UNIVERSITÀ DEGLI STUDI
DI MILANO

DIPARTIMENTO DI SCIENZE
CLINICHE E DI COMUNITÀ

Scuola di Dottorato in Scienze Biomediche Cliniche e Sperimentali

Dottorato in Statistica Biomedica XXVI ciclo

Settore scientifico disciplinare MED/01

Guarigione dal tumore al seno: un tema dibattuto

Dottoranda: Letizia Trevisi

Matricola: R09141

Tutor: Dott. Federico Ambrogi

Coordinatore del dottorato: Prof. Adriano Decarli

A.A. 2012/2013

Indice

Elenco delle tabelle

Elenco dei grafici

Introduzione

Numerosi studi hanno dimostrato che l'aspettativa di vita per gli individui con diagnosi di carcinoma mammario è aumentata nei paesi occidentali. Certamente l'attuazione di programmi di screening e i crescenti progressi nel trattamento di tale cancro sono tra i motivi della riduzione della mortalità osservata negli ultimi due decenni[1,2,3].

Considerato l'aumento della sopravvivenza per questa patologia, numerosi ricercatori hanno effettuato ulteriori studi per capire se effettivamente si può guarire, ottenendo però risultati contrastanti[4,5,6]. Infatti, attualmente, esistono prove dell'esistenza di una proporzione di guariti per alcuni sottogruppi di pazienti, ma in molti studi si osserva un eccesso di mortalità anche dopo un follow-up lungo[7].

Molti di questi studi sono stati effettuati utilizzando dati provenienti da registri tumori[8]. Il metodo standard per indagare la guarigione è mediante l'applicazione di modelli di sopravvivenza relativa, confrontando la mortalità generale osservata nella coorte di donne con cancro al seno con la mortalità generale di donne comparabili per età e coorte, dati disponibili da statistiche di mortalità nazionali[9].

Una prova della guarigione è data dall'appiattimento della curva di sopravvivenza cumulata relativa che dimostra che la probabilità di sopravvivenza delle donne con cancro al seno è simile a quella delle donne senza cancro al seno. Molti studi mostrano però un eccesso di mortalità anche dopo tempi di follow-up lunghi[10]. Il modello statistico utilizzato è il "Cure Mixture Model", che suppone a priori l'esistenza di una frazione di guariti[11]. Ovviamente se tale ipotesi è discutibile, i risultati del modello devono essere considerati con molta attenzione.

Un diverso approccio ha come oggetto lo studio della mortalità causa-specifica, applicabile solo nel caso in cui tali informazioni siano disponibili e affidabili. In questo contesto, la causa della morte di ogni paziente è classificata come "cancro" o "non cancro" quindi è possibile osservare solo una delle cause di morte. In analisi della sopravvivenza, si è in un contesto di rischi competitivi[12]. Se fosse confermata la guarigione di pazienti con cancro al seno, ci si aspetta che a partire da un certo tempo del follow-up in poi si verifichino pochi decessi dovuti al cancro. Dal momento che sono presenti varie cause concorrenti di morte, la probabilità di morte per cancro al

seno non raggiunge mai il 100%, anche per lunghissimi tempi di follow-up, in quanto i pazienti guariti moriranno per cause diverse dal cancro al seno. Un plateau nella curva della funzione d'incidenza cumulata per cancro al seno è quindi utilizzato come prova della guarigione, il che significa che solo una parte delle donne con cancro al seno morirà per la patologia in studio[13].
Dal punto di vista statistico si tratta di descrivere le funzioni d'incidenza cumulata di rischio specifico mediante una distribuzione impropria. Questo approccio è stato sviluppato nell'articolo "Direct parametric inference for the cumulative incidence function" da Jeong e Fine e denominato di stima diretta delle funzioni d'incidenza cumulata. La stima diretta si contrappone a quella indiretta in cui si stimano i rischi istantanei causa-specifici per poi risolvere, mediante una loro combinazione, la stima della funzione d'incidenza cumulata. L'approccio diretto ha il vantaggio che il modello è più facilmente interpretabile rispetto a quello per rischi istantanei causa-specifici.

Razionale del lavoro

Lo scopo di questo lavoro è quello di studiare la guarigione per cancro al seno e il ruolo dei principali fattori prognostici sulla guarigione utilizzando dati provenienti da una sperimentazione clinica controllata per la quale è oggi disponibile un lungo follow-up (30 anni) necessario da considerare per la patologia in esame. Il reclutamento dei pazienti è iniziato nel 1973 presso l'Istituto Nazionale per lo Studio e la Cura dei Tumori di Milano. Il trial confronta la mastectomia radicale e la chirurgia conservativa del seno. Per l'analisi ci si è avvalsi sia di metodi standard di sopravvivenza relativa, sia di approcci parametrici che tengono in considerazione i rischi competitivi, grazie alla buona qualità delle informazioni disponibili sulle cause di morte.

1. Metodi

1.1. Background

La sopravvivenza relativa confronta le esperienze di sopravvivenza dei pazienti in studio con quelle previste per la corrispondente coorte (appaiata per data di nascita, età e sesso) della popolazione generale. In questo modo è possibile fornire una misura della mortalità dovuto alla malattia in studio. Questi metodi sono di solito utilizzati per studi con dati raccolti in modo retrospettivo da registro dei tumori e non necessitano di alcuna informazione sulle cause di morte. La funzione di sopravvivenza relativa cumulativa indica la proporzione di soggetti ancora in vita in un determinato periodo di follow-up se il cancro al seno è l'unica possibile causa di morte, supponendo una bassa incidenza di mortalità per cancro nella popolazione e l'indipendenza tra le cause di morte. Dal punto di vista statistico, la guarigione si verifica quando la curva di sopravvivenza cumulativa relativa ha un appiattimento o, analogamente, il tasso di mortalità in eccesso è pari a zero[14,15,16].

Un modello frequentemente usato per stimare la frazione di guariti è il "Cure Mixture Model"[17]. Si presuppone che una parte dei soggetti, π, guarisca, mentre il restante, $1-\pi$, sperimenti un eccesso di mortalità rispetto alla popolazione generale. Se, invece, sono disponibili le informazioni sulle cause del decesso, è possibile applicare l'analisi dei rischi competitivi.

1.2. Modelli di mistura

I modelli di sopravvivenza “mistura” permettono di modellare il tempo all’evento in tutte quelle situazioni in cui un’unica funzione di densità parametrica è inadeguata a descrivere correttamente la presenza di differenti sottopopolazioni all’interno di una popolazione complessiva. Nel caso in esame le due sottopopolazioni sono i soggetti guariti e i soggetti non guariti.

Le due più importanti classi di modelli sviluppati sono:

- ✓ Modelli di mistura parametrici basati su funzioni di densità del tempo all’evento come la Weibull, la Lognormale e la Gamma che permettono di stimare la frazione di guarigione e il tasso di mortalità attraverso il metodo della massima verosimiglianza;

- ✓ Modelli di mistura non parametrici.

Nell'ambito della sopravvivenza per cancro l'interesse verso questi modelli è andato via via crescendo, infatti i continui progressi nei trattamenti hanno permesso di suddividere la popolazione dei pazienti in una *mistura* di sottogruppi eterogenei per rischio di morte: pazienti che muoiono per cancro e pazienti guariti che non presentano quindi un eccesso di mortalità rispetto alla popolazione generale.

Non ci si focalizza esclusivamente sul tempo alla morte ma anche sulla proporzione di guariti.

La sopravvivenza per cancro è influenzata da svariati fattori: fattori biologici come l'età, il genere, o differenti stadi al tempo della diagnosi, l'efficacia della terapia o del trattamento, facilità nell'accedere all'assistenza sanitaria.

Le variabili esplicative introdotte nel modello possono giocare un ruolo differente sulla proporzione di guariti o sul tasso di mortalità. Per quanto riguarda modelli di mistura parametrici solitamente vengono combinate la funzione logistica con funzioni di densità come la Weibull o l'esponenziale per modellare rispettivamente la frazione di guarigione e la sopravvivenza per coloro che non guariscono. I parametri vengono stimati tramite il metodo della massima verosimiglianza.

1.2.1. Specificazione del modello

In sopravvivenza relativa la funzione di sopravvivenza per tutte le cause, $S(t)$, può essere espressa come il prodotto della funzione di sopravvivenza attesa, $S^*(t)$, e la funzione di sopravvivenza relativa, $R(t)$:

$$S(t) = S^*(t)R(t)$$

come visto precedentemente su scala di hazard la relazione può essere scritta come:

$$\lambda(t) = \lambda^*(t) + \lambda_e(t)$$

Observed mortality rate = Expected mortality rate + Excess mortality rate.

L'observed mortality rate non è altro che la somma di due componenti l'expected mortality rate e l'e*xcess mortality rate* associato alla malattia d'interesse.

Quando il tasso di mortalità osservato ritorna al pari di quello atteso la curva della sopravvivenza relativa raggiunge un appiattimento (momento in cui si raggiunge la guarigione).

1.2.2. Frazione di guariti nei modelli "mistura"

I modelli di mistura vengono così chiamati perché suppongono che la popolazione studiata è una "miscela" di soggetti che sperimentano l'evento di interesse (morte per cancro al seno) e soggetti che non sperimentano mai l'evento[17].

Questo approccio consente di stimare la proporzione dei soggetti guariti e il tempo di guarigione.

La funzione di sopravvivenza per tutte le cause incorporando l'*expected mortality* è:

$$S(t) = S^*(t)\{\pi + (1-\pi)S_u(t)\}$$

dove π è la proporzione dei soggetti guariti e $S_u(t)$ è la funzione di sopravvivenza per i pazienti non guariti, che solitamente ha una funzione parametrica, come la Weibull, la lognormale o la Gamma.

Su scala hazard l'espressione diventa:

$$\lambda(t) = \lambda^*(t) + \frac{(1-\pi)f_u(t)}{\pi + (1-\pi)S_u(t)}$$

Un presupposto standard dei modelli di sopravvivenza relativa è che gli *expected* e gli *excess* mortality rate siano indipendenti. Risulta un'ipotesi ragionevole in studi su base di popolazione ad eccezione dei tumori per fumo. Solitamente la modellazione avviene a partire dal momento della diagnosi.

Il "Cure Mixture Model" presuppone quindi che al momento della diagnosi ci sia un gruppo di individui che vive senza un eccesso di mortalità rispetto alla popolazione generale.

Sposto (2002) sostiene che la separazione dei soggetti guariti e non, già al tempo t = 0 non è appropriata in un'epoca in cui il trattamento può durare molti anni e la guarigione potrebbe verificarsi in qualsiasi momento.

Sembra improbabile che ci sia un gruppo di individui guariti prima che venga somministrato qualsiasi tipo di trattamento.

Comunque ciò non compromette l'utilizzo di tale modello in quanto può adattarsi bene ai dati.

Il contributo alla funzione di verosimiglianza in un generico modello di sopravvivenza di ciascun soggetto con tempo di sopravvivenza o di censura t_i e variabile indicatrice di censura d_i è [11]:

$$lnL_i = d_i ln\big(h(t_i)\big) + ln\big(S(t_i)\big)$$

In un mixture model la log-verosimiglianza incorpora l'*expected mortality* ed è:

$$lnL_i = d_i ln\left(h^*(t_i) + \frac{(1-\pi)f_u(t_i)}{\pi + (1-\pi)S_u(t_i)}\right) + ln\big(S^*(t_i)\big) + ln\big(\pi + (1-\pi)S_u(t_i)\big)$$

$S^*(t_i)$ è indipendente dai parametri del modello di conseguenza può essere rimosso dalla verosimiglianza, quindi la funzione di verosimiglianza può essere definita per ogni distribuzione parametrica data la funzione di densità $f_u(t_i)$ e la funzione di sopravvivenza $S_u(t_i)$ per i pazienti non guariti.

Questo modello è stato anche utilizzato per ottenere misure sintetiche utili per coloro che sono "destinati a morire", anche se queste misure sono disponibili analogamente per il modello non-mixture.

1.3. Introduzione del concetto di funzione d'incidenza cumulata

Nell'ambito dell'analisi dei rischi competitivi è necessario definire due quantità fondamentali:

1. hazard causa-specifica

$$\lambda_k(t) = \lim_{\Delta \to 0} \frac{\Pr\,(t < T \leq t + \Delta, K = k | T \geq t)}{\Delta}$$

   Questa misura indica la probabilità che un evento di tipo *k* accada al tempo *t*, dato che il soggetto considerato è vivo al tempo *t*.

2. funzione d'incidenza cumulata (CCI, detta anche subdistribution, funzione di probabilità marginale, funzione d'incidenza cruda, absolute risk causa - specifica).

   Indica la probabilità cumulata di occorrenza di un evento specifico in presenza di eventi competitivi senza l'assunzione di dipendenza tra gli eventi.

$$F_k(t) = \int_0^t S(u) d\Lambda_k\,(u)$$

dove $S(t) = \Pr\,(T > t)$ indica la funzione di sopravvivenza e $\Lambda_k(t) = \int_0^t \lambda_k(u)du$ rappresenta il rischio istantaneo per la *k*-esima causa di morte.

T è il tempo alla morte e $K \,\epsilon\, (1, \dots, n_k)$ è il tipo di causa di morte, dove n_k è il numero di cause di morte di diverso tipo. Si noti che $F_k(t)$ è una funzione impropria dato che $\lim_{t\to\infty} F_k(t) = P(K = k)$ e $\sum_k F_k\,(t) = F(t)$. Per ogni funzione d'incidenza cumulata il $\lim_{t\to\infty} 1 - F_k(t)$ rappresenta la proporzione di soggetti che non sperimenteranno mai un evento di tipo *k*, ovvero la frazione di guariti per ogni tipo di evento studiato (morte per cancro o morte per altro).

Questa misura è diventata di fondamentale importanza in campo medico per le analisi di costo-efficacia in cui le probabilità di sopravvivenza sono necessarie per determinare l'utilità del trattamento.

1.4. Metodi di stima della CCI

La CCI non può essere stimata semplicemente con il metodo non-parametrico di Kaplan-Meier perciò sono stati introdotti dei metodi non parametrici "ad hoc". Uno dei vantaggi di questi approcci non parametrici è che non vi è alcuna necessità di assumere una distribuzione per la funzione d'incidenza cumulata. Quando si osserva un plateau nella CCI per cancro alla mammella, allora si può ipotizzare l'esistenza di una frazione di guariti, che in altri termini significa che non tutte le donne con cancro al seno muoiono per esso. Però è necessario un modello parametrico per sviluppare un test statistico formale per l'esistenza del plateau.
Jeong e Fine nel 2006 propongono un nuovo approccio che consiste nella parametrizzare direttamente la funzioni d'incidenza cumulata tramite una distribuzione impropria di Gompertz. Il termine "improprio" significa che la $F(t)$ non raggiunge 1 per $t \to \infty$. La distribuzione di Gompertz è definita da due parametri: τ il parametro di forma e ρ il parametro di scala, che determinano la forma del rischio di base. In particolare, se ρ assume valori negativi si ha una distribuzione impropria e la presenza di un plateau. A scopo inferenziale sono stati calcolati gli intervalli di confidenza al 95% per il parametro ρ. La quantità $e^{\frac{\tau}{\rho}}$ rappresenta la frazione di pazienti guariti (nel caso in cui ρ è negativo). Al fine di includere le caratteristiche dei pazienti nell'analisi delle CCI si è avvalsi dei transformation models.

1.5. Stima parametrica della CCI e transformation models

Utilizzare un transformation model permette di modellare la funzione d'incidenza cumulata in funzione di covariate modellando il rischio base tramite una funzione parametrica.
Il transformation model assume tale forma:

$$g_k\{(F_k(t;Z)\} = u_k(t) + Z^T\beta_k$$

dove $u_k(t)$ è una funzione monotona crescente e invertibile pari a $u_k(t) = \log_k \left[\int_0^t \lambda_{k0}(s)ds\right]$, β_k è un vettore di parametri di dimensione *P*x*1* e Z è il vettore *P*x*1* di covariate.

Le funzioni più usate per la g_k sono:

1. la funzione complementary log-log (Proportional Hazard):

$$g_k\{(F_k(t;Z)\} = log\{-log(1-F_k)\}$$

da cui si ricava

$$\log(1-F_k) = -\exp\,(u_k(t) + Z^T\beta_k)$$

$$F_k = 1 - \exp\,\{-\exp(u_k(t) + Z^T\beta_k)\}$$

2. la funzione logit (Proportional Odds):

$$g_k\{(F_k(t;Z)\} = \log\left(\frac{F_k}{1-F_k}\right)$$

da cui si ricava

$$\log\left(\frac{F_k}{1-F_k}\right) = u_k(t) + Z^T\beta_k$$

$$\left(\frac{F_k}{1-F_k}\right) = exp\{u_k(t) + Z^T\beta_k\}$$

$$F_k = exp\{u_k(t) + Z^T\beta_k\} - F_k exp\{u_k(t) + Z^T\beta_k\}$$

$$F_k + F_k exp\{u_k(t) + Z^T\beta_k\} = exp\{u_k(t) + Z^T\beta_k\}$$

$$F_k = \frac{exp\{u_k(t) + Z^T\beta_k\}}{1 + exp\{u_k(t) + Z^T\beta_k\}}$$

Con questo link il vettore dei parametri *β* è interpretabile come odds ratio.

3. la trasformazione Aranda-Ordaz (classe generale che include anche i precedenti link):

$$g_k(F_k;\alpha_k) = \log\left[\frac{(1-F_k)^{-\alpha_k} - 1}{\alpha_k}\right]$$

da cui si ricava

$$log[\{(1-F_k)^{-\alpha_k} - 1\}/\alpha_k] = u_k(t) + Z^T\beta_k$$

$$F_k = 1 - \{1 + \alpha_k \exp(u_k(t) + Z^T\beta_k)\}^{-\frac{1}{\alpha_k}}$$

a. se $\alpha_k=1$ si ha il modello ad odds proporzionali (si ritorna al link logit) infatti:

$$F_k = 1 - \{1 + \exp(u_k(t) + Z^T\beta_k)\}^{-1}$$

$$F_k = 1 - \frac{1}{1 + \exp(u_k(t) + Z^T\beta_k)}$$

$$F_k = \frac{exp\{u_k(t) + Z^T\beta_k\}}{1 + exp\{u_k(t) + Z^T\beta_k\}}$$

b. se$\alpha_k \to 0$ si ha la situazione ad Hazard Proportionali infatti:

$$F_k = 1 - \{1 + \alpha_k \exp(u_k(t) + Z^T\beta_k)\}^{-\frac{1}{\alpha_k}}$$

$$\lim_{\alpha_k \to 0}\left[1 - \left(1 + \alpha_k exp(u_k(t) + Z^T\beta_k)\right)^{-\frac{1}{\alpha_k}}\right]$$

$$= \lim_{\alpha_k \to 0}\left[1 - e^{-\frac{1}{\alpha_k}\log\left(1+\alpha_k * exp\left(u_k(t)+Z^T\beta_k\right)\right)}\right]$$

$$= \lim_{\alpha_k \to 0}\left[1 - e^{-\frac{1}{\alpha_k}\log\left(1+\alpha_k * exp\left(u_k(t)+Z^T\beta_k\right)\right)}\right]$$

$$= 1 - exp\left(-exp(u_k(t) + Z^T\beta_k)\right)$$

Poiché F_k ha un dominio compreso tra 0 e 1, mentre il modello ha un dominio che si estende su tutto R, le trasformate permettono di riassestare il dominio del modello in [0;1].

1.5.1. Distribuzione Gompertz a 2 parametri

La distribuzione di Gompertz è una distribuzione di probabilità continua introdotta da Benjamin Gompertz nel 1825. Tale distribuzione è spesso applicata per descrivere la distribuzione della durate di vita adulta sia dai demografi che dagli attuari. Anche nel campo dell'analisi della sopravvivenza la Gompertz viene utilizzata in scienze come la biologia e la gerontologia. La funzione di sopravvivenza può essere scritta come:

$$S(t) = exp\left\{-\frac{\tau}{\rho}(e^{\rho t} - 1)\right\}$$

La funzione di distribuzione cumulata è pari a:

$$F(t;\rho,\tau) = 1 - exp\left[\frac{\tau}{\rho}\{1 - \exp(\rho t)\}\right]$$

La funzione hazard è data da

$$\lambda(t) = \frac{dF(t;\rho,\tau)}{dt} \cdot \frac{1}{1 - F(t;\rho,\tau)}$$

$$\frac{dF(t;\rho,\tau)}{dt} = \tau exp\left(\rho t + \frac{\tau}{\rho} - \frac{\tau}{\rho}e^{\rho t}\right)$$

divisa per $1 - F(t;\rho,\tau)$ si ottiene

$$\lambda(t) = \frac{\tau exp\left(\rho t + \frac{\tau}{\rho} - \frac{\tau}{\rho}e^{\rho t}\right)}{exp\left[\frac{\tau}{\rho}\{1 - e^{\rho t}\}\right]} = \tau e^{\rho t}$$

I parametri coinvolti sono $\tau > 0$ e $-\infty < \rho < \infty$. Nel caso in cui ρ sia positivo la distribuzione Gompertz è propria (la sua funzione di sopravvivenza è compresa tra 0 e 1), mentre se ρ è negativo allora la distribuzione è impropria e la funzione di sopravvivenza con il tempo che tende ad infinito $S(t) = e^{\frac{\tau}{\rho}}$.

Come già introdotto in precedenza questa caratteristica rende la distribuzione Gompertz particolarmente adatta ad analisi in presenza di rischi competitivi.

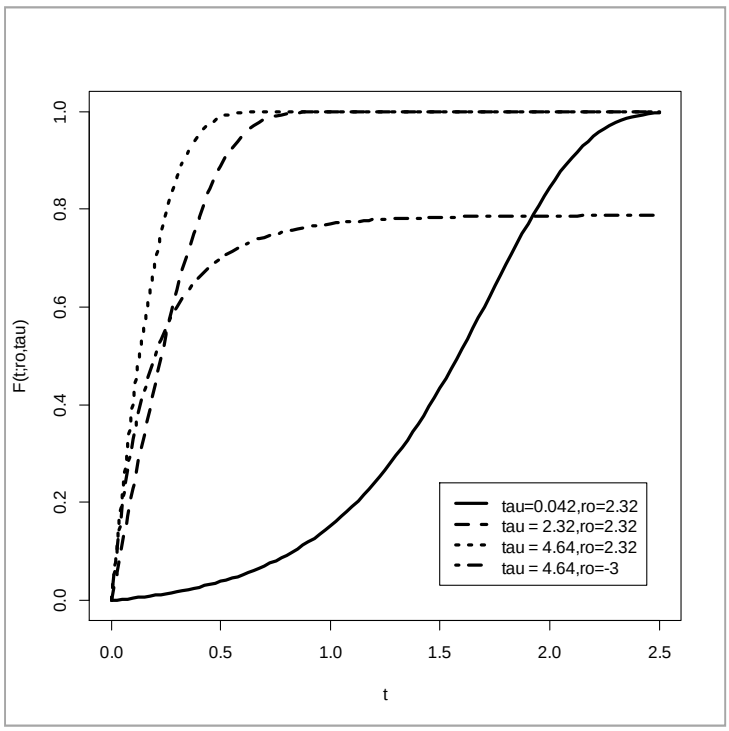


Figura 1. Funzione di distribuzione cumulativa di una Gompertz al variare del parametro τ (parametro di forma) tenendo fisso l'altro parametro (ρ = 2.32; parametro di scala) e con valore del parametro ρ negativo, i valori negativi di ρ implicano una distribuzione impropria.

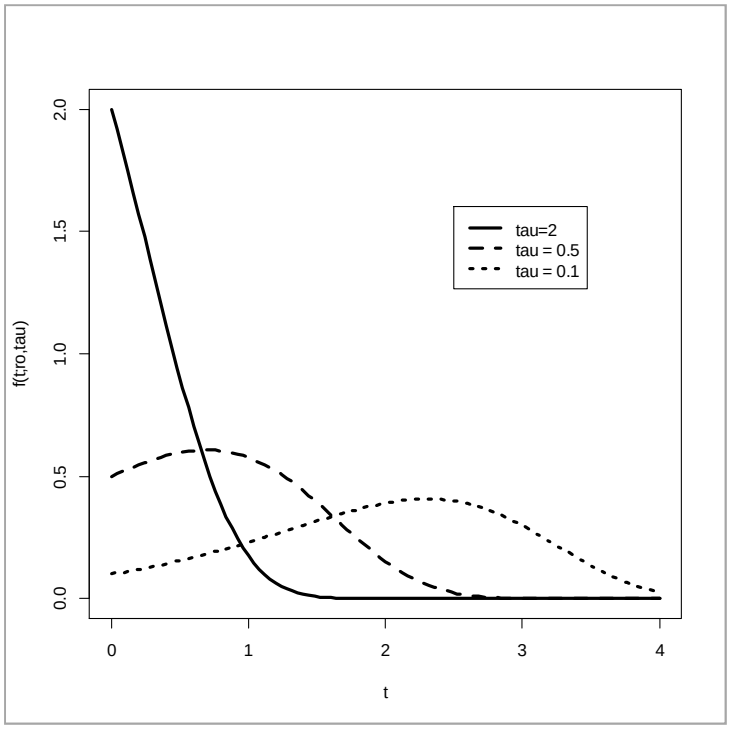


Figura 2. Funzione di densità della Gompertz al variare del parametro τ (parametro di forma) tenendo fisso il parametro di scala ($\rho = 1$).

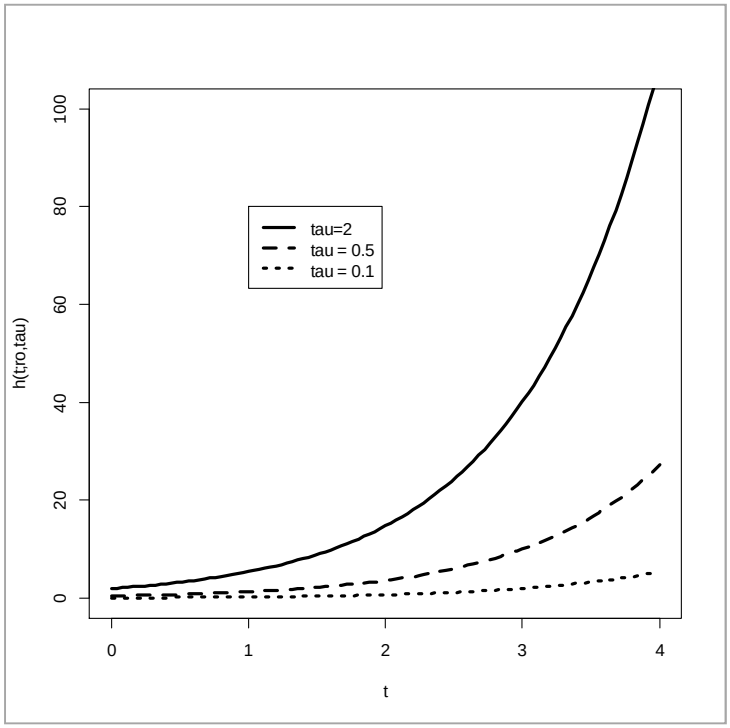


Figura 3. Funzione hazard della Gompertz al variare del parametro τ (parametro di forma) tenendo fisso il parametro di scala ($\rho = 1$).

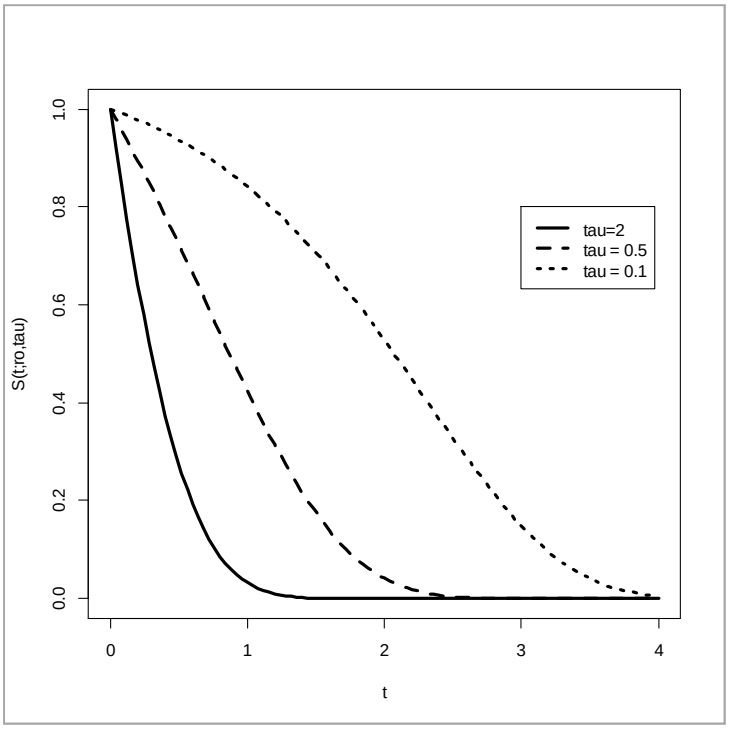


Figura 4. Funzione di sopravvivenza della Gompertz al variare del parametro τ (parametro di forma) tenendo fisso il parametro di scala (ρ = 1).

Nel caso in cui si ipotizzi come funzione hazard, l'hazard della distribuzione Gompertz ne segue che $u_k = \log_k[\int_0^t \tau_k e^{\rho_k t}(s)ds] = \log_k\left[\frac{\tau_k}{\rho_k}(e^{\rho_k t} - 1)\right]$

1. sostituendolo nella CCI con link complementary log-log si ottiene:

$$F_k = 1 - \exp\{-\frac{\tau_k}{\rho_k}(e^{\rho_k t} - 1)e^{Z^T\beta_k}\}$$

e funzione di densità

$$f_k = \tau_k \exp\left(-\frac{\tau_k}{\rho_k}(e^{\rho_k t} - 1) + \rho_k t\right)$$

2. sostituendolo nella CCI con link logit (in assenza di covariate) si ottiene:

$$F_k = \frac{\frac{\tau_k}{\rho_k}(e^{\rho_k t} - 1)}{1 + \frac{\tau_k}{\rho_k}(e^{\rho_k t} - 1)}$$

con funzione di densità pari a:

$$f_k = \frac{\partial F_k}{\partial t} = \frac{\tau_k \mathrm{e}^{\rho_k t}}{(1 + \frac{\tau_k}{\rho_k}(e^{\rho_k t} - 1))^2}$$

3. sostituendolo nella CCI con link Aranda Ordaz si ottiene:

$$F_k = 1 - \{1 + \alpha_k \frac{\tau_k}{\rho_k}(e^{\rho_k t} - 1)e^{Z^T \beta_k}\}^{-\frac{1}{\alpha_k}}$$

Se si calcola il limite della derivata della funzione cumulata d'incidenza nel caso del link Aranda-Ordaz

$$\lim_{\alpha \to 0} \frac{\partial F_k(\alpha)}{\partial t} = \lim_{\alpha \to 0} \frac{1 - \left(1 + \alpha_k e^{(u_k(t) + Z^T \beta_k)}\right)^{-\frac{1}{\alpha_k}}}{\partial t}$$

$$\lim_{\alpha \to 0} \tau_k e^{(Z\beta + \rho_k t)} \left\{1 + \alpha_k \frac{\tau_k}{\rho_k}(e^{\rho_k t} - 1)e^{(Z^T \beta_k)}\right\}^{\frac{-1-\alpha_k}{\alpha_k}} = \tau_k e^{(Z\beta + \rho_k t)} e^{-\frac{\tau_k}{\rho_k} e^{(Z^T \beta_k)}(e^{\rho_k t} - 1)}$$

Se si calcola la derivata della funzione cumulata d'incidenza nel caso del link complementary log-log (cioè con $\alpha_k \to 0$)

$$f_k = \frac{\partial F_k}{\partial t} = \frac{1 - exp\left\{-\frac{\tau_k}{\rho_k}(\mathrm{e}^{\rho_k t} - 1)e^{Z^T \beta_k}\right\}}{\partial t}$$

$$f_k = \frac{\partial F_k}{\partial t} = -exp\left\{-\frac{\tau_k}{\rho_k}(e^{\rho_k t} - 1)e^{Z^T \beta_k}\right\}$$

$$f_k = -e^{\left\{-\frac{\tau_k}{\rho_k}(\mathrm{e}^{\rho_k t} - 1)\mathrm{e}^{Z^T \beta_k}\right\}} \left(-\frac{\tau_k}{\rho_k}\right) \mathrm{e}^{Z^T \beta_k} \mathrm{e}^{\rho_k t} \rho_k$$

$$f_k = -e^{\left\{-\frac{\tau_k}{\rho_k}(\mathrm{e}^{\rho_k t} - 1)\mathrm{e}^{Z^T \beta_k}\right\}} (-\tau_k) \mathrm{e}^{Z^T \beta_k + \rho_k t}$$

$$f_k = \tau_k \mathrm{e}^{(Z^T \beta_k + \rho_k t)} e^{-\frac{\tau_k}{\rho_k} \mathrm{e}^{Z^T \beta_k}(\mathrm{e}^{\rho_k t} - 1)}$$

ne risulta che: $\lim_{\alpha \to 0} \frac{\partial F_k(\alpha)}{\partial t} = \frac{\partial F_k}{\partial t}$

1.5.2. Distribuzione Log-logistica a 4 parametri

Per modellare la funzione d'incidenza cumulata la distribuzione di Gompertz è la più adatta nel caso in cui la forma della funzione dell'hazard sia crescente o decrescente, mentre non è la più

appropriata nel caso di una forma di hazard unimodale. Sono state sviluppate delle distribuzioni parametriche specifiche, più flessibili, per analizzare i tempi di evento ed in particolare nel caso dei rischi competitivi. Una di queste è la distribuzione log-logistica a 4 parametri (un'estensione della distribuzione log-logistica a 2 parametri) che permette di tenere conto di differenti forme dell'hazard[18]. Anche in questo caso la distribuzione può essere impropria.

La funzione di sopravvivenza della distribuzione log-logistica a 2 parametri è:

$$S(t) = \frac{1}{1+\lambda t^{\tau}}$$

dove $\lambda > 0$ e $\tau > 0$ e rappresentano rispettivamente il parametro di scala e di forma. Se $\tau \leq 1$ la funzione hazard decresce monotonicamente, in caso contrario se $\tau > 1$ la funzione hazard è unimodale.

La funzione di sopravvivenza a 2 parametri appartiene ad una famiglia più ampia (distribuzioni di Hougaard) che assume la seguente forma:

$$S(t) = e^{\left\{-\frac{\upsilon\theta^{\alpha}}{\alpha}\left[\left(\frac{H}{\theta}+1\right)^{\alpha}-1\right]\right\}}$$

dove H è la funzione hazard cumulata. Se al posto di H venisse usata una funzione hazard cumulata log-logistica a 2 parametri allora si otterrebbe una nuova distribuzione impropria.

Viene utilizzata la subdistribution $\upsilon = \theta^{2-\alpha}$ per ridurre il numero dei parametri.

$$S(t;\lambda,\tau,\theta,\alpha) = e^{\left\{-\frac{\theta^2}{\alpha}\left[\left(\frac{\log(1+\lambda t^{\tau})}{\theta}+1\right)^{\alpha}-1\right]\right\}}$$

e il dominio dei parametri è il seguente: $\theta > 0$, $\lambda > 0$, $\tau > 0$,$-\infty < \alpha < \infty$.

La funzione di sopravvivenza è compresa tra 0 e 1, se $\alpha < 0$ la funzione risulta impropria.

Questa è un'importante caratteristica della CCI che non si ritrova nella distribuzione log-logistica a 2 parametri e nelle altre distribuzioni.

La funzione di distribuzione cumulata è pari a $1 - S$:

$$F_k(t;\lambda_k,\tau_k,\theta_k,\alpha_k) = 1 - e^{\left\{-\frac{{\theta_k}^2}{\alpha_k}\left[\left(\frac{log(1+\lambda_k t^{\tau_k})}{\vartheta_k}+1\right)^{\alpha_k}-1\right]\right\}}$$

La funzione hazard è:

$$h_k(t;\lambda_k,\tau_k,\theta_k,\alpha_k) = \frac{-\frac{d}{dt}S(t)}{S(t)} = \frac{\theta_k\tau_k\lambda_k t^{\tau_k-1}}{1+\lambda_k t^{\tau}}\left[\frac{\log(1+\lambda_k t^{\tau_k})}{\theta_k}+1\right]^{\alpha_k-1}$$

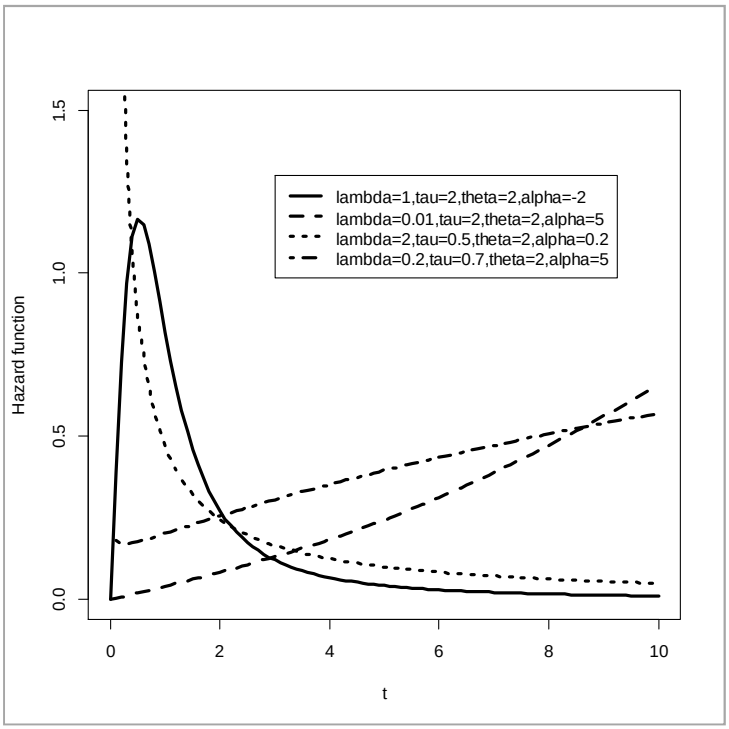


Figura 5. Funzione hazard della Log-logistica al variare dei 4 parametri.

Tale funzione è molto più flessibile rispetto a quella della distribuzione a 2 parametri, infatti può essere sia monotona crescente, che decrescente, unimodale o a forma di U.

1.6. Stime di massima verosimiglianza della CCI

Il metodo è stato proposto da Jeong e Fine nel 2006.

Siano T_i e C_i rispettivamente i potenziali tempi all'evento e i potenziali tempi di censura per l'i-esimo soggetto. Si definisce $X_i = \min(T_i, C_i)$. Per semplicità si ipotizza l'esistenza di due soli eventi competitivi: l'evento in studio $k = 1$ e l'evento competitivo $k = 2$ e un campione di dimensione *n*. Se l'evento *k*-esimo si verifica come primo evento allora $\delta_{ki} = 1$, altrimenti $\delta_{ki} = 0$ $(k = 1, 2,\ i = 1, 2, \dots, n)$. Questo metodo non assume l'indipendenza tra i diversi tipi di eventi e non richiede distribuzioni improprie.

La funzione di verosimiglianza è (in presenza anche di più di 2 eventi):

$$L_{JF} = \prod_{i=1}^{n}\left[\left\{\prod_{k=1}^{n_K} f_k(x_i, \psi_k; z_i)^{\delta_{ki}}\right\}\left\{1 - \sum_{k=1}^{n_K} F_k(x_i, \psi_k; z_i)\right\}^{1-\sum_{k=1}^{n_K}\delta_{ki}}\right]$$

Si noti come in tale verosimiglianza vengono incluse informazioni per tutti i tipi di eventi considerati e non fattorizza in parti separate per ogni tipo di evento.

Il logaritmo della funzione di verosimiglianza è:

$$log(L_{JF}) = \sum_{i=1}^{n}\left[\sum_{k=1}^{n_k}\delta_{ki} log\{f_k(x_i\,,\psi_k;z_i)\} + \left(1-\sum_{k=1}^{n_K}\delta_{ki}\right) log\left\{1-\sum_{k=1}^{n_K}F_k(x_i\,,\psi_k;z_i)\right\}\right]$$

$$k = 1,\dots,n_K$$

$f_k(x_i\,,\psi_k;z_i)$ è la funzione di densità, $F_k(x_i\,,\psi_k;z_i)$ è la funzione d'incidenza cumulata, $Z_i = z_i$ è il vettore delle covariate e ψ_kè il vettore dei parametri.

Nel caso di k=2 la verosimiglianza si riduce a:

$$\prod_{i=1}^{n} f_1(x_i\,,\psi_1;z_i)^{\delta_{1i}} f_2(x_i\,,\psi_2;z_i)^{\delta_{2i}}\{1-F_1(x_i\,,\psi_1;z_i)-F_2(x_i\,,\psi_2;z_i)\}$$

Per l'ottimizzazione della funzione di verosimiglianza si utilizza solitamente il metodo di Newton-Raphson. Le stime di massima verosimiglianza $\widehat{\psi}_k$ soddisfino alcune condizioni di regolarità, come la consistenza e l'asintoticità normale.

Lo stimatore di massima verosimiglianza della funzione d'incidenza cumulata è $F_k\left(x_i\,,\hat{\psi}_k;z_i\right)$.

Utilizzando le derivate seconde rispetto a $\psi_k$del logaritmo della funzione di verosimiglianza è possibile ottenere la matrice d'informazione osservata e applicando il delta method la matrice di varianze e covarianze di $F_k\left(x_i\,,\hat{\psi}_k;z_i\right)$ è:

$$\widehat{var}(F_k\left(x_i\,,\hat{\psi}_k;z_i\right)) = \left(\frac{\partial F_k(x_i\,,\psi_k;z_i)}{\partial\psi_k}\right)\Bigg|_{\psi_k=\hat{\psi}_k} \widehat{var}(\hat{\psi}_k)\left(\frac{\partial F_k(x_i\,,\psi_k;z_i)}{\partial\psi_k}\right)^T\Bigg|_{\psi_k=\hat{\psi}_k}$$

dove $\frac{\partial F_k(x_i\,,\psi_k;z_i)}{\partial\psi_k}$ è il vettore delle derivate prime della funzione d'incidenza cumulata per la k-esimo evento rispetto a ψ_k. La matrice $\widehat{var}(\hat{\psi}_k)$ corrisponde alla varianza di $\hat{\psi}_k$, valutata in $\hat{\psi}_1,\dots,\hat{\psi}_{n_k}$.

L'intervallo di confidenza al 95% per $F_k(t;z)$ è:

$$F_k\left(x_i\,,\hat{\psi}_k;z_i\right) \pm 1.96\ \times\sqrt{\widehat{var}(F_k\left(x_i\,,\hat{\psi}_k;z_i\right)}$$

Il delta method può essere applicato anche per stimare la varianza della frazione stimata dei guariti,$\widehat{var}(FC)$

Se la frazione di guariti stimata è $\phi(\psi)$, allora:

$$\widehat{var}(\phi(\psi)) = \left(\frac{\partial\phi(\psi)}{\partial\psi}\right)\Big|_{\psi_k=\widehat{\psi}_k} \widehat{var}(\widehat{\psi}) \left(\frac{\partial\phi(\psi)}{\partial\psi}\right)^T\Big|_{\psi_k=\widehat{\psi}_k}$$

L'approccio solitamente usato prima di quello introdotto da Jeong e Fine assume che gli eventi competitivi siano incorrelati e la verosimiglianza è il prodotto di 2 (se $k = 2$) verosimiglianze causa-specifica:

$$L_{CS} = \prod_{i=1}^{n} f_1(t_i, \psi_1)^{\delta_{1i}} f_2(t_i, \psi_2)^{\delta_{2i}} S_1(t_i, \psi_1)^{1-\delta_{1i}} S_2(t_i, \psi_2)^{1-\delta_{2i}}$$

In questo caso si ipotizza che tutte le distribuzioni siano proprie.

$$log(L_{CS}) = \sum_{i=1}^{n} \delta_{1i} log\left(f_1(t_i, \psi_1)\right) + \delta_{2i}\, log\left(f_2(t_i, \psi_2)\right) + (1-\delta_{1i}) log\left(S_1(t_i, \psi_1)\right) + (1-\delta_{2i}) log\left(S_2(t_i, \psi_2)\right)$$

dove ψ_k è il vettore dei parametri per l'evento *k*-esimo, $S_k(t_i, \psi_k)$è la funzione di sopravvivenza per l'evento *k*-esimo e $f_k(t_i, \psi_k)$è la funzione di densità per il *k*-esimo evento.
La matrice di varianza e covarianza è l'inversa della matrice di Fisher,$I^{-1}(\psi_1, \psi_2)$.

1.7. Metodi di validazione del modello e Calibrazione

I modelli di regressione multipla sono strumenti usati frequentemente in studi con outcome clinici. Questi modelli possono utilizzare sia variabili nominali che continue e sono in grado di gestire risposte con censura. Tuttavia, una loro applicazione acritica può produrre modelli che mal si adattano ai dati in studio o che mal prevedono gli effetti sui nuovi soggetti. E' fondamentale saper misurare la bontà di adattamento di un modello al fine di evitare modelli non appropriati al fenomeno che si vuole descrivere ed evitare inoltre l'overfitting.
Misurare l'accuratezza predittiva può essere difficile soprattutto per dati di sopravvivenza in presenza di censura. Esiste un indice facilmente interpretabile di discriminazione predittiva, nonché vari metodi per valutare la calibrazione delle probabilità di sopravvivenza predette[19]. Prima di utilizzare le previsioni su una nuova serie di dati, l'accuratezza predittiva deve essere valutata utilizzando il metodo bootstrap o di cross-validation.
L'accuratezza della stima prognostica è importante per molteplici aspetti.
In primo luogo perché tali stime possono essere utilizzate per informare il paziente circa gli esiti probabili della sua malattia, in secondo luogo il medico può utilizzarle come guida per ordinare

test aggiuntivi e per selezionare terapie appropriate. Un ricercatore potrebbe voler stimare l'effetto di un singolo fattore (per esempio, il trattamento somministrato) sulla prognosi in uno studio osservazionale in cui sono misurati anche molti fattori confondenti incontrollati. In questo caso l'azione contemporanea delle variabili incontrollate deve essere controllata (mantenuta costante matematicamente se si utilizza un modello di regressione) in modo che l'effetto del fattore di interesse possa essere stimato in modo efficiente ed evitando i problemi di confondimento. La stima prognostica è utile nella progettazione di studi clinici randomizzati. Sia la decisione relativa a quali pazienti randomizzare e al disegno del progetto di randomizzazione (per esempio, randomizzazione stratificata utilizzando fattori prognostici) sono facilitati dalla disponibilità di stime prognostiche accurate prima della randomizzazione. Infine, modelli prognostici accurati possono essere utilizzati per verificare differenze di beneficio terapeutico o stimare il beneficio clinico per un singolo paziente in uno studio clinico, tenendo conto del fatto che i pazienti a basso rischio devono avere benefici in termini assoluti (ovvero minore variazione di probabilità di sopravvivenza). I modelli possono essere non accurati essenzialmente per:

- ✓ una violazione d'ipotesi;
- ✓ l'omissione di predittori importanti;
- ✓ un'alta frequenza di dati mancanti e/o metodi di imputazione impropri;
- ✓ overfitting, soprattutto in piccoli dataset.

Per descrivere l'accuratezza predittiva del modello si utilizzano la calibrazione e la discriminazione. La calibrazione si riferisce al bias. Ad esempio, se la mortalità media prevista per un gruppo di pazienti simili è 0.3 e l'effettiva proporzione è 0.3 allora le previsioni sono ben calibrate. La discriminazione misura la capacità di un predittore d'individuare pazienti con risposte differenti.

E' probabile che molti modelli clinici non validati non si adattino bene ad una nuova serie di dati, perché l'overfitting è un problema comune. I metodi principali per ottenere validazioni interne sono:

- ✓ il data-splitting

  Una parte casuale, per esempio 2/3 del campione viene usata per sviluppare il modello (trasformazioni di dati, selezione delle variabili, stima dei coefficienti di regressione, etc). Il modello che si ottiene viene poi applicato al campione rimanente per le statistiche di calibrazione.

Il data-splitting è semplice, perché tutte le fasi di modellazione, che possono includere valutazioni soggettive, si fanno solo una volta.

- ✓ la cross-validalion è una ripetizione del data-splitting

  preso un campione di dati, esso viene suddiviso in sottoinsiemi alcuni dei quali vengono usati per la costruzione del modello (*training sets*) e gli altri da confrontare con le predizioni del modello (insiemi di validazione, *validation sets*). Mediando la qualità delle predizioni tra i vari insiemi di validazione si ha una misura dell'accuratezza delle predizioni.

- ✓ il bootstrap

  è un metodo alternativo di validazione interna che prevede il campionamento con reinserimento di un dato numero di campioni dal campione originale. Lo scopo è quello di ottenere stime robuste.

1.7.1. Bootstrap

La procedura *bootstrap* è stata elaborata da Bradley Efron alla fine degli anni '70 e consiste nella generazione di campioni a partire dai dati del campione originale.

Si utilizza la simulazione per scopi inferenziali, nella fattispecie in ambito frequentista. L'idea base è quella di valutare qualche proprietà, ad esempio la varianza, di uno stimatore, o di un'altra procedura statistica, attraverso il ricampionamento dal campione osservato.

Assumiamo di avere un campione $x = (x_1, \dots, x_n)$ da una distribuzione ignota F e di avere una stima, $\hat{\theta}(x)$, di un qualche parametro della popolazione θ (ad esempio θ può essere la media di F, la media del quadrato o un qualche quantile di F). Se potessimo ottenere altri campioni di numerosità n dalla popolazione (e cioè da F), potremmo calcolare diverse stime di θ (una per ogni campione) e la varianza campionaria di queste stime sarebbe la varianza dello stimatore $\hat{\theta}(X)$, $var_F\{\hat{\theta}(X)\}$. In realtà, si dispone solamente del campione osservato, quindi l'idea del bootstrap è di simulare, invece che dalla popolazione, dall'unico campione che abbiamo a disposizione.

Nel nostro caso usiamo il metodo del ricampionamento Bootstrap per valutare l'ottimismo delle stime di calibrazione. Di seguito si descrivono i passi seguiti per ottenere le stime dell'ottimismo nel caso della distribuzione di Gompertz:

1. Si sviluppa il modello di Gompertz utilizzando tutti i soggetti presenti nel trial.

2. Si calcolano le incidenze crude cumulate predette dal modello utilizzando i tempi di evento dei soggetti.
3. Si calcolano i quartili dei valori predetti ottenuti nel passo precedente.
4. Si calcolano le 4 medie dei valori predetti in ogni quantile.
5. Si ottengono le stime non-parametriche delle incidenze crude cumulate sulla base del campione originale.
6. Si calcolano incidenze crude cumulate non parametriche per ogni paziente utilizzando i tempi di evento dei soggetti.
7. Si calcolano le medie delle previsioni non parametriche in ognuno dei gruppi definiti al punto 3.
8. Si ottengono le differenze tra le medie delle previsioni parametriche e le medie non-parametriche.
9. Si estrae un campione bootstrap.
10. Si sviluppa il modello di Gompertz utilizzando il campione bootstrap.
11. Si ottengono le stime parametriche predette dal modello sulla base del campione bootstrap.
12. Si calcolano le medie dei valori predetti appena trovati sulla base dei quantili del punto 4.
13. Si sviluppa il modello non-parametrico sulla base del campione bootstrap.
14. Si calcolano le medie dei predetti appena trovati sulla base dei gruppi al punto 14.
15. Differenze tra le medie parametriche del campione bootstrap e le medie non-parametriche sempre del campione bootstrap (delta 1).
16. Si calcolano i valori predetti sulla base delle stime del modello Gompertz fatto sul bootstrap sui soggetti del campione originale.
17. Si trovano le medie dei predetti al punto precedente sui quartili del punto 4.
18. Differenza tra le medie del punto 20 e quelle del punto 8 (delta 2).
19. Si ripetono i passi dal 10 al 21 200 volte (cioè tanti quanti sono i campioni bootstrap estratti).
20. Si calcolano le medie delle differenze tra delta 1 e delta 2 per i 200 campioni (stime dell'eccesso di ottimismo del modello Gompertz).
21. Si ripetono tutti gli step sul modello Log-logistico.
22. Si ottengono le stime dell'eccesso di ottimismo del modello Log-logistico.

2. Pazienti

2.1. Disegno dello studio

Dal 1973 al maggio del 1980, 701 donne con un tumore al seno di diametro non più grande di 2 cm sono state assegnate in modo casuale o a mastectomia radicale (Halsted; 349 pazienti) o a chirurgia conservativa del seno (quadrantectomia) seguita da radioterapia sul tessuto mammario ipsilaterale (352 pazienti)[20].

La mastectomia radicale introdotta da Halsted era il trattamento scelto per il cancro al seno di qualsiasi tipo o dimensione, indipendentemente dall'età del paziente, fino all'età di 80 anni.

La mastectomia di Halsted è stato eseguita come pensata originariamente per tutto questo periodo a parte alcune modifiche, come quella di allargare l'ampiezza della dissezione per includere i linfonodi mammari interni o ridurla per risparmiare i muscoli pettorali. Non venne mai considerata in quegli anni la possibilità di tentare una procedura chirurgica che potesse conservare il seno.

Nel 1969 venne condotto uno studio randomizzato per confrontare la mastectomia radicale con una chirurgia conservativa del seno, definita "quadrantectomia", con lo scopo di valutare i metodi di diagnosi e i trattamenti del carcinoma mammario. Il reclutamento dei pazienti è iniziato nel 1973 presso l'Istituto Nazionale per lo Studio e la Cura dei Tumori (INT) di Milano. Inseguito la procedura è stata standardizzata e i dati iniziali che mostravano come i tassi di sopravvivenza della mastectomia e della chirurgia conservativa fossero molto simili furono pubblicati nel 1977 e nel 1981. La principale critica che venne mossa fu la preliminarietà dei dati, in quanto sarebbe stato necessario seguire i pazienti con piccoli tumori per un lungo periodo di tempo, anche decenni, per assicurare che la valutazione di efficacia del nuovo trattamento fosse accurata.

Sono stati reclutati soggetti con tumore al seno di diametro massimo di 2 cm (stadio T1) e senza linfonodi ascellari palpabili (N0), mentre sono stati esclusi dallo studio i pazienti con più di 70 anni o che avevano una storia pregressa di cancro.

Le pazienti sono state sottoposte inizialmente ad una biopsia in anestesia generale, e quelle che avevano un carcinoma duttale infiltrante, che non era più di 2 cm di diametro sono state stratificate a seconda dello stato menopausale e assegnate in modo casuale alla sola

mastectomia radicale (Halsted) o una quadrantectomia conservativa del seno in combinazione con la completa dissezione ascellare e radioterapia post-operatoria al tessuto mammario ipsilaterale.

A partire dal 1976 a tutti i pazienti con linfonodi ascellari positivi sono stati dati 12 cicli mensili di chemioterapia secondo il seguente calendario: 100 mg di ciclofosfamide per metro quadrato di superficie corporea al giorno per via orale per 14 giorni e 40 mg di metotrexato per metro quadrato, più 600 mg di fluorouracile per metro quadrato per via endovenosa al 1° e all'8° giorno. La chemioterapia è stata iniziata dopo 15 fino ai 30 giorni dalla mastectomia radicale e contemporaneamente alla radioterapia nel gruppo assegnato alla terapia conservativa del seno. Nessun paziente ha ricevuto il tamoxifen durante il trial. Nei primi 10 anni le pazienti sono state osservate ogni 3 mesi presso la clinica e sono state sottoposte ad un esame completo, incluse radiografie dello scheletro e radiografia toracica, ecografia epatica e mammografia, ogni anno. Successivamente, le pazienti sono stati viste una volta l'anno e sono state sottoposte ogni anno alla mammografia di routine. Ulteriori esami sono stati eseguiti ogni volta che veniva clinicamente indicato. Tre pazienti sono state perse al follow-up. Il follow-up è di 30 anni. I dati principali per tutti i pazienti sono stati registrati, aggiornati e conservati in un sistema automatizzato di dati e successivamente è stata verificata l'accuratezza dei dati.

3. Risultati

3.1. Analisi su popolazione omogenea

In Tabella 1 sono riportate le caratteristiche delle 701 donne colpite da cancro al seno incluse nel Trial Milano 1 sulla base dell'età, del numero dei linfonodi e dello stato menopausale.

Il numero di casi con pre-menopausa e peri/post-menopausa sono piuttosto equilibrati. La maggior parte delle pazienti ha un'età superiore ai 51 anni (47.93%). Delle donne in studio 181 (25.82%) hanno almeno un linfonodo positivo. Considerando le prime due classi di età (<41, 41-51 anni), le donne che sono in stato di pre-menopausa e che non hanno linfonodi positivi sono rispettivamente 70.73% e 61.57%, mentre per quanto riguarda la classe d'età > 51, il 72.62 % non ha linfonodi positivi, ma è in peri/post stato menopausale (come ci si aspettava).

Età	**Stato Menopausale**				
	Pre		Peri+Post		
	Linfo-	Linfo+	Linfo-	Linfo+	**Tot**
< 41	87 *(70.73)* *(33.72)*	32 *(26.02)* *(30.77)*	1 *(0.81)* *(0.38)*	3 *(2.44)* *(3.90)*	123 *(17.55)*
41-51	149 *(61.57)* *(57.75)*	60 *(24.79)* *(57.69)*	17 *(7.02)* *(6.49)*	16 *(6.61)* *(20.78)*	242 *(34.52)*
> 51	22 *(6.55)* *(8.53)*	12 *(3.57)* *(11.54)*	244 *(72.62)* *(93.13)*	58 *(17.26)* *(75.32)*	336 *(47.93)*
Tot	258 *(36.80)*	104 *(14.84)*	262 *(37.38)*	77 *(10.98)*	701

Tabella 1. Numero di pazienti stratificato per età, stato menopausale e numero di linfonodi positivi e percentuali di riga e di colonna.

In Figura 6 sono riportate le curve di sopravvivenza cumulate: osservata, attesa e relativa per le pazienti incluse nel Trial. La curva di sopravvivenza relativa mostra un appiattimento dopo circa

20 anni di follow-up, che dimostra un calo importante nel rischio di morte per la patologia in studio.

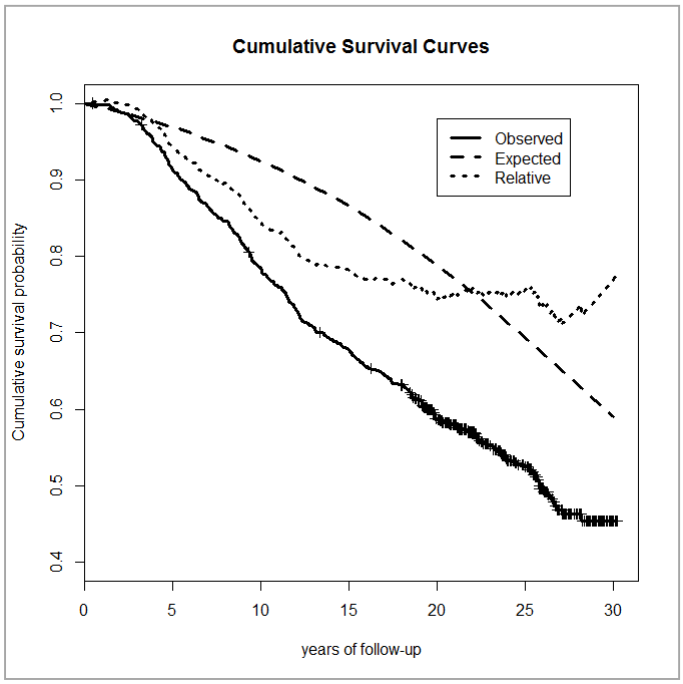


Figura 6. Curve di sopravvivenza cumulate: osservata, attesa e relativa.

In prima analisi è stata considerata una popolazione omogenea e per lo studio è stato utilizzato un "Cure Mixture Model". Secondo i risultati del modello, la frazione dei guariti è stimata intorno al 40% (IC 95%: 31%-50%). Tuttavia, dopo 30 anni, circa il 10% della popolazione non guarita non ha ancora sperimentato l'evento d'interesse. Successivamente è stato applicato un modello parametrico in grado di considerare le cause specifiche di morte.

3.2. Analisi con distribuzione Gompertz

Nella Figura 7 sono riportate le stime parametriche e non parametriche delle funzioni d'incidenza cumulata cruda per morte per cancro al seno e per le altre cause di morte. Dalle stime non parametriche si può vedere che le morti per cancro al seno si verificano per lo più durante i primi 20 anni e poi la curva inizia ad appiattirsi (raggiungendo un plateau). La curva d'incidenza per le altre cause di morte, invece, continua ad crescere nel corso del follow-up. Di conseguenza, sembra adattarsi meglio alla CCI di morte per cancro al seno una funzione impropria, considerando la presenza di plateau, mentre si adatta meglio alla CCI per altre cause di morte

una funzione propria, che non raggiunge il plateau durante il follow-up. Vi è evidenza statistica che una distribuzione di Gompertz impropria sia più adatta per le morti di cancro al seno ($\hat{\rho}$= -0.04445, IC 95%: -0. 05925; -0.02967), mentre una distribuzione propria è più appropriata per morte per altre cause ($\hat{\rho}$ = 0.05674, IC 95%: 0.04382; 0.06958). Secondo le stime ottenute per la CCI di morte per cancro al seno, la percentuale di pazienti guariti da tumore al seno è stimato intorno al 63% [IC 95%: 57.21%, 69.89%]. È da notare che a 20 anni la probabilità di morire per cancro al seno è circa il 23% [IC 95%: 21.31%; 25.53%] ed a 30 anni è del 28% [IC 95%: 25.77%; 30.97%]: dove la CCI di morte per cancro al seno non ancora raggiunto il plateau.

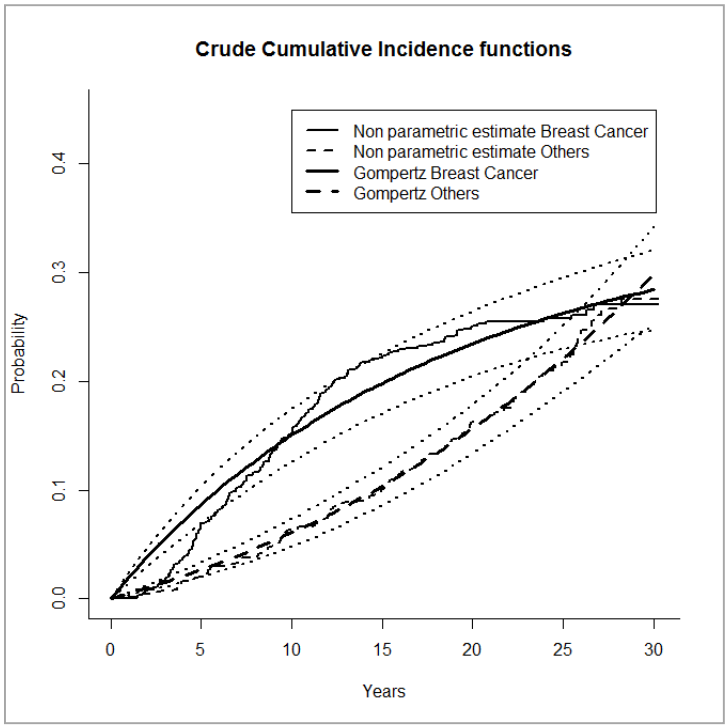


Figura 7. Stime parametriche con distribuzione Gompertz (e intervalli di confidenza al 95%) e non-parametriche delle funzioni d'incidenza cumulate di cancro al seno e di altre cause di morte in assenza di covariate per il Trial Milano 1.

Sono riportate di seguito le stime parametriche e non-parametriche della CCI per morte per cancro al seno e per altre cause per i diversi sottogruppi di pazienti, in Figura 8 per gruppi di età, in Figura 9 per stato menopausale e in Figura 10 per numero di linfonodi. Considerando la variabile età, è evidente che la CCI di morte per cause diverse dal cancro al seno è maggiore per i pazienti anziani, mentre la CCI di morte per cancro al seno è simile tra i due gruppi d'età. Le stime per la classe di età <41 non sono state ottenute (il modello non converge). Infatti, la maggior parte delle donne è morta per cancro al seno e non era possibile stimare la CCI di morte

per altre cause di morte. Le stime della CCI per il pre e peri/post stato menopausale sono simili a quelle per le classi d'età, come previsto a causa dell'associazione esistente tra lo stato menopausale e l'età. La CCI di morte per cancro al seno per le donne con linfonodi positivi è superiore a quello delle donne senza coinvolgimento dei linfonodi, si avvicina al 40% a 20 anni.

Nella Tabella 2 sono riportate le stime di ρ e di τ per la CCI di morte per cancro mammario e di altre cause per i differenti sottogruppi di pazienti. Le stime dei parametri sono molto simili a quello della popolazione generale. Per le donne in peri/post menopausa e con linfonodi negativi la frazione dei guariti è superiore a quella degli altri sottogruppi (circa 68%). Anche per questi sottogruppi di pazienti, la CCI a 30 anni di follow-up non ha ancora raggiunto il plateau stimato, ad eccezione delle pazienti con età maggiore di 51 anni. Il parametro ρ è significativamente inferiore a 0 per tutti i sottogruppi di pazienti considerati. I parametri di scala e di forma della distribuzione di Gompertz hanno lo stesso segno per tutti i sottogruppi di pazienti.

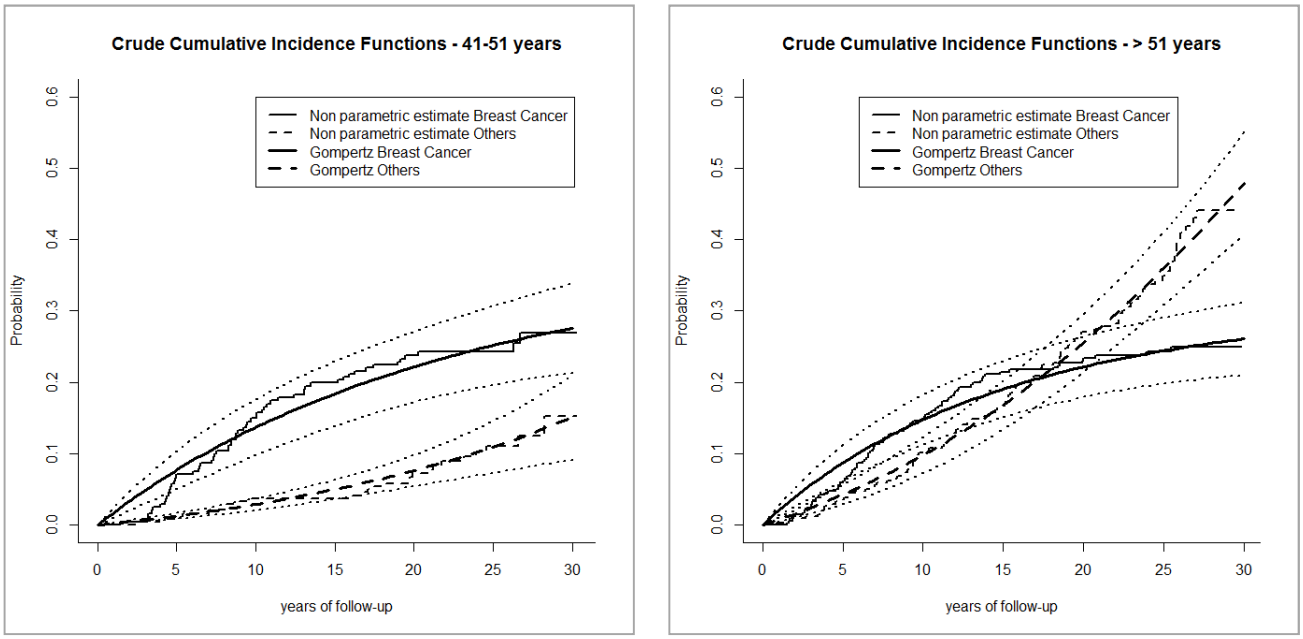
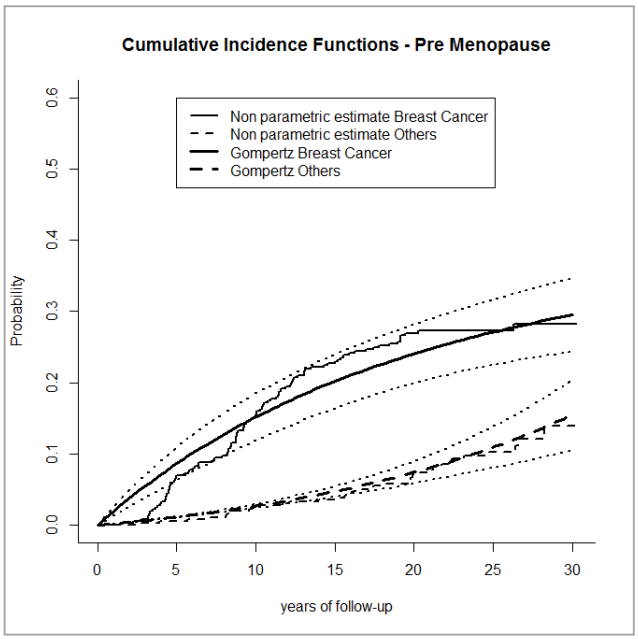


Figura 8. Stime parametriche (Gompertz e intervalli di confidenza al 95%) e non parametriche delle funzioni d'incidenza cumulate di cancro al seno e di altre cause di morte per sottogruppi di donne con differenti età.



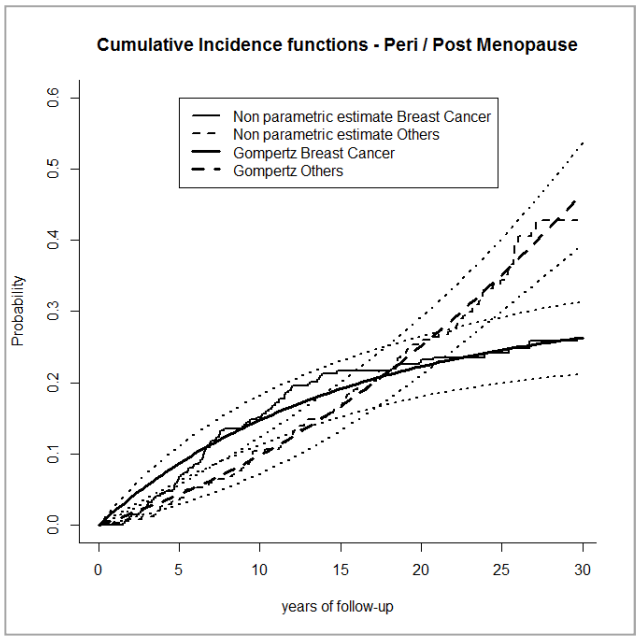
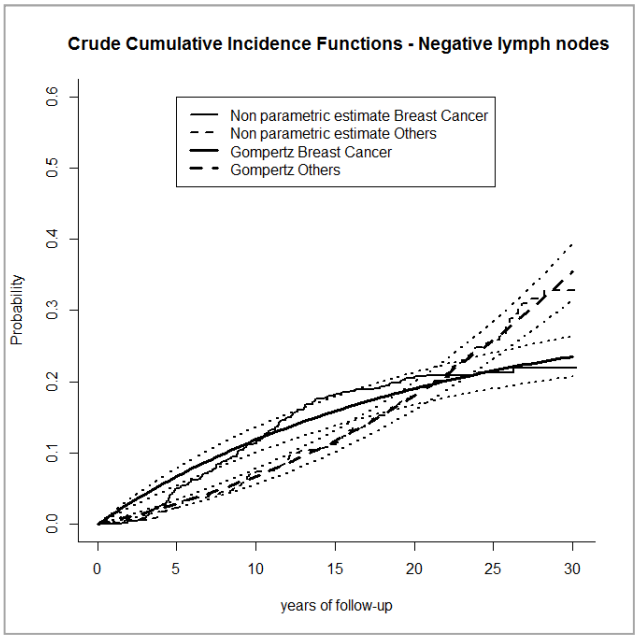


Figura 9. Stime parametriche (Gompertz e intervalli di confidenza al 95%) e non parametriche delle funzioni d'incidenza cumulate di cancro al seno e di altre cause di morte per donne in Pre and Peri/Post stato menopausale.



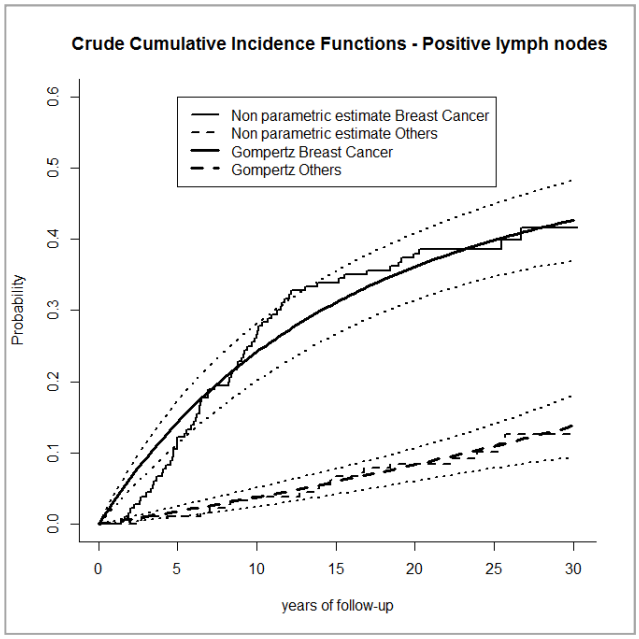


Figura 10. Stime parametriche (Gompertz e intervalli di confidenza al 95%) e non parametriche delle funzioni d'incidenza cumulate di cancro al seno e di altre cause di morte per differenti tipi di linfonodi.

	τ [IC 95%]	ρ [IC 95%]	Frazione non-guariti	% Eventi a 20 anni	% Eventi a 30 anni
Tutta la casistica	0.02015	-0.04445	36.45	23.43	28.38
	[0.01670; 0.02363]	[-0.05925; -0.02967]	[30.107; 42.789]	[21.309; 25.529]	[25.767; 30.963]
41-51 anni	0.01785	-0.03957	36.31	21.86	26.91
	[0.01012; 0.02558]	[-0.07590; -0.00323]	[18.520; 54.093]	[16.957; 26.757]	[20.809; 33.006]
> 51 anni	0.02167	-0.06421	28.64	21.65	25.05
	[0.01402; 0.02932]	[-0.09587; -0.03255]	[21.290; 35.993]	[17.468; 25.838]	[20.218; 29.874]
Pre menopausa	0.02151	-0.04404	38.64	24.87	30.10
	[0.01643; 0.02659]	[-0.06423; -0.02385]	[29.620; 47.658]	[21.869; 27.877]	[26.439; 33.760]
Peri/Post menopausa	0.02080	-0.05507	31.46	22.29	26.31
	[0.01558; 0.02602]	[-0.07760; -0.03254]	[24.743; 38.176]	[19.287; 25.292]	[22.705; 29.922]
Linfonodi Negativi	0.01529	-0.03964	32.00	19.04	23.53
	[0.01187; 0.01871]	[-0.05833; -0.02095]	[23.640; 40.370]	[16.768; 21.304]	[20.682; 26.381]
Linfonodi Positivi	0.03474	-0.04749	51.88	36.15	42.62
	[0.02545; 0.04403]	[-0.07162; -0.02336]	[39.954; 63.812]	[31.450; 40.841]	[36.980; 48.267]

Tabella 2. Stime dei parametri della distribuzione Gompertz, della frazione di non guariti e delle percentuali di eventi a 20 anni e 30 anni per tutta la popolazione e per i differenti sottogruppi di pazienti.

Infine è stato applicato un modello parametrico con distribuzione di Gompertz (proporzionale) che tiene conto dei fattori prognostici. I risultati sono riportati in Tabella 3. L'unico effetto significativo è quello dello stato linfonodale, come previsto.

	Stime	IC 95%	
τ	0.01902	0.01198	0.02606
ρ	-0.04161	-0.06245	-0.02077
41-51 anni	-0.31927	-0.71307	0.07453
> 51 anni	-0.21595	-0.73283	0.30092
Stato Menopausale	-0.00772	-0.45435	0.43890
Tipo di linfonodi	0.76199	0.46174	1.06224

Tabella 3. Stime dei parametri della Gompertz e dei coefficienti delle covariate per tutte le pazienti in studio.

3.3. Analisi con distribuzione Log-logistica a 4 parametri

Nella Figura 11 sono riportate le stime parametriche e non parametriche delle funzioni d'incidenza cumulata cruda per morte per cancro al seno e per le altre cause di morte utilizzando la distribuzione Log-logistica.

Come ci si aspettava tale distribuzione, per il maggior numero di parametri coinvolti, produce stime più simili a quelle non-parametriche rispetto a quelle ottenute tramite la distribuzione di Gompertz, come evidenziato dalla Figura 15.

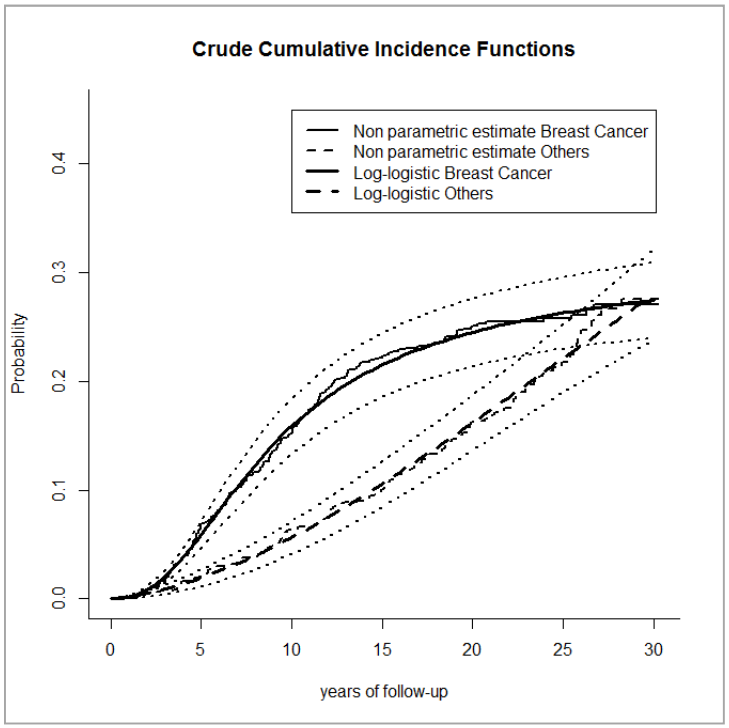
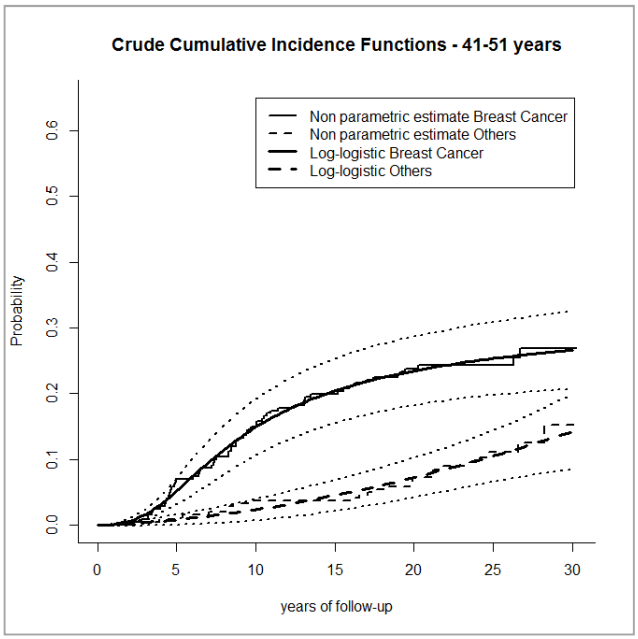


Figura 11. Stime parametriche (Log-logistica a 4 parametri) e non-parametriche delle funzioni d'incidenza cumulate di cancro al seno e di altre cause di morte in assenza di covariate per il Trial Milano 1.



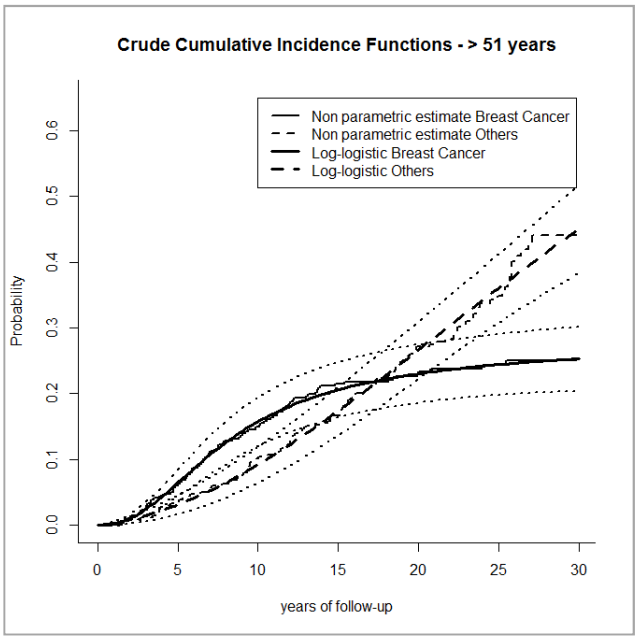
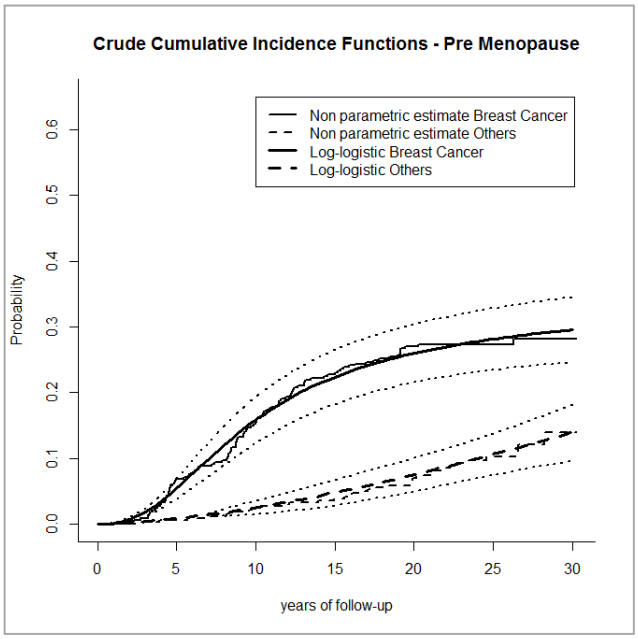


Figura 12. Stime parametriche (Log-logistica e intervalli di confidenza al 95%) e non parametriche delle funzioni d'incidenza cumulate di cancro al seno e di altre cause di morte per sottogruppi di donne con differenti età.



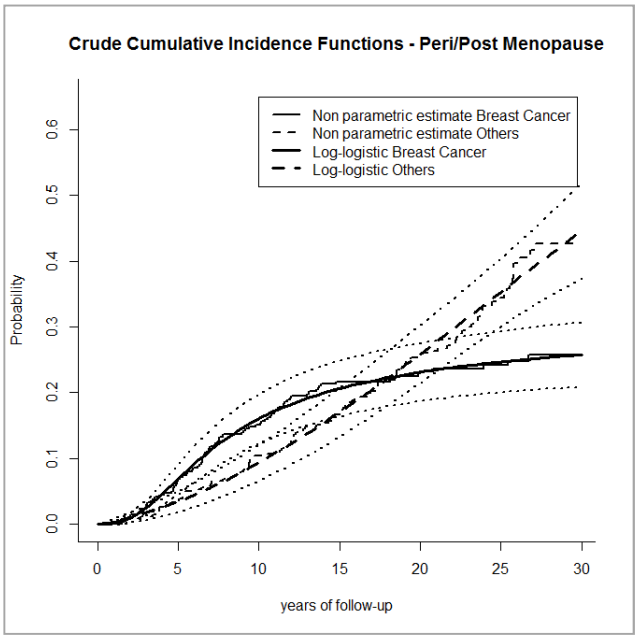
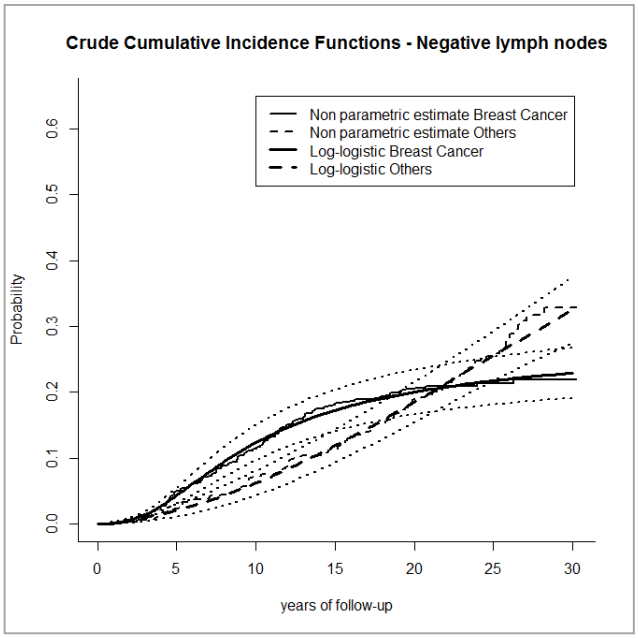


Figura 13. Stime parametriche (Log-logistica e intervalli di confidenza al 95%) e non parametriche delle funzioni d'incidenza cumulate di cancro al seno e di altre cause di morte per donne in Pre and Peri/Post stato menopausale.



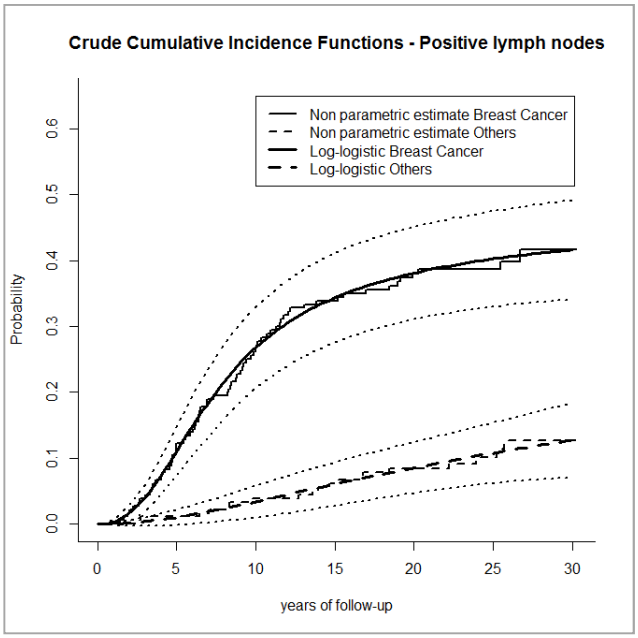


Figura 14. Stime parametriche (Log-logistica e intervalli di confidenza al 95%) e non parametriche delle funzioni d'incidenza cumulate di cancro al seno e di altre cause di morte per differenti tipi di linfonodi.

	θ [IC 95%]	λ [IC 95%]	τ [IC 95%]	α [IC 95%]	% Eventi a 20 anni	% Eventi a 30 anni
Tutta la casistica	0.54371	0.00290	2.42014	-0.58071	24.50	27.45
	[0.17752;0.90989]	[0.00189;0.00389]	[1.91423;2.92604]	[-1.98938;0.82796]	[21.388;27.619]	[23.966;30.936]
41-51 anni	0.41345	0.00223	2.68969	-0.15908	23.47	26.63
	[-0.25630;1.08319]	[0.00141;0.00306]	[1.39005;3.98934]	[-2.40063;2.08247]	[18.227;28.705]	[20.718;32.544]
> 51 anni	0.59407	0.00368	2.28992	-0.93226	23.04	25.31
	[-0.13474;1.32288]	[0.00124;0.00611]	[1.40060;3.17924]	[-4.19027;2.32574]	[18.618;27.461]	[20.421;30.191]
Pre menopausa	0.57752	0.00239	2.42396	-0.57373	25.99	29.57
	[0.15020;1.00484]	[0.00161;0.00316]	[1.90191;2.94601]	[-2.14331;0.99585]	[21.572;30.412]	[24.607;34.523]
Peri/Post menopausa	0.38534	0.00435	2.58315	-0.11385	23.11	25.78
	[-0.20620;0.97688]	[0.00044;0.00825]	[1.28979;3.87651]	[-2.09357;1.86586]	[18.685;27.533]	[20.853;30.711]
Linfonodi Negativi	0.40518	0.00247	2.51161	-0.24452	20.03	22.93
	[-0.01939;0.82975]	[0.00160;0.00333]	[1.74045;3.28277]	[-1.84474;1.35569]	[16.659;23.403]	[19.085;26.782]
Linfonodi Positivi	0.77183	0.00410	2.42038	-0.76259	38.08	41.62
	[-0.25352;1.79717]	[0.00075;0.00745]	[1.39918;3.44159]	[-4.02449;2.49932]	[31.066;45.100]	[34.072;49.176]

Tabella 4. Stime dei parametri della distribuzione Log-logistica e delle percentuali di eventi a 20 anni e 30 anni per tutta la popolazione e per i differenti sottogruppi di pazienti.

Infine è stato applicato un modello parametrico con distribuzione Log-logistica (proporzionale) che tiene conto dei fattori prognostici in studio. I risultati sono riportati in Tabella 5. Anche in questo caso l'unico effetto significativo è quello dello stato linfonodale, come previsto.

	Stime	***IC 95%***	
θ	0.02604	-0.01030	0.06237
α	1.09303	0.84435	1.34170
λ	0.01267	-0.00927	0.03460
τ	3.72999	2.41049	5.04950
41-51 anni	-0.35193	-0.75538	0.05152
> 51 anni	-0.26986	-0.79555	0.25582
Stato Menopausale	0.01691	-0.42964	0.46347
Tipo di linfonodi	0.76720	0.46579	1.06861

Tabella 5. Stime dei parametri della Log-logistica a 4 parametri e dei coefficienti delle covariate per tutte le pazienti in studio.

3.4. Risultati della validazione dei modelli

Per la validazione dei due modelli è stato utilizzato inizialmente un metodo grafico che permette di comparare direttamente le probabilità previste. Si evince come la Log-logistica si adatti molto meglio al modello non parametrico rispetto alla Gompertz per quanto riguarda la funzione d'incidenza cumulata per cancro al seno. Tale differenza è meno evidente per la funzione d'incidenza cumulata per altre cause in cui i due modelli sembrano quasi sovrapporsi.

Utilizzando il ricampionamento bootstrap, si è misurato l'ottimismo relativo alla bontà di adattamento dei due modelli (Gompertz e Log-logistico).

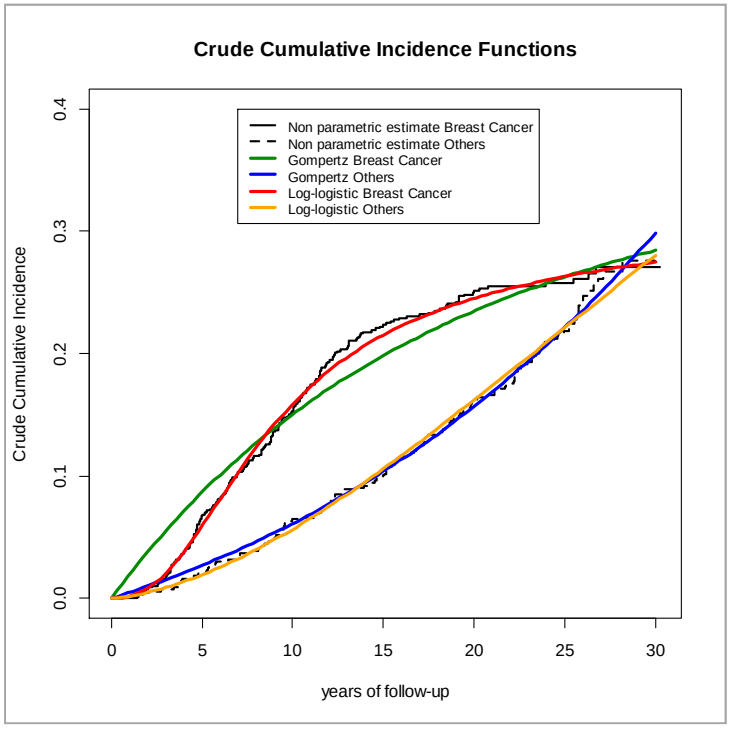


Figura 15. Confronto tra stime parametriche e non-parametriche delle funzioni d'incidenza cumulate di cancro al seno e di altre cause di morte utilizzando la distribuzione Gompertz e la distribuzione Log-logistica (a 4 parametri) in assenza di covariate per il Trial Milano 1.

In Tabella 7 sono riportate le stime dell'eccesso di ottimismo che permettono di ottenere le stime corrette che misurano la bontà di adattamento dei due modelli proposti.

Stime della bontà di adattamento del modello	***1° quartile***	***2° quartile***	***3° quartile***	***4° quartile***
Gompertz	0.1050439	0.2100250	0.2490752	0.2725887
Log-logistica	0.0916883	0.2236553	0.2545517	0.2683634

Tabella 6. Stime della bontà di adattamento del modello senza correzione su 200 campioni bootstrap senza tener conto dei fattori prognostici.

Stime dell'eccesso di ottimismo	***1° quartile***	***2° quartile***	***3° quartile***	***4° quartile***
Gompertz	0.00031269	0.00068845	0.00339261	-0.00501566
Log-logistica	-0.00121025	-0.00302220	0.00043523	-0.00522357

Tabella 7. Stime dell'eccesso di ottimismo calcolate su 200 campioni bootstrap senza tener conto dei fattori prognostici.

In Figura 16 sono riportate le medie delle probabilità predette dai due modelli all'interno di sottogruppi di soggetti (sottogruppi individuati sulla base dei quartili) confrontate con quelle del modello non-parametrico. Successivamente le medie del modello Gompertz e Log-logistico vengono corrette con le stime dell'eccesso di ottimismo che sono ottenute su 200 campioni bootstrap.

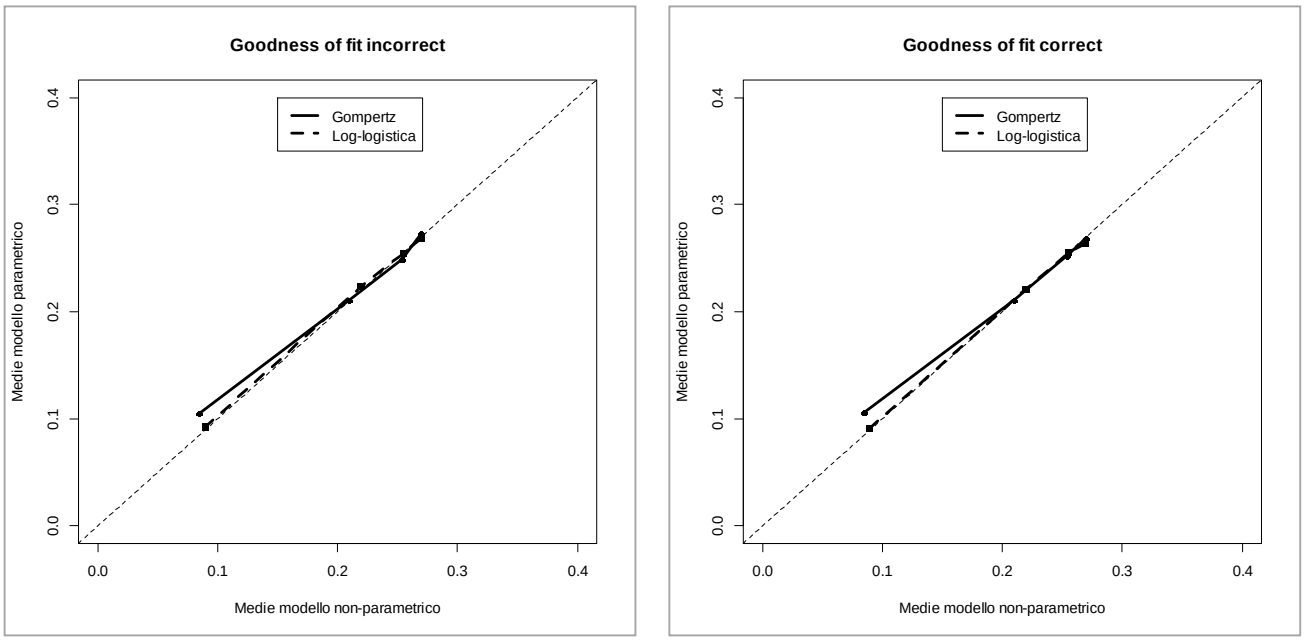


Figura 16. Rette di calibrazione per i due modelli (non corrette e corrette sulla base delle stime dell'eccesso di ottimismo).

4. Conclusioni

La guarigione dal cancro al seno è un argomento molto dibattuto a partire dal lavoro "La curabilità del cancro al seno" di Duncan e Kerr del 1976[21]. L'approccio statistico standard per studiare il problema è quello di cercare un appiattimento nella curva di sopravvivenza relativa cumulata che confronta la mortalità della coorte di donne con cancro al seno con la mortalità della popolazione generale e successivamente applicare un "Cure Mixture Model". Dal momento che la curva di sopravvivenza relativa non raggiunge un vero e proprio appiattimento, la frazione di guariti stimata dal modello è essenzialmente ottenuta attraverso un processo di estrapolazione[22]. Pertanto i tempi di guarigione assumono valori che vanno oltre l'intervallo di tempo di follow-up disponibile nello studio. Sulla base di questo approccio, l'effettiva esistenza di un gruppo di pazienti guarite è discutibile[23].

Un altro approccio considera direttamente la causa specifica di morte e ricerca un plateau nella CCI di morte per cancro al seno (approccio applicato in questo lavoro). Per verificare l'esistenza di un plateau viene utilizzato un modello parametrico in grado di parametrizzare la funzione d'incidenza cumulata tramite una distribuzione impropria[24]. Sulla base delle analisi condotte sul Trial Milano 1 si conclude che le stime ottenute utilizzando una distribuzione impropria della Gompertz si adattano meglio alle stime non parametriche rispetto a quelle ottenute tramite distribuzioni di mistura parametriche. La parametrizzazione diretta era già stata applicata nel carcinoma mammario[13,24] per studiare la prima recidiva, mentre in questo lavoro è stata applicata per modellare le diverse cause di morte[25]. L'applicazione della parametrizzazione diretta della CCI attraverso una distribuzione parametrica rivela una prova dell'esistenza di una frazione di guariti per cancro al seno, infatti, dalle stime ottenute, si nota come la maggior parte delle morti dovute al cancro si concentrano nei primi 20 anni di follow-up. Tuttavia, tra i 20 e i 30 anni, la CCI continua ad aumentare, anche se a un ritmo molto più lento. Questi risultati sono in linea con i risultati precedenti[26]. Gli autori hanno detto: " ... la prognosi delle donne con cancro al seno che sono sopravvissute per 20 anni si avvicina alla normalità ma non la raggiunge mai". Infatti, la percentuale di pazienti guariti è stimata intorno al 63% (mentre di morire per cancro al seno è del 37%), ma viene raggiunta dopo un periodo di follow-up molto più lungo di quello disponibile. A 30

anni la percentuale stimata di decessi è del 28%, e a 40 anni è del 31%. Non è corretto effettuare un processo di estrapolazione del modello al di là del tempo di follow-up osservato. Tutte queste considerazioni mettono in dubbio la possibilità quindi di affermare che esista realmente una guarigione per il cancro mammario. In conclusione, anche se vi è evidenza statistica della presenza di una frazione di guariti, in termini pratici, sembra che dopo 30 anni di follow-up la guarigione non possa ancora essere rivendicata.

5. Problemi Computazionali

5.1. Funzione Optim

Per l'ottimizzazione della funzione di verosimiglianza il software R utilizza la funzione `Optim()` che include cinque metodi per la minimizzazione della funzione:

- ✓ Nelder-Mead
- ✓ BFGS
- ✓ CG
- ✓ SANN
- ✓ L-BFGS-B.

Il metodo che viene utilizzato di default è il metodo Nelder-Mead ideato da John Nelder e Roger Mead nel 1965 ed è un metodo che non fa uso della derivata della funzione da ottimizzare. BFGS, CG, SANN e L-BFGS-B utilizzano invece i gradienti, ma con modalità differenti.
SANN (ideato da Belisle nel 1992) è un metodo di Simulates Annealing. Nell'utilizzo di tali procedure è importante controllare le stime della derivata seconda nel punto di ottimo.
La funzione `Optim` può essere impiegata in modo ricorsivo.

5.1.1. Sann

Il metodo SANN costituisce un approccio, spesso imperfetto della ricerca di un ottimo globale per la funzione considerata. SANN e altri approcci che sono contenuti in R sono approcci di tipo stocastico. Si caratterizzano per essere meno soggetti rispetto agli algoritmi tradizionali a finire il processo di ricerca in un minimo locale.
Il metodo di ottimizzazione "SANN" (Simulated Annealing) è una variante del metodo fornito da Belisle (1992) ed appartiene alla classe dei metodi stocastici di ottimizzazione globale. E' relativamente lento. E' in grado però di lavorare anche con funzioni non-differenziabili. Questa implementazione si avvale della funzione "Metropolis" per la probabilità di accettazione. Per impostazione predefinita, il punto successivo candidato è generato da un kernel gaussiano di

Markov. Nel caso in cui venga fornita la funzione per generare un nuovo punto candidato, il metodo "SANN" è in grado di risolvere problemi di ottimizzazione combinatoria. Uno handicap del metodo "SANN" è che dipende in modo critico dai parametri iniziali che vengono forniti. E' particolarmente indicato quando si è in presenza di una superficie molto "irregolare" e in questo caso permette di ottenere dei buoni risultati.

5.1.2. Nelder-Mead

La tecnica Nelder-Mead è stata proposta da John Nelder & Roger Mead (1965) ed è un algoritmo che minimizza una funzione obiettivo senza necessariamente far uso delle derivate. E' una tecnica di ottimizzazione non lineare ed è la più utilizzata per l'efficienza dimostrata soprattutto per problemi di piccole dimensioni.

Questo metodo si sposta nello spazio delle soluzioni tramite simplessi, cioè figure geometriche.

L'idea che sta alla base dell'algoritmo è quella di cercare di espandere il simplesso se si trovano valori buoni della funzione obiettivo e contrarlo se non se ne trovano. Questo algoritmo è una tecnica euristica, nel senso che non è possibile assicurare la convergenza globale della sequenza prodotta, salvo alcuni casi specifici. Infatti si dimostra che si ha convergenza ad un punto stazionario per funzioni strettamente convesse con una sola variabile, mentre, nel caso di funzioni strettamente convesse con due variabili, si dimostrano risultati di convergenza più deboli. In particolare, sono noti contro-esempi di problemi a 2 variabili, in cui la sequenza generata dal metodo converge ad un punto che non è un punto stazionario.

Tuttavia, questo algoritmo, sebbene non caratterizzato da proprietà teoriche di convergenza, si è rivelato in pratica molto efficiente, in particolare per la soluzione di problemi di dimensioni non superiori alle dieci variabili. Tale metodo è perciò presente in varie librerie standard di ottimizzazione.

5.1.3. BFGS

Il metodo Broyden-Fletcher-Goldfarb-Shanno (BFGS) è un metodo iterativo per la risoluzione di problemi non vincolati di ottimizzazione non lineare.

Il metodo BFGS approssima il metodo di Newton, una classe di tecniche "*hill-climbing*" di ottimizzazione che cerca un punto stazionario di una funzione. Per tali problemi, una condizione

necessaria per l'ottimalizzazione è che il gradiente sia pari a zero. Sia per il metodo di Newton che per il BFGS non è necessaria la convergenza a meno che la funzione ha uno sviluppo quadratico di Taylor in prossimità di un ottimo. Questi metodi si servono sia delle derivate prime che delle derivate seconde. Il BFGS ha dimostrato di avere buone prestazioni anche per ottimizzazioni non regolari.

Nei metodi quasi-Newton, la matrice Hessiana di derivate seconde non deve essere valutata direttamente ma è approssimata utilizzando aggiornamenti specifici per le valutazioni del gradiente. I metodi quasi-Newton sono una generalizzazione del metodo secante per trovare la soluzione della derivata in caso di problemi multidimensionali. In questi casi l'equazione secante non specifica una soluzione unica, e metodi quasi-Newton differiscono nel modo di vincolare la soluzione. Il metodo BFGS è uno dei più utilizzati di questa classe. Di uso comune è L-BFGS, che è una versione di BFGS che è particolarmente adatto per problemi con un gran numero di variabili (> 1000). La variante BFGS-B gestisce problemi di ottimizzazione vincolata.

5.2. Funzione BBoptim

La funzione `BBoptim()` permette di utilizzare in sequenza diverse tecniche di ottimizzazione e di provare in modo automatico una griglia di stime iniziali. L'ottimizzazione della funzione di verosimiglianza è un passaggio molto delicato nell'implementazione del modello parametrico.

Spesso accade che l'algoritmo utilizzato non converga in modo esplicito. In altre situazioni la convergenza sembra essere ottenuta ma le stime di varianza dei parametri non sono ammissibili.

Nel primo caso è utile poter provare una serie di stime iniziali da cui far partire il processo di ottimizzazione ed anche più algoritmi. La funzione `BBoptim` si è rivelata utile nell'implementazione del bootstrap in quanto ha permesso di risolvere in modo automatico l'ottimizzazione sui diversi campioni estratti.

Appendice

Sopravvivenza relativa

In molti casi esiste un problema di affidabilità dei certificati di morte, infatti, è comune l'uso di codifiche generiche come arresto cardiocircolatorio risultando quindi una generale sottostima delle altre cause di morte per i malati di tumore.
In questi casi l'analisi della sopravvivenza con endpoint causa specifica di morte non può essere usata a causa della scarsa qualità dei dati. Per valutare l'impatto della mortalità dovuta al tumore si può ricorrere alle tecniche di sopravvivenza relativa. Queste tecniche permettono di paragonare l'esperienza di sopravvivenza di una coorte in studio con quella attesa sulla base dei tassi di mortalità della popolazione generale. In questo modo si è in grado di stimare l'incremento della probabilità di morte dovuta ad una certa causa rispetto alla mortalità generale.
Tali metodi sono molto usati per i registri tumori.
La funzione di sopravvivenza relativa cumulata è definita come:

$$r(t) = \frac{S_O(t)}{S_P(t)}$$

dove $S_O(t)$ rappresenta la sopravvivenza osservata (in cui tutti i decessi sono considerati eventi) e $S_P(t)$ la sopravvivenza attesa di un gruppo paragonabile della popolazione generale con una bassa incidenza del cancro in studio, stimata sulla base delle tavole di sopravvivenza della popolazione (gruppo paragonabile = sottogruppo della popolazione abbinata ai pazienti rispetto ai principali fattori che possono influire sulla sopravvivenza, nel caso in esame età, sesso e anno di diagnosi). $r(t)$ può assumere qualsiasi numero non negativo e spesso è minore di 1. La funzione di sopravvivenza relativa dipenderà certamente anche da alcune caratteristiche del soggetto come: età, stato menopausale, numero di linfonodi positivi e dimensione iniziale del tumore.

Stima della sopravvivenza osservata

Con il termine sopravvivenza osservata ci si riferisce alla probabilità di sopravvivenza dei soggetti inclusi nello studio, nel nostro caso, donne con una diagnosi di tumore mammario. La funzione di sopravvivenza è in generale definita da:

$$S(t) = Pr(T > t) = 1 - F(t)$$

e si stima con il metodo di Kaplan-Meier, ed indica la probabilità che un soggetto sopravviva almeno fino al tempo *t*.

Stima della sopravvivenza attesa

La sopravvivenza attesa si definisce come la probabilità di sopravvivenza che un paziente avrebbe avuto utilizzando le statistiche di mortalità della popolazione generale e calcolando quindi la probabilità di sopravvivenza attesa di una persona "simile", per genere, età e coorte di nascita.

Per la stima della sopravvivenza attesa vi sono tre metodi [3] che differiscono per la durata per la quale gli individui sono considerati a rischio:

1) ***Ederer I***: ogni soggetto viene considerato a rischio indefinitamente (anche dopo la fine dello studio); il momento in cui il paziente sperimenta l'evento o viene considerato censurato non influisce sulla sopravvivenza attesa. La sopravvivenza attesa viene calcolata come media al tempo *t*:

$$S_P(t) = \sum_{i=1}^{n} \frac{S_{Pi}(t)}{n}$$

dove $S_{Pi}(t)$ è la sopravvivenza attesa di un soggetto della popolazione generale che ha, alla data della diagnosi, le stesse caratteristiche della life table del paziente *i*.

2) ***Ederer II*** (**Condizionale**): un soggetto è considerato a rischio solo fino a quando il corrispondente paziente sperimenta l'evento o è censurato.

$$S_P(t) = \frac{\sum_{i=1}^{n} Y_i(t) S_{Pi}(t)}{\sum_{i=1}^{n} Y_i(t)}$$

dove $Y_i(t) = 1$ se il paziente è a rischio di sperimentare l'evento al tempo *t* e 0 altrimenti.

3) ***Hakulinen***: se il tempo di sopravvivenza di un paziente è censurato, tale censura viene fatta anche per il corrispondente individuo della popolazione generale; però se un paziente muore, il corrispondente soggetto della popolazione generale viene considerato a rischio sino al termine dello studio.

$$S_P(t) = \frac{\sum_{i=1}^{n} C_i(t) S_{Pi}(t)}{\sum_{i=1}^{n} C_i(t)}$$

dove $C_i(t) = 1$ se t è minore o uguale al tempo massimo durante il quale il paziente può essere osservato e 0 altrimenti. Questo metodo tiene in conto della censura informativa a causa di fattori che influenzano la sopravvivenza e contemporaneamente il follow-up potenziale.

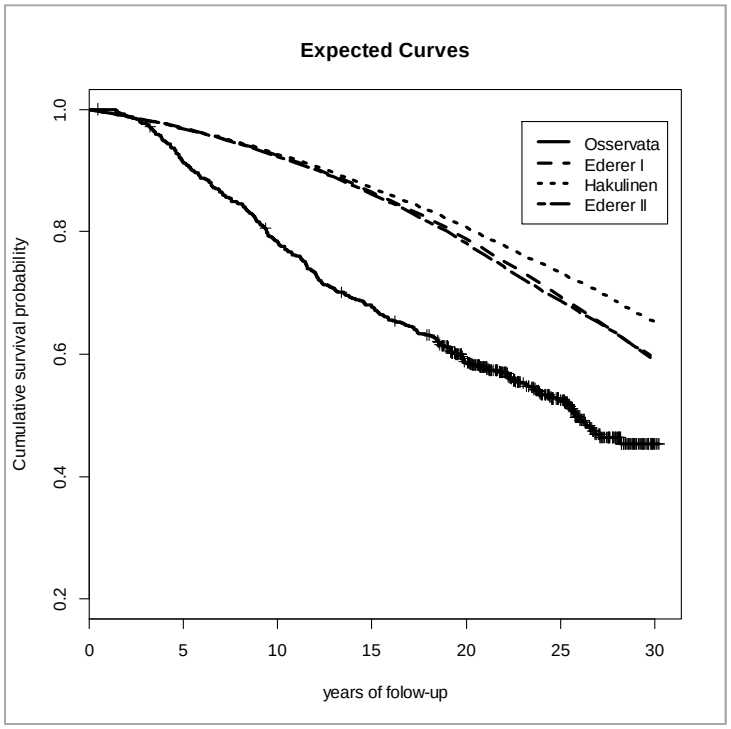
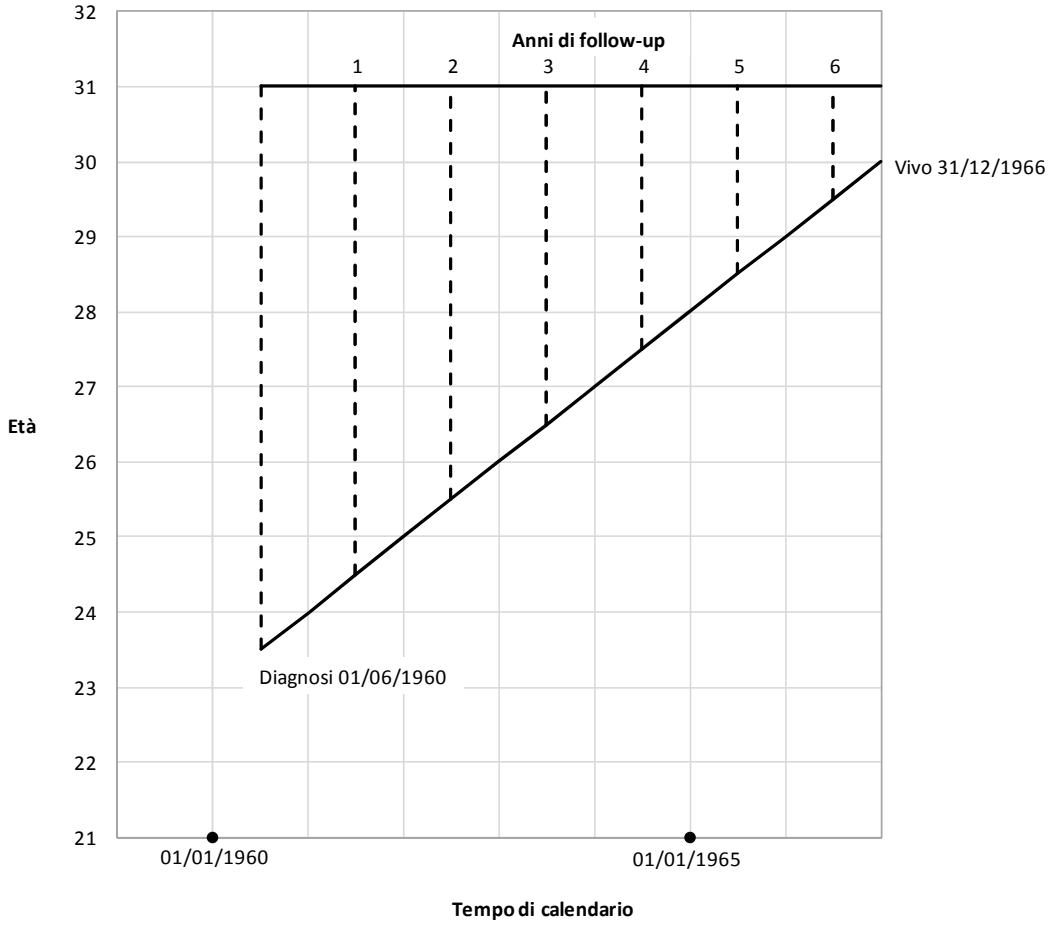


Figura 17. Curve di sopravvivenza attesa cumulata con i tre metodi: Ederer I, Hakulinen e Ederer II e la curva di sopravvivenza generale osservata

Dati necessari per il calcolo del numero di morti attesi e per la stima della sopravvivenza relativa



Per calcolare il numero di morti attesi si procede come segue: ogni soggetto della coorte in studio è rappresentato da un segmento obliquo come mostra il grafico (diagramma di Lexis), dove sull'asse orizzontale si trovano gli anni di calendario, mentre sull'asse verticale l'età del soggetto.
Ad esempio ad un paziente 23-enne viene diagnosticato un tumore il 1 giugno del 1960 e tale paziente viene seguito fino al 31 dicembre del 1966 anno in cui è ancora vivo.
Contribuisce 7/12 anni persona per l'anno 1960 (dato che i mesi in cui il soggetto contribuisce sono solo 7), mentre un anno persona per ogni anno dal 1961 al 1966.
Per ottenere il numero di morti attesi si moltiplica il tasso riferito ad una data età e un dato anno di calendario per il numero di anni-persona sotto osservazione nella coorte e si sommano per tutti i gruppi di età e per tutti gli anni di calendario.

Per il calcolo della sopravvivenza relativa è necessaria l'informazione sulla probabilità di sopravvivenza della popolazione generale di riferimento (le cosiddette "life tables") che verrà confrontata con la sopravvivenza delle pazienti incluse nella casistica in analisi.

Le informazioni relative al follow-up dei pazienti inclusi nello studio deve contenere necessariamente le seguenti variabili:

- data di inizio dell'osservazione (es. data di diagnosi del tumore al seno);
- data di termine dello studio, o alternativamente data dell'evento o di perdita al follow-up;
- stato alla data di uscita: 0=vivo o perso al follow-up, 1=decesso (status);
- codice identificativo del paziente (id);
- eventuali variabili di stratificazione.

Il file "life tables" deve contenere le probabilità di sopravvivenza della popolazione generale di riferimento, stratificate per tutte le variabili da cui solitamente dipende la sopravvivenza attesa.

I dati a disposizione sono dati relativi a 3 trial clinici condotti su pazienti con cancro al seno reclutati in un periodo totale compreso tra il 1973 e il 1989.

Per ottenere le "life table" della popolazione italiana è possibile rivolgersi o agli uffici nazionali di statistica, o se è necessario avere tali tavole in un formato specifico come quello richiesto da R si può fare riferimento a specifici siti web che forniscono tavole per vari paesi in un formato uniforme. Uno di questi siti è il "human mortality database" (HMD, http://www.mortality.org) e include 26 paesi tra cui l'Italia.

Di seguito viene riportata una parte di tavola relativa alle sole donne che è stata utilizzata nel presente lavoro. Le colonne incluse sono:

- Year: anno considerato;
- Age: età dei soggetti;
- m(x): tasso di morte tra l'età *x* e l'età *x*+*n* (in questo caso *n*=1);
- q(x): probabilità di morte tra l'età *x* e l'età *x*+*n*;
- a(x): durata media di sopravvivenza fra le età x e x + n per le persone che muoiono nell'intervallo;
- l(x): numero di soggetti che sopravvivono all'età *x*, assumendo l(0)=100.000;
- d(x): numero di soggetti che muoiono tra l'età *x* e l'età *x*+*n*;
- L(x): numero di anni persona vissuti tra l'età *x* e l'età *x*+*n*;
- T(x): numero di anni rimanenti dopo l'età *x* ;
- e(x): aspettativa di vita (in anni) all'età *x*.

Year	Age	m(x)	q(x)	a(x)	l(x)	d(x)	L(x)	T(x)	e(x)
…	…	…	…	…	…	…	…	…	…
1900	0	0.18212	0.16284	0.35	100000	16284	89415	4180475	41.8
1900	1	0.06743	0.06523	0.5	83716	5461	80985	4091060	48.87
1900	2	0.04746	0.04636	0.5	78255	3628	76441	4010074	51.24
1900	3	0.03122	0.03074	0.5	74627	2294	73479	3933634	52.71
1900	4	0.02029	0.02009	0.5	72332	1453	71606	3860154	53.37
1900	5	0.01379	0.01369	0.5	70879	971	70394	3788549	53.45
1900	6	0.00896	0.00892	0.5	69909	624	69597	3718154	53.19
1900	7	0.00553	0.00551	0.5	69285	382	69094	3648557	52.66
1900	8	0.00341	0.0034	0.5	68903	235	68786	3579463	51.95
1900	9	0.0026	0.00259	0.5	68669	178	68580	3510677	51.12
1900	10	0.00293	0.00293	0.5	68491	200	68390	3442097	50.26
1900	11	0.0035	0.00349	0.5	68290	238	68171	3373707	49.4
1900	12	0.00397	0.00396	0.5	68052	270	67917	3305536	48.57
1900	13	0.00446	0.00445	0.5	67782	302	67631	3237619	47.76
1900	14	0.00474	0.00473	0.5	67481	319	67321	3169987	46.98
1900	15	0.00499	0.00498	0.5	67161	335	66994	3102666	46.2
1900	16	0.0055	0.00548	0.5	66827	366	66644	3035672	45.43
1900	17	0.00588	0.00586	0.5	66461	389	66266	2969028	44.67
1900	18	0.00631	0.00629	0.5	66071	415	65863	2902762	43.93
1900	19	0.00667	0.00664	0.5	65656	436	65437	2836899	43.21
1900	20	0.00675	0.00673	0.5	65219	439	65000	2771462	42.49
1900	21	0.0069	0.00688	0.5	64781	445	64558	2706462	41.78
1900	22	0.00742	0.00739	0.5	64335	475	64098	2641904	41.06

1900	23	0.00786	0.00783	0.5	63860	500	63610	2577806	40.37
1900	24	0.00788	0.00785	0.5	63360	498	63111	2514196	39.68
1900	25	0.00775	0.00772	0.5	62862	485	62620	2451085	38.99
1900	26	0.00752	0.00749	0.5	62377	467	62143	2388466	38.29
1900	27	0.00755	0.00752	0.5	61910	466	61677	2326322	37.58
1900	28	0.00772	0.00769	0.5	61444	473	61208	2264645	36.86
1900	29	0.00791	0.00788	0.5	60972	480	60731	2203437	36.14
1900	30	0.00808	0.00805	0.5	60491	487	60248	2142706	35.42
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…

La colonna denominata "q(x)" (cioè la probabilità di morte tra l'anno *x* e *x + 1*) è quello che ci interessa. Tuttavia, i requisiti dell' HMD sono tali che solo i paesi in cui la registrazione della morte e i dati di censimento sono completi vengano inclusi.

Un altro sito è "human lifetable database" (HLD: http://www.lifetable.de/) è una collezione ancora più grande costruita da individui o enti che utilizzano svariate tecniche. Ci sono tavole di 38 paesi, e gli intervalli di tempo sono in molti casi più lunghi di quelli in HMD (in Francia, si risale fino all'anno 1806). Le tavole, tuttavia, non sono direttamente paragonabili dato che sono in una varietà di formati e sono state calcolate utilizzando diverse tecniche. Anche in questo caso, la maggior parte delle tavole sono divise per sesso, età e tempo di calendario e può essere scaricato in formato .txt.

In questo lavoro si è fatto riferimento alle "life tables" del sito HMD.

Bibliografia

1. Sant M, Francisci S, Capocaccia R, Verdecchia A, Allemani C, Berrino F: Time trends of breast cancer survival in Europe in relation to incidence and mortality. Int J Cancer, 119: 2417–2422, 2006.
2. Holleczek B, Arndt V, Stegmaier C, Brenner H: Trends in breast cancer survival in Germany from 1976 to 2008 - A period analysis by age and stage.Cancer Detect Prev, 35: 399–406, 2011.
3. Taylor R, Davisa P, Boyages J: Long-term survival of women with breast cancer in New South Wales. Eur J Cancer, 39:215-222,2003.
4. Joensuu H, Toikkanen S: Cured of breast cancer? J ClinOncol,13:62-69,1995.
5. Langlands A, Pocock SJ, Kerr GR, Gore SM: Long-term survival of patients with breast cancer: a study of the curability of the disease.BMJ, 2: 1247-1251, 1979.
6. Gamel JW, Meyer JS, Feuer E, Miller BA: The impact of stage and histology on the long-term clinical course of 163.808 patients with breast carcinoma. Cancer, 77: 1459–1464, 1996.
7. Louwman WJ, Klokman WJ, Coebergh JWW: Excess mortality from breast cancer 20 years after diagnosis when life expectancy is normal. Br J Cancer;84:700-703, 2001.
8. De Angelis R, Capocaccia R, Hakulinen T, Soderman B, Verdecchia A: Mixture models for cancer survival analysis: application to population- based data with covariates.Stat Med, 18: 441-454, 1999.
9. Hakulinen T: On long-term relative survival rates.J Chronic Dis, 30: 431-43, 1977.
10. Lambert PC: Modeling of the cure fraction in survival studies. The Stata Journal, 3: 1-25, 2007.
11. Singhal MK, Raina V: Cure from breast cancer, not quite yet but getting there? AnnOncol, 20: 1291-1292, 2009.
12. Marubini E, Valsecchi MG: Estimation of survival probabilities. In: Analyzing Survival Data from Clinical Trials and Observational Studies, 41-74, John Wiley & Sons Ltd, Chichester, 1995.
13. Jeong J, Fine JP: Parametric regression on cumulative incidence function. Biostatistics, 8: 184-196, 2007.
14. Pohar M, Stare J: Relative survival analysis in R. Computer Methods and Programs in Biomedicine, 81: 272–278, 2006.

15. Pohar M, Stare J: Making relative survival analysis relatively easy. Computers in biology and medicine, 37: 1741–1749, 2007.

16. Pohar M, Stare J, Esteve J: On Estimation in Relative Survival. Biometrics, 68: 113–120, 2012.

17. Lambert PC, Thompson J R, Weston CL, Dickman PW: Estimating and modeling the cure fraction in population-based cancer survival analysis. Biostatistics, 8: 576–594, 2006.

18. Shayan Z, Taghi SM, Zare N: A parametric method for cumulative incidence modeling with a new four-patameter log-logistic distribution. Theoretical Biology and Medical Modelling, 8:43, 2011.

19. Harrell F, Lee K, Mark D: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, Statistics in Medicine, 15:361-387, 1996.

20. Veronesi U, Cascinelli N, Mariani L, Greco M, Saccozzi R, Luini A, Aguilar M, Marubini E: Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. NEngl J Med, 347:1227-1232, 2002.

21. Duncan W, Kerr GR: The curability of breast cancer.BMJ, 2:781–783, 1976.

22. Dickman P W: Estimating and modelling relative survival, Workshop on Statistical Methods for Cancer Patient Survival, 2009.

23. Rutqvist L E, Wallgren A, Nilsson B:Is Breast Cancer a Curable Disease? A Study of 14,731 Women With Breast Cancer From the Cancer Registry of Norway.Cancer, 53:1793-1800, 1984.

24. Jeong J, Fine JP: Direct parametric inference for the cumulative incidence function. Appl. Statist, 55: 187-200, 2006.

25. Benichou, J, Gail M. H: Estimates of absolute cause-specific risk in cohort studies. Biometrics, 46, 813–826, 1990.

26. Hibberd AD, Horwood LJ, Wells JE: Long term prognosis of women with breast cancer in New Zealand: study of survival to 30 years. BMJ, 286:1777-1779, 1983.

27. http://noi-italia2013.istat.it/, accessed September 2013.