Metodi

Methods

Fabrizio Stracci,[1] Claudio Sacchettini[2]

[1] *Registro tumori umbro di popolazione, Dipartimento di igiene e sanità pubblica, Università di Perugia*
[2] *Registro tumori toscano, Unità operativa di epidemiologia clinica e descrittiva, Centro per lo studio e la prevenzione oncologica, Istituto scientifico della Regione Toscana, Firenze*

Corrispondenza: *Fabrizio Stracci, Registro tumori umbro di popolazione, Dipartimento di igiene e sanità pubblica, Università di Perugia, via del Giochetto 06100 Perugia*

Riassunto

L'analisi dei *trend* nella banca dati dell'Associazione Italiana Registri Tumori ha avuto prevalentemente un carattere descrittivo. Il periodo di analisi è stato definito dal 1986 al 1997. Nell'ambito di questo periodo sono stati inclusi i nove registri che presentavano almeno dieci anni di attività (*pool* AIRT). Per tre registri sono stati stimati, in base ai dati osservati, i dati di incidenza per gli anni mancanti. Per questo *pool* di registri sono stati calcolati tassi standardizzati (popolazione europea) sia di incidenza che di mortalità con il metodo diretto. E' stata applicata la cosiddetta analisi dei *joinpoint* per identificare punti temporali di cambiamento del *trend*. I *trend* sono stati riassunti attraverso le stime dei cambiamenti annuali percentuali (EAPC). Sono stati calcolati anche tassi età-specifici e tassi specifici per coorti di nascita.

Introduzione

La finalità di questa prima presentazione degli andamenti o *trend* temporali osservati dai Registri Tumori Italiani è eminentemente descrittiva. Questa analisi dei *trend* di incidenza e mortalità è volta all'individuazione di andamenti prevalenti, di eventuali variazioni nel periodo in studio, e a fornire una valutazione della intensità di variazione del fenomeno, piuttosto che a confrontare l'andamento per area geografica o a formulare previsioni. La preferenza per i dati osservati piuttosto che per dati stimati è stata motivata dalla copertura solo parziale da parte dei registri e non rappresentativa della popolazione italiana.[1] In questa linea si collocano il calcolo degli indicatori per un insieme di registri (*pool AIRT*) e la presentazione dei *trend* di incidenza e mortalità standardizzate per età.

Criteri di inclusione/esclusione

La scelta degli indicatori presentati, dei metodi analitici e dei criteri di inclusione dei casi è stata dettata dai seguenti principi generali:

1. presentazione di indicatori osservati e riduzione al minimo delle procedure di stima e, dunque, del numero di assunti necessari;
2. semplicità e intuitività della presentazione consentita dall'impiego di indicatori di popolazione osservati e ottenuta mediante un largo impiego di figure;
3. individuazione del periodo e delle aree in studio sulla base dei dati di incidenza disponibili.

Abstract

The trend analysis of the data base of the Italian Network of Cancer Registries has had a descriptive approach. The period of study has been from 1986 to 1997. Within this period we have included only nine Registries that were active for at least ten years (pool AIRT). For three Registries few lacking incidence years have been estimated according to their observed data. For the pool of Registries standardised (standard = European population) incidence and mortality rates have been computed with the direct method. The so called joinpoint analysis has been used to detect temporal points of trend change. Trends have been summarised by means of estimated annual percent change (EAPC) of the rate. Age-specific and birth cohort-specific rates have been also computed.

Introduction

The presentation of the cancer trends observed by the Italian Cancer Registries has mainly a descriptive approach.
The present incidence and mortality cancer trends analysis is mainly aimed to evidence the main trends, to detect changes occurring over the study time, and to quantify the extent of variation rather than to carry out geographical comparisons or to forecast future events.
A preference for observed data rather than for estimated once was due to the partial coverage of the Italian population by Cancer Registries.[1]
Several indicators were computed for a group of registries (pool AIRT), e.g. age-standardised incidence and mortality rates.

Inclusion / exclusion criteria

The choice of indicators, the methodology and the inclusion/exclusion criteria agreed with the following general rules:

1. *preference for measured indicators and limited use of estimates (and assumptions);*
2. *simple and friendly presentation of measured population-data with the support of a wide iconography;*
3. *period and area inclusion on the basis of the available data.*

The analysis included malignant cancers incidence in subjects 15+ years old. All the registries active for at least 10 years were included. This seems similar to select only complete

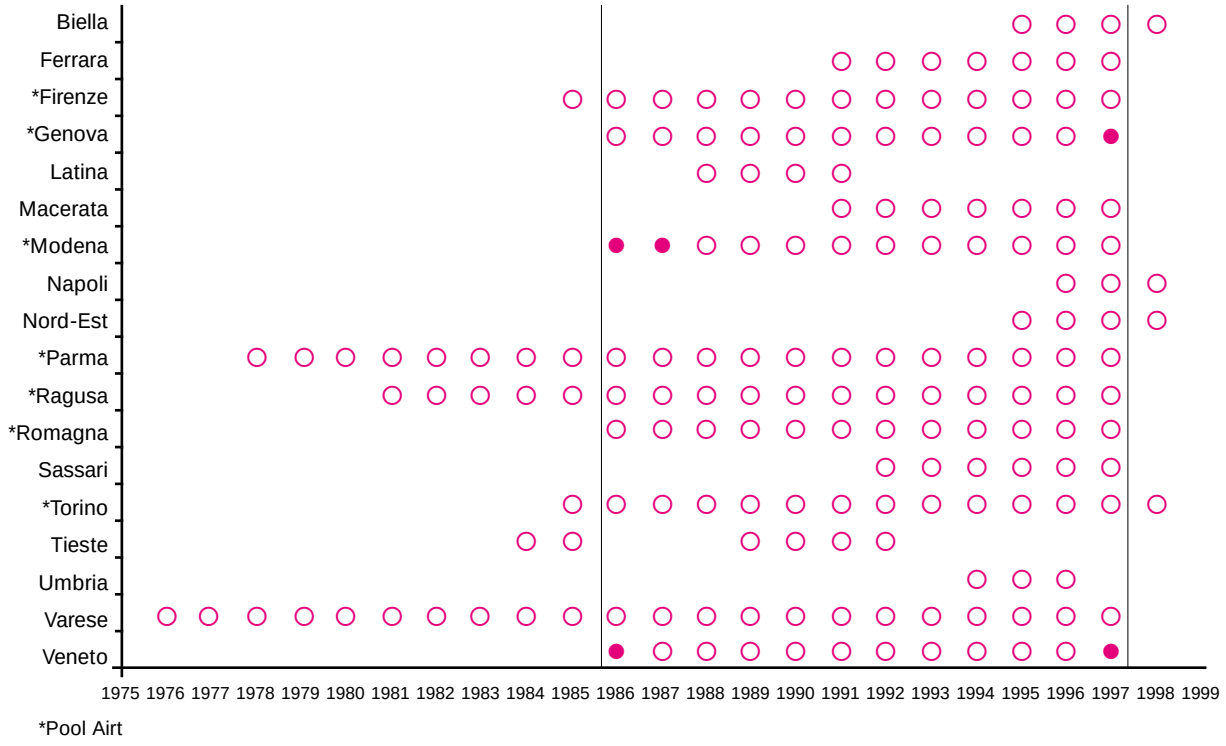

Figura 1. Registrazione dell'incidenza di tumori maligni per anno e Registro. Il periodo considerato è quello tra le linee verticali e i pallini pieni indicano i tassi di incidenza stimati.

Figure 1. Cancer registration by Registry and year. The period included in the study is the one between the vertical lines; the full circles stand for estimated data.

Nell'analisi sono stati considerati i tumori maligni invasivi incidenti in individui ≥15 anni. Sono state incluse nello studio le aree con disponibilità di un periodo di registrazione dei tumori maligni di almeno 10 anni, considerato come periodo minimo di osservazione per lo studio del *trend*. Sebbene la scelta dell'esclusione possa sembrare analoga all'analisi dei soli casi completi in uno studio con dati mancanti e, dunque, implicare uno spreco di informazione, in realtà, l'inclusione di tutti i dati disponibili avrebbe comportato un radicale cambiamento di approccio. Il periodo di osservazione per i Registri presenti nella banca dati AIRT varia da un minimo di tre a un massimo di ventidue anni (media 9.7, DS 5.83). Al momento dell'ingresso nel *pool* di un registro con un diverso livello di incidenza si modifica il tasso complessivo e, dunque, si crea un *trend* apparente; una distorsione che impedisce di formare un *pool* con i dati disponibili, tanto più che in Italia l'incidenza di molti tumori mostra una marcata variabilità geografica. Per poter includere tutti i registri si sarebbe dunque reso necessario includere in un modello di regressione l'effetto dei Registri come variabile di aggiustamento o ricorrere in misura estensiva a stime, per arrivare a una base dati completa oppure ancora rinunciare ad una analisi congiunta (vedi anche l'articolo su *La rete dei Registri Tumori italiani* in questa monografia).

data in a study with missing data, and therefore it may cause an information wasting. On the other hand, the inclusion of all the available date would have caused a radical different methodological approach.
The period of activity of the Registries included in the AIRT database ranges from three to twenty-two years (mean 9.7, SD 5.83).
When a registry with a different incidence level is included in a set of registries it may cause a change in the overall incidence rate, and an artificial trend may become evident. This source of possible bias has hampered to include all the available data, also considering the variability of incidence in different Italian geographic areas. The choice of including all the Registries would have made necessary to use a regression model with the effect of each Registry as an adjustment variable, or to make a lot of estimates or, finally to decide not to carry out a combined analysis (see the paper 'The Italian Registries Network' in the present monograph for further details).
We made some simulation and according to their results we decided to restrict our study to a pool of registries.
It would be interesting, in the future, to evaluate similar and different trends, for different cancer sites, for each Italian Registry.

Alcune simulazioni hanno orientato la scelta verso l'analisi di un *pool* ridotto di registri storici. Un'analisi focalizzata sulla variabilità dei *trend* nei diversi registri e sulla valutazione dell'esistenza di *trend* temporali comuni ai registri italiani per le diverse patologie tumorali costituisce una interessante prospettiva di studio.

Registri inclusi nello studio e periodo di osservazione

L'archivio AIRT include diciotto Registri Tumori ma solo nove di essi hanno fornito dati di incidenza per un periodo ≥ 10 anni (Figura 1) e sono stati quindi inclusi nello studio (*pool* AIRT). I nove registri di lungo corso coprivano nel 1997 una popolazione di circa 8.000.000 distribuita in modo diseguale tra Nord (21%), Centro (19%) e Sud (2%) d'Italia. I periodi per i quali erano disponibili i dati di incidenza necessari sono risultati non sovrapponibili per differenze nell'anno di inizio della serie temporale o nella tempestività di completamento dell'incidenza per gli anni più recenti (Figura 1). Il periodo in studio è stato definito dal 1986 al 1997. Si è reso necessario ricorrere ad una stima dell'incidenza per i due registri che non avevano ancora fornito dati relativi al 1997 e lo stesso criterio empirico (80% di registri del *pool* attivi) è stato applicato agli anni di ingresso con la conseguenza che è stato possibile includere nello studio gli anni 1987 e 1986 a prezzo di stimare l'incidenza per uno e due registri rispettivamente. I dati mancanti sono stati stimati utilizzando il modello di regressione log-lineare tipo *joinpoint* descritto in seguito ma adattato alla serie temporale dei singoli registri con dati mancanti anziché all'intero *pool*.

Indicatori

Per il *pool* dei registri sono riportati il numero di casi (n), i tassi standardizzati di incidenza e di mortalità per tumori maligni separatamente per sede e sesso e anno. I nuovi casi di malattia per anno, sesso e classe d'età sono stati forniti dai registri tumori e il numero dei decessi dall'ISTAT. I tassi sono stati standardizzati con metodo diretto:

$$T_{stand} = \frac{\sum_i (T_i \text{ pop stand}_i)}{\sum_i (\text{pop stand}_i)}$$

in cui T_i è il tasso specifico per la classe d'età *i* (pari al rapporto tra gli eventi registrati in un anno nella classe d'età *i* e una stima degli anni-persona nella stessa classe d'età), espresso per 100.000 abitanti, $T_i = 100.000 * n_i/p_i$. Gli anni-persona per classe d'età relativi alle aree in studio sono stati stimati utilizzando popolazioni fornite dall'ISTAT.
Per la standardizzazione è stata utilizzata la popolazione europea *standard* (Figura 2). La sopravvivenza relativa a uno, tre e cinque anni per i tumori maligni in studio è riportata separatamente per i casi incidenti nei periodi 1985-89 e 1990-94. La sopravvivenza relativa è il rapporto tra la sopravvi-

Included Registries and period of observation.

The AIRT database included eighteen Cancer Registries, among which only nine have incidence data for a period of ≥10 years (Figure 1); the latter have been included in the present study (pool AIRT).
These nine registries correspond in 1997 to a population of about 8,000,000 inhabitants and they represented mainly the Northern (21%) and Central (19%) than the Southern (2%) Italian population.
Time periods for which there were available data from long lasting Registries were not overlapping according to different starting and ending years (Figure 1). The study period was defined since 1986 to 1997. Therefore, for three Registries, according to measured data in each Registry, we have estimated incidence data for 1986 and 1987 (one Registry) and 1997 (two Registries). Lacking data have been estimated by means of a joinpoint log-linear regression model implemented on the temporal trend of the Registry for which data have to be computed.

Indicators

For the pool of Registries, the number of cases (n), the standardised incidence and mortality rates for each sex, cancer site and year have been presented. Incident cases were based on Registries data and mortality data were retrieved from ISTAT (National Institute of Statistics).
Rates have been age-standardised with the direct method:

$$T_{stand} = \frac{\sum_i (T_i \text{ pop stand}_i)}{\sum_i (\text{pop stand}_i)}$$

T_i means the age-specific rate for the age-class i *(the rate between the event occurred in one year in the age-group* i *and an estimate of the person-years in the same age-group) per 100,000 inhabitants, $T_i = 100,000 * n_i/p_i$.*
Person-years for each age-class in the analysed areas have been estimated by means of the ISTAT populations.
The standard was the European population (Figure 2).
Relative survival at one, three and five years after diagnosis, for the analysed malignant cancers, has been presented for incident cases in 1985-1989 and in 1990-1994.
Relative survival is the ratio between the observed survival in cancer patients and that expected among subjects of the same age and sex in the general population and it estimates the specific effect on survival of the disease on study. Further details are available in Rosso et al.[2]

Trend analysis

An intuitive approach to trend has been presented by means of a figure with standardised incidence and mortality rates by year. The same information, with the number of events, is also presented in a table. At the bottom of such table there are further information on the number of joinpoints, the year of

venza osservata per gli ammalati di tumore maligno e la sopravvivenza attesa per individui della stessa età e sesso appartenenti alla popolazione generale e stima l'effetto specifico della malattia sulla sopravvivenza. Ulteriori dettagli sul metodo di calcolo possono essere reperiti in Rosso *et al.*[2]

Analisi del *trend*

Un grafico che riporta i tassi standardizzati di incidenza e mortalità per anno fornisce una informazione intuitiva sui rispettivi *trend*. L'informazione è inoltre riportata in forma tabulare assieme al numero di eventi. Ulteriori spiegazioni richiedono i valori riportati in fondo alla tabella relativi a numero di *joinpoint*, anno e EAPC ovvero alla stima del cambiamento percentuale medio annuo dei tassi. Per avere una nozione sintetica del segno della variazione temporale dei tassi e della intensità di variazione annuale si è adottato un modello di regressione che assume la linearità dei *trend* (più precisamente del logaritmo dei tassi). In realtà non vi è motivo per cui l'andamento dei tassi debba essere lineare e, in effetti, assumere un *trend* lineare è spesso ingiustificato. Tuttavia ogni curva può essere approssimata localmente in modo soddisfacente mediante un segmento lineare purché di lunghezza appropriata. Il modello impiegato si basa, appunto, sull'individuazione dei segmenti lineari che meglio si adattano ai tassi osservati, rendendo minima la somma dei quadrati delle distanze dei punti dai segmenti stessi.[3] Il massimo numero di segmenti in cui è scomposto il *trend* è limitato dal numero *k* di *joinpoint* arbitrariamente fissato prima dell'analisi. Per *k*=2 *joinpoint* il *trend* può essere rappresentato al massimo da (*k*+1)=3 segmenti con diversa pendenza. Il *joinpoint* è il punto di giunzione, l'anno che individua una variazione del *trend*; di nuovo si tratta di un'utile approssimazione in quanto è improbabile che incidenza o mortalità varino bruscamente in un anno definito. Il modello log-lineare *joinpoint* può essere rappresentato in un'unica equazione come segue:

$$\ln(T_{stand}) = ß_0 + ß_1 x_i + \ _1(x_i - \ _1)^+ + \ _2(x_i - \ _2)^+ + \ _i^{(k)}$$

Nell'equazione exp($ß_0$) è la stima basata sul modello del tasso standardizzato per il *pool* di registri nel 1986. è un termine di errore che rappresenta la variabilità casuale delle misure. Nel caso di un modello senza *joinpoint* l'equazione si riduce all'equazione di una retta con intercetta $ß_0$ e pendenza $ß_1$. I termini tipo $_i(a)^+$ rappresentano, invece, la variazione della pendenza della retta per il secondo e terzo segmento in presenza di *joinpoint* e si annullano per gli anni precedenti l'anno *joinpoint* ($a \leq 0$). *k* è il numero massimo di variazioni consentite ma il numero di *joinpoint* di volta in volta selezionato per ogni sede tumorale varia da 0 a 2 in base ad un processo di semplificazione del modello. Il modello con il maggior numero di *joinpoint* fornisce il migliore adattamento ma è desiderabile utilizzare una procedura per

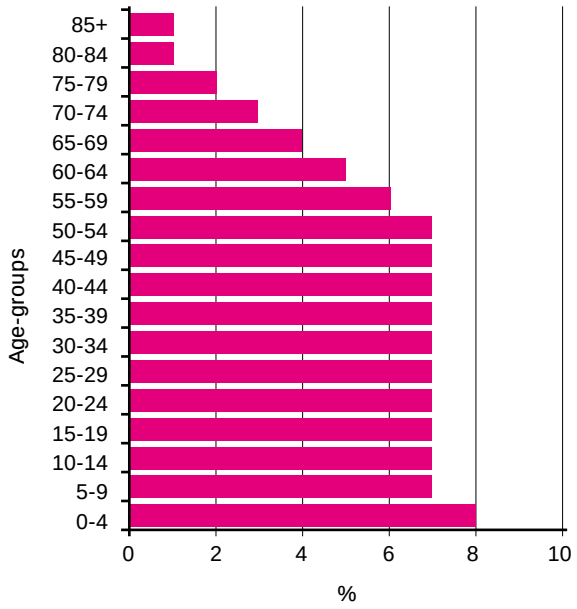

Figura 2. Popolazione europea standard.

Figure 2. European standard population.

*joinpoint and on the estimated annual percent change in rates (EAPC). To summarise the direction and the intensity of the trend we used a regression model which assumes the linearity of the trend (of the logarithm of the rates). The assumption of linearity may not be always justified. However, each trend may be considered linear in a sufficiently short period of time. The final model is based on linear segments connected at joinpoints that represent the best fit of the observed data.[3] The maximum number of segments for each trend is limited by the maximum number of joinpoint stated at the beginning of the analysis (*k*). For k=2 joinpoints the trend may be described by a maximum of (k+1) = 3 segments with different slop. The joinpoint is the point that links two different lines, it represents the year when the trend changes. This is a simplification because it is unlikely that a trend changes sharply exactly on one specific year. The joinpoint log-linear model may be described by the following equation:*

$$\ln(T_{stand}) = ß_0 + ß_1 x_i + \ _1(x_i - \ _1)^+ + \ _2(x_i - \ _2)^+ + \ _i^{(k)}$$

*Exp($ß_0$) is the estimate based on the model of the standardised rate for the pool of Registries in the year 1986. represents the random variability in the measures. A model without joinpoints is represented by a straight line with intercept $ß_0$ and slope $ß_1$. $_i(a)^+$ represents, the change in slope for the second, or third segment, if there are any joinpoints; they are zero for the years before the year of the joinpoint (*a≤0*). k is the maximum number of possible changes; the number of*

semplificare il modello senza perdere informazioni importanti sui *trend*.[4] L'algoritmo confronta dapprima il miglior modello con due *joinpoint* con il modello senza *joinpoint*: se la variabilità residua non è significativamente diversa (<0.05) per i due modelli allora non possiamo concludere che il modello complesso sia migliore del modello senza *joinpoint* e la procedura passa allora al confronto tra il modello con un *joinpoint* e quello senza *joinpoint*. Se invece la somma dei quadrati dell'errore è significativamente inferiore per il modello più complesso allora la procedura viene ripetuta confrontando il modello a due *joinpoint* con il modello con un solo *joinpoint*. Una volta selezionato il modello è possibile fornire una misura della variazione annua percentuale (EAPC, Estimated Annual Percent Change) a partire dalle pendenze di ogni segmento lineare utilizzando la formula:

$$EAPC = 100 \times (e^{ß} - 1)$$

e il relativo intervallo di confidenza al 95% è ottenuto applicando la formula agli estremi dell'intervallo di confidenza per la pendenza dei segmenti lineari piuttosto che alla pendenza stessa:

$$ICi = 100 \times (e^{ß - (t\,[n-p\,;\,0.025]\,x\,ES)} - 1)$$

e

$$ICs = 100 \times (e^{ß + (t\,[n-p\,;\,0.025]\,x\,ES)} - 1)$$

in cui ES è l'errore standard di *ß* e *t* è il quantile della distribuzione t di Student per (n-p) gradi di libertà, con p=2k+2. Un limite del modello basato sui tassi standardizzati è che non tiene conto di tendenze specifiche per età che possono essere discordanti. Il grafico dei tassi di incidenza e mortalità specifici per classe d'età possono essere d'aiuto alla valutazione della bontà del modello *joinpoint*. Infine il *trend* della composizione percentuale secondo la base delle diagnosi di tumore maligno (ad esempio istologica) fornisce elemento a speculazioni sul ruolo della variazione della qualità dei dati nel determinare i *trend* di incidenza, mortalità e sopravvivenza.

La presente analisi si riferisce ai dati dell'Associazione Italiana Registri Tumori (*pool* AIRT) relativi ai seguenti registri: Registro Tumori del Piemonte e della Valle d'Aosta, Registro Tumori del Veneto, Registro Tumori della Provincia di Modena, Registro Tumori della Provincia di Parma, Registro Tumori della Provincia di Ragusa, Registro Tumori della Romagna, Registro Tumori Lombardia - Provincia di Varese, Registro Tumori Regione Liguria, Registro Tumori Toscano, e riguardano il periodo 1986-1997.

joinpoints changes for each cancer sites from 0 to 2 according to the simplest model that best fits the observed data. The model with the maximum number of joinpoints usually fits the data better than the others, but the choice is for the simplest model which fits enough well the data without losing any relevant information.[4] The algorithm firstly compares the best model - with two joinpoints - with the one without joinpoints. If the residual variability between the two models is not significantly different (<0.05) we can not state that the more complex model is better than the simpler one. Then, the comparison is between the model with one joinpoint against the one without joinpoints. In case the sum of square errors is significantly below for the more complex model the comparison is between the model with two against the one with only one joinpoint. When the model is defined, it is possible to measure the Estimated Annual Percent Change of the rate (EAPC) for each line of the trend, according to the following formula:

$$EAPC = 100 \times (e^{ß} - 1)$$

the 95% confidence intervals are computed as follows:

$$CIl = 100 \times (e^{ß - (t\,[n-p\,;\,0.025]\,x\,ES)} - 1)$$

e

$$CIu = 100 \times (e^{ß + (t\,[n-p\,;\,0.025]\,x\,ES)} - 1)$$

where ES is the standard error of ß and t *is the quantile of a Student's t distribution with (n-p) degrees of freedom, with p=2k+2.*
A limit of this model based on standardised rates is that it does not consider age-specific trends.
The figures showing age-specific incidence and mortality rates trends may help to evaluate the goodness of the join.point model.
Finally, time trends of the basis of cancer diagnosis (e.g. histology) may help the evaluation of possible changes in data quality that may have caused changes in incidence, mortality, and survival trends.

The present analysis refers to the following registries of the Italian Network of Cancer Registries (pool AIRT): Registro Tumori del Piemonte e della Valle d'Aosta, Registro Tumori del Veneto, Registro Tumori della Provincia di Modena, Registro Tumori della Provincia di Parma, Registro Tumori della Provincia di Ragusa, Registro Tumori della Romagna, Registro Tumori Lombardia - Provincia di Varese, Registro Tumori Regione Liguria, Registro Tumori Toscano, and to the period 1986-1997.

Bibliografia - References

1. Falcini F, Paci E, La Rosa F, *et al.*, La rete italiana dei Registri Tumori. In: Zanetti R, Gafa' L, Pannelli F, Conti E, Rosso S, eds, *Il cancro in Italia. I dati di incidenza dei Registri Tumori, 1993-1998*. Roma, Il Pensiero Scientifico, 2002. Vol. 3.
2. Rosso S, Casella C, Crocetti E *et al.*, eds, Sopravvivenza dei casi di tumore in Italia negli anni novanta: i dati dei Registri Tumori, *Epidemiol Prev*, 2001; 25(suppl 3): 1-375.
3. Lerman PM. Fitting Segmented Regression Models by Grid Search, *Applied Statistics*, 1980; 29: 77-84.
4. Kim H J, Fay M P, Feuer E J, Midthune D N. Permutation tests for joinpoint regression with applications to cancer rates, *Statistics in Medicine*, 2000; 19: 335-351.