Procedure di record linkage in epidemiologia: uno studio multicentrico italiano

Record-linkage procedures in epidemiology: an Italian multicentre study

Carla Fornari,[1] Fabiana Madotto,[1] Moreno Demaria,[2] Anna Romanelli,[3] Pasquale Pepe,[3] Mauro Raciti,[3] Valeria Tancioni,[4] Francesco Chini,[4] Paolo Trerotoli,[5] Nicola Bartolomeo,[5] Gabriella Serio,[5] Giancarlo Cesana,[1] Giovanni Corrao[6]

[1] Centro di studio e ricerca sulla patologia cronico-degenerativa negli ambienti di lavoro, Dipartimento di medicina clinica e prevenzione, Facoltà di medicina e chirurgia, Università degli studi di Milano Bicocca
[2] Epidemiologia ambientale, ARPA Piemonte
[3] CNR, Istituto di fisiologia clinica, Sezione di epidemiologia e ricerca sui servizi sanitari
[4] Laziosanità, Agenzia di sanità pubblica, Regione Lazio
[5] Dipartimento di scienze biomediche ed oncologia umana, Facoltà di medicina e chirurgia, Università degli studi di Bari
[6] Dipartimento di statistica, Facoltà di scienze statistiche, Università degli studi di Milano Bicocca

Corrispondenza: Carla Fornari, Centro di studio e ricerca sulla patologia cronico-degenerativa negli ambienti di lavoro, Dipartimento di medicina clinica e prevenzione, Facoltà di medicina e chirurgia, Università degli studi di Milano Bicocca. Villa Serena 6° piano, via Pergolesi 33, 20052 Monza, Italy; tel. 039 2333097/8; fax 039-365378; e-mail: carla.fornari@unimib.it

Riassunto

Obiettivo: confrontare le caratteristiche operative di procedure di record linkage (RL) utilizzate in diverse realtà italiane e di una tecnica probabilistica standard per l'integrazione di archivi di dati sanitari.
Disegno: appaiamento dell'archivio delle schede di dimissione ospedaliera (SDO) e di quello anagrafico degli assistiti o dei residenti, disponibili presso centri di servizio epidemiologico localizzati in diverse regioni italiane. Le procedure di RL utilizzate da ciascun centro, una procedura deterministica esatta e una procedura probabilistica sono applicate selezionando le SDO per infarto miocardico acuto e diabete. Un controllo manuale di un campione estratto casualmente dall'archivio delle SDO ha permesso di stimare sensibilità e specificità delle procedure. Tassi di ospedalizzazione annuali, standardizzati per genere ed età, sono calcolati al fine di valutare il possibile impatto della tecnica di RL adottata su misure di interesse epidemiologico.
Setting: comuni di Pisa e Roma e Regioni Piemonte e Puglia.
Partecipanti: popolazioni assistite o residenti al 31 dicembre 2003 e corrispondenti archivi delle SDO registrate nell'anno 2004.
Outcome principali: misure di accuratezza di procedure di RL per l'appaiamento di banche dati sanitarie.
Risultati: la qualità dei dati, assai differente in ogni archivio, influenza il criterio decisionale della procedura probabilistica. Pertanto è stato individuato un criterio standard che garantisse un valore predittivo positivo di almeno il 98%. La procedura probabilistica individua in media l'11% di coppie in più rispetto a quella deterministica esatta; risulta inoltre paragonabile, o migliore, delle procedure utilizzate dai singoli centri in termini di sensibilità. Dal confronto dei tassi standardizzati di ospedalizzazione ottenuti con le procedure del centro e probabilistica emergono differenze di direzione variabile nei diversi centri.
Conclusione: l'utilizzo del RL deterministico esatto appare limitato alle situazioni in cui sono disponibili codici univoci di identificazione di buona qualità. La tecnica di RL probabilistico proposta risulta paragonabile a quella usualmente adottata dai centri quando questi implementano un controllo di qualità dei dati o una revisione manuale dei risultati ottenuti. Se questo non accade, la tecnica usualmente adottata dai centri comporta errori sistematici di direzione ed entità non note.
(*Epidemiol Prev* 2008; 32 (3) suppl 1: 79-88)
Parole chiave: record linkage, epidemiologia, database amministrativi sanitari, Italia.

Abstract

Objective: *to compare record linkage (RL) procedures adopted in several Italian settings and a standard probabilistic RL procedure for matching data from electronic health care databases.*
Design: *two health care archives are matched: the hospital discharges (HD) archive and the population registry of four Italian areas. Exact deterministic, stepwise deterministic techniques and a standard probabilistic RL procedure are applied to match HD for acute myocardial infarction (AMI) and diabetes mellitus. Sensitivity and specificity for RL procedures are estimated after manual review. Age and gender standardized annual hospitalization rates for AMI and diabetes are computed using different RL procedures and compared.*

Setting: *municipalities of Pisa and Roma, and Regions of Puglia and Piemonte.*
Participants: *residents in the considered areas on 31 December 2003 and corresponding episodes of hospitalization in the same areas during 2004.*
Main outcome measures: *measures of accuracy of RL procedures to match health care administrative databases.*
Results: *data quality varies among archives and affects the decision rule of the probabilistic procedure. A unique decision rule was therefore adopted by means of choosing a positive predictive value of at least 98% for all the considered areas. The number of matched pairs identified with the probabilistic procedure is on average more then 11% greater than the number identified with the deterministic procedure. Sensitivity of probabilistic RL is similar or greater than that of other procedures. Differences between annual standardized hospitalization rates computed with stepwise deterministic RL and the standard probabilistic RL procedure vary among areas.*
Conclusion: *exact deterministic RL works well when unique identifiers and high quality data are available. The probabilistic procedure here proposed works as well as semi-deterministic RL when the latter implements a quality control of data or a manual review of final results. Otherwise, deterministic or semi-deterministic procedures imply classification errors of unknown size and direction.*
(*Epidemiol Prev* 2008; 32 (3) suppl 1: 79-88)
Key words: *record linkage, epidemiology, electronic health care databases, Italy.*

Introduzione

Il record linkage (RL) è lo strumento di elezione per l'integrazione delle informazioni provenienti da diverse sorgenti di dati.[1] La principale distinzione tra le tecniche di RL disponibili è tra quelle di natura deterministica (euristica) e quelle cosiddette probabilistiche.
Le tecniche di RL deterministico utilizzano una serie di regole basate sull'accordo esatto dell'insieme delle caratteristiche (campi) che costituiscono la chiave identificativa di un individuo. Il più semplice e intuitivo tra le tecniche deteministiche, il metodo esatto, prevede che due record provenienti da diverse sorgenti si riferiscano allo stesso individuo se l'intera chiave identificativa coincide perfettamente. Nella stessa categoria rientrano le procedure semi-deterministiche (o *stepwise*), caratterizzate da una sequenza di passi in cui la concordanza è valutata su un sottoinsieme di campi identificativi. Sebbene le tecniche deterministiche siano le più utilizzate, la critica principale che a esse si muove è la loro dubbia capacità di riconoscere un appaiamento in condizione di incertezza.[2]
Le tecniche probabilistiche, formalizzate da Fellegi e Sunter,[3] sono tutt'oggi basate sulle intuizioni e sulle prove empiriche di Newcombe.[4,5] Secondo queste tecniche nessun accordo o disaccordo singolo tra i campi identificativi è sufficiente per stabilire l'appaiamento, o il non appaiamento, di due record. Il criterio decisionale si basa sulla capacità discriminante e sull'attendibilità dei singoli campi identificativi.[6]
Il processo di RL può comportare errori di appaiamento che possono influenzare i risultati dello studio.[7-9] A oggi, le pubblicazioni che valutano l'effetto degli errori di RL sulla validità delle misure epidemiologiche sono scarse. La maggior parte di esse riporta dati relativi all'insieme di campi identificativi in grado di minimizzare gli errori di specifiche procedure deterministiche.[10-13] I risultati di tali studi, tuttavia, risultano scarsamente generalizzabili a contesti diversi da quelli da cui sono generati. Le tecniche probabilistiche sono più promettenti soprattutto sotto questo punto di vista, visto che, basandosi sulle caratteristiche dei campi identificativi e sulla qualità dei dati che li compongono, comportano criteri decisionali legati alla quota di errore accettabile nel contesto applicativo.[14-20] Tuttavia, soprattutto in Italia, si rende necessario lo sviluppo di esperienze in grado di diffonderne e consolidarne l'uso.
Gli studi epidemiologici pubblicati che non prevedono la raccolta ad hoc dei dati, ma sfruttano le potenzialità informative delle banche dati elettroniche, sono sempre più numerosi.[21] Studi effettuati in Svezia, Norvegia e Danimarca, nazioni con un consolidato codice univoco di identificazione del cittadino, utilizzano tecniche deterministiche per l'integrazione dei dati. Un panorama più vario si osserva nei paesi in cui manca, o non è ancora consolidato, l'utilizzo di una chiave di identificazione univoca. I ricercatori di Regno Unito, Giappone, Canada e Stati Uniti utilizzano varie tecniche di RL in relazione alla qualità del dato, con maggiore propensione per le tecniche probabilistiche.[21,22] In Italia si prediligono tecniche deterministiche, nonostante l'utilizzo della chiave univoca non sia ancora consolidato.[23-25]
Questo studio mette a confronto le tecniche di RL utilizzate in diverse realtà italiane per l'integrazione di archivi di dati sanitari e valuta le caratteristiche operative di una tecnica probabilistica standard messa a punto per l'utilizzo in epidemiologia. Si è valutata l'accuratezza dei metodi e la realistica implementazione della procedura probabilistica in base alle risorse informatiche disponibili nei centri di servizio epidemiologico partecipanti. Infine, è stato studiato l'impatto dei diversi metodi adottati su alcune misure di interesse epidemiologico.

Metodi

Centri partecipanti

Hanno partecipato allo studio i seguenti quattro centri di servizio epidemiologico:

■ CNR di Pisa, Istituto di fisiologia clinica, Sezione di epidemiologia e ricerca sui servizi sanitari, Reparto di bioingegneria e informatica medica;
■ Laziosanità, Agenzia di sanità pubblica, Regione Lazio;
■ Università degli studi di Bari, Facoltà di medicina e chirurgia, Dipartimento di scienze biomediche e oncologia umana;
■ Epidemiologia ambientale, ARPA Piemonte.
Sebbene l'analisi dei dati sia stata direttamente effettuata da ciascuno centro, gli aspetti metodologici, computazionali e le verifiche di validità sono stati curati da un unico centro presso l'Università degli studi di Milano Bicocca.

Archivi utilizzati
L'attenzione è stata concentrata sull'appaiamento di due banche dati automatizzate:
■ l'archivio anagrafico comunale, o quello degli assistiti, aggiornato al 31 dicembre 2003;
■ l'archivio delle schede di dimissione ospedaliera (SDO) dell'anno 2004.
Dall'archivio delle SDO sono stati estratti i ricoveri per infarto miocardico acuto (SDO-IMA), utilizzando il criterio proposto dal gruppo di lavoro AIE-SISMEC per i casi ospedalizzati riportato in questo numero della rivista, e per diabete mellito (SDO-diabete), ICD-9 250 riportato in almeno una delle diagnosi di dimissione, includendo anche i ricoveri in day hospital.
I centri piemontese e pugliese dispongono di archivi completi a livello regionale, mentre Laziosanità dispone dell'archivio anagrafico limitato ai soli residenti nel comune di Roma. Per il centro toscano entrambi gli archivi sono riferiti ai soli residenti nella città di Pisa (tabella 1).

Chiave identificativa
Per l'esecuzione delle procedure di RL è stata considerata una chiave identificativa univoca composta da: nome, cognome, data di nascita (divisa in tre campi distinti: giorno, mese e anno) e comuni di nascita e di residenza (codice Istat). Prima di procedere all'applicazione di queste tecniche, i campi che compongono la chiave identificativa sono stati normalizzati, utilizzando un algoritmo standard distribuito ai centri.

Record linkage deterministico
Ogni centro partecipante adotta una procedura di RL le cui caratteristiche dipendono dalle proprie esigenze, esperienze e disponibilità computazionali. Tali tecniche sono prevalentemente di tipo semideterministico, a esclusione del centro pugliese, che utilizza il metodo esatto con chiave identificativa diversa da quella sopra riportata (tabella 1).
In generale, le tecniche semideterministiche utilizzate dai centri collaborativi consistono in una serie di passi (step) di appaiamento di tipo deterministico esatto in cui la chiave identificativa viene ridotta, sottraendone alcuni campi, o parte di campi. Queste tecniche si basano su una serie di regole decisionali sempre meno restrittive, applicate alle coppie di record non appaiate nei passi precedenti. La decisione di escludere un campo dalla chiave identificativa dipende dall'esperienza e dalla logica del ricercatore. Le procedure stepwise utilizzate dai centri sono infatti molto diverse tra loro sia per i campi identificativi utilizzati, sia per il numero di passi che le caratterizzano (tabella 1).

Una breve nota sul linkage probabilistico
L'appaiamento probabilistico mira a minimizzare la probabilità di errori di linkage legati alla qualità dei dati, quali errori di trascrizione e dati incompleti. La teoria statistica prevede che per ogni campo identificativo vengano calcolati:
■ la probabilità che il campo concordi dato che i record confrontati si riferiscono allo stesso individuo (m); tale probabilità viene definita attendibilità;
■ la probabilità che il campo concordi dato che i due record si riferiscono a individui diversi (u), il cui complementare all'unità è definito potere discriminante ($1-u$).
Potere discriminante e attendibilità, ignoti a priori, sono rispettivamente calcolati dalla stima campionaria della probabilità di accordo casuale e mediante un algoritmo iterativo, noto come algoritmo EM (*expectation-maximization algorithm*).[26] A ogni coppia di record è quindi associato un peso complessivo w calcolato come funzione del potere discriminante e dell'attendibilità dei campi identificativi. Secondo la teoria classica, il criterio decisionale consiste nell'individuare due valori soglia di w. La soglia superiore si riferisce al valore corrispondente alla probabilità di appaiare due record che nella realtà non appartengono allo stesso individuo (α, probabilità di falsi positivi). La soglia inferiore si riferisce al valore corrispondente alla probabilità di non appaiare due record che appartengono allo stesso individuo (β, probabilità di falsi negativi). Risulta quindi intuitivo che le soglie vengano scelte in base agli errori α e β che si è disposti ad accettare. Tra le due soglie, tuttavia, rimane un intervallo di incertezza che comporta il controllo manuale delle corrispondenti coppie di record.
La procedura probabilistica consta di uno o più passi in relazione alle dimensioni degli archivi da confrontare. Nel caso di archivi di grandi dimensioni l'esecuzione in un unico passo di RL potrebbe risultare incompatibile con le risorse di calcolo disponibili, e comunque richiedere tempi di esecuzione eccessivamente lunghi. Il bloccaggio dei file consente di limitare l'insieme dei confronti da analizzare senza influenzare i risultati dell'analisi. Esso consiste nel suddividere gli archivi in sottoinsiemi esclusivi ed esaustivi in base alle modalità di uno o più campi identificativi ed effettuare i confronti all'interno di ogni sottoinsieme. Più

	Puglia	Pisa	Roma	Piemonte
anagrafica 31 Dicembre 2003	3.801.120	91.212	2.834.419	4.313.028
genere				
uomini	1.841.278 (48,4%)	43.207 (47,4%)	1.348.475 (47,6%)	2.080.494 (48,2%)
donne	1.959.842 (51,6%)	48.005 (52,6%)	1.484.566 (52,4%)	2.232.533 (51,8%)
missing	0 (0,0%)	0 (0,0%)	1.378 (0,0%)	1 (0,0%)
classi d'età				
0-19	864.875 (22,8%)	13.544 (14,8%)	480.137 (16,9%)	714.281 (16,6%)
20-39	1.124.780 (29,6%)	26.483 (29,0%)	809.244 (28,6%)	1.199.791 (27,8%)
40-59	987.781 (26,0%)	24.874 (27,3%)	804.566 (28,4%)	1.207.250 (28,0%)
60-79	665.276 (17,5%)	20.864 (22,9%)	597.789 (21,1%)	969.252 (22,5%)
80-99	154.152 (4,1%)	5.433 (6,0%)	135.193 (4,77%)	220.780 (5,1%)
100 e più	3.594 (0,1%)	11 (0,0%)	2.934 (0,1%)	1.618 (0,0%)
missing	662 (0,0%)	3 (0,0%)	4.556 (0,2%)	56 (0,0%)
ricoveri IMA 2004^	5.409	204	10.002§	8.668
ricoveri diabete 2004*	67.845	1.183	75.122§	44.587
caratteristiche hardware	processore Dual Core 2 Ghz, 2 GB RAM	processore 1 Ghz, 512 MB RAM	processore 1 Ghz, 512 MB RAM	processore Dual Core 2,8 Ghz, 1 GB RAM
procedura di RL centro	deterministico esatto cognome, nome, sesso, data di nascita, comune e provincia di nascita	stepwise 3 passi + controllo manuale cognome, nome, data e comune di nascita	stepwise 4 passi cognome, nome, sesso, data e comune di nascita	stepwise 30 passi codice fiscale dichiarato + codice fiscale ricostruito dopo controllo in base ad algoritmi di assonanza

^ ricoveri con diagnosi principale di infarto miocardico acuto (ICD-9 = 410) o con diagnosi principale di una presunta complicanza non evitabile (complicanza dell'IMA che si presume aumenti il rischio di morte del paziente indipendentemente dal trattamento) dell'infarto e IMA in diagnosi secondarie. Le diagnosi principali presunte complicanze dell'infarto sono i codici ICD-9 = 427.1, 427.41, 427.42, 427.5, 428.1, 429.5, 429.6, 429.71, 429.79, 429.81, 518.4, 780.2, 785.51, 414.10, 423.0; *hospital discharge with principal diagnosis of acute myocardial infarction (ICD-9 = 410) or of non avoidable myocardial infarction complications if reported with ICD-9 = 410 code in secondary diagnoses. Myocardial infarction complications are identified by ICD-9 codes = 427.1, 427.41, 427.42, 427.5, 428.1, 429.5, 429.6, 429.71, 429.79, 429.81, 518.4, 780.2, 785.51, 414.10, 423.0*

* ricoveri con codice ICD-9 = 250 in almeno una delle diagnosi di dimissione; *hospital discharge with ICD-9 = 250 code in at least one of the discharge diagnoses*

§ SDO a livello regionale; *regional archive of hospital discharges*

Tabella 1. Caratteristiche dei centri partecipanti e degli archivi utilizzati.

Table 1. Profile of participating centres and their administrative database.

passi con diversi campi di bloccaggio permettono di non perdere possibili appaiamenti.

Applicazione della procedura probabilistica

Il RL probabilistico utilizzato in questa applicazione consiste di due passi di linkage svolti in parallelo. Visto che più ricoveri possono appartenere a un unico individuo, ma non l'opposto, a ogni passo è previsto un appaiamento di tipo uno (archivio anagrafico) a molti (archivio SDO). I due passi sono definiti da diversi campi di bloccaggio in modo da permettere il completo confronto tra i campi scelti come chiave: comune e anno di nascita in un passo e giorno e mese di nascita e comune di residenza nell'altro.

Per entrambi i passi il criterio decisionale è quello che consiste nella scelta della soglia superiore corrispondente al minor peso complessivo positivo osservato. Dal confronto tra i risultati dei due passi una coppia è stata definita appaiata se il peso in entrambi superava quello della corrispondente soglia, non appaiata se non la superava in entrambi. In caso di discordanza è stato utilizzato un criterio più complesso la cui trattazione supera la natura intuitiva di questa nota.

Al fine di valutare l'influenza del criterio decisionale sui risultati, per ogni passo sono state individuate dieci soglie di decisione, oltre a quella iniziale, ottenendo così 121 possibili insiemi di coppie appaiate. Tali soglie sono state determinate sulla base di intervalli dei pesi complessivi di uguale ampiezza. E' stato quindi effettuato un controllo manuale delle coppie appaiate in modo da verificare il valore predittivo positivo (VPP) al variare del criterio decisionale.

	Puglia			Pisa			Roma			Piemonte		
IMA 2004												
	n_C = 1.197			n_C = 204			n_C = 1.333			n_C = 1.306		
	N_A	SE	SP	N_A	SE	SP	N_A	SE	SP	N_A	SE	SP
RLP	4.164	0,997 (0,990; 0,999)	0,962 (0,926; 0,983)	132	0,929 (0,864; 0,969)	0,969 (0,879; 0,977)	4.725	0,962 (0,942; 0,977)	0,986 (0,972; 0,994)	7.356	0,999 (0,994; 0,999)	0,899 (0,842; 0,941)
RLD	3.915	0,937 (0,917; 0,954)	1,000 (0,983; 1,000)	125	0,893 (0,820; 0,944)	1,000 (0,934; 1,000)	3.880	0,802 (0,764; 0,837)	1,000 (0,994; 1,000)	5.991	0,829 (0,802; 0,854)	1,000 (0,979; 1,000)
RLC	3.636	0,878 (0,852; 0,901)	1,000 (0,983; 1,000)	144	0,979 (0,932; 0,997)	0,891 (0,772; 0,961)	4.330	0,895 (0,864; 0,920)	1,000 (0,994; 1,000)	7.281	0,993 (0,985; 0,977)	0,962 (0,920; 0,985)
Diabete 2004												
	n_C = 1.503			n_C = 1.183			n_C = -			n_C = 1.486		
	N_A	SE	SP	N_A	SE	SP	N_A	SE	SP	N_A	SE	SP
RLP	57.904	0,998 (0,993; 1,000)	0,859 (0,805; 0,903)	854	0,986 (0,974; 0,994)	1,000 (0,986; 1,000)		-	-	34.490	0,999 (0,994; 1,000)	0,958 (0,926; 0,978)
RLD	54.120	0,968 (0,954; 0,978)	1,000 (0,984; 1,000)	802	0,926 (0,904; 0,945)	1,000 (0,986; 1,000)	-	-	-	28.568	0,843 (0,817; 0,867)	1,000 (0,988; 1,000)
RLC	50.230	0,908 (0,888; 0,926)	1,000 (0,984; 1,000)	878	0,984 (0,971; 0,992)	0,918 (0,877; 0,949)	-	-	-	34.153	0,998 (0,992; 1,000)	0,961 (0,931; 0,980)

n_C: numerosità campione SDO estratto per la stima di SE e SP; *sample extracted from hospital discharge archive to estimate SE and SP*
N_A: numero di eventi catturati; *number of events identified by record linkage procedures*
SE: sensibilità; *sensitivity*
SP: specificità; *specificity*
RLP: RL probabilistico con VPP pari al 98%; *probabilistic RL with 98% of PPV*
RLD: RL deterministico; *deterministic RL*
RLC: procedura di RL adottata dal centro; *RL adopted by participating centre*

Tabella 2. Numero di eventi catturati dalle procedure di record linkage e corrispondenti caratteristiche operative.

Table 2. Number of events identified, sensitivity and specificity of record linkage procedures.

Confronto tra le tecniche di record linkage

Le caratteristiche operative (sensibilità e specificità) delle tecniche di RL sono state stimate effettuando un controllo manuale su un campione casuale estratto dall'archivio SDO.

La numerosità del campione è stata stabilita in base alla dimensione dell'archivio, e tollerando che la stima del parametro sia compresa nell'intervallo ±2,5% con una probabilità $1-\alpha = 0,95$.

A causa della ridotta dimensione dell'archivio SDO del centro toscano, si è deciso di lavorare sull'intero campione.

Tassi di ospedalizzazione annuali per IMA e diabete, standardizzati per genere ed età (rispetto alla popolazione italiana Istat all'1 gennaio 2004) sono stati calcolati utilizzando le diverse tecniche di RL.

Un programma per il calcolo automatico

La procedura probabilistica è stata scritta e implementata in una macro SAS al fine di applicare un metodo uniforme tra centri (allegato 1, pagina 84). La macro comprende i principali e più rilevanti sviluppi delle teorie probabilistiche di RL: l'algoritmo EM, la correzione dei pesi in base al numero di caratteri e alla frequenza dei campi identificativi e alla presenza di dati mancanti.[27,28] Tale macro è adattabile alle diverse situazioni che si possono riscontrare nell'applicazione pratica: diverso numero e tipologia di campi identificativi, utilizzo o meno del bloccaggio dei file, diverso numero di campi utilizzabili per il bloccaggio, possibilità di introdurre a priori i parametri della distribuzione (m, u), possibilità di introdurre una determinata soglia di decisione o di crearne più di una e utilizzo dell'algoritmo di assegnazione lineare della somma nel caso l'ap-

Allegato 1

%RL_MIS

```
(dbA=,dbB=,          /*Nomi dei due dataset da appaiare.*/
PASSO=,              /*Indicare il numero del passo di linkage.*/
CORRISP=,            /*1= se relazione "UNO A UNO"(un record del dbA può essere appaiato con un solo record del
                     dbB e viceversa).
                     2= se relazione "UNO A MOLTI"(un record del dbA può essere appaiato con più record del dbB).*/
Nstringa=,           /*Numero di variabili stringa, se non ve ne sono Nstringa=0*/
stringaA=, stringaB=, /*Nome delle variabili stringa in dbA e dbB separate da un + e riportate nello stesso ordine.
                     Es:stringaA=nomeA+indirizzoA stringaB=nomeB+indirizzoB*/
Nconfro=,            /*Numero di variabili di confronto non stringa, se non ve ne sono Nconfro=0*/
confroA=, confroB=,  /*Nome delle variabili di confronto non stringa in dbA e dbB separate da un + e riportate nello
                     stesso ordine.*/
Nbloc=,              /*Numero di variabili di bloccaggio, se non ve ne sono Nbloc=0.*/
blocA=, blocB=,      /*Nome delle variabili di bloccaggio in dbA e dbB separate da un + e riportate nello stesso ordi-
                     ne.*/
                     /* Livelli d'errore */
SOGLIE=,             /*SI=Indicare i livelli d'errore prescelti; NO=Individua il primo peso positivo come soglia supe-
                     riore.*/
PROVE=,              /*Numero di soglie decisionali aggiuntive.*/
alfa=,               /*Probabilità di avere falsi positivi P=(M/U).*/
beta=,               /*Probabilità di avere falsi negativi P=(U/M).*/
STIMA_M=,            /*SI=Stimare le probabilità M per le variabili di confronto; NO=Non stimare.*/
M=,                  /*Se STIMA_M=NO inserire i valori delle M separate da un + e riporle seguendo l'ordine di in-
                     serimento (stringa e confro).*/
STIMA_U=,            /*SI=Stimare le probabilità U per le variabili di confronto; NO=Non stimare.*/
U=,                  /*Se STIMA_U=NO inserire i valori delle U separate da un + e riporle seguendo l'ordine di inse-
                     rimento (stringa e confro).*/
                     /*Variabili per la stima di M ed U.*/
CampioU=1000,        /*Numerosità del campione per la stima delle probabilità U 1000*1000*/
                     /*Valori iniziali per l'ALGORITMO EM che stima le probabilità M.*/
M_iniz=0.9,
p=0.01,
diff=0.001);
```

paiamento degli archivi sia di tipo uno a uno.[29,30] L'operatore deve solamente inserire le informazioni richieste secondo le modalità specificate tra i simboli /* */.

Risultati

La tabella 1 mostra che ogni centro partecipante ha contribuito allo studio con archivi di dimensione e qualità molto diverse tra loro. Inoltre, i centri differivano in relazione sia ai campi identificativi sia alle modalità con le quali il RL era normalmente effettuato. Infine, le risorse hardware disponibili in ogni centro, anch'esse diversificate, hanno limitato in alcuni casi la possibilità di effettuare alcune prove relative alle procedure probabilistiche che richiedono maggior tempo di esecuzione. Per esempio, la procedura

Puglia

Passo 2

	0	1	2	3	4	5	6	7	8	9	10
0	4.393 93,9	4.344 95,0	4.344 95,0	4.339 95,0	4.294 95,4	4.286 95,4	4.283 95,4	4.276 95,4	4.260 95,4	4.260 95,4	4.260 95,4
1	4.295 95,8	4.164 98,8	4.164 98,8	4.159 98,9	4.114 99,3	4.106 99,3	4.103 99,3	4.096 99,3	4.080 99,3	4.080 99,3	4.080 99,3
2	4.289 95,9	4.158 98,9	4.158 98,9	4.153 99,0	4.108 99,4	4.100 99,5	4.097 99,5	4.090 99,5	4.074 99,5	4.074 99,5	4.074 99,5
3	4.283 95,9	4.152 98,9	4.152 98,9	4.139 99,2	4.092 99,6	4.084 99,7	4.081 99,7	4.074 99,7	4.058 99,7	4.058 99,7	4.058 99,7
4	4.272 95,9	4.141 98,9	4.141 98,9	4.128 99,2	4.078 99,7	4.070 99,7	4.067 99,8	4.060 99,8	4.044 99,8	4.044 99,8	4.044 99,8
5	4.246 96,1	4.115 99,1	4.115 99,1	4.102 99,4	4.051 99,9	4.014 100,0	4.010 100,0	4.003 100,0	3.987 100,0	3.987 100,0	3.987 100,0
6	4.246 96,1	4.115 99,1	4.115 99,1	4.102 99,4	4.051 99,9	4.014 100,0	4.010 100,0	4.003 100,0	3.987 100,0	3.987 100,0	3.987 100,0
7	4.234 96,1	4.103 99,1	4.103 99,1	4.090 99,4	4.039 99,9	4.002 100,0	3.983 100,0	3.520 100,0	3.504 100,0	3.504 100,0	3.504 100,0
8	4.211 96,1	4.080 99,1	4.080 99,1	4.067 99,4	4.016 99,9	3.979 99,9	3.960 100,0	3.468 100,0	1.619 100,0	1.619 100,0	1.619 100,0
9	4.211 96,1	4.080 99,1	4.080 99,1	4.067 99,4	4.016 99,9	3.979 99,9	3.960 100,0	3.468 100,0	1.619 100,0	1.619 100,0	1.619 100,0
10	4.211 96,1	4.080 99,1	4.080 99,1	4.067 99,4	4.016 99,9	3.979 99,9	3.960 100,0	3.468 100,0	1.619 100,0	1.619 100,0	1.619 100,0

Pisa

Passo 2

	0	1	2	3	4	5	6	7	8	9	10
0	149 94,0	132 98,5	132 98,5	132 98,5	132 98,5	132 98,5	132 98,5	132 98,5	130 98,5	130 98,5	130 98,5
1	149 94,0	132 98,5	132 98,5	132 98,5	132 98,5	132 98,5	132 98,5	132 98,5	130 98,5	130 98,5	130 98,5
2	149 94,0	132 98,5	132 98,5	128 100,0	128 100,0	128 100,0	128 100,0	128 100,0	126 100,0	126 100,0	126 100,0
3	149 94,0	132 98,5	132 98,5	128 100,0	128 100,0	128 100,0	128 100,0	128 100,0	126 100,0	126 100,0	126 100,0
4	149 94,0	132 98,5	132 98,5	128 100,0	128 100,0	128 100,0	128 100,0	128 100,0	126 100,0	126 100,0	126 100,0
5	149 94,0	132 98,5	132 98,5	128 100,0	128 100,0	128 100,0	128 100,0	128 100,0	126 100,0	126 100,0	126 100,0
6	149 94,0	132 98,5	132 98,5	128 100,0	128 100,0	128 100,0	110 100,0	110 100,0	62 100,0	62 100,0	62 100,0
7	149 94,0	132 98,5	132 98,5	128 100,0	128 100,0	128 100,0	110 100,0	110 100,0	62 100,0	62 100,0	62 100,0
8	149 94,0	132 98,5	132 98,5	128 100,0	128 100,0	128 100,0	110 100,0	110 100,0	62 100,0	62 100,0	62 100,0
9	149 94,0	132 98,5	132 98,5	128 100,0	128 100,0	128 100,0	110 100,0	110 100,0	62 100,0	62 100,0	62 100,0
10	149 94,0	132 98,5	132 98,5	128 100,0	128 100,0	128 100,0	110 100,0	110 100,0	62 100,0	62 100,0	62 100,0

Roma

Passo 2

	0	1	2	3	4	5	6	7	8	9	10
0	6.458 73,8	6.451 73,9	6.429 74,1	6.402 74,4	6.344 75,1	6.323 75,3	6.291 75,6	6.239 75,4	6.239 75,4	6.178 75,4	6.157 75,3
1	6.262 76,1	6.255 76,2	6.233 76,5	6.206 76,8	6.148 77,5	6.127 77,7	6.095 78,0	6.043 77,9	6.043 77,9	5.982 77,8	5.961 77,8
2	5.931 80,4	5.924 80,5	5.902 80,8	5.875 81,1	5.817 81,9	5.796 82,2	5.764 82,5	5.712 82,4	5.712 82,4	5.651 82,4	5.630 82,4
3	5.669 84,0	5.662 84,1	5.640 84,5	5.613 84,9	5.554 85,7	5.533 86,0	5.501 86,4	5.449 86,3	5.449 86,3	5.388 86,4	5.367 86,4
4	5.367 88,7	5.360 88,9	5.338 89,2	5.311 89,7	5.248 90,6	5.224 91,0	5.192 91,4	5.140 91,3	5.140 91,3	5.079 91,5	5.058 91,5
5	5.211 91,4	5.204 91,5	5.182 91,9	5.155 92,4	5.092 93,4	5.065 93,8	5.024 94,2	4.972 94,2	4.972 94,2	4.911 94,4	4.890 94,4
6	5.141 92,6	5.134 92,7	5.112 93,1	5.085 93,6	5.022 94,6	4.995 95,0	4.930 95,4	4.876 95,4	4.876 95,4	4.815 95,6	4.794 95,7
7	5.050 93,6	5.043 93,8	5.021 94,2	4.994 94,7	4.931 95,8	4.904 96,2	4.837 96,6	4.341 96,3	4.341 96,3	4.249 96,5	4.228 96,6
8	5.050 93,6	5.043 93,8	5.021 94,2	4.994 94,7	4.931 95,8	4.904 96,2	4.837 96,6	4.341 96,3	4.341 96,3	4.249 96,5	4.228 96,6
9	4.938 95,3	4.931 95,4	4.909 95,8	4.882 96,4	4.819 97,5	4.792 97,9	4.725 98,3	4.229 98,3	4.229 98,3	3.705 98,4	3.683 98,4
10	4.938 95,3	4.931 95,4	4.909 95,8	4.882 96,4	4.819 97,5	4.792 97,9	4.725 98,3	4.229 98,3	4.229 98,3	3.705 98,4	3.683 98,4

Piemonte

Passo 2

	0	1	2	3	4	5	6	7	8	9	10
0	7.356 98,1	7.343 98,3	7.335 98,3	7.305 98,4	7.287 98,4	7.096 98,4	7.096 98,4	6.983 98,3	6.809 98,3	6.809 98,3	6.809 98,3
1	7.269 99,3	7.253 99,5	7.245 99,5	7.215 99,5	7.197 99,5	7.006 99,6	7.006 99,6	6.893 99,6	6.719 99,6	6.719 99,6	6.719 99,6
2	7.257 99,3	7.233 99,6	7.224 99,7	7.194 99,7	7.176 99,7	6.985 99,7	6.985 99,7	6.872 99,7	6.698 99,7	6.698 99,7	6.698 99,7
3	7.247 99,4	7.223 99,7	7.210 99,8	7.180 99,8	7.162 99,8	6.971 99,9	6.971 99,9	6.858 99,9	6.684 99,9	6.684 99,9	6.684 99,9
4	7.234 99,4	7.210 99,7	7.197 99,8	7.166 99,8	7.148 99,8	6.957 99,9	6.957 99,9	6.844 99,9	6.670 99,9	6.670 99,9	6.670 99,9
5	7.162 99,5	7.138 99,7	7.125 99,8	7.094 99,9	7.074 99,9	6.785 100,0	6.785 100,0	6.672 100,0	6.498 100,0	6.498 100,0	6.498 100,0
6	7.162 99,5	7.138 99,7	7.125 99,8	7.094 99,9	7.074 99,9	6.785 100,0	6.785 100,0	6.672 100,0	6.498 100,0	6.498 100,0	6.498 100,0
7	7.138 99,5	7.114 99,7	7.101 99,8	7.070 99,9	7.050 99,9	6.712 100,0	6.712 100,0	6.227 100,0	6.053 100,0	6.053 100,0	6.053 100,0
8	7.085 99,5	7.061 99,7	7.048 99,8	7.017 99,9	6.997 99,9	6.659 100,0	6.659 100,0	5.470 100,0	3.079 100,0	3.079 100,0	3.079 100,0
9	7.085 99,5	7.061 99,7	7.048 99,8	7.017 99,9	6.997 99,9	6.659 100,0	6.659 100,0	5.470 100,0	3.079 100,0	3.079 100,0	3.079 100,0
10	7.085 99,5	7.061 99,7	7.048 99,8	7.017 99,9	6.997 99,9	6.659 100,0	6.659 100,0	5.470 100,0	3.079 100,0	3.079 100,0	3.079 100,0

Figura 1. Numero di coppie appaiate e VPP (%) al variare delle soglie decisionali nei due passi della procedura probabilistica (SDO-IMA).

Figure 1. Number of matched record pairs and PPV (%) at cut-off value defined in the two steps of probabilistic record linkage procedure (HD-AMI).

probabilistica è stata eseguita in tutti i centri partecipanti a esclusione di Laziosanità, in cui non è stata possibile l'integrazione con le SDO-diabete.

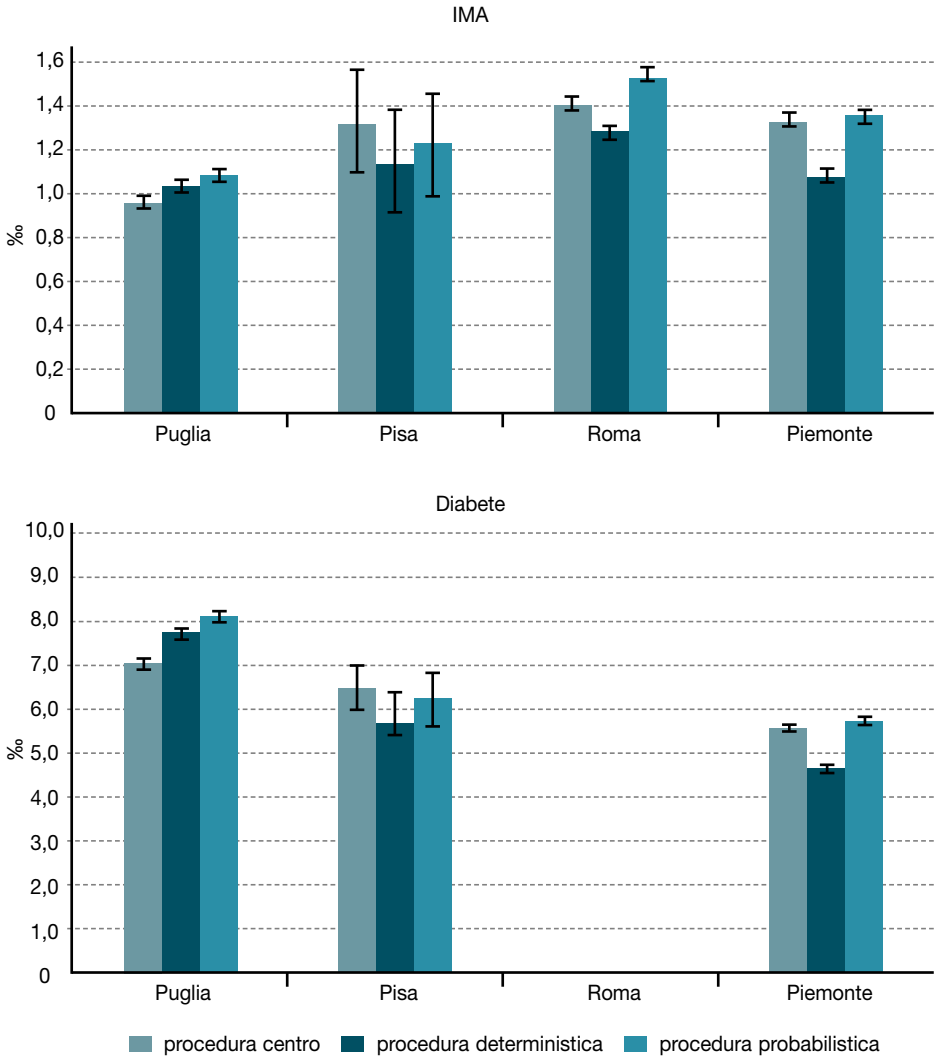
In figura 1 sono riportati, per ogni centro, il numero di coppie appaiate e il corrispondente VPP per ogni coppia di soglie riferite ai due passi di RL probabilistico applicato all'integrazione tra l'archivio anagrafico e quello SDO-IMA. Le soglie più restrittive (caratterizzate da peso più elevato per entrambi i passi e i cui risultati sono riportati nell'angolo inferiore destro delle tabelle), a meno di scarsa qualità dei dati, permettono di ottenere le stesse coppie appaiate osservate dall'applicazione del RL deterministico esatto. Il loro VPP è quindi per definizione del 100%. Viceversa, le soglie più permissive (caratterizzate da peso più basso per entrambi i passi e i cui risultati sono riportati nell'angolo superiore sinistro delle tabelle) permettono l'identificazione di un maggior numero di coppie appaiate, ma a discapito della riduzione del VPP, che risulta compreso tra 73% (Roma) e 98% (Piemonte). Bisognerebbe osservare che il numero delle coppie appaiate con il criterio più permissivo è circa 2-3 volte superiore rispetto a quello delle coppie appaiate con il criterio più restrittivo.

Per individuare un criterio uniforme non è stato possibile utilizzare le stesse soglie per i centri partecipanti in quanto il VPP della procedura è direttamente influenzato dalla qualità dei dati, risultata assai differente in ogni archivio. Si è quindi deciso di utilizzare come criterio decisionale standard la coppia di soglie che garantisse un VPP di almeno il 98%.

La tabella 2 confronta, per ogni centro, il numero di casi

Figura 2. Tasso annuale di ospedalizzazione per 1.000 anni/persona (e corrispondenti intervalli di confidenza al 95%) standardizzato per genere ed età.

Figure 2. Age and gender standardized annual rate of hospitalization per 1,000 person/year (and 95% confidence interval).



Standardizzazione rispetto alla popolazione italiana all'1 gennaio 2004 (dati Istat); *rates are standardized on the 2004 Italian population (Istat)*

individuati e le corrispondenti caratteristiche operative di ogni procedura di RL. La procedura probabilistica appaia in media l'11% di coppie in più rispetto a quella deterministica. Il valore aggiunto della procedura probabilistica rispetto a quella adottata dai singoli centri è pressoché nullo per Pisa e Piemonte. Per lo stesso motivo, i livelli di sensibilità e specificità delle procedure del centro piemontese e pisano sono paragonabili a quelle ottenute con la tecnica probabilistica. Viceversa, la sensibilità della procedura probabilistica risulta migliorata rispetto a quella specifica del centro a Roma e in Puglia.

La figura 2 confronta i tassi standardizzati di ospedalizzazione per le due malattie in studio calcolati con le tre procedure di RL. Come atteso, il metodo esatto fornisce stime inferiori a quelle ottenute con le altre metodologie, a esclusione del centro pugliese, che adotta una tecnica deterministica esatta con diversa chiave identificativa. Viceversa, dal confronto tra la procedura del centro e quella probabilistica emergono differenze meno marcate e di direzione variabile tra i centri.

Discussione

Dal confronto tra le performance delle tecniche di linkage nei quattro centri italiani è emerso che:

- il RL deterministico esatto è caratterizzato dai più bassi livelli di sensibilità e il suo utilizzo appare limitato alle situazioni in cui sono disponibili codici univoci di identificazione di buona qualità;
- la tecnica di RL probabilistico qui proposta risulta paragonabile a quella usualmente adottata dai centri quando questi implementano una revisione manuale dei record non appaiati o un controllo sulla qualità dei campi identificativi. Se questo non accade, la tecnica usualmente adottata dai centri comporta errori sistematici di direzione ed entità non note.

La performance delle tecniche probabilistiche è strettamente legata alla qualità dei dati disponibili. Criteri decisionali basati su livelli di errore α e β, che a priori sono considerati accettabili, generano stime di validità eterogenea e quindi non confrontabili. La soluzione qui proposta è quella di utilizzare un criterio decisionale basato su un livello di predittività positiva uniforme considerato a priori accettabile (nell'applicazione il VPP scelto è del 98% per tutti i centri). In queste condizioni i tassi di ospedalizzazione forniti dalla tecnica probabilistica qui proposta sono in alcuni casi caratterizzati da una validità sovrapponibile rispetto a quelli forniti dalla tecnica adottata dal centro. Questo è vero per i centri che dispongono di dati di buona qualità e per quelli che implementano una revisione manuale della qualità dei dati. Tuttavia, l'utilizzo della tecnica probabilistica qui proposta ha il vantaggio di fornire tassi caratterizzati dallo stesso livello di incertezza sistematica, consentendo quindi il confronto non distorto tra centri.

La stima della predittività positiva comporta la revisione manuale delle coppie appaiate, quindi l'impiego di risorse non sempre disponibili. In letteratura sono stati proposti alcuni algoritmi per la stima automatica del VPP associato alla procedura probabilistica.[31,32] Attualmente stiamo lavorando nel tentativo di validare tali tecniche nel contesto italiano ed eventualmente aggiornare la macro in questa direzione.

I tempi di implementazione delle differenti tecniche di RL dipendono dalla dimensione degli archivi da integrare e dalle caratteristiche hardware, RAM e spazio libero su disco. La procedura probabilistica è la più complessa e richiede quindi maggior tempo di esecuzione a parità di risorse informatiche. Nello specifico, la procedura probabilistica è stata implementata nei vari centri di servizio epidemiologico quando questi disponevano di risorse hardware sufficienti. L'efficienza della macro potrebbe essere migliorata per l'applicazione della procedura probabilistica ad archivi dati completi o, ancor meglio, implementata in linguaggi informatici più efficienti, in quanto le tecniche di bloccaggio dei file non sempre sono in grado di ridurre la dimensione dei dati che si stanno analizzando.

Tra le iniziative del gruppo di lavoro tese a fare maggiore chiarezza sull'appropriatezza d'uso delle tecniche di RL probabilistiche nel contesto dell'epidemiologia italiana, si segnalano quelle relative alla validazione di tali tecniche utilizzando gold standard di riferimento di validità nota. A tal fine sono stati avviati due progetti che prevedono l'integrazione e il confronto tra due registri con base di popolazione (Registro dialisi del Lazio e Registro cardiovascolare di Pisa) e i pertinenti archivi SDO.

Un ulteriore aspetto che richiede ulteriori approfondimenti riguarda l'utilizzo di tecniche che diano garanzie di rispetto delle leggi sulla tutela della privacy. Come visto, l'esperienza qui presentata ha utilizzato dati identificativi personali, ma questo può essere non compatibile con il rispetto della riservatezza. In letteratura sono state proposte alcune tecniche di RL che utilizzano codici identificativi criptati che il gruppo di lavoro sta approfondendo.[33,34]

Conflitti di interesse: nessuno.

Bibliografia

1. Dunn HL. Record linkage. *Am J Public Health* 1946; 36: 1412-16.
2. Scheuren F. Linking health records: human rights concerns. In: Proceeding of an international workshop and exposition: record linkage techniques; 20-21 March 1997; Arlington, USA. National Academic Press, Washington DC 1999.
3. Fellegi IP, Sunter A. A theory of record linkage. *JASA* 1969; 64: 1183-210.
4. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. *Science* 1959; 30: 954-59.
5. Newcombe HB, Kennedy JM. Making maximum use of the discriminatory power of identifying information. In: Communications of ACM 1962; 5: 563-66.
6. Newcombe HB. Handbook of record linkage, methods for health and statistical studies, administration and business. Oxford University Press, New York City 1988.
7. Brenner H, Schmidtmann I, Stegmaier C. Effects of record linkage errors on registry-based follow-up studies. *Stat Med* 1997; 16(23): 2633-43.
8. Brenner H, Schmidtmann I. Effects of record linkage errors on disease registration. *Methods Inf Med* 1998; 37(1): 69-74.
9. Brenner H, Schmidtmann I. Determinants of homonym and synonym rates of record linkage in disease registration. *Methods Inf Med* 1996; 35(1): 19-24.
10. Schnell R, Bachteler T, Bender S. Record linkage using error prone strings. Proceedings of the joint statistical meeting. *ASA* 2003; S: 3713-17.
11. Grannis SJ, Overhage JM, McDonald C. Real world performance of approximate string comparators for use in patient matching. *Medinfo* 2004; 11(Pt 1): 43-47.
12. Maizlish NA, Herrera L. A record linkage protocol for a diabetes registry at ethnically diverse community health centres. *JAMIA* 2005; 12: 331-37.
13. Scales DC, Guan J, Martin CM, Redelmeier DA. Administrative data accurately identified intensive care unit admissions in Ontario. *J Clin Epidemiol* 2006; 59(8): 802-07.
14. Newgard CD. Validation of probabilistic linkage to match de-identified ambulance records to a state trauma registry. *Acad Emerg Med* 2006; 13: 69-75.
15. Jaro M. Probabilistic linkage of large public health data files. *Stat Med* 1995; 14: 491-98.
16. Cook LJ, Olson LM, Dean JM. Probabilistic record linkage: relationships between file sizes, identifiers and match weights. *Methods Inf Med* 2001; 40: 196-203.
17. Nitsch D, Morton S, De Stavola BL et al. How good is probabilistic record linkage to reconstruct reproductive histories? Results from the Aberdeen children of the 1950s study. *BMC Medical Research Methodology* 2006; 6: 15.
18. Ramsay CR, Campbell MK, Glazener CM. Linking community health index and Scottish morbidity records for neonates. The Grampian experience. *Health Bulletin* (Edinburgh) 1999; 57(1): 70-75.
19. Shannon HS, Jamieson E, Walsh C et al. Comparison of individual follow-up and computerised record linkage using the Canadian mortality data base. *Can J Public Health* 1989; 80: 54-57.
20. The West of Scotland coronary prevention study group. Computerised record linkage compared with traditional patient follow-

up methods in clinical trials and illustrated in a prospective epidemiological study. *J Clin Epidemiol* 1995; 48(12): 1141-52.
21. Sorensen HT, Sabroe S, Olsen J. A Framework for evaluation of secondary data sources for epidemiological research. *Intl J Epidemiol* 1996; 25: 435-42.
22. Machado CJ. A literature review of record linkage procedures focusing on infant health outcomes. *Cad Saùde Pùblica* 2004; 20(2): 362-71.
23. Brocco S, Visentin C, Fedeli U et al. Monitoring the occurrence of diabetes mellitus and its major complications: the combined use of different administrative databases. *Cardiovasc Diabetol* 2007; 6: 5.
24. Franceschi S, Dal Maso L, Pezzotti P et al. Cancer and AIDS registry linkage study. Incidence of AIDS-defining cancers after AIDS diagnosis among people with AIDS in Italy, 1986-1998. *J Acquir Immune Defic Syndr* 2003; 34(1): 84-90.
25. Corrao G, Zambon A, Bertù L et al. Lipid lowering drugs prescription and the risk of peripheral neuropathy: an exploratory case-control study using automated databases. *J Epidemiol Community Health* 2004; 58(12): 1047-51.
26. Jaro MA. Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *JASA* 1989; 84: 414-20.
27. Winkler WE. Frequency based matching in the Fellegi–Sunter model of record linkage. ASA Proceedings of the section on survey research methods 1989: 788-93.
28. Winkler WE. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. ASA Proceedings of the section on survey research methods 1990: 354-9.
29. Cooper L, Steinberg D. Methods and application of linear programming. WB Saunders, Philadelphia 1974.
30 Burkard RE, Derigs U. Assignment and matching problems: solution methods with FORTRAN-programs. In: Lecture Notes in Economics and Mathematical System, n. 184. Springer-Verlag, New York 1981.
31. Blakely T, Salmond Clare. Probabilistic record linkage and a method to calculate the positive predictive value. *Intl J Epidemiol* 2002; 31: 1246-52.
32. Grannis SG, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. AMIA Symposium proceedings 2003: 259-63.
33. Quantin C, Binquet C, Allaert FA et al. Decision analysis for the assessment of a record linkage procedure. *Methods Inf Med* 2005; 44: 72-9.
34. Churches T, Christen P. Some methods for blindfolded record linkage. *BMC Medical Informatics and Decision Making* 2004; 28: 4-9.