH/NCI/NCI/CSB, August, 1996.
21. Dershaw DD. Mammographic screening of the high-risk woman. *American Journal of Surgery* 2000; **180**:288–289.
22. Woloshin S, Schwartz LM, Byram SJ, Sox HC, Fischhoff B, Welch HG. Women's understanding of the mammography screening debate. *Archives of Internal Medicine* 2000; **160**:1434–1440.
23. Adami HO, Malker B, Holmberg L, Persson I, Stone B. The relation between survival and age at diagnosis in breast cancer. *New England Journal of Medicine* 1986; **315**:559–563.
24. Sant M, Gatta G, Micheli A, Verdecchia A, Capocaccia R, Crosignani P, Berrino F. Survival and age at diagnosis of breast cancer in a population-based cancer registry. *European Journal of Cancer* 1991; **27**:981–984.
25. Clemmesen J. *Statistical Studies in the Aetioligy of Malignant Neoplasms.* 1: *Review and Results.* Munskgaard: Copenhagen, 1965.
26. Micheli A, Yancik R, Krough V, Verdecchia A, Sant M, Capocaccia R, Berrino F. Contrast in Cancer Prevalence in Connecticut, Iowa, and Utah. *Cancer* 2002; **95**, in print.
27. Byrne J, Kessler LG, Devesa SS. The prevalence of cancer among adults in the United States. *Cancer* 1992; **69**:2154–2159.
28. Merril R, Capocaccia R, Feuer EJ. Cancer prevalence estimates based on tumour registry data in the Surveillance, Epidemiology, and End Results (SEER) Program. *International Journal of Epidemiology* 2000; **29**:197–207.
29. Miller A, Feuer EJ, Hankey FB. The increasing incidence of breast cancer since 1982: relevance of early detection. *Cancer Causes and Control* 1991; **2**:67–74.
30. Feuer EJ, Wun LW. How much of the recent rise in breast cancer incidence can be explained by increases in mammography utilization? *American Journal of Epidemiology* 1992; **136**:1423–1436.

