L'ANGOLO DELLA STATISTICA

Interpretare i modelli prognostici multivariati: il modello logistico

Catherine Klersy, Luigia Scudeller

G Ital Aritmol Cardiostim 2002;4:205-209

Servizio di Biometria ed Epidemiologia Clinica, Direzione Scientifica, IRCCS Policlinico San Matteo, Pavia

Negli ultimi anni i modelli prognostici sono stati pubblicati con frequenza crescente nella letteratura medica internazionale. Parallelamente lo sviluppo dei software statistici, con interfaccia sempre più intuitiva, ha reso appetibile la definizione di propri modelli prognostici. Tuttavia il percorso accademico di ognuno non sempre ha incluso una formazione in campo statistico ed epidemiologico, rendendo da una parte difficile e magari incompleta l'interpretazione di tali modelli, dall'altra pericoloso l'utilizzo indiscriminato del software.

Scopo di questa nota è di illustrare il significato e l'interpretazione dei modelli logistici utilizzati per definire la prognosi.

Perché utilizzare un modello prognostico multivariato?

Una ricerca scientifica ben pianificata dovrebbe porsi come obiettivo la dimostrazione che una certa caratteristica della nostra coorte sia effettivamente un fattore di rischio non identificato prima. Ciò implica, da una parte, la necessità di dimostrare che la presenza di questo fattore di rischio effettivamente si associ con la malattia/l'evento che stiamo studiando e, dall'altra, che sia un fattore di rischio addizionale rispetto a quelli già noti dalla letteratura, anche alla luce delle caratteristiche della popolazione. I modelli prognostici multivariati permettono di rispondere a questi quesiti: sarà possibile calcolare una misura dell'associazione fra fattore di rischio e "outcome" e verificare che questa associazione si mantenga anche dopo avere tenuto conto delle altre caratteristiche del paziente.

Inoltre, da un modello di regressione logistica è possibile costruire un indice prognostico sintetico che permetta di classificare i pazienti in base alle caratteristiche di presentazione, in categorie di rischio crescenti, con probabilità crescenti note di realizzarsi dell'evento.

Infine, alcune (rare) volte ci si può trovare a valutare una malattia per cui non sono noti dei fattori di rischio, ma esiste una serie di diverse ipotesi. In questo caso, i modelli prognostici multivariati permettono di identificare possibili fattori prognostici indipendenti. Sono fattori che si associano con la malattia in studio, anche dopo avere tenuto conto (indipendentemente/a pari-

tà) di tutte le altre caratteristiche del paziente. È necessario essere consapevoli che, non partendo da un'ipotesi scientifica specifica, si tratta di un'analisi esplorativa e non conclusiva, che può essere fonte di ipotesi per studi successivi.

Che cos'è un modello logistico?

Il modello logistico appartiene alla famiglia dei modelli di regressione per un "outcome" (o variabile dipendente) binario. Una variabile dipendente binaria può avere 2 valori, generalmente codificati come 0 per un outcome negativo (evento assente) e come 1 per un outcome positivo (evento presente). "Essere/non essere affetti da fibrillazione atriale", "essere/non essere in vita 5 anni dopo la diagnosi di tumore (dove tutti i pazienti sopravvissuti hanno un follow-up di almeno 5 anni)" sono esempi di variabili dipendenti binarie in medicina.

I modelli per variabili dipendenti binarie permettono di esplorare come ogni variabile esplicativa (o indipendente) influenzi la probabilità che l'evento in studio si verifichi. Sono modelli non lineari, nel senso che l'associazione fra probabilità dell'evento e variabili esplicative non è lineare. La probabilità che l'evento si verifichi è data da:

$$\Pr(y = 1 \mid x) = \frac{\exp(\alpha + \beta x_1 + \beta x_2 + \beta x_3)}{1 + \exp(\alpha + \beta x_1 + \beta x_2 + \beta x_3)}$$

Dove x_1, x_2 e x_3 sono le caratteristiche del paziente (covariate) inserite nel modello. Questa formulazione è equivalente alla seguente, in cui l'effetto delle covariate viene associato non con la probabilità dell'evento, ma con "l'odd", cioè con il rapporto fra probabilità di evento e probabilità di non evento.

$$\ln\left[\frac{\Pr(y = 1 \mid x)}{1 - \Pr(y = 1 \mid x)}\right] = \alpha + \beta_1 x + \beta_2 x + \beta_3 x$$

Il primo termine dell'uguaglianza, il logaritmo dell'odd, è detto "logit", da cui il termine "modello logistico". L'associazione fra il logit e le variabili esplicative è lineare, contrariamente alla formulazione precedente, ed è più facile da modellare. Questo spiega la diffusione di tali modelli.

Probabilità e odd sono legati dalla relazione seguente:

$$p = \frac{odd}{1 + odd}$$

Come valutare se il modello è applicato correttamente?

Applicare un modello statistico a un set di dati richiede che siano soddisfatte determinate condizioni. Nel caso del modello logistico, queste sono:

- la variabile di outcome deve essere binaria;
- il disegno dello studio deve essere appropriato: studio trasversale; studio caso-controllo; studio longitudinale con tempi di follow-up uguali per tutti i soggetti;
- la relazione della variabile con il logit deve essere lineare (nel caso contrario vanno utilizzate trasformate della variabile originale);
- se necessario, sono stati inseriti termini di interazione (per tenere conto che l'effetto di una variabile non è lo stesso nelle categorie di una seconda: per esempio l'effetto dell'età sulla probabilità di comparsa dell'evento è diverso nei maschi e nelle femmine).

Dopo avere costruito il modello va verificato:

- che le variabili inserite nel modello effettivamente spieghino i dati (p-value globale del modello <0,05);
- che il modello sia efficace nel descrivere la variabile di outcome (bontà di adattamento del modello): esistono diverse statistiche che permettono di quantizzare tale fenomeno e che andrebbero valutate e riportate (varianza spiegata, pseudo R2, criterio di informazione di Akaike, criterio di informazione bayesiano, ecc.);
- se esistono outlier e/o osservazioni influenzanti, attraverso l'esame dei residui. I residui sono le differenze fra gli outcome predetti dal modello e gli outcome osservati nello studio. Gli outlier sono quei casi con ampi residui, cioè i casi in cui il modello predice male l'outcome; le osservazioni influenzanti sono quelle che hanno un effetto importante sugli effetti stimati.

Quali e quante variabili in un modello logistico?

La scelta delle variabili da includere nel modello, oltre a quella su cui si basa l'ipotesi scientifica da dimostrare, deve essere guidata dalla conoscenza del problema studiato; è consuetudine includere età e sesso, nonché i fattori di rischio già noti dalla letteratura e che potrebbero influenzare la relazione fra il nuovo fattore prognostico e la probabilità dell'evento.

Esistono due limiti principali che riducono l'efficienza del modello: la presenza di collinearità (variabili correlate fra di loro) e l'eccessivo numero di covariate (causa del fenomeno di "overfitting"). Nel primo caso si può scegliere una delle due variabili correlate, o inserire alternativamente una loro combinazione; nel secondo caso si può utilizzare una regola pragmatica che limita il rapporto fra covariate e casi a 1:10.

Una nota merita anche il problema dei dati mancanti: in un modello multivariato, è sufficiente che una delle variabili del modello abbia un dato mancante perché il caso venga eliminato dall'analisi, con conseguente perdita di potenza dello studio, anche importante se i dati mancanti sono numerosi.

Come interpretare i coefficienti di un modello di regressione logistica?

Generalmente i risultati derivanti dall'applicazione di un modello logistico vengono presentati sotto forma di "odds ratio" (OR) e del relativo intervallo di confidenza al 95%. L'OR rappresenta il rapporto fra quoziente delle probabilità di evento/non evento nei pazienti con il fattore di rischio e lo stesso quoziente nei pazienti senza il fattore di rischio.

Per eventi rari (<5-10%), l'OR rappresenta un'approssimazione del rischio relativo. Il valore dell'OR risponde alla domanda: "Di quanto moltiplico il rischio (espresso come odd), quando aumento di 1 unità il valore della mia variabile di interesse?". L'intervallo di confidenza dell'OR permette da una parte di capire con quanta precisione l'OR è stato stimato (ampiezza dell'intervallo), dall'altra di valutare l'entità minima dell'effetto (limite inferiore dell'intervallo di confidenza per fattori che aumentano il rischio; limite superiore dell'intervallo di confidenza per fattori protettivi). Infine, permette di valutare indirettamente se il test statistico effettuato sulla variabile, per saggiare l'ipotesi nulla di effetto assente, è significativo al livello del 5% (OR = 1 non incluso nell'intervallo). Un OR >1 corrisponde a un aumentato rischio; un OR <1 corrisponde a un rischio ridotto (effetto protettivo). Gli studi clinici controllati tipicamente indagano trattamenti che riducono la proporzione di eventi, si avranno quindi OR <1. Gli studi epidemiologici cercano invece di identificare fattori di rischio, e quindi quelli con un OR >1.

Esemplifichiamo con uno studio fittizio, dove l'ipotesi scientifica di interesse sarà di verificare l'associazione fra età (anni) ed evento, oppure fra sesso (maschi vs femmine – 0/1) ed evento, o alternativamente fra classe funzionale (3 categorie – 0/1/2) ed evento (Tabella I). L'interpretazione dell'OR sarà leggermente diversa per ognuna delle 3 variabili, che sono rispettivamente di tipo continuo, categorico – 2 categorie e categorico-ordinale – 3 categorie:

- per ogni incremento di un anno di età, il rischio di evento viene moltiplicato per un fattore 1,04 e l'aumento minimo sarà di 1,00 (con una fiducia del 95%, a parità di sesso e classe funzionale);
- i maschi hanno un rischio di evento 3,2 superiore a quello delle femmine; l'aumento minimo (con una fiducia del 95%) sarà di 2,1, a parità di età e classe funzionale. Alternativamente le femmine hanno un rischio di evento pari a 1/3,2 = 0,31 quello dei maschi;
- la classe funzionale rappresenta un fattore di rischio (p globale <0,05), a parità di età e sesso: il rischio di evento per i pazienti in classe 1 è 1,4 volte il rischio dei pazienti in classe 0 (riferimento), mentre i pazienti in classe 2 hanno un rischio 2,9 maggiore di quelli in classe 0. Il rischio verrà moltiplicato per un minimo di 1,1 per i pazienti di classe 1 e di 2,1 per i pazienti di classe 2 (con una fiducia del 95%).

Invece di usare un fattore moltiplicativo per esprimere un aumento del rischio, è possibile utilizzare la variazione percentuale, che può essere calcolata dalla precedente come

$$\Delta\% = 100*[OR-1]$$

e quindi riportare le descrizioni corrispondenti alle precedenti (sempre a parità di sesso e classe funzionale, età e classe funzionale, età e sesso rispettivamente):

- per ogni incremento di un anno di età, il rischio di evento cresce del 4%;
- i maschi hanno un rischio di evento del 220% superiore a quello delle femmine o, alternativamente, le femmine hanno una riduzione di rischio del 69% rispetto ai maschi;
- la classe funzionale rappresenta un fattore di rischio (p globale <0,05): il rischio di evento per i pazienti in classe 1 è superiore del 40% al rischio dei pazienti in classe 0 (riferimento), mentre i pazienti in classe 2 hanno un rischio superiore del 190% rispetto a quelli in classe 0.

Dobbiamo tuttavia ricordare che la definizione di modello logistico partiva dall'esigenza di verificare come la variabile in studio influenzasse la probabilità che l'evento si verificasse.

Sarà quindi possibile calcolare e mettere in grafico la probabilità di evento predetta dal modello, per range di valori della variabile di interesse, fissati i valori delle altre covariate. Per esempio, ponendo a 0 sesso e classe funzionale, quindi considerando la popolazione di femmine in classe 0, potremo valutare come varia la proba-

TABELLA I Modello logistico multivariato (Model p = 0,001)

Evento 1/0	OR	Intervallo confidenza 95%	Note
Età in anni	1,04	1,00-1,08	
Sesso (M vs F)	3,2	2,1-6,5	
Classe		p globale per classe = 0,023	
1/0	1,4	1,1-2,6	
2/0	2,9	2,1-5,3	

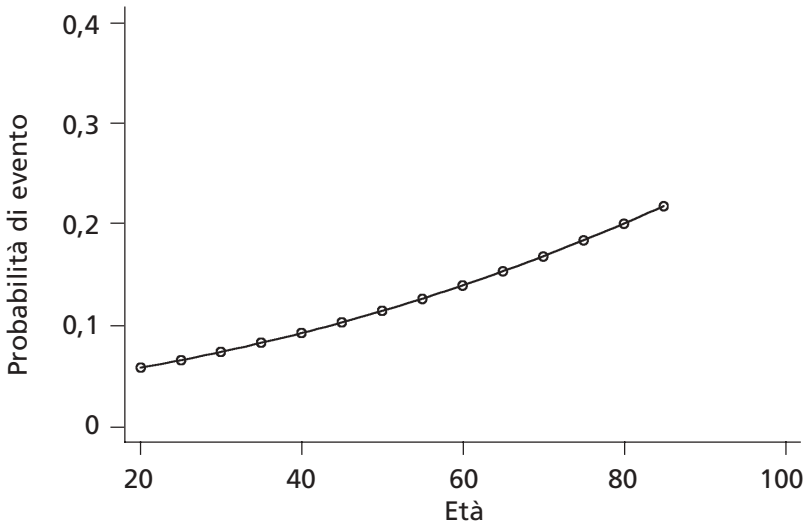

FIGURA 1 Andamento della probabilità di evento in base all'età, calcolata dal modello logistico.

bilità di evento per ogni 5 anni di età per età fra 20 e 60 anni (Figura 1).

Validazione del modello logistico

La capacità di ottenere un modello che spieghi i dati raccolti non esaurisce il problema della ricerca; infatti uno degli scopi dello sviluppo di un modello prognostico è la possibilità di fornire previsioni attendibili per individui futuri. È noto dalla letteratura che modelli prognostici tendono a dare delle stime eccessivamente ottimistiche dell'effetto di un fattore di rischio. Diventa quindi necessario valutare l'accuratezza predittiva (validità) di un modello per poterlo applicare a casi futuri. Questa viene stimata mediante 2 indici:

- una misura di calibrazione del modello;
- una misura di discriminazione del modello.

La calibrazione del modello fornisce un'indicazione di quanto la predizione legata al modello sia attendibile e di quanto sia il rumore dovuto a una cattiva specificazione del modello stesso; in particolare si osserverà più rumore se è stato inserito un numero eccessivo di covariate nel modello (overfitting). Per calibrare il modello viene calcolata una misura chiamata coefficiente di "shrinkage" γ, che permette di quantificare il rumore ($=1-\gamma$) e ricalibrare il modello per predizioni future (moltiplicando i coefficienti di regressione per γ).

La discriminazione del modello è la sua capacità di separare pazienti con risposte diverse. Viene misurata mediante indici che valutano la correlazione fra outcome predetti e outcome osservati (o proporzione di concordanti), l'indice di concordanza c, l'indice D di Somer e altri.

Conclusioni

Il modello logistico rappresenta uno strumento potente per la definizione di modelli prognostici. Permette di quantizzare la forza di associazione fra un fattore di rischio e un outcome, ma anche la probabilità di realizzarsi dell'evento. Come tutti i modelli statistici, le risposte che vengono fornite saranno attendibili solo se il modello è applicato avvedutamente. Infine, la capacità predittiva di un modello per casi futuri può e deve essere calcolata mediante apposite misure di accuratezza.

Bibliografia

1. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996:15:361-387.
2. Bland J, Altman DG. The odds ratio. *BMJ* 2000;320:1468.
3. Hosmer DW, Lemeshow S. *Applied logistic regression*, 2d ed. New York: John Wiley & Sons, 2000.
4. Long JS, Freese J. *Regression models for categorical dependent variables using Stata*. College Station, TX. A Stata press Publication, Stata Corporation, 2001.
5. Bender R, Grouven U. Logistic regression models used in medical research are poorly presented. *BMJ* 1996;313:628.
6. Klersy C. Effetto statisticamente rilevante ed effetto clinicamente rilevante. *G Ital Aritmol Cardiostim* 2001;4:145-148.
7. Klersy C. I test multipli. *G Ital Aritmol Cardiostim* 2001;1:23-25.

Indirizzo per la corrispondenza

Catherine Klersy
Servizio di Biometria ed Epidemiologia Clinica
Direzione Scientifica
IRCCS Policlinico San Matteo
27100 Pavia
e-mail: klersy@smatteo.pv.it