Obiettivi, strumenti e metodi per un utilizzo epidemiologico di archivi sanitari elettronici correnti in diverse aree italiane

Objectives, tools and methods for an epidemiological use of electronic health archives in various areas of Italy

Lorenzo Simonato,[1] Ileana Baldi,[2] Daniela Balzi,[3] Alessandro Barchielli,[3] Giuseppe Battistella,[4] Cristina Canova,[1] Giulia Cesaroni,[5] Giovanni Corrao,[6] Francesca Collini,[7] Susanna Conti,[8] Giuseppe Costa,[9] Moreno Demaria,[10] Carla Fornari,[11] Annunziata Faustini,[5] Claudia Galassi,[2] Roberto Gnavi,[12] Andrea Inio,[13] Fabiana Madotto,[11] Enrica Migliore,[2,14] Giada Minelli,[8] Michele Pellizzari,[15] Mariangela Protti,[16] Anna Romanelli,[16] Antonio Russo,[17] Mario Saugo,[15] Valeria Tancioni,[18] Roberta Tessari,[1,13] Alice Vianello,[1] Maria Angela Vigotti[19]

[1]Dipartimento di medicina ambientale e sanità pubblica, Università di Padova
[2]Servizio di epidemiologia dei tumori, ASO S. Giovanni Battista, CPO Piemonte e Università di Torino
[3]Unità di epidemiologia, Azienda sanitaria 10, Firenze
[4]Servizio di statistica ed epidemiologia, UOC controllo di gestione, Azienda ULSS 9, Treviso
[5]Dipartimento di epidemiologia, ASL RME, Roma
[6]Dipartimento di statistica. Facoltà di scienze statistiche, Università degli studi di Milano Bicocca
[7]Osservatorio di qualità, Agenzia regionale di sanità della Toscana
[8]Ufficio di statistica, CNESPS, Istituto superiore di sanità, Roma
[9]Dipartimento di sanità pubblica e microbiologia, Università di Torino
[10]Epidemiologia ambientale, ARPA Piemonte
[11]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
[12]Servizio di epidemiologia, ASL TO3, Regione Piemonte
[13]Unità di epidemiologia, Dipartimento di prevenzione, Azienda ULSS 12 Veneziana
[14]Unità di pneumologia, CPA-ASL TO2, Torino
[15]Servizio epidemiologico ULSS 4 Alto Vicentino, Thiene (Vi)
[16]Sezione di epidemiologia e ricerca sui servizi sanitari, IFC-CNR, Pisa
[17]Servizio di epidemiologia, ASL Città di Milano, Milano
[18]Laziosanità Agenzia di sanità pubblica
[19]Dipartimento di biologia, Università di Pisa

Riassunto

La disponibilità di archivi sanitari elettronici (ASE) correnti è andata aumentando in maniera considerevole negli ultimi due decenni. Allo scopo di confrontare le esperienze disponibili e verificare l'applicazione di procedure standardizzate di integrazione di ASE, nel 2005 si è costituito un gruppo di lavoro nazionale dell'Associazione italiana di epidemiologia (AIE) e della Società italiana di statistica medica ed epidemiologia clinica (SISMEC), in particolare per definire algoritmi per la stima della frequenza di alcune delle maggiori patologie che affliggono la popolazione italiana. La produzione di questo volume nasce dalla comune decisione di rendere disponibili in maniera esauriente i risultati e i metodi utilizzati per produrre tali stime. Viene presentato inoltre uno studio per confrontare le procedure di record linkage mediante l'utilizzo di una tecnica probabilistica standard in diverse aree italiane.

Undici centri, distribuiti in cinque regioni italiane, per una popolazione totale di 11.932.026 persone, hanno esplorato l'utilizzo di cinque fonti sanitarie correnti (certificati di morte, schede di dimissione ospedaliera, prescrizioni farmaceutiche, esenzioni da ticket e referti anatomopatologici) per un totale di 21.374.426 di record (anno 2003), per la stima di alcune categorie diagnostiche: diabete, cardiopatia ischemica, infarto miocardico acuto (IMA), ictus acuto, bronchite polmonare cronico-ostruttiva (BPCO), asma, malattia polmonare cronico-ostruttiva (MPCO). Sono stati elaborati e applicati algoritmi standard per identificare i casi (prevalenti/incidenti) delle patologie studiate e sono stati standardizzati i metodi per la stima della loro frequenza.

I centri partecipanti presentavano una notevole disomogeneità nella disponibilità temporale dei dati, nelle dimensioni e nell'apparente qualità degli archivi. Non sono state riscontrate particolari difficoltà nell'elaborazione e nell'applicazione degli algoritmi patologia-specifici. I risultati delle stime di frequenza delle singole categorie diagnostiche sono commentati nei singoli capitoli di questo volume.

L'insieme del lavoro svolto sottolinea la necessità di adottare metodi standardizzati nell'utilizzo degli ASE in considerazione della variabilità della qualità e completezza degli archivi selezionati e della difficoltà oggettiva a standardizzare le procedure di record linkage nei vari centri. Il pregio maggiore di questo lavoro è avere eliminato la variabilità generata dall'uso di algoritmi diversi.

(*Epidemiol Prev* 2008; 32(3) suppl 1: 5-14)

Parole chiave: stime di frequenza, archivi elettronici, record linkage

Abstract

The availability of Electronic Health Archives (EHA) has increased remarkably over the last twenty years.
As part of a joint project of the Italian Association of Epidemiology (AIE) and the Italian Association of Medical Statistics and Clinical Epidemiology (SISMEC), a workgroup of experts was set up in 2005 with the aim of comparing various experiences and of standardizing the procedures by which electronic sources can be integrated. In particular, the workgroup's aim was to estimate the frequency of certain major diseases using standard algorithms applied to EHA.
This volume is published with the purpose of making available in a common publication the methods and the results obtained. The results from a multicentre study using a standard approach to probabilistic record-linkage procedures are also included in a specific chapter.
Eleven Italian centres from five Italian regions with an overall population of 11,932,026 collected and treated more than 21,374,426 records (year 2003) from five electronic information sources: death certificates, hospital discharge records (including outpatient discharges), drug prescriptions, tax- exemptions, and pathology records in order to estimate the frequency of the following diseases: diabetes, ischemic heart diseases, acute myocardial infarction, stroke, asthma, chronic obstructive pulmonary disease, obstructive lung diseases. For each pathology a specific algorithm was developed and used by all centres for the identification of the prevalent/incident cases of the selected diseases. Standardized methods were used to estimate the rates.
The results confirm the need for a common standard approach to produce estimates based on EHA, considering the variability of the quality and of the completeness of the archives, and the difficulties of standardizing record-linkage operations in the various centres. The main achievement of this work was the elimination of the variability due to the use of different algorithms to identify cases using EHA.
(*Epidemiol Prev* 2008; 32(3) suppl 1: 5-14)

Keywords: *frequency estimates, electronic archives, record linkage*

Razionale

Lo sviluppo delle tecniche di miniaturizzazione dell'informazione ha reso la quantità di dati archiviabili in ogni settore in pratica illimitata. Di conseguenza, si sta ponendo in maniera nuova la questione della lettura e interpretazione delle informazioni archiviate, per evitare che formino capientissimi, ma sterili, magazzini elettronici non utilizzati.
In vari settori (economia, informazione e comunicazione, trasporti, sicurezza) l'investimento in ricerca ha portato all'introduzione di algoritmi che supportano i processi di risoluzione di problemi specifici.
In campo sanitario gli esempi più frequenti riguardano la diagnostica strumentale e per immagini, e la ricerca su sistemi di risoluzione diagnostica con l'obiettivo di sostituire, o affiancare, il processo di valutazione e sintesi clinica.
Un minor numero di esperienze sono invece in atto nel settore sanitario a livello di popolazione, nonostante la crescente domanda di prestazioni sanitarie richieda inevitabilmente informazioni complete e aggiornate sullo stato di salute della popolazione potenzialmente utili anche per la programmazione sanitaria. Vi sono tutt'oggi notevoli potenzialità di sviluppo dell'utilizzo a fini epidemiologici di enormi archivi alimentati dai presidi ospedalieri e dalle aziende sanitarie.
Una delle esperienze più interessanti di utilizzo di archivi elettronici a fini epidemiologici è stata realizzata nel campo della registrazione dei tumori. In Friuli-Venezia Giulia gli archivi sanitari, gestiti in un'unica wharehouse centralizzata, sono utilizzati come base per costruire il sistema di registrazione dei tumori della Regione, successivamente allargata alle popolazioni delle Provincie autonome di Trento e Bolzano.[1] Questo metodo, cosiddetto automatico (automated cancer registration, ACR), dapprima sperimentato nella Regione Veneto,[2] è costruito su un sistema di concordanze e compatibilità, formalizzato in un algoritmo di limitate dimensioni, e quindi facilmente esportabile e riproducibile. Esperienze simili sono state realizzate anche in altri paesi europei.[3]
Da questa prima esperienza, ormai consolidata, nasce l'esigenza di estendere l'utilizzo di tali strumenti alla stima della frequenza di altre patologie nella popolazione per migliorare le conoscenze sullo stato della salute ai fini della gestione e della programmazione della sanità pubblica.
In ambito sanitario, la programmazione riveste un'importanza cruciale per lo sviluppo del sistema: le risorse sempre più limitate e i bisogni della popolazione, tendenzialmente crescenti, impongono di affrontare le questioni in modo globale, con raziocinio e competenza, basandosi su informazioni certe.
Intervenire sullo stato di salute della popolazione significa combattere le malattie pianificando attività di cura, prevenzione e promozione della salute.
Le potenzialità informatiche sviluppatesi negli ultimi 25 anni hanno reso possibile l'archiviazione di un'enorme quantità di informazioni a scopi prevalentemente amministrativi o economici (per esempio, rimborsi fra aziende sanitarie, calcolo dei DRG eccetera) e proprio per questo raccolte con continuità. L'importanza di questi archivi dal punto di vista epidemiologico risiede nella presenza, accanto a quelle anagrafiche, anche di informazioni concernenti l'ambito più strettamente diagnostico. Rientrano in queste fonti gli archivi di mortalità, delle schede di dimissione ospedaliera, dei referti di anatomia patologica, gli archivi delle prescrizioni farmaceutiche, delle visite specialistiche eccetera.

La principale registrazione a livello nazionale è quella della mortalità. La prassi di archiviazione di dati sanitari è comunque diffusa, almeno in parte, presso tutte le aziende sanitarie, e sottoposta a specifici regolamenti regionali; gli archivi sono generalmente a sé stanti, organizzati e gestiti separatamente.
La potenzialità e l'innovazione rappresentate dall'impiego di questi archivi in epidemiologia sono legate al passaggio a una loro gestione integrata: le informazioni raccolte attraverso un canale informativo possono così essere controllate sulla base dell'incrocio con gli altri archivi e la stessa portata informativa delle basi di dati si accresce notevolmente.
La sfida che si presenta ora in epidemiologia è dunque lo sfruttamento di questo prezioso apparato informativo: le prospettive di utilizzo dei dati raccolti con sistematicità, se organizzati in sistemi integrati, sono infatti molteplici. In questo campo sono già attive altre esperienze tra cui, a livello ministeriale, il progetto Mattoni che affronta il tema degli strumenti informativi nell'ambito del Servizio sanitario nazionale,[4] il progetto SiVEAS-Sistema nazionale di verifica e controllo sull'assistenza sanitaria,[5] il progetto ARNO,[6,7] lo studio ReClust Record Linkage System dell'Istituto Mario Negri Sud.[8]
A livello nazionale, in un primo informale incontro di esperti del settore tenutosi a Firenze nell'aprile 2004, venne auspicata la standardizzazione dei metodi di utilizzo degli archivi elettronici nel definire la frequenza delle patologie a livello di popolazione e si propose la creazione di un gruppo di lavoro (GdL) dell'Associazione italiana di epidemiologia (AIE) su questo tema.
Questo GdL venne formalizzato in occasione del Congresso dell'AIE tenutosi a Pisa nel 2005 e si arricchì successivamente della collaborazione con la Società italiana di statistica medica ed epidemiologia clinica (SISMEC) costituendo il Gruppo nazionale di lavoro AIE/SISMEC sull'utilizzo epidemiologico degli archivi sanitari elettronici (ASE).
Alla base dell'attività del GdL stava la necessità di conoscere e censire le esperienze in corso di uso epidemiologico degli ASE, e di coordinare e condividere tra diversi gruppi di ricerca queste conoscenze. Infatti, solo la sinergia di competenze e ricerche metodologiche può consentire di evitare che singole esperienze portino a soluzioni diverse sempre più difficilmente ricongiungibili. Viceversa, non si è volutamente affrontato l'aspetto della normativa sulla tutela dei dati personali, anche per la necessità di competenze di tipo legislativo non presenti nel GdL.
Il GdL ha affrontato tre aspetti diversi dell'utilizzo di ASE: le operazioni di record linkage (RL) fra diversi archivi; le stime di frequenza di malattia a livello di popolazione; l'importanza dell'uso di covariate socioeconomiche.
Il sottogruppo per l'esame delle procedure di record linkage probabilistico, coordinato da Giovanni Corrao, si è proposto per la sperimentazione e la simulazione dei metodi conosciuti per l'integrazione delle basi di dati a disposizione dell'epidemiologia.
Il sottogruppo per la produzione di stime di frequenza di popolazione di categorie diagnostiche, coordinato da Lorenzo Simonato, si è posto l'obiettivo di esplorare le possibilità di uso degli ASE per la stima di alcune patologie rilevanti, da consolidare nella prospettiva di un'eventuale istituzionalizzazione delle procedure con estensibilità a livello nazionale.
Il terzo sottogruppo, coordinato da Giuseppe Costa, si è posto l'obiettivo di studiare l'utilizzo di covariate socioeconomiche individuate tramite archivi elettronici per arricchire gli studi di epidemiologia analitica. Per questo terzo aspetto l'attività si è basata soprattutto su questioni metodologiche e sulle esperienze esistenti,[9] non essendo ovviamente attuabile un contributo originale su dati non ancora disponibili.
Il coordinamento del gruppo, formato da Giovanni Corrao, Giuseppe Costa e Lorenzo Simonato, ha cercato di coinvolgere i centri più attivi in campo nazionale, pur restando aperto a contributi spontanei provenienti sia dall'ambito universitario sia dal Servizio sanitario nazionale. La partecipazione non va comunque considerata esaustiva di tutte le esperienze in atto su questo tema.
E' stata preziosa la collaborazione con il Centro nazionale di epidemiologia, sorveglianza e promozione della salute (CNESPS) che, per conto del Centro per il controllo delle malattie (CCM), ha seguito i lavori del GdL grazie alla partecipazione di Susanna Conti e Giada Minelli.
Ricordiamo che anche la riunione di primavera 2007 dell'AIE ha trattato questo tema in collaborazione con il CNESPS.
Lo scopo principale dell'attività del GdL era confrontare le esperienze disponibili e verificare l'applicazione di procedure standardizzate di integrazione dei dati da archivi sanitari elettronici correnti, in particolare per stimare la frequenza di alcune categorie diagnostiche.
La produzione di questo volume nasce dalla comune decisione di rendere disponibili in maniera esauriente i risultati e i metodi utilizzati per produrre tali stime. I risultati presentati sono il frutto di circa due anni di attività del GdL. Essi hanno lo scopo di fornire agli operatori del Servizio sanitario nazionale un esempio organizzato di utilizzo standardizzato di archivi sanitari elettronici.
In particolare viene presentato e discusso il lavoro del sottogruppo per la produzione di stime di frequenza di alcune voci nosologiche in diverse aree italiane (capitoli 2-8) e di quello per l'esame delle procedure di record linkage probabilistico (capitolo 9), illustrando la metodologia seguita, i risultati ottenuti e le criticità emerse, che richiedono ulteriori approfondimenti.

I risultati ottenuti vanno valutati con cautela e non devono essere considerati conclusivi in considerazione del carattere esplorativo che inevitabilmente qualifica gli algoritmi elaborati e qui utilizzati, della mancanza di validazione sistematica attraverso il confronto con fonti indipendenti, e della ristretta finestra temporale durante la quale è stato possibile utilizzare le fonti correnti.
Gli algoritmi qui presentati non vanno quindi considerati come modelli definitivi proposti dalle due associazioni scientifiche, ma come il risultato di una prima fase di razionalizzazione dell'approccio metodologico.

Materiali e metodi

I materiali e metodi presentati in questo paragrafo si riferiscono a quelli comuni utilizzati per le stime della frequenza di alcune malattie (vedi capitoli da 2 a 8), mentre i metodi utilizzati per il record linkage sono descritti in maggior dettaglio nel capitolo 9.
Dopo aver condiviso i risultati delle diverse precedenti esperienze condotte dai centri partecipanti nell'utilizzo di archivi elettronici, sono stati identificati gli obiettivi di un lavoro comune definendo quale apporto conoscitivo, ma soprattutto metodologico, ci si prefiggeva di fornire.
Si tratta essenzialmente di «catturare» i casi di una malattia incrociando le informazioni di più archivi sanitari elettronici correnti, senza ricorrere alla raccolta attiva della casistica. Questo metodo si basa su alcuni assunti:

- di norma, ogni contatto tra assistito e servizio sanitario è registrato in un archivio specifico e riporta informazioni codificate, che possono permettere di risalire alla motivazione del contatto;
- la maggior parte delle diagnosi di malattia è effettuata sulla base dei risultati ottenuti da esami specifici e segue un processo logico semplice.

Partendo da queste considerazioni, si è esplorata la possibilità di definire un caso mediante un algoritmo decisionale che ripercorre i contatti che un paziente ha con il servizio sanitario per la diagnosi e il trattamento della malattia (ricoveri, uso farmaci eccetera).
L'efficacia di questo approccio è una funzione della probabilità che le prestazioni di cui fruisce un paziente siano registrate e correttamente codificate negli ASE, ed è quindi variabile a seconda delle caratteristiche cliniche della malattia, della sua gravità e dell'accessibilità al servizio sanitario.
E' evidente che ad archivi elettronici molto incompleti e/o con bassa qualità delle informazioni contenute non devono essere applicate le metodologie qui descritte.

Centri partecipanti allo studio

I centri, le aree e le dimensioni delle popolazioni coinvolte sono presentate nella tabella 1.
Le aree incluse nello studio presentano una notevole disomogeneità per quanto riguarda le dimensioni delle popolazioni coinvolte e, come emerge nella presentazione delle singole patologie, anche per i periodi di disponibilità degli archivi. La numerosità maggiore si ha nelle regioni Lazio e Toscana, che rappresentano da sole più di due terzi della popolazione in studio.
Le aree incluse nello studio non costituiscono un campione rappresentativo della popolazione italiana (obiettivo che comunque non rientrava tra gli obiettivi del GdL), ma so-

Centro	Ambito territoriale (area)	Uomini	Donne	Totale
AULSS 12 Veneziana	Venezia-Mestre, Marcon, Quarto d'Altino, Cavallino-Treporti (Venezia)	144.618	159.326	303.944
ASL Città di Milano^	Milano	627.241	697.581	1.324.822
ULSS 9 Treviso*	Treviso e Sud Provincia (Treviso)	195.595	200.935	396.530
ULSS 4 Alto Vicentino	Thiene e comuni dell'alto vicentino (Thiene)	88.434	91.008	179.442
ASL 5 Torino, CPO Piemonte, CPA ASL 4	Torino	409.384	455.367	864.751
CNR Pisa	Pisa	42.241	46.723	88.964
ASL 10 Firenze	Firenze e 33 comuni della provincia (Firenze)	368.936	406.840	775.776
Agenzia regionale sanità Toscana (ARS)	Regione Toscana	1.691.051	1.825.245	3.516.296
ASL Roma E§	Roma	1.186.748	1.329.918	2.516.666
Laziosanità	Regione Lazio	2.466.028	2.679.777	5.145.805
ASL Taranto	Taranto	95.826	104.610	200.436
totale		5.718.177	6.213.849	11.932.026

^ centro che ha contribuito unicamente alle stime delle due patologie neoplastiche; *centre contributing only to the two neoplastic pathologies*
* popolazione al 31/12/2005; *population at 31/12/2005*
§ popolazione al censimento 2001; *population at census 2001*

Tabella 1. Distribuzione delle popolazioni, per sesso e centro. Anno 2003.

Table 1. Distribution of populations, by sex and centre. Year 2003.

no tuttavia in grado di rivelare le diversità nella disponibilità di dati a livello nazionale.

Procedure di record linkage

Per record linkage si intende la procedura utilizzata per determinare se due record, appartenenti a due diversi set di dati, si riferiscono a uno stesso individuo.
Le procedure di record linkage tra fonti utilizzate dai vari centri sono tutte di tipo deterministico, ovvero basate sull'accordo esatto dell'insieme delle caratteristiche (campi) che costituiscono la chiave identificativa di un individuo. Nel caso in cui quest'ultima sia composta da più campi identificativi sono state applicate procedure semideterministiche (*stepwise*) caratterizzate da una sequenza di passi in cui la concordanza esatta è valutata su un sottoinsieme di campi identificativi. Tipo e numero di campi identificativi, sequenza e numero di passi variano tra centri. Il criterio deterministico esatto è utilizzato quando è disponibile un unico campo identificativo. I centri prevedono generalmente l'allineamento con l'anagrafe sanitaria o comunale (tabella 2).
Tale disomogeneità di metodi di record linkage è giustificata dalla mancanza, soprattutto in Italia, di esperienze che siano in grado di diffondere e consolidare l'uso di una metodologia rispetto a un'altra. Nel capitolo 9 si valuta l'effetto della disomogeneità di metodo su misure di frequenza e si valida una tecnica probabilistica per l'utilizzo in epidemiologia.
Nella tabella 3 sono descritte in maggior dettaglio le procedure di record linkage utilizzate dai centri.

Categorie diagnostiche indagate

Le categorie diagnostiche proposte e discusse dai centri partecipanti sono state numerose. I criteri addottati per la selezione comprendevano la rilevanza della loro frequenza in Italia, l'esistenza di fonti correnti utilizzabili per la definizione di un algoritmo di stima, l'interesse di un numero sufficiente di centri all'elaborazione dei dati, e una valutazione positiva dei risultati nella prospettiva di un loro utilizzo da parte degli operatori sanitari.
Va anche ricordato che nella maggioranza dei centri erano già state condotte esperienze di utilizzo delle fonti correnti che hanno costituito l'indispensabile background per le decisioni prese.
A conclusione di questo processo di elaborazione e valutazione sono state escluse alcune categorie diagnostiche, come la schizofrenia, per la quale non è stato individuato un algoritmo condiviso tra i partecipanti del gruppo.
In questo volume, nei capitoli da 2 a 8 sono presentate le seguenti categorie diagnostiche: diabete, cardiopatia ischemica, infarto miocardico acuto, ictus acuto, broncopneumopatia cronico-ostruttiva (BPCO), asma, malattie polmonari cronico-ostruttive (MPCO), inclusive di asma e BPCO.

Procedure	Centri
allineamento con anagrafe sanitaria o comunale	tutti
linkage deterministico esatto unico campo identificativo (codice fiscale, codice sanitario o loro parti)	ULSS 4 Alto Vicentino, ASL 5 Torino, ASL 10 Firenze, Regione Toscana, ASL Roma E (broncopneumopatia cronico-ostruttiva)
linkage semi-deterministico più campi identificativi	altri centri
revisione manuale sui non appaiati	ASL Città di Milano, ULSS 9 Treviso, CNR Pisa, ASL Taranto, CPO Piemonte

Tabella 2. Schema di procedure di record linkage.

Table 2. Record linkage procedure chart.

Si è inoltre condotto un esercizio di stima e di validazione su due patologie neoplastiche (tumori della mammella e del colon retto) in aree nelle quali erano operanti Registri tumori. Per la peculiarità della rilevazione dell'incidenza dei tumori in Italia (sistema da tempo consolidato, ampiezza della rete di rilevazione dei Registri tumori, disponibilità di metodologie di stima a partire dai dati rilevati eccetera), e data la caratteristica ancora esplorativa delle stime effettuate dal GdL, la patologia oncologica non è stata inclusa in un capitolo specifico, ma sintetizzata in appendice.
Per ciascun gruppo nosologico è stato incaricato un referente, scelto tra i partecipanti al GdL in base all'interesse e alle esperienze in materia, con funzione di responsabile del coordinamento dei centri coinvolti nel processo di stima, della raccolta dei dati e della discussione dei risultati.
Per ogni categoria diagnostica, un numero variabile di centri si è impegnato ad applicare gli algoritmi stabiliti per confrontare poi i risultati ottenuti in contesti diversi sotto il profilo della popolazione e dell'organizzazione degli archivi (tabella 4).

Metodi utilizzati per le stime

Gli archivi elettronici correnti utilizzati per le stime sono quelli contenenti le cause di morte (CM), le schede di dimissione ospedaliera, con i ricoveri ordinari e in regime di day hospital e comprensive della mobilità passiva regionale ed extraregionale (SDO), le prescrizioni farmaceutiche (PF), le esenzioni ticket (ET), e i referti di anatomia patologica (AP), utilizzati esclusivamente per la stima delle due patologie neoplastiche. Poiché la disponibilità delle basi dati nei diversi centri non risaliva allo stesso anno, si è stabilita una data di partenza comune, affinché nel confronto fra le stime ottenute non si dovesse tener conto di eventuali trend temporali e del diverso apporto delle fonti nel tempo, dimensioni difficilmente quantificabili e interpretabili.

Centro	Tipo di record linkage
AULSS 12 Veneziana	Allineamento di ciascuna fonte con anagrafe sanitaria attraverso 5 combinazioni di chiavi a cascata: K1 tessera sanitaria, cognome, nome, data di nascita K2 tessera sanitaria, cognome, data di nascita K3 tessera sanitaria, nome, data di nascita K4 tessera sanitaria, cognome, nome K5 cognome, nome, data di nascita
ASL Città di Milano	Allineamento con l'anagrafe del comune di Milano attraverso procedure di record linkage che prevedono un controllo anche sull'anagrafe degli assistiti della Lombardia. Procedura a cascata con rimozione dei record linkati attraverso le seguenti combinazioni di chiavi: • codice fiscale completo (codice fiscale dell'anagrafe assistiti partendo dal codice sanitario per il periodo antecedente al 1999) • cognome e nome (rimuovendo spazi e caratteri speciali), data di nascita e comune di nascita • codice fiscale composto dai primi 15 caratteri • cognome e nome (rimuovendo spazi e caratteri speciali), data di nascita • sulla base residua dei record non linkati revisione manuale con revisione anche dell'anagrafica delle cartelle cliniche relative ai ricoveri non linkati.
ULSS 9 Treviso	Linkage di ciascuna fonte all'anagrafe assistiti aziendale e regionale attraverso procedure deterministiche a cascata di questo tipo: K1 codice fiscale K2 codice sanitario K3 nome+cognome+data nascita+comune di nascita K4 nome+cognome+data nascita (con controlli manuali)
ULSS 4 Alto Vicentino	Linkage di ciascuna fonte con anagrafe sanitaria con tessera sanitaria.
Comune di Torino (ASL 5, CPO Piemonte, CPA-ASL 4)	v Per stima IMA, ictus acuto, cardiopatia ischemica e diabete Tutti gli archivi sanitari utilizzati, tranne l'archivio di mortalità, sono stati linkati con l'anagrafe del comune di Torino per aggiornarli alla data di riferimento (in questo caso il 31.7.2003), usando come chiave di linkage il codice fiscale. L'abbinamento con l'anagrafe ha consentito di recuperare per le osservazioni abbinate nei vari archivi anche il codice identificativo personale dell'anagrafe, che è stato poi utilizzato come chiave di linkage fra gli archivi. L'archivio di mortalità già contiene il numero identificativo personale dell'anagrafe, che è stato usato come chiave di abbinamento fra questi due archivi. L'abbinamento fra l'archivio delle prescrizioni farmaceutiche e anagrafe non è immediato in quanto l'archivio delle prescrizioni farmaceutiche contiene come identificativo personale soltanto il codice di tessera sanitaria (a 11 caratteri). L'archivio delle prescrizioni farmaceutiche è stato pertanto preliminarmente abbinato con l'archivio Banca assistibili regionale (di fonte Asl, basata sulla procedura di scelta e revoca del medico di base) per ottenere per ogni soggetto il codice fiscale. ❖ **Per stima patologie tumorali** La validazione di entrambi gli algoritmi è avvenuta mediante confronto tra i casi selezionati dalle schede di dimissione ospedaliera e i dati nominali del Registro tumori del Piemonte (città di Torino) corrispondenti alle patologie in esame. L'approccio utilizzato è quello di abbinamento esatto e prevede l'applicazione in serie di due chiavi di linkage a cascata: • il codice fiscale, originale o ricostruito • sesso, primi 4 caratteri del cognome (senza spazi, apostrofi e accenti), codice Istat del comune di nascita e data di nascita.

Tabella 3. Procedure di record linkage utilizzate dai centri.

Nella tabella 5 sono illustrate le numerosità degli archivi elettronici per area geografica, anno (periodo 2002-2004) e fonte informativa.
In relazione alle caratteristiche delle fonti informative utilizzate e delle patologie studiate, per alcune malattie è stato possibile stimare misure di incidenza, per altre misure di prevalenza.

Per quanto riguarda la prevalenza, si è scelto di calcolarla in due modi differenti a seconda delle caratteristiche della malattia e della conseguente diversa necessità di ricorso a prestazioni sanitarie:
■ definendola come numero di malati presenti nella popolazione nell'anno t, calcolati sulla base delle fonti riferite a quell'anno;

Centro	Tipo di record linkage
	La procedura automatizzata è stata poi integrata da una componente di abbinamento manuale. ❖ **Per stima bronco pneumopatia cronico-ostruttiva, asma, malattie polmonari cronico-ostruttive** Il record linkage per allineare gli archivi (prescrizioni farmaceutiche, schede di dimissione ospedaliera, esenzioni ticket per patologia e registro di mortalità) è stato eseguito utilizzando come chiave univoca il codice fiscale (16 lettere). Successivamente è stato effettuato un ulteriore controllo, utilizzando come chiave di linkage il cognome, il nome e la data di nascita dei soggetti (dove presente l'informazione).
CNR Pisa	❖ **Linkage tra anagrafe comunale e fonti con 4 chiavi: nome, cognome, data di nascita, comune di nascita in due fasi:** fase 1: linkage automatico: corrispondenza esatta delle 4 chiavi e appaiamento per discordanza di una delle chiavi numeriche di linkage fase 2: linkage decisionale con risoluzione manuale dei soggetti non linkati in fase 1
ASL 10 Firenze	❖ **Linkage con anagrafe per quanto riguarda la farmaceutica, mentre non sono previste validazioni dei soggetti attraverso l'incrocio con l'anagrafe sanitaria per gli archivi delle schede di dimissione ospedaliera e delle cause di morte.** La chiave di linkage utilizzata per identificare i soggetti e per linkare i diversi flussi è stato il codice fiscale a 11 cifre.
Agenzia regionale sanità Toscana	La chiave di linkage utilizzata per identificare i soggetti e per interfacciare i diversi flussi è stato il codice fiscale completo.
ASL Roma E	❖ **Per stima infarto miocardico acuto** Linkage tra anagrafe comunale e schede di dimissione ospedaliera attraverso 7 combinazioni di chiavi a cascata: K1: cognome, nome, sesso, data e luogo di nascita K2: codice fiscale K3: codice fiscale con l'esclusione dei 4 caratteri che indicano il luogo di nascita K4: codice fiscale con l'esclusione dei 2 caratteri che indicano il mese di nascita K5: codice fiscale con l'esclusione dei 2 caratteri che indicano il giorno di nascita, cognome e sesso K6: codice fiscale con l'esclusione dei 2 caratteri che indicano l'anno di nascita, cognome e sesso K7: codice fiscale con l'esclusione dei 3 caratteri che indicano il nome, cognome e sesso ❖ **Per stima broncopneumopatia cronico-ostruttiva** Linkage tra fonti (schede di dimissione ospedaliera e cause di morte) attraverso codice fiscale; linkage fonti e anagrafe comunale non effettuato
Laziosanità	Record linkage deterministico con codice fiscale ricalcolato per le SDO. Per gli archivi gestiti dall'Agenzia sanità pubblica della Regione Lazio 7 passi: K1: cognome, nome, data di nascita, luogo di nascita e sesso K2: codice fiscale ricalcolato K3: cognome, nome, data di nascita, luogo di nascita K4: cognome, nome, luogo di nascita, sesso K5: cognome, nome, data di nascita, sesso K6: cognome, data di nascita, luogo di nascita, sesso K7: nome, data di nascita, luogo di nascita, sesso
ASL Taranto	Linkage tra anagrafe assistiti e fonti con 3 chiavi: nome, cognome, data di nascita in 2 fasi: Fase 1: linkage automatico: corrispondenza esatta delle 4 chiavi e appaiamento per discordanza di una delle chiavi numeriche di linkage Fase 2: linkage decisionale con risoluzione manuale dei soggetti non linkati in fase 1.

Table 3. Record linkage procedures used by centres.

utilizzando alcune fonti (SDO, ET) nell'anno di stima e in un intervallo temporale precedente (prevalenza longitudinale); in questo modo si aumenta la sensibilità della stima per quelle patologie croniche per le quali gli accessi ad alcuni servizi sanitari (in particolare episodi di ricovero ospedaliero) avvengono solitamente con cadenza superiore ai 12 mesi.

Da un punto di vita concettuale la prevalenza è rappresentata dai pazienti affetti da una specifica malattia in un determinato momento di rilevazione (per convenzione, l'anno di calendario nelle patologie croniche), indipendentemente da quando la malattia è stata diagnosticata. Spesso i flussi informativi correnti non sono in grado di identificare tutti i casi prevalenti presenti in una popolazione. Ciò

Categoria diagnostica	Centri partecipanti
diabete	Comune di Torino (ASL 5) ULSS 4 Alto Vicentino ASL 10 Firenze AULSS 12 Veneziana
cardiopatia ischemica	Comune di Torino (ASL 5) ASL 10 Firenze AULSS 12 Veneziana Ulss 9 Treviso
IMA	Comune di Torino (ASL 5) CNR Pisa ASL Taranto ASL Roma E ASL 10 Firenze AULSS 12 Veneziana
ictus acuto	Laziosanità Comune di Torino (ASL 5) AULSS 12 Veneziana ARS Toscana
BPCO	ASL Roma E Comune di Torino (CPO Piemonte) ASL 10 Firenze CNR Pisa ASL Taranto AULSS 12 Veneziana
asma	ASL 10 Firenze Comune di Torino (CPO Piemonte) AULSS 12 Veneziana
MPCO	ASL 10 Firenze Comune di Torino (CPO Piemonte, CPA) AULSS 12 Veneziana

Tabella 4. Centri partecipanti, per categorie diagnostiche indagate.

Table 4. Participating centres, by studied pathology.

si verifica in particolare per le condizioni morbose con un'elevata quota di pazienti trattati ambulatorialmente. Per questo motivo, quando è disponibile una fonte di rilevazione sufficientemente esaustiva abbiamo utilizzato la fonte informativa nell'anno di stima (per esempio, uso di farmaci nell'asma bronchiale e nelle malattie polmonari cronico-ostruttive). Viceversa, in altri casi abbiamo ritenuto che le fonti informative delle prestazioni sanitarie usufruite nell'anno di stima fossero insufficienti ad assicurare una buona completezza nell'identificazione dei casi prevalenti. Pertanto abbiamo incluso anche i pazienti trattati in anni precedenti (in particolare per i ricoveri ospedalieri) e vivi all'inizio dell'anno di stima della prevalenza (per esempio, diabete, BPCO, cardiopatia ischemica). In considerazione della massima disponibilità degli archivi utilizzati comune ai centri coinvolti nello studio, abbiamo considerato i dati dell'anno di stima e quelli dei quattro anni precedenti.
Da un punto di vista concettuale l'incidenza è rappresentata dai nuovi casi di malattia in un determinato anno (ovvero diagnosticati per la prima volta in un paziente). Questa condizione è difficilmente verificabile nel caso dell'utilizzo di archivi sanitari correnti, in quanto i dati sono disponibili solo per periodi relativamente recenti e comunque non sono tali da coprire l'intera vita dei pazienti. In questo studio, per definire un caso «incidente» abbiamo introdotto il limite convenzionale dell'assenza di ricoveri per la patologia in oggetto nei 60 mesi precedenti. Il limite di 5 anni è stato introdotto in considerazione della massima disponibilità degli archivi utilizzati comune ai centri coinvolti nello studio. Tale intervallo dovrebbe inoltre essere in grado di identificare la maggior parte dei casi prevalenti.
Le stime prodotte per le patologie indagate sono così suddivise:

- incidenza annuale per IMA e ictus acuto;
- prevalenza annuale per asma e MPCO;
- prevalenza annuale, utilizzando longitudinalmente le fonti SDO ed ET per cardiopatia ischemica, diabete e BPCO.

Per quanto concerne la fonte SDO, nel calcolo dell'incidenza è stata utilizzata come data dell'evento la data di ammissione, mentre nel calcolo della prevalenza si è considerata la data di dimissione; questa scelta è motivata dal fatto che per il calcolo dell'incidenza interessa il nuovo evento, al momento del suo manifestarsi, mentre per il calcolo della prevalenza interessano i soggetti che permangono malati nel periodo, quindi coloro che nella diagnosi di dimissione presentano ancora la malattia.
E' stato utilizzato anche l'archivio della mobilità passiva (dell'intero territorio nazionale) per non escludere dalla stima i casi appartenenti alla popolazione in studio che hanno usufruito di servizi presso strutture afferenti ad altre aziende sanitarie.
Le stime di frequenza prodotte sono relative al periodo 2002-2004.
Nel calcolo dei tassi annuali, al denominatore si è utilizzata la popolazione residente al 30 giugno dell'anno, ricavata da fonte Istat; nel calcolo del tasso di prevalenza ottenuto utilizzando longitudinalmente alcune fonti, la data di riferimento per la popolazione e il calcolo dell'età dei soggetti è l'1 gennaio dell'anno di stima.
Per il calcolo dei tassi standardizzati, la popolazione di riferimento è quella italiana al censimento del 2001 (fonte Istat); le classi di età sono quinquennali fino alla classe «85 e più».
Per ogni categoria diagnostica è stato calcolato il contributo di ciascuna fonte alla definizione dei casi, distinguendo un «contributo esclusivo», determinato da soggetti identificati come malati per mezzo di quell'unica fonte, e «contributo assoluto», come numero di casi alla cui identificazione ha contribuito la fonte considerata.
Per costruzione, la somma dei contributi esclusivi per le diverse fonti e dei soggetti identificati dalla combinazione di più di una fonte è pari al 100% dei casi individuati; la

Area	Anno	CM	SDO	PF	ET
Venezia	2002	3.479	64.176	2.230.841	10.255
	2003	3.637	63.098	2.097.957	28.597
	2004	3.486	58.763	2.184.712	39.476
Treviso	2002	3.186	72.000	2.170.478	4.152
	2003	3.197	72.252	2.106.432	10.640
	2004	3.140	72.374	2.212.190	6.067
Thiene	2002	1.717	32135	-	7.173
	2003	1.639	32411	1.036.740	13.547
	2004	1.603	31.170	1.091.277	10.609
Torino	2002	9.326	173.307	6.411.534	57.565
	2003	10.226	171.049	6.403.242	40.440
	2004	9.068	173.846	6.827.731	41.228
Firenze	2003	9.417	157.831	6.190.000	13.337
Pisa	2001	1.043	17.402	-	-
	2002	1.015	16.792	-	-
	2003	1.030	15683	-	-
Regione Toscana	2002	41.389	775.915	-	-
	2003	43.047	744.034	-	-
	2004	39.735	737.619	-	-
Roma	2002	25.179	651.506	-	-
	2003	26.559	681.582	-	-
	2004	24.492	706.696	-	-
Regione Lazio	2002	49.225	1.343.174	-	-
	2003	52.580	1.388.170	-	-
	2004	49.361	1.456.435	-	-
Taranto	2002	1.549	53.119	-	-
	2003	1.822	52.495	-	-
	2004	1.695	51.337	-	-
totale	2002	94.676	2.334.281	10.812.853	21.580
	2003	104.905	3.378.605	17.834.371	56.545
	2004	129.754	3.260.187	12.315.910	56.152

CM: cause di morte; *causes of death*
SDO: schede di dimissione ospedaliera; *hospital discharges*
PF: prescrizioni faramaceutiche; *drug prescriptions*
ET: esenzione ticket; *health-tax exemption*

Tabella 5. Numero di record degli archivi, per area.

Table 5. Number of records of the archives, by area.

somma dei contributi assoluti è invece ≥100%. I contributi esclusivi e assoluti sono stati calcolati stratificando per fasce di età, sesso e anno di stima.
Per IMA e ictus celebrale si sono inoltre calcolati i tassi di mortalità annuale come misura ulteriore di confronto tra i centri e di validazione delle stime di incidenza ottenute.
Nella tabella 6 sono presentate le fonti utilizzate per la stima dei 7 gruppi nosologici le cui stime sono presentate nei successivi capitoli (2-8).
Per queste patologie sono stati discussi e testati più algoritmi varianti sia per la combinazione di fonti, sia per il numero di eventi rilevato in ciascuna fonte nell'unità temporale considerata. Questo aspetto è particolarmente rilevante per le prescrizioni farmaceutiche, sia perché una prescrizione isolata per una malattia cronica può essere indicativa di errori nel codice identificativo individuale riportato sulla ricetta medica, sia perché sulla base della frequenza di uso di alcuni farmaci si possono individuare livelli di gra-

Categoria diagnostica	Fonti
diabete	SDO, PF, ET
cardiopatia ischemica	CM, SDO, PF, ET
IMA	CM, SDO
ictus acuto	CM, SDO
asma	CM, SDO, PF, ET
BPCO	CM, SDO
MPCO	CM, SDO, PF, ET

CM: cause di morte; *causes of death*
SDO: schede di dimissione ospedaliera; *hospital discharges*
PF: prescrizioni faramaceutiche; *drug prescriptions*
ET: esenzione ticket; *health-tax exemption*

Tabella 6. Fonti utilizzate per le stime delle categorie diagnostiche indagate.

Table 6. Data sources used for the studied pathologies estimates.

Centro	Residenti	Linkage			Incidenza		Prevalenza longitudinale			Prevalenza	
		K1	K+	Man	IMA	ictus acuto	cardiopatia ischemica	diabete	BPCO	asma	MPCO
AULSS 12 Veneziana	303.944		*		*	*	*	*	*	*	*
ASL Città di Milano^	1.324.822		*	*							
ULSS 9 Treviso	396.530		*	*			*				
ULSS 4 Alto Vicentino	179.442	*						*			
ASL 5 Torino,	864.751	*			*	*	*	*			
CPO Piemonte, CPA ASL 4			*	*					*	*	*
CNR Pisa	88.964		*	*	*				*		
ASL 10 Firenze	775.776	*			*		*	*	*	*	*
Agenzia regionale sanità Toscana (ARS)	3.516.296	*				*					
ASL Roma E	2.516.666	*	*		*				*		
Laziosanità	5.145.805		*			*					
ASL Taranto	200.436		*	*	*				*		
totale	11.932.026	7.852.931	8.282.526	2.518.626	4.750.537	9.830.796	1.984.124	2.123.913	4.750.537	1.944.471	1.944.471

^ centro che ha contribuito unicamente alle stime delle due patologie neoplastiche; *centre contributing only to the two neoplastic pathologies*
K1 unico campo identificativo; *one key only*
K+ più campi identificativi; *multiple keys*
Man revisione manuale dei non appaiati; *manual verification of not limked*

Tabella 7. Tabella riassuntiva: dimensione della popolazione, tipo di linkage e patologie studiate, per i centri partecipanti.

Table 7. Summarizing table: population size, type of linkage and studied patholgies, by participating centres.

vità diversi di malattia. Gli effetti dell'utilizzo di diversi algoritmi sulle stime sono stati valutati per arrivare alla scelta di quelli che offrivano una maggior affidabilità.
Nei capitoli specifici 2-8 non sono presentati esaustivamente i risultati dei diversi algoritmi testati, ma sono illustrate e discusse alcune analisi di sensibilità effettuate mettendo a confronto scelte differenti (per esempio, aggiungendo o togliendo codici di patologia, oppure valutando le conseguenze dell'utilizzo della sola diagnosi principale della SDO o anche delle diagnosi secondarie).
Le informazioni sui contributi assoluti ed esclusivi delle fonti utilizzate possono inoltre aiutare a definire le frazioni confermate di malati (per esempio, quelli che presentano nello stesso anno più di una fonte con codice della malattia) e le frazioni che possono aumentare la sensibilità della stima, mettendone però in discussione la specificità.
La tabella 7 fornisce un quadro complessivo delle operazioni di linkage e di stima per ciascun centro coinvolto con le relative dimensioni di popolazione.

Bibliografia

1. Cancer incidence in five continents. Vol VIII. IARC Scientific publications no. 155, Lyon 2002.
2. Simonato L, Zambon P, Rodella S et al. (1996). A computerised cancer registration network in the Veneto region, North-east of Italy: a pilot study. *Br J Cancer*. 1996; 73. 1436-39.
3. Automated data collection in cancer registration. IARC Technical reports no. 32, Lyon 1998.
4. www.assr.it/mattoni/mattone_index.htm
5. www.ministerosalute.it/programmazione/lea/sezOrgani.jsp?label=siveas
6. http://osservatorioarno.cineca.org/rapporti.htm
7. De Rosa M, Berti A, Covezzoli A et al. Progetto ARNO. Osservatorio sulla prescrizione farmaceutica. 2000 CINECA, Centro di calcolo interuniversitario dell'Italia Nord-Orientale.
8. Monte S, Macchia A, Romero M et al. L'uso di antidepressivi in pazienti anziani con scompenso cardiaco: analisi epidemiologica di outcome clinici. Atti della XXXI Riunione annuale dell'Associazione italiana di epidemiologia. 17-19 ottobre 2007, Ostuni (Brindisi).
9. Costa G, Spadea T, Cardano M. Diseguaglianze di salute in Italia. Parte II (Fonti informative, classificazioni e misure per il monitoraggio delle diseguaglianze sociali in Italia). *Epidemiol Prev* 2004; 28 (3) suppl: 115-61.