Codifica automatica di variabili patologiche tumorali tramite la ricerca di stringhe di testo nei referti anatomo-patologici. L'esperienza del Registro tumori toscano

Automatic coding of pathologic cancer variables by the search of strings of text in the pathology reports. The experience of the Tuscany Cancer Registry

Emanuele Crocetti, Claudio Sacchettini, Adele Caldarella, Eugenio Paci

UO Epidemiologia clinica e descrittiva, Centro per lo studio e la prevenzione oncologica, Firenze

Corrispondenza: *Emanuele Crocetti, UO Epidemiologia clinica e descrittiva, Centro per lo studio e la prevenzione oncologica, via di San Salvi 12, 50135 Firenze; e-mail: e.crocetti@cspo.it*

Riassunto

Si presentano i risultati di uno studio condotto sulla casistica del Registro tumori toscano (RTT) per la valutazione di un sistema automatico di lettura di variabili patologiche nel testo libero dei referti anatomo-patologici. Sono stati utilizzati i dati relativi ai casi incidenti negli anni 2000 (n. 6.297) e 2001 (n. 6.291) per i quali era disponibile un referto anatomo-patologico su supporto informatico. Il sistema è basato su *query* e funzioni eseguibili con il programma «Access» di Windows. La concordanza fra quanto inserito dagli operatori del RTT secondo il metodo tradizionale manuale e quanto prodotto dalla procedura automatica è stata valutata tramite la statistica kappa di Cohen. Le variabili oggetto dello studio sono: sede (per la quale si è ottenuta, per i casi del 2001, una kappa, fra codifica manuale e automatica, di 0,87), morfologia (kappa=0,75), classi morfologiche di Berg (kappa=0,87), comportamento (kappa=0,70), grado di differenziazione cellulare (kappa=0,90), Gleason (kappa=0,90), focalità (kappa=0,86), lateralità (kappa=0,36), pT (kappa=0,92), pN (kappa=0,76), pM (kappa=0,28), numero di linfonodi repertati (kappa=0,69), numero di linfonodi positivi (kappa=0,70), spessore secondo Breslow (kappa=0,94), livello di Clark (kappa=0,91) e stadio secondo Dukes (kappa=0,74). Il sistema di lettura automatica di stringhe offre vantaggi in termini di rapidità e semplicità di acquisizione di informazioni, per questo motivo se ne auspica l'applicazione nella pratica dei Registri tumori.

(*Epidemiol Prev* 2005; 29(1): 57-60)

Parole chiave: registro tumori, codifica automatica, stringa di testo, referto patologico

Abstract

The present study evaluates the application of an automatic system for variables coding by means of strings reading in the text of the pathology reports, in the database of the Tuscany Cancer Registry.

Incidence data for the years 2000 (n. 6297) and 2001 (n. 6291) for subjects for whom computerised pathology reports were available were included. The system is based on Queries (SQL language) linked to Functions (Visual Basic for Applications) that work on Windows Access. The agreement between original data inputted by the registrars and variables coded by means of automatic reading has been evaluated by means of Cohen's kappa. The following variables were analysed: cancer site (kappa = 0.87 between «manual» and automatic coding, for cases incident in the year 2001), morphology (kappa=0.75), Berg's morphology groups (kappa=0.87), behaviour (kappa=0.70), grading (kappa=0.90), Gleason (kappa=0.90), focality (kappa=0.86), laterality (kappa=0.36), pT (kappa=0.92), pN (kappa=0.76), pM (kappa=0.28), number of lymph nodes (kappa=0.69), number of positive lymph nodes (kappa=0.70), Breslow thickness (kappa=0.94), Clark level (kappa=0.91), Dukes (kappa=0.74).

The system of automatic reading of strings allows to collect a very huge amount of reliable information and its use should be implemented by the Registries.

(*Epidemiol Prev* 2005; 29(1): 57-60)

Key words: cancer registry, automated coding, string of text, pathology report

Introduzione

I Registri tumori di popolazione sono strumenti per la rilevazione delle nuove diagnosi tumorali in popolazioni definite. Obiettivo dei Registri tumori è l'esaustività (almeno teorica) della rilevazione dell'incidenza, che viene perseguito utilizzando più fonti informative, fra le quali le principali sono le schede di dimissione ospedaliera, i certificati di morte e i referti anatomo-patologici.

Tradizionalmente, la registrazione dei tumori prevedeva che operatori rivedessero tutta la documentazione clinica relativa a un singolo soggetto e a un singolo tumore per estrarre le informazioni fondamentali sulla nuova diagnosi tumorale.

Nel corso del tempo, con lo sviluppo di fonti informative informatizzate, si sono sviluppati anche in Italia, Registri tumori automatici,[1] dove *software* dedicati definiscono, tramite alberi decisionali, i casi incidenti, almeno per una quota della casistica. Allo stesso tempo nei Registri tradizionali la possibilità di effettuare incroci automatici tra varie fonti informative sta mo-

dificando la tradizionale gestione manuale delle informazioni.
La raccolta di variabili relative alla neoplasia prevede un *set* minimo, definito sulla base delle pubblicazioni internazionali dell'Agenzia internazionale per la ricerca sul cancro di Lione, che comprende la data di incidenza, la sede topografica, la morfologia, il comportamento e la base della diagnosi.[2]
Un crescente interesse per i percorsi diagnostico-terapeutici, stimolato anche da studi multicentrici europei, sta portando i registratori a estendere la quantità delle informazioni raccolte comprendendo anche quelle relative al percorso terapeutico, al trattamento, a specifiche caratteristiche istologiche e allo stadio di diffusione della malattia alla diagnosi.[3]
Questo interesse verso nuove variabili obbliga a un continuo cambiamento del formato di raccolta dei casi e rappresenta un crescente aggravio del carico di lavoro.
Il presente studio intende verificare la possibilità di limitare la raccolta manuale delle informazioni ottenibili dai referti anatomo-patologici, utilizzando un sistema di algoritmi per l'identificazione automatica di stringhe nel testo libero della diagnosi anatomo-patologica, e di verificare la concordanza tra i codici attribuiti con questa modalità con quelli definiti secondo la procedura tradizionale.

Metodi

Si è utilizzata la casistica del Registro tumori toscano (RTT) relativa agli anni di incidenza 2000 e 2001, raccolta e codificata secondo un sistema tradizionale che prevede la lettura e l'interpretazione della documentazione clinica da parte di operatori sanitari esperti.[4]
Sono stati utilizzati i casi incidenti per i quali era disponibile, su supporto informatizzato, un referto anatomo-patologico (n. 6.297 nel 2000 e n. 6.291 nel 2001). Nel 2001 i referti patologici in forma informatizzata hanno rappresentato circa il 73% del totale dei referti. A questi referti, identificati in base a una chiave univoca che prevede il codice ospedaliero, il codice del servizio di anatomia patologica e il codice del referto, si è proceduto ad applicare un sistema di algoritmi per la lettura automatica di stringhe di testo direttamente dal referto anatomo-patologico.
Il sistema è composto da *query* (in linguaggio SQL) che richiamano funzioni (redatte nel programma «Visual Basic for Applications»); le *query* sono eseguibili con il programma «Access» di Windows. Il sistema è stato creato sulla casistica dell'RTT relativa all'anno 2000 con un processo di creazione e incremento di stringhe sulla base di quanto presente nel materiale originale utilizzato per definire le variabili dell'RTT.
Il confronto fra codifica manuale e automatica è stato condotto, per ogni variabile, dove fosse stato attribuito un codice da parte degli operatori dell'RTT.
Le variabili oggetto dello studio sono state:

- sede (3-digit) secondo la ICDO-1 [5]
- morfologia (4-digit) secondo la ICDO-1 [5]
- classi di morfologia. Dodici gruppi morfologicamente omogenei secondo la definizione di Berg [6]
- comportamento: secondo la ICDO-1 (0=benigno, 1=incerto, 2=in situ, 3=maligno primitivo, 6=maligno metastatico)[5]
- il grado di differenziazione cellulare - grading (1= ben differenziato, 2= mediamente differenziato, 3= scarsamente differenziato, 4=indifferenziato, 5=sconosciuto)
- il Gleason (per i tumori prostatici)
- la focalità (multifocale, non multifocale)
- lateralità (D=destro, S=sinistro, B=bilaterale)
- il T, N e M patologico
- il numero di linfonodi repertati
- il numero di linfonodi positivi
- lo spessore secondo Breslow (per i melanomi)
- il livello di Clark (per i melanomi) (livelli I-V)
- lo stadio secondo Dukes (per i tumori colo-rettali).

Il confronto tra codici attribuiti dai registratori e i codici attribuiti automaticamente è stato effettuato utilizzando, oltre la concordanza osservata, la statistica kappa di Cohen che permette di escludere la concordanza attesa per effetto del caso.[7]
La misura della concordanza offerta dalla statistica kappa varia da 0, quando la concordanza osservata è quella attesa in base al caso, a 1, quando c'è una concordanza perfetta. Per i valori intermedi è stata proposto che valori di kappa superiori a 0,75 siano indicativi di una eccellente concordanza, mentre valori inferiori a 0,40 rappresentino una concordanza debole; valori compresi tra 0,75 e 0,40 possono rappresentare una concordanza abbastanza buona.[8]
Nel caso di discordanza fra codifica tradizionale e automatica, una anatomo-patologa (AC) ha provveduto a una revisione dei referti per verificare quale fosse la diagnosi corretta. E' stata calcolata la percentuale di casi discordanti per i quali il risultato automatico fosse quello corretto.

Risultati

Nella tabella 1 è riportata, per la casistica degli anni 2000 e 2001 e per le diverse variabili oggetto dello studio, la concordanza osservata fra l'attribuzione dei codici secondo il tradizionale sistema di codifica manuale e quella secondo la lettura automatica delle variabili con l'utilizzo dell'algoritmo per la ricerca di stringhe prefissate di testo.
I risultati della lettura delle stringhe risultano molto buoni in termini di concordanza con quanto raccolto con la procedura manuale per la gran parte delle variabili oggetto dello studio.
La statistica kappa ha valori eccellenti (≥75%) in entrambi gli anni per la sede, la morfologia, le classi morfologiche di Berg, il pT, il pN, il grading, il Gleason, la focalità, i livelli di Clark, lo spessore di Breslow e per i casi del 2000 anche per Dukes, il numero di linfonodi e il numero di linfonodi positivi.
La statistica kappa ha mostrato una concordanza abbastanza buona (0,40-0,74) per il numero di linfonodi, per il numero di linfonodi positivi e per il Dukes dei casi 2001, mentre è risultata scarsa in entrambi gli anni per la definizione della

lateralità e per la presenza di metastasi. Nella tabella 1 è riportata la percentuale di casi discordanti per i quali la codifica automatica tramite stringhe è risultata corretta. Per alcune variabili, in caso di discordanza fra codifica manuale e tramite stringhe, l'informazione definita manualmente era più spesso corretta, come per sede, morfologia, classi morfologiche di Berg, comportamento, Clark, lateralità, pT, numero di linfonodi e numero di linfonodi postivi. Mentre per Dukes, focalità, Gleason, grading, Breslow e pM è risultata più spesso corretta la codifica automatica.

Discussione

L'informatizzazione dei sistemi informativi sanitari determina la necessità di modificare procedure già codificate nel corso del tempo e di codificarne di nuove che tengano conto delle potenzialità offerte dallo sviluppo tecnologico.
Le richieste di dati relativi a una migliore definizione dello stadio alla diagnosi e di altre variabili diagnostiche, terapeutiche e prognostiche rappresentano per gli operatori dei Registri un impegno gravoso che è destinato ad aumentare nel tempo. Inoltre, per quanto esaustiva sia la raccolta di dati, questa rischia di essere sempre incompleta rispetto a esigenze sede-specifiche e alla crescente disponibilità di indicatori prognostici offerta, per esempio, dallo sviluppo delle indagini immunoistochimiche. In questo senso la possibilità di utilizzare la ricerca delle stringhe di testo nei referti anatomo-patologici da un lato riduce il carico di lavoro degli operatori, dall'altro aumenta – teoricamente senza limiti – le informazioni che possono essere raccolte, semplicemente adattando i criteri di ricerca e uscendo dai vincoli imposti da un tracciato record *standard*. Per esempio, nella casistica patologica del 2001 sono state ricercate stringhe relative ad alcuni marcatori immuno-istochimici, sono stati identificati 503 casi per i quali era disponibile il Ki67 e 275 casi con l'informazione sul c-erbB-2; sebbene non sia possibile valutare l'esaustività di questa raccolta, il sistema permette, in tempi trascurabili, la raccolta di molte centinaia di informazioni che possono rappresentare la base per una rivalutazione del materiale disponibile.
Questo studio ha dimostrato che per molte delle variabili che vengono raccolte dall'RTT la concordanza fra sistema automatico e manuale è molto elevata, tanto da far ritenere che l'utilizzo del sistema automatico possa diventare d'uso corrente. In particolare c'è stata una concordanza particolarmente elevata per le variabili pT, pN, Clark, Gleason, focalità, Breslow e grading con valori di kappa superiori al 75%, così come per sede e morfologia. Valori più bassi di concordanza, ma sempre con kappa indicativi di una concordanza buona, sono stati ottenuti per le variabili numero di linfonodi e numero di linfonodi positivi, per le quali spesso non è indicato nel referto un valore complessivo che deve essere ricavato tramite la lettura automatica e la somma di conteggi di singole stazioni linfonodali. La concordanza è particolarmente elevata per queste variabili, per le quali i valori possibili sono limitati, ben codificati e quindi facilmente prevedibili. Per esempio la concordanza sul comportamento (dati 2001) aumenta riducendo le categorie a due (non infiltranti e infiltranti), con una kappa che passa da 0,70 a 0,80. In questo senso deve essere considerata particolarmente elevata una concordanza del 77,2% come quella ottenuta per la morfologia (kappa 0,75), che è stata calcolata in una matrice di molte centinaia di valori per lato (codici da 8000 a 9990). I risultati sono molto migliori quando per la morfologia si è considerata la concordanza fra le dodici classi morfologiche di Berg. [6]
Un risultato mediocre si è ottenuto per la variabile relativa alla presenza di metastasi. La definizione di questa condizione è legata a una valutazione clinica complessiva del paziente e può derivare dai referti patologici solo quando questi sono relativi a sedi metastatiche. Si tratta pertanto di un risultato atteso che non esprime un limite del metodo ma l'inadeguatezza della fonte informativa.
La procedura di lettura automatica è stata creata sulla casistica dei referti dell'anno 2000 partendo da stringhe consuete ma includendo anche stringhe molto particolari (per esempio legate a errori di digitazione da parte dei patologici nella stesura nel referto anatomo-patologico) che con scarsa probabilità si sarebbero ripetute in anni successivi; questo giustifica il fatto che, in linea generale, la concordanza sia leggermente migliore per il 2000 rispetto al 2001.
Fra la casistica del 2000 e quella del 2001 si è osservata una consistente riduzione della concordanza per la definizione dello stadio Dukes; ciò ha portato a una rivalutazione della casistica e a identificare un errore sistematico di un codificatore dell'RTT nel codificare questa variabile nell'anno 2001. Questo evidenzia le potenzialità della procedura anche come sistema per verificare la completezza e la correttezza della raccolta di informazioni specifiche e il loro andamento nel tempo.
Nell'ambito dei Registri tumori la codifica manuale delle informazioni è tradizionalmente ritenuta il metodo di riferimento. In questo senso la concordanza osservata ha lo scopo di fornire un'impressione (parziale in quanto condizionata dalla concordanza legata al caso, dalla dispersione della variabile e quindi dal numero di classi considerate, eccetera) su quanto l'utilizzo del sistema automatico costi in termini di completezza della raccolta. Ma è necessario sottolineare che la codifica manuale non è esente da errori e le discrepanze osservate tra procedura manuale e automatica possono originare da limiti sia dell'una sia dell'altra.
I limiti della procedura manuale sono sostanzialmente legati alla sua natura, che implica la trascrizione di dati da un supporto (il referto) a un altro (il formato record del registro). Essendo fatta manualmente la trascrizione di molte variabili, ripetuta per migliaia di casi, è soggetta a errori di lettura, di valutazione, di trascrizione, di battitura che possono solo in parte essere previsti da sistemi di allerta per l'inserimento di valori insoliti. D'altra parte il sistema automatico trova i suoi limiti nelle situazioni per le quali è prevista un'integra-

Variabile	n. casi 2000	Concordanza 2000	Kappa 2000	n. casi 2001	Concordanza 2001	Kappa 2001	% di casi discordanti,
sede (3 digit)	6.297	91,2	0,90	6.291	88,8	0,87	44,1
morfologia (4 digit)	6.277	81,7	0,80	6.291	77,2	0,75	15,7
classi morfologiche^	6.297	92,4	0,90	6.291	90,6	0,87	18,4
comportamento	6.297	93,4	0,75	6.291	92,3	0,70	17,7
Clark	165	99,4	0,99	157	93,6	0,91	10,0
Dukes	586	94,9	0,94	584	78,6	0,74	77,6
focalità	6.297	99,5	0,86	6.291	99,4	0,86	78,3
Gleason	360	96,1	0,95	413	92,3	0,90	65,6
grading	6.297	93,9	0,90	6.291	93,6	0,90	71,8
lateralità	1.230	55,0	0,38	1.092	54,1	0,36	40,3
T (pTNM)	2.206	97,2	0,97	2.287	93,0	0,92	46,8
N (pTNM)	1.580	84,1	0,76	1.708	84,8	0,76	36,3
M (pTNM)	6.297	94,6	0,37	478	22,0	0,28	70,6
n. linfonodi	1.733	88,9	0,89	1.902	70,0	0,69	45,0
n. linonodi positivi	1.868	94,1	0,91	2.013	80,2	0,70	44,6
Breslow	168	96,4	0,96	152	94,1	0,94	100,0

* secondo la procedura automatica, in base alla revisione del referto anatomo-patologico; ^secondo la classificazione di Berg.[6]

Tabella 1. Registro Tumori Toscano. Concordanza fra la definizione di alcune variabili fra l'attribuzione dei codici secondo il tradizionale sistema di codifica manuale e quello di lettura di stringhe di testo dai referti anatomo-patologici. Sono riportati il numero di casi, la concordanza osservata, il valore della statistica Kappa di Cohen, per gli anni 2000 e 2001 e per quest'ultimo la percentuale fra i casi discordanti di informazione «corretta» secondo le stringhe.

Table 1. Tuscany Cancer Registry. Agreement between the traditional 'manual' definition of several variables and the automatic system of variables reading by means of strings of text in the pathology reports. Number of cases, observed agreement, Cohen's Kappa, for the years 2000 and 2001 and for the latter also the percentage of correct values from strings among discordant cases.

zione di informazioni più complesse; questo può essere il caso della definizione della lateralità in referti che contengono informazioni relative a più di un tumore (per esempio mammella destra e sinistra), situazioni nelle quali la procedura automatica non è in grado di attribuire a ciascun tumore le variabili specifiche. Allo stesso modo la lettura automatica incontra difficoltà quando più referti anatomo-patologici si riferiscono a uno stesso tumore; è il caso in cui vi sia il referto di una biopsia alla quale fa seguito il referto del pezzo operatorio: in questo caso è necessario che la chiave di incrocio si riferisca al documento più ricco di informazioni.

Una distinzione è necessaria fra variabili «di base» che servono a definire il tumore (sede, morfologia e comportamento), e le altre «di arricchimento» che descrivono la diffusione alla diagnosi e l'iter diagnostico-terapeutico. Per le prime i risultati della lettura delle stringhe, seppur buoni o molto buoni, possono ancora non essere ritenuti sufficienti per le necessità di precisione dei registratori, considerando che per esempio il referto anatomo-patologico non è sempre la miglior fonte di informazione sulla sede e che una definizione della sede a tre cifre senza sottosede, pur essendo quanto viene abitualmente registrato dai registri automatici, può essere considerata non sufficiente.[1]

In conclusione il cambiamento delle fonti informative è una realtà della quale i registri tumori devono cogliere le potenzialità. In questo senso questo studio apre la prospettiva alla possibilità di spostare parte del tempo lavorativo degli operatori di registro da una fase di «ricopiatura» di informazioni, tanto semplice da poter essere sostituita - con risultati molto buoni – da una procedura automatizzata, a una fase di revisione della qualità della casistica, con un vantaggio in termini di qualità del lavoro e qualità del prodotto. In questo senso è auspicabile che, almeno per le variabili «di arricchimento», ovvero quelle non strettamente legate alla definizione del caso (sede, morfologia, comportamento) questo sistema entri nella pratica dei Registri.

Bibliografia

1. Falcini F, Paci E, La Rosa F, Marani E, De Lisi V, Tonini G, Zanetti R. La rete dei registri tumori In: Zanetti R, Gafà L, Pannelli F, Conti E, Rosso S, eds. *Il Cancro In Italia I dati dei Registri Tumori 1993-1998.* Roma , Il Pensiero Scientifico Editore, 2002. Vol. 3.
2. Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB,eds, *Cancer Incidence in Five Continents*. Lyon, IARC Scientific Publications n.155, 2002. Vol VIII.
3. Sant M, Gatta G. *The EUROCARE database*. IARC Scientific Publications n.132, 1995, pp.15-31.
4. Paci E, Crocetti E, Miccinesi G, Benvenuti A, Intrieri T, Sacchettini C, Giovannetti L. Tuscany Cancer Registry. In: Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB, eds, *Cancer Incidence in Five* Continents. Lyon, IARC Scientific Publications n.155, 2002. Vol VIII, pp.362-63.
5. WHO. Classificazione Internazionale delle Malattie per l'Oncologia ICD-0. Milano, Edi-Ermes, 1983.
6. Berg JW. Morphologic classification of human cancer. In: Scottenfeld D, Fraumeni J Jr, eds, *Cancer epidemiology and prevention*, 2nd ed. New York, Oxford University Press, 1996.
7. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 1960;* 20: 37-46.
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159-74.