SISMEC

Società Italiana di Statistica Medica
ed Epidemiologia Clinica

II Convegno Nazionale

INFORMAZIONE CONOSCENZA DECISIONE

CONTRIBUTI LIBERI

a cura di
Stefano Calza e Adriano Decarli

Con il patrocinio di:
Università degli Studi di Brescia
Facoltà di Medicina e Chirurgia

Brescia 1-4 Ottobre 2003

USO DELLE CURVE ROC CON DATI DI SOPRAVVIVENZA IN SOTTOGRUPPI DETERMINATI GENETICAMENTE

The use of ROC curves with survival data in genetically determined subgroups

ALBERTO MORABITO[a], EMANUELA MORENGHI, MONICA FERRARONI, GIOVANNI RADAELLI, FABIO MACCIARDI[1]

Dipartimento di Medicina Chirurgia e Odontoiatria, Unità di Statistica Medica, Ospedale San Paolo, [1]Dipartimento di Biologia e Genetica per le Scienze Mediche, Università degli Studi di Milano, Milano.

Premessa

Nel contesto degli studi epidemiologici, uno degli scopi principali nell'ambito delle associazioni genetiche è identificare e mappare il(i) gene(i) responsabile(i) di una data malattia. I criteri per identificare le associazioni si basano sull'individuazione degli alleli dei geni candidati o marker preferibilmente trasmessi con la malattia. Questa individuazione è interpretata come una dimostrazione indiretta che un allele non ancora identificato del locus, e responsabile per la malattia, cosegrega con il corrispondente allele associato del marker, dando luogo a un aplotipo. In queste assunzioni, l'analisi di associazione è basata sull'esistenza di *linkage disequilibrium*. Un semplice approccio per esaminare le associazioni genetiche è basato sugli studi caso-controllo. In tal caso, sono state tradizionalmente usate procedure statistiche basate sull'analisi chi-quadrato. Tuttavia, queste analisi sovente non sono soddisfacenti.

Obiettivo

L'obiettivo di questo studio è di proporre un nuovo metodo statistico per lo studio delle asssociazioni genetiche, basato sulle curve *Receiver Operating Characteristic* (ROC) dipendenti dal tempo, suggerite inizialmente da Heagerty et al. [1]. Il metodo sarà illustrato analizzando dati derivati dal *Framingham Heart Study*. Specificatamente saranno esaminate le potenziali relazioni tra il glucosio nel sangue e la sopravvivenza rispetto alla presenza/assenza di un dato allele.

Materiali e Metodi

Metodologia statistica

Denotiamo con X una covariata continua, e assumiamo che più elevati valori di X siano maggiormente suggestivi di una data malattia. Sia D lo stato (binario) di malattia. La curva ROC per X è la funzione monotona in [0,1], [{P (X > c | D = 0), P (X > c | D = 1)}, c $\in(-\infty, \infty)$]. Geometricamente la curva ROC rappresenta, nel piano cartesiano ortogonale, la sensibilità in funzione di (1-specificità), al variare dei possibili valori della variabile X. Una introduzione formale alla teoria delle curve ROC può essere trovata, ad esempio, in Thompson e Zucchini [2] e Begg [3]. Le curve ROC per test diagnostici continui possono essere stimate usando stime empiriche delle funzioni $S_0(c) = P(X > c \mid D = 0)$, e $S_1(c) = P(X > c \mid D = 1)$. Supponiamo che lo stato di malattia

[a] Corrispondenza: Alberto Morabito, Dipartimento di Medicina Chirurgia e Odontoiatria, Ospedale San Paolo, via A.di Rudini' 8 – 20142 Milano. Email: Alberto.Morabito@unimi.it

dipenda dal tempo. Il metodo proposto da Heagerty et al. [1] per la stima di una curva ROC dipendente dal tempo si basa sull'uso diretto del teorema di Bayes, utilizza l'approccio di Kaplan-Meier e lo stimatore di Akritas [4].

Più precisamente, sia T_i il *tempo di morte*, C_i il *tempo censurato di morte*, e $Z_i = \min(T_i,C_i)$ il tempo di follow up. Poniamo $d_i=1$ se $T_i \leq C_i$, e $d_i=0$ se $T_i>C_i$. Consideriamo un processo di conteggio, $D_i(t)=1$ se $T_i \leq t$, e $D_i(t)=0$ se $T>t$. Denotiamo con $S(t)$ la funzione di sopravvivenza. Sia $S(t \mid X > c)$ la funzione di sopravvivenza condizionata da $X > c$. Una stima di $S(t)$ è data dallo stimatore di Kaplan-Meier (KM),

$$\hat{S}_{KM}(t) = \prod_{\substack{s \in \tau_n \\ s \leq t}} \left(1 - \frac{\sum_j I_{(Z_j=s)} \delta_j}{\sum_j I_{(Z_j \geq s)}} \right)$$

dove τ_n denota l'insieme dei valori di Z_i per gli eventi osservati $\delta_i = 1$, e I la funzione caratteristica. Una stima della sensibilità e della specificità è facilmente derivabile combinando lo stimatore KM e la funzione di distribuzione della covariata X. Abbiamo,

$$\text{Sensibilità} = \hat{P}_{KM}(X > c \mid D(t) = 1) = \frac{\left(1 - \hat{S}_{KM}(t \mid X > c)\right)\left(1 - \hat{F}_X(c)\right)}{1 - \hat{S}_{KM}(t)}$$

$$\text{Specificità} = \hat{P}_{KM}(X \leq c \mid D(t) = 0) = \frac{\hat{S}_{KM}(t \mid X \leq c)\hat{F}_X(c)}{\hat{S}_{KM}(t)}$$

dove $F_X(c) = \frac{1}{n} \sum_i I_{(X_i \leq c)}$. Poiché lo stimatore KM non garantisce la monotonicità delle funzioni che esprimono la sensibilità e la specificità, qui, come anche suggerito da Heargerty et al. [1], utilizzeremo lo stimatore di Akritas, ossia,

$$\hat{S}_{\lambda_n}(c,t) = \frac{1}{n} \sum_i \hat{S}_{\lambda_n}(t \mid X = X_i) I_{(X_i > c)}$$

dove $S_{\lambda_n}(t \mid X = X_i)$ è uno stimatore KM pesato, e λ_n un parametro indicante metà dell'ampiezza del range interquartile, centrato su X_i. Il confronto tra curve ROC è effettuato costruendo gli intervalli di confidenza tramite bootstrap [1].

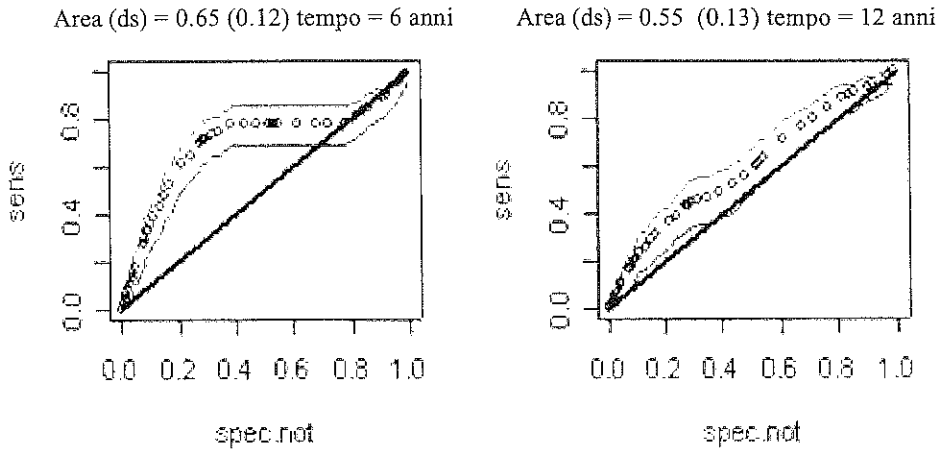
Applicazione a dati reali

La metodologia è stata applicata a dati tratti dal *Framingham Heart Study,* per saggiare la sensibilità e la specificità del glucosio nel sangue nel predire la sopravvivenza, rispetto alla presenza/assenza dell'allele 242 (marker 23, cromosoma 1). La scelta di analizzare il tratto 187-218 cM sul cromosoma 1, ed in particolare l'allele 242, è basata sui risultati di Meigs et al. [5]. Il database di Framingham è costituito da 4692 individui, (2348 maschi, 2344 femmine), 1213 appartenenti alla *Original cohort*, 1672 alla *Offspring cohort e* 1807 al *founders group*. Per semplicità, l'analisi è stata condotta sui 2885 individui con dati di glucosio non mancanti in almeno un controllo.

Risultati

La Figura presenta i risultati principali dell'analisi. Ciascuna curva ROC mostra, a differenti tempi di follow up (anni), per diversi livelli *c* di glucosio nel sangue: (sens) la proporzione stimata di individui sopravviventi aventi glucosio nel sangue > *c*, e (spec.not) la proporzione stimata di individui sopravviventi con glucosio nel sangue ≤ *c*. Le bande superiore ed inferiore rappresentano il 90^{mo} e il 10^{mo} percentile, rispettivamente. Una curva ROC sopra la diagonale rappresenta una sopravvivenza descrescente al crescere dei valori di glucosio nel sangue.

Assenza dell'allele 242



Come appare dal confronto tra le due curve ROC, l'allele 242 (marker 23) mostra abilità nel predire la sopravvivenza.

Conclusioni

I risultati numerici suggeriscono che l'approccio delle curve ROC dipendenti dal tempo può essere utile per individuare, nelle situazioni reali, differenze tra sottogruppi determinati geneticamente. Tuttavia, altri studi sono necessari per chiarire le proprietà statistiche teoriche di questa metodologia.

Bibliografia

1. Heagerty PJ, Lumley T, Pepe MS: Time–dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000, 56: 337-344.
2. Thompson ML, Zucchini W: On the statistical analysis of ROC curves. Stat Med 1989, 8: 1277-1290.
3. Begg CB: Advances in statistical methodology for diagnostic medicine in the 1980's. Stat Med 1991, 10: 1887-1895.
4. Akritas MG: Nearest neighbor estimation of a bivariate distribution under random censoring. Ann Stat 1994, 22: 1299-1327.
5. Meigs JB, Panhuysen CI, Myers RH, Wilson PW, Cupples LA: A genome-wide scan for loci linked to plasma levels of glucose and HbA (1c) in a community-based sample of Caucasian pedigrees: The Framingham Offspring Study. Diabetes 2002, 51: 833-840.