Metodi statistici

Statistical methods

Introduzione

La prevalenza per tumore è la proporzione di individui che vivono in una determinata popolazione con una pregressa diagnosi di tumore, indipendentemente da quanto questa sia lontana nel tempo. La prevalenza è espressa sia come numero di individui prevalenti nella popolazione sia come proporzione di individui prevalenti sul totale della popolazione considerata. La prevalenza è un indicatore della domanda sanitaria complessiva dei pazienti oncologici ed è un indicatore complesso perché è il risultato dei pregressi andamenti temporali dell'incidenza e della sopravvivenza per tumore, ma anche delle tendenze della mortalità generale nella popolazione considerata.

La pubblicazione sino a ora disponibile sulla prevalenza per tumore nelle aree dei registri italiani è stata prodotta dallo studio ITAPREVAL[1] che ha calcolato la prevalenza al 31 dicembre 1992.

Lo scopo dello studio attuale è di fornire informazioni aggiornate al 1°gennaio 2006 sulla prevalenza dei malati oncologici nelle aree coperte dai Registri di popolazione, che nel frattempo si sono significativamente ampliate, utilizzando la banca dati dell'Associazione italiana dei registri tumori (AIRTUM).[2]

Questa sezione è interamente dedicata alla descrizione delle procedure metodologiche utilizzate per il calcolo della prevalenza.

Metodi

La ***prevalenza completa*** indica il numero o la proporzione di tutti i soggetti in vita con una pregressa diagnosi di tumore, indipendentemente dalla data della diagnosi. La prevalenza viene calcolata a una certa data di riferimento (***data indice***), la data in cui si valuta lo stato in vita dei pazienti.

La prevalenza in un'area coperta da registrazione viene calcolata utilizzando direttamente i dati di incidenza e di stato in vita raccolti dai registri tumore di popolazione (RT), i quali rilevano tutte le nuove diagnosi nel territorio di loro pertinenza. La prevalenza interamente basata sui dati dei registri si definisce ***prevalenza osservata***.

Nella quasi totalità dei dati disponibili, la prevalenza osservata è ***incompleta*** perché i RT possono rilevare soltanto le diagnosi di tumore che si sono verificate *dopo* l'avvio della registrazione. Di conseguenza, la prevalenza osservata dai RT rappresenta una quota tanto maggiore della prevalenza completa, quanto più lungo è il tempo della loro osservazione; soltanto i RT attivi da 40-50 anni rilevano una prevalenza osservata virtualmente com-

Introduction

Cancer prevalence is defined as the proportion of persons in a population who, during their lifetime, have been diagnosed with cancer, regardless of the date of diagnosis, and are still alive at a defined reference date. Prevalence is expressed both as the number of individuals prevalent in the population, as well as the proportion of individuals prevalent out of the total population considered. Prevalence is a complex measure, depending on cancer incidence, survival and on general mortality trends in the considered population. It represents the overall health care demand of cancer patients, including those who may be considered cured and require only a few additional health care resources.

The latest available data on cancer prevalence in Italian cancer registries were provided by the ITAPREVAL Study[1] with a reference date of 31 December 1992. The aim of the present study is to provide updated information on cancer prevalence for the reference date 1 January 2006 in the areas covered by Italian population-based cancer registries (CR) and in Italy by using information available in the database of the Italian Cancer Registries Association (AIRTUM).[2] This chapter describes the methodology used in the study to calculate prevalence.

Methods

Complete prevalence *indicates the number and proportion of individuals with a previous cancer diagnosis, irrespective of the diagnosis date. Prevalence is estimated at a given point in time, called the* ***index date****.*

In areas covered by cancer registration systems, prevalence can be directly computed from incidence and life status data collected by registries on their target population. The prevalence indicator, when entirely based on incidence and follow-up data collected by a registry, is called ***observed prevalence****.*

In almost all available data, observed prevalence is necessarily ***incomplete****, since it refers to the number and proportion of cases diagnosed after the start of the registry activity. Consequently, the observed prevalence of Italian registries represents a percentage of the complete prevalence as much greater as their observation time is longer; only registries with 40-50 years of follow-up can detect a virtually complete observed prevalence. Italian registries started in different years from 1978 onwards and their observation periods range from 7 to 28 years.*

pleta. In Italia i RT sono stati avviati a più riprese a partire dal 1978 e il loro tempo di osservazione oscilla tra i 7 e i 28 anni. La prevalenza completa in questo studio è stata stimata in parte dalla prevalenza osservata e in parte, per i periodi antecedenti all'inizio della registrazione, attraverso fattori correttivi (***indici di completezza***) che consentono di stimare la *frazione* di prevalenza completa *non osservata* da un dato registro. Tali indici variano in funzione della lunghezza del periodo di registrazione e sono specifici per sede tumorale, perché dipendono da incidenza e sopravvivenza della neoplasia. Infatti, a parità di anni di osservazione, l'incompletezza della prevalenza osservata per le diverse sedi oncologiche è tanto maggiore quanto maggiore è l'incidenza e/o quanto migliore è la prognosi negli anni precedenti all'avvio della registrazione.

La ***prevalenza a durata limitata*** indica invece il numero o la proporzione di pazienti che hanno ricevuto una diagnosi di tumore negli anni precedenti, per esempio 2, 5 o 10 anni, e permette di differenziare i bisogni sanitari degli individui prevalenti in relazione alla durata della malattia. Gli indici di completezza della prevalenza consentono di stimare sia la prevalenza completa sia la prevalenza a durata limitata per quelle frazioni che superano il periodo massimo di osservazione di un registro.

Sono stati inclusi nello studio i RT che avevano un periodo di registrazione di almeno 5 anni e con ultimo anno di incidenza non antecedente al 2003. L'analisi è stata effettuata su 24 RT (tabella 1) con anni di incidenza e fine follow-up sullo stato in vita variabili, il che ha reso necessario mettere a punto delle procedure di stima specifiche per allineare la prevalenza a una comune data indice. I risultati di prevalenza presentati nello studio si riferiscono ai casi diagnosticati entro il 2005 e vivi al 1° gennaio 2006; infatti per la quasi totalità dei registri (19 su 24) l'incidenza è disponibile almeno fino all'anno 2005 e la data di valutazione dello stato in vita è sempre successiva al 2005.

Per i 5 registri con ultimo anno di incidenza variabile tra 2003 e 2004, la prevalenza è stata calcolata nelle rispettive date indice (Sassari,Trento e Veneto al 1° gennaio 2005; Genova e Varese al 1° gennaio 2004) e poi proiettata al 1° gennaio 2006 con una specifica procedura descritta nel seguito.

Il periodo di osservazione dei 24 RT italiani varia da un minimo di 7 anni a un massimo di 28. La prevalenza osservata, a durata limitata e completa, è stata prima ricavata singolarmente per ciascun RT e poi sommata per la costruzione dei valori aggregati per macro-area o nazionali.

La prevalenza a durata limitata è stata calcolata a **2, 5, 10, 15 e 20 anni** dalla diagnosi e corrisponde esattamente alla prevalenza osservata per durate inferiori al periodo massimo di osservazione del registro, mentre è stata stimata attraverso gli indici di completezza per durate superiori. Per esempio nel caso del RT del Friuli Venezia Giulia, dove la massima durata di malattia osservata è pari a 11 anni, i risultati di prevalenza a 2, 5, e 10 anni dalla diagnosi sono *osservati*, mentre quelli a 15 e 20 anni sono stati stimati applicando gli indici di completezza alla prevalenza osservata a 11 anni dalla diagnosi.

Estimation of complete prevalence was based partly on observed prevalence and partly, for the period before the start of registration, by modelling a quantity, called the ***completeness index****, that allows the estimation of the fraction of complete prevalence not observed in the recorded data. These indices vary based on the length of the registration period and are specific for tumour sites, since they depend on the incidence of neoplasia as well as on survival. In fact, on a par with years of observation, the incompleteness of prevalence observed for the different cancer sites is as much greater as incidence and/or survival levels are higher in the years preceding the start-up of registration.*

Limited duration prevalence *indicates the number or proportion of patients who have received a diagnosis of a tumour in previous years, for example 2, 5, or 10 years, and helps to differentiate the needs of health services for patients according to the duration of the illness. Completeness indices for prevalence allow estimating both complete prevalence as well as limited duration prevalence, for those fractions above the maximum period of observation for a Registry. Limited duration prevalence is entirely observed for time periods shorter than the observation period of a given registry, but must be modelled for longer periods.*

This study includes those Registries with at least a five-year registration period and the last year of incidence not before 2003. The analysis was performed on 24 contributing registries (Table 1) with variable years of incidence observation and final follow-ups on the survival status, which has made it necessary to fine-tune the specific estimation procedures to align prevalence with a common index date.

Prevalence results presented in this study refer to cases diagnosed up to 2005 and living on 1 January 2006. Indeed, for almost all registries (19 out of 24), the common latest year for incidence data is 2005 and the follow-up closing date is always later than 2005. For five registries with incidence data ending in 2003 or 2004, complete prevalence is computed at 1 January 2004 (Genoa and Varese) or 1 January 2005 (Sassari, Trento and Veneto) and subsequently projected to 1 January 2006 through specific procedures described in the following.

The observation time period ranges from a minimum 7 years to 28 years. Observed, limited duration and complete prevalence was first computed for each single registry and then summed up to derive values by macro-area or nationally.

Limited duration prevalence was calculated at ***2, 5, 10, 15 and 20 years*** *from diagnosis and corresponds exactly to the observed prevalence for durations lower than the maximum observation period for a registry, while for greater durations it was estimeted via completeness indices. For example, in the case of the Friuli Venezia Giulia Registry, where the maximum duration of illness observed was 11 years, the results of prevalence at 2, 5, and 10 years from diagnosis were observed, while those at 15 and 20 years were estimated by applying completeness indices to the prevalence observed at 11 years from diagnosis.*

Cancer Registry	Period of diagnosis considered in the study	Maximum duration of observed prevalence	Date of last known vital status	*Total number of incident cases	Population covered 01.01.2006	Proportion censored before 01.01.2006
North-West						
Biella (province)	1995-2005	11	2008	14 030	187 584	0.1%
Genova (municipality)	1986-2003	18	2006	79 452	612 700	0.0%
Milano (municipality)	1999-2005	7	2009	63 062	1 304 087	3.0%
Sondrio (province)	1998-2005	8	2007	8 651	179 428	0.3%
Torino (municipality)	1985-2005	21	2006	107 543	900 742	6.1%
Varese (province)	1980-2003	24	2007	93 511	843 250	1.2%
North-East						
Alto Adige (province of Bolzano)	1995-2005	11	2009	25 593	482 650	0.5%
Ferrara (province)	1991-2005	15	2007	37 342	352 384	1.0%
Friuli Venezia Giulia (region)	1995-2005	11	2007	92 399	1 208 278	0.4%
Modena (province)	1988-2005	18	2007	64 041	665 320	0.7%
Parma (province)	1978-2005	28	2007	66 116	418 444	0.9%
Reggio Emilia (province)	1996-2005	10	2008	27 743	497 920	0.6%
Romagna (provinces of Ravenna, Forlì-Cesena)	1991-2005	15	2006	70 685	743 282	0.0%
Trento (province)	1995-2004	10	2006	24 355	502 478	0.5%
Veneto (13 Local Health Units)	1990-2004	15	2007	143 041	1 790 294	0.3%
Centre						
Firenze Prato (province)	1985-2005	21	2007	143 359	1 190 515	0.4%
Latina (province)	1990-2005	16	2008	28 066	524 533	3.7%
Umbria (region)	1994-2005	12	2007	60 404	867 878	0.2%
South and Islands						
Napoli (local health unit)	1996-2005	10	2008	18 180	558 348	2.1%
Palermo (female breast only, province)	1999-2005	7	2007	4 459	642 294	6.4%
Ragusa (province)	1981-2005	25	2006	24 382	307 422	0.5%
Salerno (province)	1996-2005	10	2007	42 338	1 090 336	3.0%
Sassari (province)	1992-2004	13	2006	25 392	467 747	0.0%
Siracusa (province)	1999-2005	7	2008	11 209	398 178	0.0%
POOL				**1 275 353**	**16 093 795**	**1.3%**

* All sites except non-melanoma skin cancer

Table 1. Italian Cancer Registries included in the study. Period of diagnosis and maximum observation period, date of last follow up, total number of incident cases, population covered and proportion of incident cases censored alive before 01.01.2006

Le operazioni di calcolo specifiche per singolo RT, sede tumorale, sesso, ed età dei pazienti si possono schematizzare in 5 distinte fasi:

1 Conteggio del numero di casi incidenti osservati in vita alla data indice.

2 Stima del numero di casi osservati persi al follow-up che ci si attende siano ancora in vita alla data indice.

3 Stima degli indici di completezza, con cui si ricostruisce il numero atteso di casi prevalenti diagnosticati prima dell'inizio dell'attività di registrazione. Tali indici sono specifici per sede, sesso, età e lunghezza di registrazione e non per registro.

4 Applicazione degli indici di completezza alla prevalenza osservata calcolata in 1 e 2 per stimare la prevalenza completa e a durata limitata.

5 Stima della prevalenza al 1° gennaio 2006 per i 5 RT con ultimo anno di incidenza antecedente il 2005, attraverso una proiezione lineare.

The number of prevalent cases was calculated by cancer site, cancer registry area, gender, age at the prevalence date, and time from diagnosis in five different phases.

1 The number of incident cases observed by the registry and still alive at the index date was calculated by the counting method.

2 The number of cases observed by the registry, subsequently lost to follow-up and expected to be alive at the same date, was estimated.

3 The parameters of the completeness index were estimated in order to provide the expected number of prevalent cases diagnosed before the start of registration activity in each registry.

4 Complete prevalence and limited duration prevalence for periods longer than the registry's activity period were calculated by the completeness index.

5 Prevalence at the common index date of 1 January 2006 was then estimated by linear projection for the five registries with data ending prior to this date.

Le fasi 1 e 2 sono state effettuate attraverso il software SEER*Stat,[3] le fasi 3 e 5 utilizzando il pacchetto statistico SAS e la fase 4 con il software COMPREV.[4] Di seguito sono forniti maggiori dettagli sulle procedure usate in ciascuna fase.

Fase 1. Conteggio dei casi prevalenti osservati

La prevalenza osservata per singolo RT è stata calcolata con il metodo del conteggio diretto. Tale metodo consiste nell'enumerare i casi inclusi nel registro che risultano in vita alla data indice, considerando però il loro invecchiamento e ottenendo quindi anche risultati in funzione della ***età raggiunta*** alla data indice. Per ogni sede tumorale e RT è stata quindi calcolata la prevalenza osservata disaggregata per sesso e classi quinquennali d'età alla data indice. La prevalenza osservata è stata calcolata in anni di durata della malattia fino al numero massimo di anni di osservazione di ciascun registro.
I soggetti con diagnosi di più tumori primari (***tumori multipli***) sono stati conteggiati una sola volta nella prevalenza per il *complesso di tutte le neoplasie*. Invece nella prevalenza specifica per sede tumorale i soggetti con tumori multipli sono conteggiati in ciascuna delle sedi che sono state loro diagnosticate.
La data indice corrisponde al 1° gennaio 2006 per tutti i registri, tranne i 5 RT con dati di incidenza al 2003 o 2004 le cui date indice sono rispettivamente il 1° gennaio 2004 o il 1° gennaio 2005.

Fase 2. Prevalenza dei casi persi al follow-up

Il numero di casi incidenti vivi alla data di riferimento è una quantità direttamente osservabile. Tuttavia, una frazione di casi può essere persa al follow-up o censurata prima della data indice (i cosiddetti casi persi). Per ciascuno di questi individui, assumendo che siano vivi alla data della censura, è possibile stimare lo stato in vita alla data indice sulla base della sopravvivenza attesa nella coorte di pazienti con follow-up completo e con determinanti della prognosi simili al soggetto perso (per esempio sesso ed età). Il software SEER*Stat permette di scegliere le *variabili di appaiamento* rilevanti per costruire le tavole di sopravvivenza per il recupero dei casi persi. Nello studio si sono utilizzate tavole di sopravvivenza specifiche per ***sesso***, ***età alla diagnosi*** (3 classi: 0-64, 65-74 e 75+), ***sede tumorale*** e ***periodo di diagnosi*** (6 categorie: 1978-1979, 1980-1984, 1985-1989, 1990-1994, 1995-1999, 2000-2005).
La prevalenza osservata è stata quindi calcolata sommando il numero di casi persi, *stimati* vivi, alla data indice, al numero di casi prevalenti effettivamente osservati vivi alla data indice. Nel nostro studio la percentuale di casi persi al follow-up (tabella 1) è in media dell'1,3% e solo per 5 RT eccede il 3%. La maggior parte di questi casi (59%) è stata persa nei due anni antecedenti la fine del periodo di follow-up, cioè tra il 2004 e il 2005. Di conseguenza, lo stato in vita è stato estrapolato per un periodo relativamente breve e l'indicatore della prevalenza osservata non risulta particolarmente sensibile alla scelta delle variabili di appaiamento delle tavole di sopravvivenza. Per il

*Steps 1 and 2 were carried out with the SEER*Stat software,[3] steps 3 and 5 with the SAS statistical package and step 4 with the COMPREV software.[4]*
More details on procedures used for each step are given below.

Step 1. The counting method

The counting method was used to estimate observed prevalence from incidence and follow-up registry data. It consists in simply enumerating patients with data included in the registry file, those known to be alive at the index date, while still considering their ageing and so also obtaining results in relation to the ***age reached*** *by the index date.*
*Persons with more than one primary tumour (**multiple tumours**) were counted in each of the cancer specific prevalence estimates but included just once in all malignant cancer estimates.*
For each cancer and registry, observed prevalence was obtained by gender, five-year age class, and time elapsed since diagnosis. The prevalence index date is 1 January 2006 or the end of the most recent year of incidence and follow-up data. Prevalence duration was set to the maximum number of years of registration available at the index date.

Step 2. Prevalence of cases lost to follow-up

The number of incident cases alive at the index date is a directly observable quantity. However, a fraction of cases may be lost to follow-up or censored before the index date. For each of these individuals, assuming they were alive at the time of censor, the probability of being alive at the prevalence date was estimated from the life table of patient cohorts matched ***by gender, age class at diagnosis*** *(three levels: 0-64, 65-74, 75+),* ***cancer site, and period of diagnosis*** *(six levels: 1978-79, 1980-84, 1985-89, 1990-94, 1995-99, 2000-05). Observed prevalence was then computed by summing up lost cases estimated alive to the number of patients actually observed as alive on the index date.*
In our study, the percentage of cases lost to follow-up (Table 1) is on average 1.3%, and exceeds 3% only for five registries. The majority of these cases was lost in the last two years of the follow-up period (59%), i.e., between 2004 and 2005. Consequently, their vital status had to be extrapolated for a relatively short period and the observed prevalence indicator resulted scarcely sensitive to the choice of the matching criterion for constructing life tables. For instance, using the four matching variables for all cancers produces an increase of about 1.5% of lost cases estimated alive with respect to using a single life table for the entire cohort. In any case, the four matching variables guarantee an adequate approximation of life expectancy for all lost cases.

Step 3. Completeness Index Estimation

The fraction of non-observed prevalent cases was estimated using corrective factors, called ***completeness indices****, based on the application of incidence estimation models and relative survival, specific for tumour site, gender and age.*[5-7]
Complete prevalence at age x is composed of all incident cases

complesso di tutti i tumori includere sesso, età, periodo di diagnosi e sede tumorale tra le variabili di appaiamento dei persi, comporta un incremento del 1.5% del numero di casi persi stimati in vita al 1° gennaio 2006, rispetto a usare un'unica tavola di sopravvivenza per l'intera coorte. Ad ogni modo le quattro variabili utilizzate nello studio garantiscono un'adeguata approssimazione della speranza di vita di tutti i casi persi.

Fase 3. Stima degli indici di completezza

La frazione di casi prevalenti non osservati è stata stimata attraverso fattori correttivi, denominati ***indici di completezza***, che si basano sull'applicazione di modelli di stima di incidenza e sopravvivenza relativa, specifici per sede tumorale, sesso ed età.[5-7]

La prevalenza completa all'età x è costituita da tutti i casi incidenti diagnosticati a una certa età t (t<x) che sono sopravvissuti alla malattia fino a età x, ovvero che sono sopravvissuti per (x-t) anni. Un registro di popolazione attivo da L anni potrà osservare solo i casi con durata di malattia pari a x-L anni.

La prevalenza totale *attesa* N(*x*) si può dunque decomporre in due componenti: una ***osservata*** (durate tra 0 e x-L anni) e una ***non osservata*** (durate comprese tra x-L+1 e x anni), secondo la seguente relazione:

$$N(x) = N_L^{obs}(x) + N_L^{unobs}(x) = \sum_{t=x-L}^{x} I(t)S(t,\ x-t) + \sum_{t=0}^{x-L-1} I(t)S(t,\ x-t) \qquad (1)$$

dove I(t) è l'incidenza della malattia all'età t e S(t, x-t) è la sopravvivenza relativa all'età x per una persona a cui è stato diagnosticato un tumore all'età t. L'indice di completezza per uno specifico tumore è definito come il rapporto tra i valori attesi della prevalenza osservata $N^{obs}_L(x)$ per un determinato periodo di osservazione L, e la prevalenza completa attesa N(*x*), ovvero:

$$R_L(x) = \frac{N_L^{obs}(x)}{N(x)} = \frac{\sum_{t=x-L}^{x} I(t)S(t, x-t)}{\sum_{t=0}^{x} I(t)S(t, x-t)} \qquad (2)$$

Per stimare l'indice di completezza è necessario stimare a partire dai dati dei RT entrambe le funzioni di sopravvivenza e incidenza incluse nella relazione (2).

In Italia la stima disponibile degli indici di completezza della prevalenza risale all'inizio degli anni Novanta[1] e non riflette gli attuali andamenti né di incidenza né di sopravvivenza. Una stima aggiornata di tali indici è disponibile per i dati SEER statunitensi ed è incorporata nel software COMPREV. Il quadro epidemiologico dei tumori negli Stati Uniti presenta però differenze non trascurabili rispetto a quello italiano e nello studio si è optato per stimare sistematicamente gli indici di completezza per tutte le neoplasie con dati AIRTUM aggiornati. Gli indici sono stati stimati utilizzando i dati di incidenza e sopravvivenza dell'insieme dei nove registri ***(Pool trend)*** con il massimo periodo

*diagnosed at age t (t<x) surviving up to age x, thus for x-t years. A population registry active since L years can observe only prevalent cases with a disease duration lower than x-L. The expected complete prevalence N(x) can therefore be broken down into two components, one **observed** (durations from 0 to x-L) and one **unobserved** (durations from x-L and x), according to the following relationship:*

$$N(x) = N_L^{obs}(x) + N_L^{unobs}(x) = \sum_{t=x-L}^{x} I(t)S(t,\ x-t) + \sum_{t=0}^{x-L-1} I(t)S(t,\ x-t) \qquad (1)$$

where I(t) is the incidence of the disease at age t and S(t,x-t) is the relative survival at age x for a person who has been diagnosed with cancer at age t. The completeness index is defined as the ratio between the expected observed prevalence $N^{obs}_L(x)$ and the expected complete prevalence N(x) for a specific cancer and observation time L, that is:

$$R_L(x) = \frac{N_L^{obs}(x)}{N(x)} = \frac{\sum_{t=x-L}^{x} I(t)S(t, x-t)}{\sum_{t=0}^{x} I(t)S(t, x-t)} \qquad (2)$$

*Both survival and incidence functions in equation (2) must be estimated from available data of cancer registries. An updated estimate of completeness indices is not available in Italy because the last estimation is referred to 1992.[1] The COMPREV software includes indices computed from SEER data, but US cancer profiles are different from Italian patterns, both for incidence and survival. For these reasons, a systematic estimation was performed of the completeness indices by cancer site using the most current AIRTUM data available. The pool of nine cancer registries with longer common registration periods was used for this purpose. The registries are: Florence, Genoa, Modena, Parma, Ragusa, Romagna, Torino, Varese and Veneto (**Pool trend**), all active in the second half of the 1980s.*

Modelling the incidence function

*The incidence function describes the relationship between age and risk of developing and being diagnosed of cancer, as measured along the life span of each birth cohort present in the population at the prevalence date. In the study, two different models were tested: **exponential** and **polynomial**.*

The first models the logit of incidence with an exponential relationship to age and was developed according to the multistage theory of carcinogenesis:[8]

$$I(x,k) = [1 + \exp-(a_k + b\ \log(x))]^{-1} \qquad (3)$$

where I(x,k) is incidence at age x for birth cohort k.

*In the second model, in accordance with previous studies,[9] a sixth degree polynomial on age was used **for each site**:*

di registrazione comune, ovvero i registri di: Firenze, Genova, Modena, Parma, Ragusa, Romagna, Torino, Varese e Veneto tutti attivi nella seconda metà degli anni Ottanta.

Modello per la funzione di incidenza

La funzione di incidenza descrive la relazione tra età e rischio di sviluppare una diagnosi di tumore, misurata lungo la vita di ogni coorte di nascita presente nella popolazione alla data indice. Nello studio sono stati testati due diversi modelli: un modello di tipo **esponenziale** e uno **polinomiale**.

Il primo prevede una relazione di tipo esponenziale tra incidenza ed età, così come proposto dalla teoria multistadio della carcinogenesi,[8] secondo la relazione:

$$I(x,k) = [1 + \exp - (a_k + b \log(x))]^{-1} \quad (3)$$

dove I(x,k) è l'incidenza all'età alla diagnosi x per la coorte di nascita k.

Il secondo modello prevede invece, in accordo con studi precedenti,[9] che l'incidenza vari con l'età secondo una funzione logistica, avente per argomento un polinomio di sesto grado nell'età alla diagnosi:

$$I(x,k) = \left\{1 + \exp - \left[a_k + \sum_{i=1}^{6} b_i \cdot \left(\frac{x - x_0}{m}\right)^i\right]\right\}^{-1} \quad (4)$$

dove la coorte di nascita k, è stata inserita come variabile categorica attraverso il parametro a_k, per aggiustare i trend del rischio di ammalarsi tra le differenti coorti di nascita.

Per tutte le sedi tumorali i esaminate, il modello polinomiale (4), più flessibile del modello esponenziale, ha mostrato il miglior adattamento ai dati di incidenza osservati, in particolare per classi di età più anziane dove si osserva quasi sempre solo un lieve aumento o perfino una diminuzione del rischio rispetto alle età anziane immediatamente precedenti. La bontà di adattamento ai dati per ogni sede, è stata valutata sia con il confronto grafico di tassi osservati e stimati, che con il criterio di informazione di Akaike (AIC).[10,11]

I parametri della funzione di incidenza sono stati stimati mediante la procedura logistica di SAS modellizzando i tassi di incidenza grezzi registrati nel periodo 1985-2004 dai 9 RT italiani del Pool trend. I dati di incidenza sono stati stratificati per sede tumorale, sesso, classe di età quinquennale (0-4, 5-9, ..., 80-84, 85 e più) e coorte di nascita quinquennale (<1899, 1900-1904, ..., 2000-2004). Due esempi di confronto tra tassi di incidenza età-specifici osservati e stimati con il modello polinomiale sono presentati nelle figure 1 e 2.

Modello per la funzione di sopravvivenza

La funzione di sopravvivenza relativa, per ciascuna sede e sesso, è stata parametrizzata attraverso un **modello misto** con ipotesi di **guarigione.** I modelli misti consentono di distinguere due

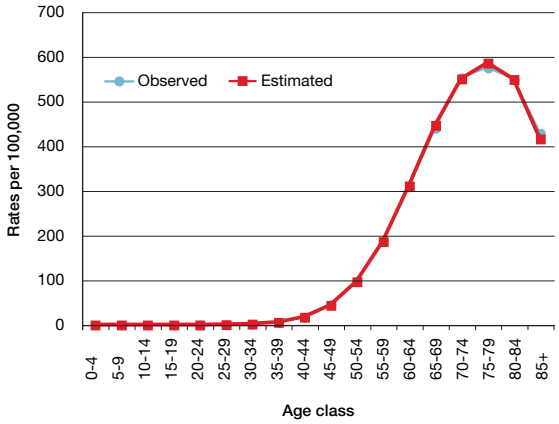

Figure 1. Incidence rates of lung cancer in men by age at diagnosis in the area covered by the 9 long-term CRs. Polynomial model estimates compared to observed rates.

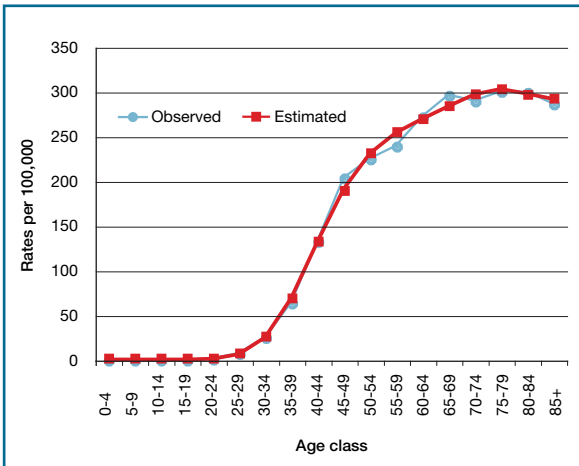

Figure 2. Incidence rates of breast cancer in women by age at diagnosis in the area covered by the 9 long-term CRs. Polynomial model estimates compared to observed rates.

$$I(x,k) = \left\{1 + \exp - \left[a_k + \sum_{i=1}^{6} b_i \cdot \left(\frac{x - x_0}{m}\right)^i\right]\right\}^{-1} \quad (4)$$

where the birth cohort covariate was also included, together with age, in the incidence function as a categorical variable to adjust for risk trends across the different birth cohorts.

However, CR data show that model (4) is more flexible and adequate than model (3) in estimating incidence rates in the oldest age group, in which rates are only slightly increasing or even decreasing with age for almost all cancer sites.

Parameters of the incidence function were estimated through the SAS logistic procedure by fitting raw incidence rates of patients registered between 1985 and 2004 by the nine Italian CRs in the Pool trend. Incidence data were categorized according to cancer site, gender, five-year age intervals (0-4, 5-9, …, 80-84, 85+), and birth cohort (<1899, 1900-1904, … , 2000-2004).

sottopopolazioni distinte di pazienti: quelli che muoiono per cause diverse dal tumore e quelli che muoiono per il tumore. I primi presentano la stessa aspettativa di vita della popolazione generale e si possono ritenere '*guariti*'. I secondi hanno un eccesso di rischio rispetto alla popolazione generale che li porterà a morire per tumore dopo un certo numero di anni dalla diagnosi e vengono definiti casi '*fatali*'. La probabilità cumulativa di sopravvivere al tumore (sopravvivenza relativa) fino all'età x per un paziente diagnosticato nell'anno y a età t, viene definita secondo la seguente relazione:

$$S(x, y, d) = \left[C + (1-C)\exp(-\lambda d)^{\gamma}\right]^{\exp[\beta_1(t-t_0)+\beta_2(y-y_0)]} \quad (5)$$

dove C è la proporzione di casi guariti; (1-C) è la proporzione di casi fatali per i quali si assume che la distribuzione del tempo alla morte segua una distribuzione di ***tipo Weibull***, dove d è il tempo di follow-up (d=x-t), λ, e γ sono rispettivamente i parametri di scala e forma; β_1 e β_2 esprimono rispettivamente l'effetto dell'età alla diagnosi e dell'anno di diagnosi nella funzione di sopravvivenza relativa di tutti i pazienti. Il valore t_0 rappresenta l'età alla diagnosi presa come riferimento (65 anni), mentre y_0 è l'anno di diagnosi di riferimento (1992).
Per alcune sedi tumorali (tumore della tiroide, linfoma di Hodgkin, mieloma multiplo, leucemia, leucemia linfatica acuta/cronica, leucemia mieloide acuta/cronica), si è ottenuto un migliore adattamento ai dati osservati, usando una distribuzione per il tempo alla morte dei casi fatali di ***tipo esponenziale*** piuttosto che Weibull, e stimando i parametri del modello specifici per età, oltre che per sede e sesso.
I parametri della funzione di sopravvivenza relativa sono stati stimati utilizzando i dati di sopravvivenza relativa estratti dai nove RT italiani (Pool trend) relativi a pazienti diagnosticati tra il 1985 e il 2002. Per ogni sede, la sopravvivenza relativa è stata stratificata in base a sesso, periodo di diagnosi (1985-1987, 1988-1990, 1991-1993, 1994-1996, 1997-1999, 2000-2002) e classe di età.
Per la maggior parte delle sedi, le classi di età alla diagnosi sono state definite dagli intervalli 15-44, 45-54, 55-64, 65-74 e 75 e più anni. Fanno eccezione i tumori del cervello e sistema nervoso centrale, il linfoma di Hodgkin, il linfoma non-Hodgkin, le leucemie, e il gruppo «tutte le sedi», per i quali è stata inclusa anche la classe 0-14 anni. Mentre per la leucemia linfatica cronica e la leucemia mieloide cronica sono state definite solo le classi d'età 15-44 e 45 e più.
Sono stati esclusi dall'analisi i casi diagnosticati incidentalmente solo all'autopsia, attraverso i certificati di morte e i casi con follow-up passivo. È stato preso in esame un periodo di follow-up di 21 anni, con intervalli semestrali. I parametri C, λ, γ, β_1 e β_2 sono stati stimati utilizzando la procedura NLIN presente in SAS (stime non mostrate). La bontà di adattamento del modello di sopravvivenza è stata valutata mediante confronto dei tassi di sopravvivenza stimati e osservati (esempi nelle figure 3 e 4).

The goodness of fit of the various incidence models was assessed by Akaike Information Criterion (AIC) [10,11] *as well as by visual comparison between estimated and observed rates. Two examples of comparison between observed and modelled age-specific incidence rates are shown in Figures 1 and 2.*

Modelling the survival function

The relative survival function for each site and gender was parameterized by means of a ***mixture 'cure-model'****. This model assumes that a fraction of cancer patients – designated as "cured" (C) – should be exposed to the same mortality rates as the general population; whereas, the remaining fraction (1-C) – the "fatal cases" – will die of the disease with a cumulative relative survival following a Weibull distribution. So the cumulative relative survival probability up to age x of a patient diagnosed at age t and year y is assumed to be:*

$$S(x, y, d) = \left[C + (1-C)\exp(-\lambda d)^{\gamma}\right]^{\exp[\beta_1(t-t_0)+\beta_2(y-y_0)]} \quad (5)$$

where d is follow-up time d=x-t , λ and γ are, respectively, the scale and shape parameters of the ***Weibull*** *distribution used to evaluate the specific differential mortality of fatal cases; β_1 and β_2 express the effects of 'age at diagnosis' and 'year of diagnosis' respectively, on the relative survival function of all patients. The value t_0 is a reference age at diagnosis (65 years) and y_0 is a reference year of diagnosis (1992).*
For some cancer sites (thyroid, Hodgkin's lymphoma, multiple myeloma, leukaemia, acute/chronic lymphatic leukaemia, acute/chronic myeloid leukaemia), a better fit was obtained by using an ***exponential*** *rather than Weibull distribution for time-to-cancer death of fatal cases, removing the age effect, and estimating age-specific model parameters.*
The models' parameters were estimated by fitting relative survival rates of patients diagnosed between 1985 and 2002 reported by the nine Italian registries of the Pool trend. Relative survival for any specific site was organized according to gender, period of diagnosis (1985-1987, 1988-1990, 1991-1993, 1994-1996, 1997-1999, 2000-2002), and age class. For most sites, age classes were defined by intervals 15-44, 45-54, 55-64, 65-74, and 75+. Exceptions were made for brain and other central nervous system cancers, Hodgkin's lymphoma, non-Hodgkin lymphoma, leukaemia, and all sites, which also included the 0-14-year class, and chronic lymphatic leukaemia and chronic myeloid leukaemia which considered only two age groups, 15-44 and 45+.
Death certificates only, cases incidentally diagnosed at autopsy only, and cases not actively followed-up, were excluded from the analysis. A 21-year follow-up period was considered with 6-month intervals. The parameters C, λ, γ, β_1 and β_2 were then estimated using the SAS NLIN procedure (estimates not shown). The goodness of fit of the modelled survival function was evaluated by visual comparison of estimated and observed survival rates (Figures 3 and 4).

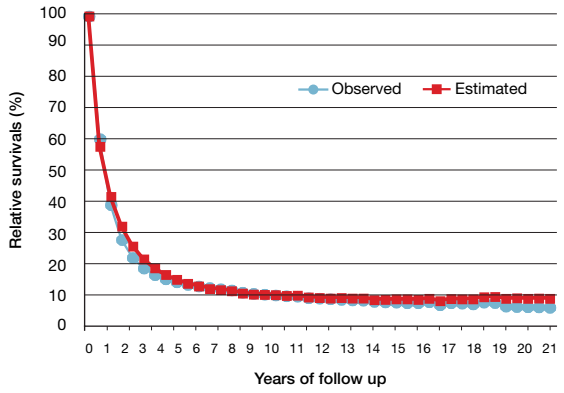

Figure 3. Relative survival model of lung cancer in men by follow up time (years) in the area covered by the 9 long-term CRs. Mixture cure model estimates compared to observations.

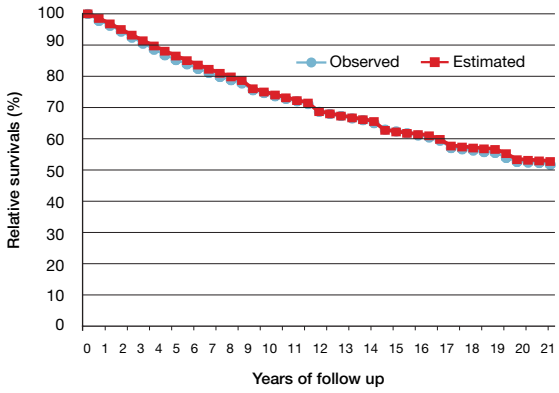

Figure 4. Relative survival of breast cancer in women by follow up time (years) in the area covered by the 9 long-term CRs. Mixture cure model estimates compared to observations.

Calcolo dell'indice di completezza

Il vettore dei parametri stimati nei modelli di incidenza (intercetta e 6 coefficienti del polinomio di età del modello polinomiale) e sopravvivenza (parametri 1-C, λ, γ, β_1 e β_2, anno ed età alla diagnosi di riferimento), così come i corrispondenti elementi della matrice di covarianza stimata, hanno fornito le informazioni necessarie per la stima degli indici di completezza.

Per calcolare gli indici si è utilizzato il software COMPREV, che in corrispondenza di parametri di incidenza e sopravvivenza forniti come dati di ingresso, calcola in uscita gli indici di completezza della prevalenza a data indice e lunghezza di registrazione fissate dall'utente. Per ogni sede e sesso è stato calcolato l'indice di completezza $\mathbf{R_L(x)}$ specifico per classe d'età quinquennale raggiunta (0-4, 5-9, ..., 85 e più anni), corrispondente a ogni possibile lunghezza del periodo di osservazione nei 24 RT dello studio (da 7 fino a 28 anni). La data indice è stata fissata al 1° gennaio 2006 per ottenere indici di completezza validi per 19 dei 24 RT esaminati. Per i restanti 5 RT sono stati calcolati indici per la correzione della prevalenza al 1° gennaio 2004 o 2005.

L'indice di completezza rappresenta la percentuale di completezza della prevalenza di durata limitata e varia tra lo 0 e il 100%, a seconda dell'aggiustamento introdotto nella prevalenza osservata in ciascun registro. Valori vicini a 0 indicano un basso livello di completezza, e quindi un'elevata correzione applicata alla prevalenza osservata. L'indice cresce all'aumentare della lunghezza di osservazione, al diminuire dell'età alla prevalenza e al diminuire della sopravvivenza, poiché una prognosi sfavorevole deprime il numero di sopravvissuti nel lungo periodo. La tabella 2 mostra i valori dell'indice di completezza calcolato per tutti i tumori al variare di sesso, classi di età e durata di osservazione. La figura 5 mostra per esempio i valori dell'indice per il tumore colorettale. Gli indici di completezza per ciascuna sede, sesso, classe di età e lunghezza del periodo di osservazione sono disponibili sul sito web dell'AIRTUM.

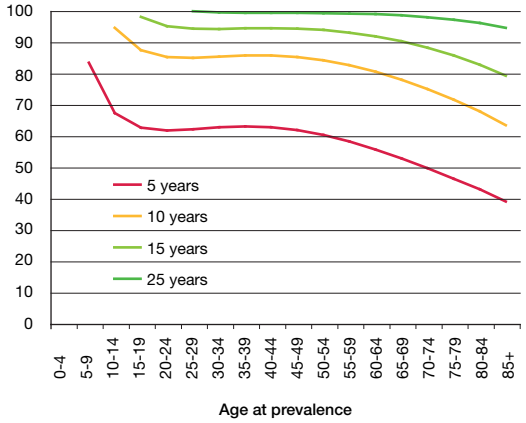

Figure 5. Completeness index (%) at 01.01.2006 for colorectal cancer in men by age at prevalence and length of the registration period (5,10,15,25 years).

Calculation of the completeness index

The vector of estimated incidence (intercept and six coefficients for the age polynomial model) and survival model parameters (1-C, λ, γ, β_1 and β_2, reference period and age), as well as the corresponding elements of the estimated covariance matrix, provided the information needed for the estimation of completeness index. The COMPREV software was used to calculate, for each site and gender, the set of age specific estimates (age groups 0-4, 5-9, …, 85+) of the completeness indices $R_L(x)$ corresponding to the observation length L of each registry (from 7 to 28 years). The index date was set at 1 January 2006 for 19 out of the 24 CRs and at January 2004 or 2005 for the remaining 5 registries.

The completeness index represents the percent of completeness of limited-duration prevalence. The index varies between 0 and 1, depending on the adjustment introduced in the prevalence observed by the registry. Values close to 0 indicate a low level

Fase 4. Prevalenza completa e a durata limitata

La prevalenza completa per ciascun registro, sede, sesso e classe di età è stata ottenuta dividendo la prevalenza osservata massima del registro per il corrispondente indice di completezza secondo la relazione (2):

$$N(x) = \frac{N_L^{obs}(x)}{R_L(x)}$$

dove L è la massima lunghezza di osservazione e x la classe di età. La prevalenza a durata limitata a 2,5,10,15, e 20 anni dalla diagnosi corrisponde esattamente alla prevalenza osservata per durate inferiori al periodo massimo di osservazione del registro, mentre è stata stimata attraverso gli indici di completezza per durate superiori.

Se L è il periodo massimo di osservazione del registro (per esempio 11 anni) la prevalenza per una durata limitata L* più estesa (per esempio 15 o 20 anni) si può ottenere invertendo i termini della relazione (2):

$$N_{L^*}(x) = N(x)R_{L^*}(x) = N_L^{obs}(x)\frac{R_{L^*}(x)}{R_L(x)} \qquad (6)$$

In questo modo sono state ottenute per tutti i registri stime coerenti della prevalenza completa e della prevalenza a durata limitata, indipendentemente dalla lunghezza del periodo di osservazione di ciascun registro.

Gli intervalli di confidenza (IC) al 95% per la prevalenza sono stati ottenuti nell'ipotesi che i casi prevalenti (N) seguano una distribuzione di Poisson. I corrispondenti intervalli per le proporzioni di prevalenza (P=N/pop dove pop indica la popolazione al 1° gennaio 2006) sono derivati di conseguenza:

of completeness, and therefore a large correction to be applied to the observed prevalence. The completeness index increases with increasing length of follow-up, for younger ages, and lower survival rates, since a worse prognosis leads to a restricted number of long-term survivors. Table 2 shows completeness index values calculated for all tumours by gender, age class and length of observation. For example, Figure 5 shows, index values for colorectal cancer. Completeness indices for each site, gender, age class and length of observation period are available at the AIRTUM web-site.

Step 4. Complete and limited duration prevalence

Observed prevalence, by registry, site, gender and age class, was then divided by the corresponding completeness index to obtain the complete prevalence according the relation (2):

$$N(x) = \frac{N_L^{obs}(x)}{R_L(x)}$$

where L is the maximum observation length of the registry and x is age class.

Limited-duration prevalence at 2, 5, 10, 15 and 20 years corresponds exactly to observed prevalence for durations smaller than the maximum observation period, for longer durations it is derived from completeness indices; L being the maximum observation length of a registry (for example 11 years), the prevalence for a higher limited duration L (for example 15 or 20 years), can be derived by inverting the terms in equation (2):*

$$N_{L^*}(x) = N(x)R_{L^*}(x) = N_L^{obs}(x)\frac{R_{L^*}(x)}{R_L(x)} \qquad (6)$$

Age class	Male					Female				
	L=5	L=10	L=15	L=20	L=25	L=5	L=10	L=15	L=20	L=25
0-4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
05-09	0.733	1.000	1.000	1.000	1.000	0.627	1.000	1.000	1.000	1.000
10-14	0.470	0.856	1.000	1.000	1.000	0.357	0.759	1.000	1.000	1.000
15-19	0.381	0.667	0.909	1.000	1.000	0.326	0.564	0.836	1.000	1.000
20-24	0.352	0.594	0.781	0.940	1.000	0.364	0.569	0.720	0.895	1.000
25-29	0.353	0.575	0.734	0.857	0.961	0.424	0.630	0.748	0.837	0.939
30-34	0.374	0.590	0.731	0.833	0.911	0.472	0.692	0.802	0.865	0.912
35-39	0.408	0.624	0.755	0.840	0.901	0.494	0.729	0.841	0.898	0.930
40-44	0.448	0.668	0.791	0.865	0.913	0.491	0.738	0.859	0.917	0.947
45-49	0.487	0.712	0.828	0.893	0.932	0.470	0.725	0.858	0.923	0.955
50-54	0.517	0.746	0.859	0.917	0.949	0.440	0.697	0.842	0.918	0.956
55-59	0.534	0.769	0.880	0.934	0.962	0.407	0.660	0.815	0.903	0.950
60-64	0.537	0.778	0.890	0.944	0.970	0.376	0.621	0.782	0.881	0.937
65-69	0.526	0.773	0.892	0.948	0.974	0.352	0.586	0.747	0.853	0.920
70-74	0.501	0.755	0.883	0.945	0.974	0.336	0.558	0.715	0.825	0.898
75-79	0.468	0.725	0.865	0.937	0.971	0.327	0.540	0.691	0.800	0.877
80-84	0.430	0.685	0.838	0.922	0.965	0.322	0.529	0.674	0.780	0.857
85+	0.396	0.643	0.804	0.901	0.954	0.314	0.518	0.661	0.764	0.841

Table 2. Completeness index R by sex, age, and length of observation period (L).
All sites except non-melanoma skin cancer (ICD-10 C00-43, C45-96, D09.0, D30.3, D41.4)

$$\text{IC 95\% di } P = P \pm 1.96\sqrt{\frac{P(1-P)}{pop}} \quad (7)$$

La prevalenza completa e a durata limitata per il pool dei registri AIRTUM e per le macro-aree Nord-ovest, Nord-est, Centro e Sud sono stati ottenuti sommando le stime specifiche per registro dei casi prevalenti e rapportando alla popolazione corrispondente per ottenere le proporzioni di prevalenza.
La figura 6 mostra il diagramma di flusso delle fasi di calcolo della prevalenza sin qui descritte.

Fase 5. Proiezioni

Per i 5 RT con dati di incidenza anteriori al 2005, gli indicatori di prevalenza sono stati estrapolati al 1° gennaio 2006 con proiezioni a breve termine in modo da allineare l'informazione a quella disponibile per la maggior parte dei registri. La proporzione di casi prevalenti è stata proiettata alla comune data indice del 1° gennaio 2006 attraverso un modello di regressione lineare rispetto all'anno di calendario. Per limitare lo scostamento dalla linearità, la base di proiezione è stata limitata all'ultimo triennio a disposizione. Si sono quindi utilizzati i trend delle stime di prevalenza 01.01.2002 - 01.01.2004 per i registri di Genova e Varese e 01.01.2003 - 01.01.2005 per i registri di Sassari,Trento e Veneto. La proiezione, di due anni per i primi e di un solo anno per i secondi, è stata effettuata separatamente per ciascuno dei 5 registri oltreché per sede tumorale, sesso, classe di età e durata della malattia. Questa procedura è stata validata sui dati osservati del Registro di Firenze-Prato, simulando una interruzione dei dati di incidenza al 2003 e proiettando la prevalenza al 2006. Le differenze tra valori osservati e proiettati sono risultate mediamente abbastanza contenute. Per esempio per la prevalenza a 15 anni dalla diagnosi lo scostamento medio assoluto su tutte le combinazioni sesso, età e sede tumorale è di 10,6 casi per 100. 000, mentre lo scostamento percentuale medio è del 2,9% per tutti i tumori, del 2% per i tumori della mammella e della prostata.

Risultati

In tabella 3 è riportata la lista e la definizione delle 46 sedi tumorali per cui sono stati stimati i parametri di incidenza e di sopravvivenza, insieme al numero totale di casi prevalenti stimati nei diversi passi della procedura descritta precedentemente.
I risultati dettagliati per ciascuna sede considerata sono presentati e commentati nella sezione principale di questa monografia. In questo paragrafo sono fornite solo alcune considerazioni sull'impatto dei diversi passi della procedura di stima.
Complessivamente per tutti i tumori maligni (eccetto gli epiteliomi della cute) si sono considerati 1.275.353 casi incidenti osservati dai Registri tumori nel periodo 1978-2005.
La stima della prevalenza completa nel pool dei RT è di 672.875 casi, diagnosticati a qualsiasi data e vivi alla data di

In this way, consistent estimates of complete and 2, 5, 10, 15, and 20-year limited duration prevalence counts and proportions were obtained for all registries.
95% confidence intervals (CI) for the number of prevalent cases (N) were obtained under the hypothesis that prevalent cases follow a Poisson distribution. Corresponding approximate 95% CI for prevalence proportions P (P=N/pop where pop is population at 01.01.2006) were derived consequently as:

$$\text{95\% CI of } P = P \pm 1.96\sqrt{\frac{P(1-P)}{pop}} \quad (7)$$

Complete and limited duration prevalence for the pool of AIRTUM registries and all macro-areas were then calculated by summing up the relevant registry specific estimates.
Figure 6 shows the flow chart of the calculation steps from 1 to 4 described above.

Step 5. Projections

For the five CRs with missing incidence data in the most recent years, prevalence indicators were estimated at the beginning of 2005 or 2004, as reported in the previous paragraph.
In order to have a common prevalence date for all the considered populations, short-term projections were carried out for the data from these registries. For each registry, cancer site, gender, age class, and time from diagnosis, the number of prevalent cases was projected to the common date of 1 January 2006 by means of a linear regression model with the calendar year as an independent variable. To limit department from linearity, only the three most recent available years were used as basis for projections: 2002-2004 for Genoa and Varese, and 2003-2005 for Sassari, Trento and Veneto. This projection procedure was validated with data from the Firenze-Prato registry by simulating incidence data truncation at 2003, and comparing projected and observed prevalence on 2006. The differences were on average quite limited. For instance, for prevalence at 15 years from diagnosis the average absolute difference on all age, gender, and site combinations was 10.6 cases per 100,000. The average percentage deviation is 2.9% for all tumours, 2% for breast and prostate tumours.

Results

Complete prevalence was computed for 46 cancer sites. For each of them, Table 3 reports the number of prevalent cases estimated by the different steps in the above-described procedures. The detailed results are presented in the main section of this monograph. In this paragraph, only some consideration is given about the relative impact of the different steps in the estimation procedure.
For all cancers combined, we considered the data collected from 1,275,353 patients diagnosed in the period 1978-2005. Complete prevalence was estimated at 672,875 cases ever diagnosed of cancer and alive at the index date. Of these, 481,690 (72% of the complete prevalence) were known to be alive at the index

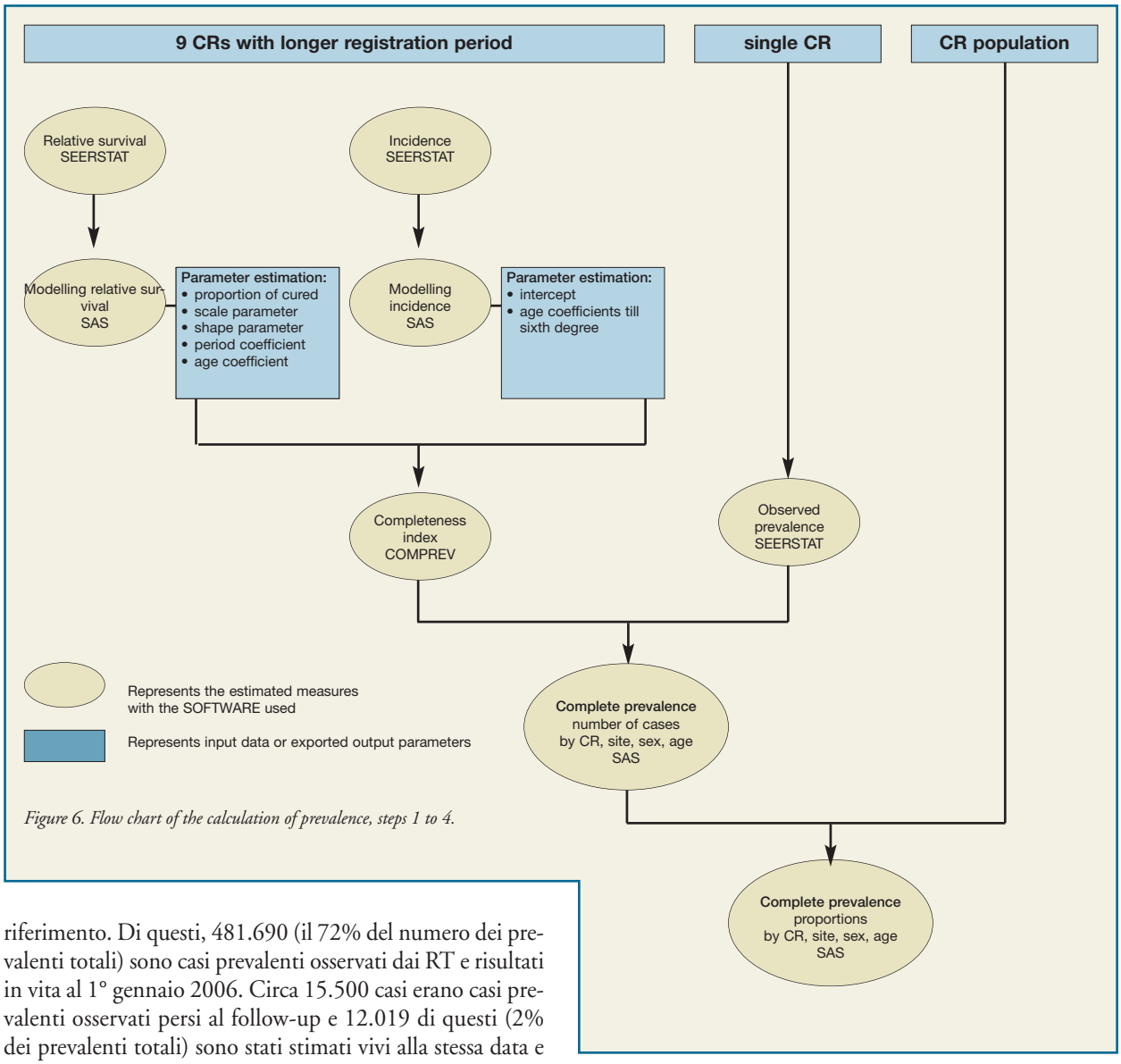

Figure 6. Flow chart of the calculation of prevalence, steps 1 to 4.

riferimento. Di questi, 481.690 (il 72% del numero dei prevalenti totali) sono casi prevalenti osservati dai RT e risultati in vita al 1° gennaio 2006. Circa 15.500 casi erano casi prevalenti osservati persi al follow-up e 12.019 di questi (2% dei prevalenti totali) sono stati stimati vivi alla stessa data e aggiunti alla prevalenza completa.
Per ultimi, si sono stimati 179.166 casi prevalenti (27% dei prevalenti totali) non osservati dai RT partecipanti perché diagnosticati prima dell'inizio dell'attività di registrazione.
Il contributo dei casi persi al follow-up stimati vivi è piccolo e simile per tutte le sedi (varia tra lo 0 e il 3%), dato che dipende essenzialmente dalle procedure di follow-up dei RT.
Al contrario, la proporzione di casi prevalenti, non registrati ma stimati attraverso l'applicazione di modelli statistici e dell'indice di completezza è risultata molto diversa tra le sedi: per i tumori con alta sopravvivenza e/o incidenza nelle più giovani età, come cervice uterina (57%), ossa (60%), testicolo (50%), vagina e vulva (84%) la proporzione è risultata elevata; per le sedi a povera prognosi come pancreas (15%), fegato (15%), polmone (20%) la proporzione di casi prevalenti non registrati è risultata invece piuttosto bassa.

date 1 January 2006. About 15,500 cases were lost to follow-up, and 12,019 of them (2%) were estimated to be alive at the same date and added to the complete prevalence.
Finally, 179,166 prevalent cases (27%) were estimated, not observed by the participating registries because they were diagnosed before the start of registration. The contribution of cases lost to follow-up and estimated alive is small and similar for all the considered sites, since it depends essentially on the registries' follow-up procedures. On the contrary, the proportion of unobserved prevalence estimated by modeling and the completeness index was very different across cancer sites. It was highest for cancers with high survival and/or youngest ages at incidence, such as cervix uteri (57%), bone (60%), testis (50%), and vagina and vulva (84%), but low for poor prognosis cancers, such as pancreas (15%), liver (15%), and lung (20%).

Validazione degli indici di completezza

Per validare gli indici di completezza della prevalenza sono stati utilizzati i dati dei tre registri con periodo di osservazione più lungo: Parma (28 anni), Ragusa (25 anni) e Varese (24 anni).
Per ciascuno dei 3 RT è stata calcolata la prevalenza osservata alla durata massima (28, 25 e 24 anni rispettivamente) e la corrispondente stimata applicando gli indici di completezza

Validation of completeness indices

In order to validate the completeness indices, we used observed prevalence for the long-established registries of Parma (28 years), Ragusa (25 years), and Varese (24 years).
For each of these three registries, we estimated limited-duration prevalence for the maximum observation period, based on artificially truncated series of data and the corresponding com-

ICD-10 Site	Incident cases 1978-2005	Cases alive 01.01.2006	Lost cases estimated alive 01.01.2006	Estimated total prevalent cases 01.01.2006
C00-14, C30-32 Head and neck	51 971	20 718	480	32 199
C01-02 Tongue	5 723	2 025	56	2 780
C03-06 Mouth	6 829	2 338	56	3 183
C07-08 Salivary glands	2 724	1 245	58	3 279
C09-10 Oropharynx	4 226	1 326	41	1 781
C11 Nasopharynx	1 875	720	28	1 250
C12-13 Hypopharinx	2 997	674	17	886
C15 Oesophagus	10 826	1 054	17	1 324
C16 Stomach	77 450	14 942	278	21 873
C17 Small intestine	3 486	1 192	24	1 469
C18-C21 Colon-rectum	167 581	68 579	1 441	90 666
C18 Colon	113 990	47 613	1 015	62 746
C19-21 Rectum	53 591	21 395	434	28 432
C22 Liver	40 313	4 962	70	5 896
C23-24 Gallbladder	18 091	1 921	38	2 527
C25 Pancreas	36 745	2 412	35	2 893
C30-31 Nasal cavities	1 851	699	11	939
C32 Larynx	21 402	9 970	187	16 096
C33-34 Trachea, bronchus and lung	156 168	17 495	308	22 222
C40-41 Bone	2 585	1 206	37	3 100
C43 Skin melanoma	26 454	18 264	504	25 611
C45 Mesothelioma	4 504	495	9	601
C46 Kaposi's sarcoma	2 851	1 353	127	1 919
C47,49 Connective and soft tissue	6 841	3 384	87	6 347
C50 Female breast	171 003	114 285	2 647	166 907
C51-52 Vulva and vagina	5 208	1 977	39	12 932
C53 Cervix uteri	12 088	6 445	232	15 614
C54 Corpus uteri	27 946	17 602	456	27 128
C56 Ovary	20 590	6 976	180	10 964
C60 Penis	1 410	706	12	1 048
C61 Prostate	106 803	60 876	1 165	69 897
C62 Testis	5 682	4 896	198	10 252
C64-66,68 Kidney and other urinary organs	39 590	19 242	356	26 693
C67, D09.0, D30.3, D41.4 Bladder	88 817	44 595	1 841	62 866
C69 Choroidal melanoma	1 292	678	16	944
C70-72 Brain and central nervous system	20 231	3 645	111	8 062
C73 Thyroid	20 205	16 534	415	22 968
C81 Hodgkin lymphoma	7 793	5 458	182	11 696
C82-85,96 Non-Hodgkin lymphoma	44 303	19 774	520	28 438
C88-90 Multiple myeloma	18 371	5 290	100	6 333
C91-95 Leukaemias	32 460	10 016	257	14 063
C91.0 Acute lymphoid leukaemia	3 120	1 401	51	3 461
C91.1 Chronic lymphoid leukaemia	10 981	4 631	93	5 922
C92.0 Acute myeloid leukaemia	7 997	1 211	47	1 645
C92.1 Chronic myeloid leukaemia	4 706	1 352	36	1 661
All sites but non-melanoma skin cancer	1 275 353	481 690	12 019	672 875

Table 3. Total number of incident cases, number of cases known to be alive, number of lost case estimated alive and estimated total number of prevalent cases (complete prevalence) by carcer sites, 01.01.2006 in cancer registry areas.

alla prevalenza osservata troncata fittiziamente a 10, 15 e 20 anni. Per esempio per il RT di Parma è stata calcolata la prevalenza stimata a 28 anni a partire dalla prevalenza osservata a 10, 15, e 20 anni tramite la relazione (6).
La prevalenza stimata è poi stata confrontata con quella osservata. Sono state analizzate per uomini e donne separatamente soltanto le sedi con almeno 50 casi prevalenti osservati nel periodo di osservazione massimo.
Gli scostamenti percentuali relativi si riducono al crescere della lunghezza d'osservazione e sono quindi risultati massimi per le stime ricostruite a partire da soli 10 anni di osservazione. La media degli scostamenti in valore assoluto sulle sedi esaminate oscilla tra 4,6% (10 anni) e 1,5% (20 anni) per Parma, tra 4,4% e 1,4% per Varese e tra 7,3% e 2% per Ragusa. Per il complesso di tutti i tumori maligni gli scostamenti relativi assoluti sono al di sotto del 4% a 15 e 20 anni e al di sotto del 7% a 10 anni in tutti e tre i Registri.
Per il tumore della tiroide nelle donne si sono riscontrati sistematicamente nei tre RT scostamenti tra 10% e 15% a 10 anni. L'aumento considerevole di sopravvivenza e incidenza del tumore della tiroide negli ultimi decenni, con modulazioni diverse per area geografica, potrebbe essere all'origine del non perfetto adattamento ai dati osservati nei singoli registri. Scostamenti superiori al 10% (prevalentemente per il valore a 10 anni) sono stati riscontrati in altre sedi (linfoma di Hodgkin, encefalo, melanoma della cute, tessuti molli, leucemie, vescica e linfoma non Hodgkin) ma non in modo sistematico per i tre registri.

Stime nazionali e regionali della prevalenza

La stima del numero totale di casi prevalenti di tumore per sede e sesso in *Italia* è stata ottenuta come somma dei casi prevalenti in ciascuna delle 4 macro-aree italiane, stimati moltiplicando le proporzioni della prevalenza completa per sede, sesso, età, anni dalla diagnosi e macro-area per le corrispondenti popolazioni residenti nelle quattro macro-aree italiane al 1°gennaio 2006. La stima della proporzione di prevalenza in Italia così costituita differisce necessariamente dalla proporzione di prevalenza stimata per il Pool dei Registri AIRTUM, perchè i livelli di prevalenza e la copertura di registrazione sono piuttosto diversificati tra Sud e Centro-nord.
La stima del numero totale di soggetti prevalenti per sede e sesso per le regioni a copertura parziale di registrazione è stata ottenuta analogamente moltiplicando le proporzioni della prevalenza specifica per sede, sesso, età e macro-area pertinente alla regione interessata per la popolazione residente regionale al 01.01.2006 (tabella 2 p. 19).
Le stime nazionali e ragionali presuppongono che la popolazione delle aree coperte sia rappresentativa di quella residente nelle aree non coperte dai Registri tumori.
In *Appendice* (pp. 180-181) sono riportate le popolazioni per regione e macro-area al 1 gennaio ricavate dal sito di statistiche demografiche dell'ISTAT http://demo.istat.it/pop2006/index.html e usate nelle stime.

pleteness indices. For the Parma registry, for instance, we calculated the 28-year prevalence for each cancer site based on 10, 15, and 20 year observed prevalence and completeness index by the relation (6).
Estimated prevalence was then compared to that observed. Only cancers with at least 50 prevalent cases in the longest period were analysed.
The percentage relative differences decrease with increasing observation length, so they were highest for estimates derived from only ten years registration. The average of the absolute differences on all cancer sites analysed ranged between 4.6% (10 years) and 1.5% (20 years) for the Parma registry, 4.4% and 1.4% for Varese, and 7.3% and 2% for Ragusa. For all malignant cancers (except non-melanoma skin cancers), the absolute relative differences were below 4%, when prevalence was truncated at 15 or 20 years, and below 7% at 10 years for all the three CRs. Discrepancies between 10% -15% were systematically obtained for all the three registries for thyroid cancer in women. The substantial increase of incidence and survival for this neoplasm in the last decades, varying by geographical area, may explain the reason for the sub-optimal adaptation to cancer registry data. Discrepancies higher than 10% (especially at 10 years) were found in other cancer sites (Hodgkin's lymphoma, brain, skin melanoma, connective and soft tissues, leukaemia, urinary bladder, and non-Hodgkin lymphoma) but not systematically in all three registries used for validation.

National and regional estimates of prevalence

The estimated complete number of prevalent subjects by cancer site and gender in Italy was derived by summing up the number of prevalent cases in each of the four Italian macro-areas, which were derived by multiplying site, gender, age and macro-area specific complete prevalence proportions by the corresponding resident population size at 1 January 2006. The prevalence proportion in Italy estimated in this way necessarily diverges from the prevalence proportion estimated for the Italian Pool of CRs, because prevalence levels and registration coverage between Center-North and South are quite different. The regional estimated complete number of prevalent subjects by cancer site and gender in the regions with partial registration coverage was derived in the same way, by multiplying the site, gender, age and macro-area specific complete prevalence proportions by the corresponding 1 January 2006 regional population (Table 2, p. 19).
National and regional estimates assume that populations living in areas covered by a cancer registry are representative of population living in areas not covered by registries.
Appendix *(p. 180-181) reports the regional and macro-area populations on 1 January 2006 used in estimates and downloaded from the Italian National Institute of Statistics (ISTAT) website (http://demo.istat.it/pop2006/index.html).*

References - Bibliografia

1. Micheli A, Francisci S, Krogh V, Rossi AG, Crosignani P. Cancer prevalence in Italian cancer registry areas: the ITAPREVAL study. ITAPREVAL Working Group. *Tumori* 1999; 85(5): 309-69.
2. AIRTUM Working Group, I trend dei tumori negli anni duemila (1998-2005). Cancer trend (1998-2005) *Epidemiol Prev* 2009; 33(4-5) suppl. 1: 1-168.
3. Surveillance Research Program. National Cancer Institute SEER*Stat Software, http://seer.cancer.gov/seerstat/
4. Surveillance Research Program. National Cancer Institute COMPREV Software, http://srab.cancer.gov/comprev/
5. Capocaccia R, De Angelis R. Estimating the completeness of prevalence based on cancer registry data. *Stat Med* 1997;16(4): 425-40.
6. Corazziari I, Mariotto A, Capocaccia R. Correcting the completeness bias of observed prevalence. *Tumori* 1999; 85(5): 370-81.
7. Capocaccia R, Colonna M, Corazziari I, De Angelis R, Francisci S, Micheli A, Mugno E; EUROPREVAL Working Group. Measuring cancer prevalence in Europe: the EUROPREVAL project. *Ann Oncol* 2002; 13(6): 831-9.
8. Holford TR, Zhang Z, McKay LA. Estimating age, period and cohort effects using the multistage model for cancer. *Stat Med* 1994; 15; 13(1): 23-41.
9. Merrill RM, Capocaccia R, Feuer EJ, Mariotto A. Cancer prevalence estimates based on tumour registry data in the Surveillance, Epidemiology, and End Results (SEER) Program. *Int J Epidemiol* 2000; 29(2): 197-207.
10. Akaike, H. (1973). I*nformation theory and an extension of the maximum likelihood principle. In 2nd International Symposium on Information Theory.* Ed. B. N. Petrov and F. Csaki, pp. 267-81. Budapest. Akademiai Kiado. (Reproduced (1992) in *Breakthroughs in Statistics* 1, Ed. S. Kotz and N. L. Johnson, pp. 610-24. New York: Springer-Verlag.).
11. Akaike, H. (1974). *A new look at the statistical model identification.* IEEE Trans. Auto. Control AC-19, 716-23.