



Identificazione, sperimentazione e validazione di alcuni indicatori di qualità delle attività sanitarie e socio-sanitarie del territorio

Roma 14, 15 settembre 2006

ASSR AGENZIA per i SERVIZI SANITARI REGIONALI

UNIVERSITÀ DEGLI STUDI DI ROMA
U **TuVagata**

Record linkage

Cristina Mazzali
Università degli Studi di Milano

- **Introduzione al record linkage**
- **Il record linkage probabilistico – teoria**
- **Il record linkage probabilistico – metodi di implementazione**
- **Casi di studio**

Introduzione al record linkage

Cos'è il record linkage
Applicazioni
Il contesto
Le origini
Il linkage deterministico
Il linkage probabilistico
Altre tecniche correlate al record linkage

Alcuni fenomeni si possono conoscere e analizzare solo legando informazioni contenute in banche dati diverse tra loro e magari nate per soddisfare scopi differenti.

Esempi in ambito sanitario:

Fenomeno della mortalità ospedaliera e dopo ricovero → legame tra la banca dati dei ricoveri e quella delle schede di morte

Analisi della continuità assistenziale di soggetti affetti da particolari patologie → legame tra banche dati dei ricoveri, dell'ambulatoriale, dei farmaci e delle schede di morte

Introduzione al record linkage - *Cos'è il record linkage*

Il record linkage è un processo di confronto di record da due o più sorgenti di dati con l'obiettivo di determinare quali coppie di record rappresentino la stessa entità nel mondo reale.

Introduzione al record linkage - *Cos'è il record linkage*

Durante la trascrizione, la battitura e la memorizzazione di dati come nomi o indirizzi l'introduzione di errori e variazioni è inevitabile.



- problemi nell'esecuzione del linkage
- incompletezza/incertezza dei risultati

Introduzione al record linkage - *Cos'è il record linkage*

Può accadere che le banche dati da legare **non contengano lo stesso codice identificativo** dei soggetti.

Le informazioni, identificative dei soggetti, presenti in entrambe le banche dati possono essere costituite da **dati non corretti o mancanti**



Necessità di tecniche per condurre a termine l'appaiamento nonostante la presenza di errori dei record appaiati

Introduzione al record linkage - *Cos'è il record linkage*

Se due record concordano su tutte le variabili, ed è improbabile che ciò sia avvenuto per caso, il livello di sicurezza che il link sia corretto sarà alto.

Se la maggior parte delle variabili discordano ci saranno pochi dubbi che il link non sia corretto.



Per le situazioni intermedie bisogna trovare un metodo per predire se il link sia vero o falso.

Introduzione al record linkage - *Cos'è il record linkage*

Le tecniche di record linkage possono essere utilizzate anche nella **ricerca di duplicati** all'interno di una base dati, cioè dei record che si riferiscono allo stesso soggetto.

In questi termini sta avendo grande diffusione per quei processi di pulizia dei dati molto diffusi in ambiente commerciale.



In ambito sanitario potrebbe essere applicato per esempio alla pulizia dell'anagrafe assistiti di un'azienda ospedaliera, o di altri enti. Oppure alla ricerca di ricoveri ripetuti nella base dati dei ricoveri.

Introduzione al record linkage - *Cos'è il record linkage*

Sono molteplici gli esempi di applicazioni delle tecniche di record linkage a studi in ambito sanitario:

Newman T.B. e Brown A.N.

“Use of commercial record linkage software and vital statistics to identify patient deaths”

J. Am. Med. Inform. Assoc., 4, 233-237, 1997

Zingmond D.S. et al.

“Linking hospital discharge and death records – accuracy and sources of bias”

Journal of Clinical Epidemiology, 57, 21-29, 2004

Liu S. e Wen S.

“Development of record linkage of hospital discharge data for the study of neonatal readmission”

Chronic Diseases in Canada, 20(3), 2000

Introduzione al record linkage - Applicazioni

The West of Scotland Coronary Prevention Study Group,

“Computerized Record Linkage: Compared with Traditional patient Follow-up in Clinical Trials and Illustrated in Prospective Epidemiological Study”

Journal of Clinical Epidemiology, 48(12), 1995

Howe G.R. e Lindsay J.

“A generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies”

Computer and Biomedical Research, 1981

Lewis G. e Sloggett A.

“Suicide, deprivation, and unemployment: record linkage study”

BMJ, 1998

Roos L.L. et al.

“Record Linkage Strategies, Outpatient Procedures, and Administrative Data”

Medical Care, 1996

Introduzione al record linkage - Applicazioni

Symposium on Health Data Linkage – Australia 2002

Clapton W. et al.

“Data linkage and the South Australian Cancer Registry”

Sundararajan V. et al.

“Rates and patterns of participation in cardiac rehabilitation in Victoria”

Finn J.

“The use of linked ambulance data to estimate the effect of comorbidity on determinants and outcomes of out-of-hospital cardiac arrest in Perth, Western Australia”

Semmens J. et al.

“Trends of Cataract Surgery and Post-Operative Endophthalmitis in Western Australia (1980-1998): A Population Based Study”

Introduzione al record linkage - Applicazioni

Symposium on Health Data Linkage – Australia 2002

Roos N. et al.

“Inequalities in child health: Bringing together new data sources to assess the roles of family, community, education and health care”

Andrews N.

“The value of linked data for research into surveillance and adverse events”

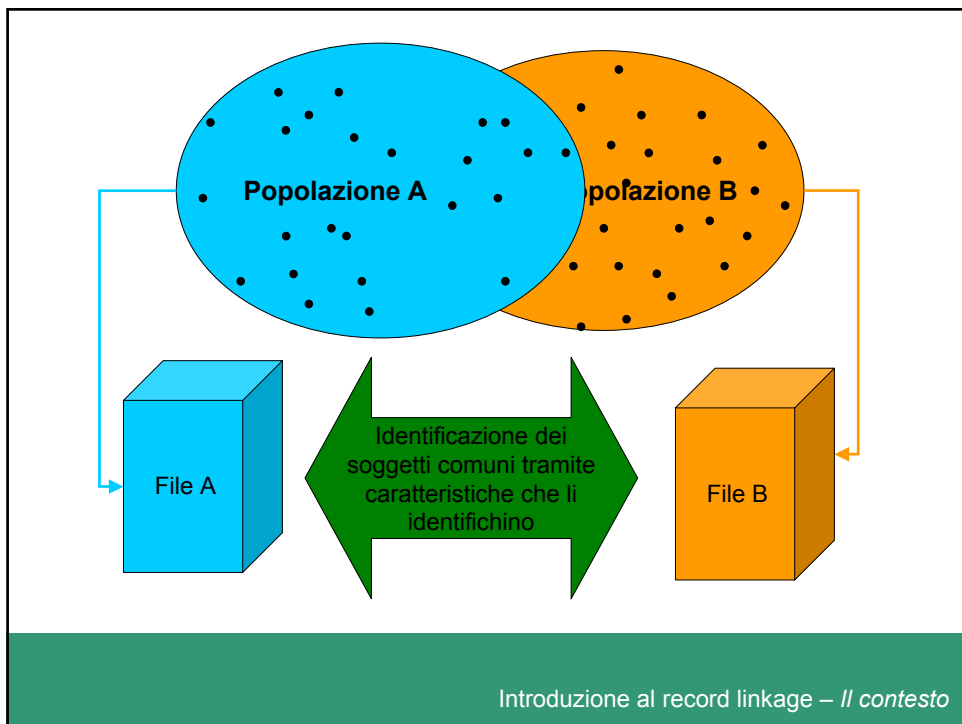
Bramedl K. et al.

The effect of locational and social disadvantage on utilization and outcomes of health care: Cardiovascular disease”

Hobbs M. et al.

“Applications of record linkage in cardiovascular disease: Monitoring trends and incidence, determinants and outcome of coronary heart disease and stroke”

Introduzione al record linkage - Applicazioni



Le prime forti richieste in questo campo vennero negli anni '50 con la diffusione dei computer e la possibilità di memorizzare le informazioni mediche



necessità di legare record medici, presi in differenti tempi e posti, in modo che la nuova informazione fosse correttamente legata al precedente record medico per la stessa persona.

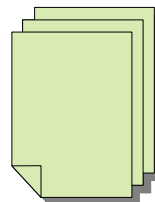
Inizialmente i linkage erano effettuati utilizzando procedure manuali basate su regole e decisioni ad hoc.



Alcuni impiegati:

- rivedevano liste stampate,
- ottenevano informazioni aggiuntive quando necessarie,
- prendevano decisioni di linkage

File erano ordinati alfabeticamente per nome o indirizzo per semplificare il processo di revisione. Per grandi file, i match potevano essere separati da numerose pagine di printout, in modo che diversi match potevano andare persi.



Introduzione al record linkage – *Le origini*

Il giudizio umano giocava un ruolo importante nel decidere quali record legare tra loro



casi borderline erano spesso risolti, non ricorrendo a un insieme di regole esplicite, ma dando i casi a uno o più 'esperti' ritenuti i più qualificati a giudicare secondo regole stabilite empiricamente

Introduzione al record linkage – *Le origini*

Con la diffusione dei computer si fecero tentativi per valutarne potenzialità e affidabilità per record linkage su grande scala

Primi lavori su fattibilità di applicare record linkage probabilistico su dati esistenti di morbilità e mortalità (Newcombe 1959)



Uso empirico dei rapporti di frequenza per quantificare il “grado di matching”

Introduzione al record linkage – *Le origini*

Fine anni '60 maggiore supporto teorico per metodi automatici di record linkage:

- Nathan 1967
- Tepping 1968
- D'Andrea Du Bois 1969
- Fellegi e Sunter 1969

Introduzione al record linkage – *Le origini*

Vantaggi del linkage automatizzato rispetto a quello manuale:

- supervisione centrale del processo
- migliore controllo di qualità
- maggiore velocità
- risultati più consistenti e riproducibili

Il record linkage automatizzato moderno consente di raggiungere risultati non inferiori a quelli raggiunti da impiegati esperti

Introduzione al record linkage – *Le origini*

Nonostante la maggior diffusione dei computer e i risultati teorici raggiunti negli anni '80 e '90 si continuò ad utilizzare metodi di linkage euristici e ad hoc

- poche persone il cui principale interesse professionale fosse il record linkage, molti dei lavori applicati fatti da persone che risolvevano i problemi di matching forse per la prima volta.
- statistici coinvolti molto tardi nel processo di matching, spesso dopo che i file erano già stati combinati e nuovi file creati.

Introduzione al record linkage – *Le origini*

Il match esatto (deterministico, all-or-none) genera link che sono basati sulla concordanza delle variabili di identificazione selezionate nei due record

quando le stringhe delle variabili a confronto concordano esattamente, i record sono considerati da combinare, altrimenti sono considerati come riferiti a persone diverse

Introduzione al record linkage – *Il linkage deterministico*

Il linkage deterministico fallisce quando si combinano record che contengono anche solo piccoli errori od omissioni:

- errori di comprensione dovuti alla pronuncia
- errori di editing
- variazioni nei nomi
- date di nascita, codici dei comuni troncati, incompleti o mancanti

Introduzione al record linkage – *Il linkage deterministico*

La condizione richiesta per un record linkage deterministico è che i record di entrambi i file contengano una variabile o caratteristica di persona od oggetto che sia idealmente:

- universalmente disponibile
- fissa
- facilmente registrata
- unica per l'individuo
- facilmente verificabile.

Introduzione al record linkage – *Il linkage deterministico*

Nella più semplice versione di match esatto l'output del match è chiaro: i due record sono appaiati o non lo sono.

Questa tattica semplifica la metodologia di record linkage, rendendola più pratica per lavori veloci e per realizzare processi rapidi su piccoli sistemi di calcolo.



Per confrontare record che contengono molte variabili per le quali c'è la possibilità di una bassa frequenza di errori, c'è un'altra versione di matching esatto, che "rilassa" i criteri di match.

In questa il numero di variabili che concordano è usato per determinare se i record debbano essere appaiati: si tratta di un match quasi-esatto.

Introduzione al record linkage – *Il linkage deterministico*

Stepwise Deterministic Strategy (Gomatam et al., 2002)

* I record sono legati in sequenze di più passi ognuno dei quali decide lo stato di linkage della coppia di record considerando concordanza esatta su differenti sottoinsiemi di identificatori.

* I passi che sono implementati per primi nella procedura usano insiemi di identificatori che sono considerati più affidabili di quelli nei passi successivi

* La scelta della sequenza può essere fatta tenendo conto delle conoscenze a priori di esperti.

Introduzione al record linkage – *Il linkage deterministico*

I metodi probabilistici sono stati sviluppati per dati e file che contengono errori e omissioni, e per i quali non è disponibile un identificatore unico, universale e di alta qualità.

E' un processo di raccolta e valutazione di informazioni per determinare se due record si riferiscono alla stessa persona

Non si richiede più un match esatto.

I match parziali sono valutati con modalità più sofisticate di quelle utilizzate dalle diverse strategie di linkage deterministico.

Introduzione al record linkage – *Il linkage probabilistico*

Nella forma standard del linkage probabilistico si calcola la probabilità che ci sia concordanza tra record che sono correttamente appaiati.

Si calcola anche la probabilità che ci sia un accordo per caso o come esito di errori tra record che non sono appaiati.

Vengono quindi calcolati dei pesi come rapporto tra queste due probabilità sulla base dei quali definire una coppia un linkage o meno.

Introduzione al record linkage – *Il linkage probabilistico*

Quando utilizzare il record linkage probabilistico?

In assenza di un identificatore univoco e universale

se si verificano errori a causa di una cattiva raccolta di informazioni, o perché i dati sono stati trascritti o inseriti in modo errato

se parte dei record sono mancanti o codificati in modo diverso; e i dati sono mancanti in modo casuale o sistematico

allora il linkage probabilistico sembra da preferirsi

Introduzione al record linkage – *Il linkage probabilistico*

- metodi di intelligenza artificiale
 - reti bayesiane
 - clustering
- metriche di similarità
- modelli bayesiani
- fuzzy matching