Algoritmi di ricerca
e
integrazione tra basi di dati

Michela Franchini

Sez. di Epidemiologia e Ricerca Sui Servizi Sanitari, IFC-CNR

michela.franchini@ifc.cnr.it

ARCHES Electronic health databases as a source of reliable information for effective health policy
Ricerca Finalizzata 2010 (RF-2010-2315604)

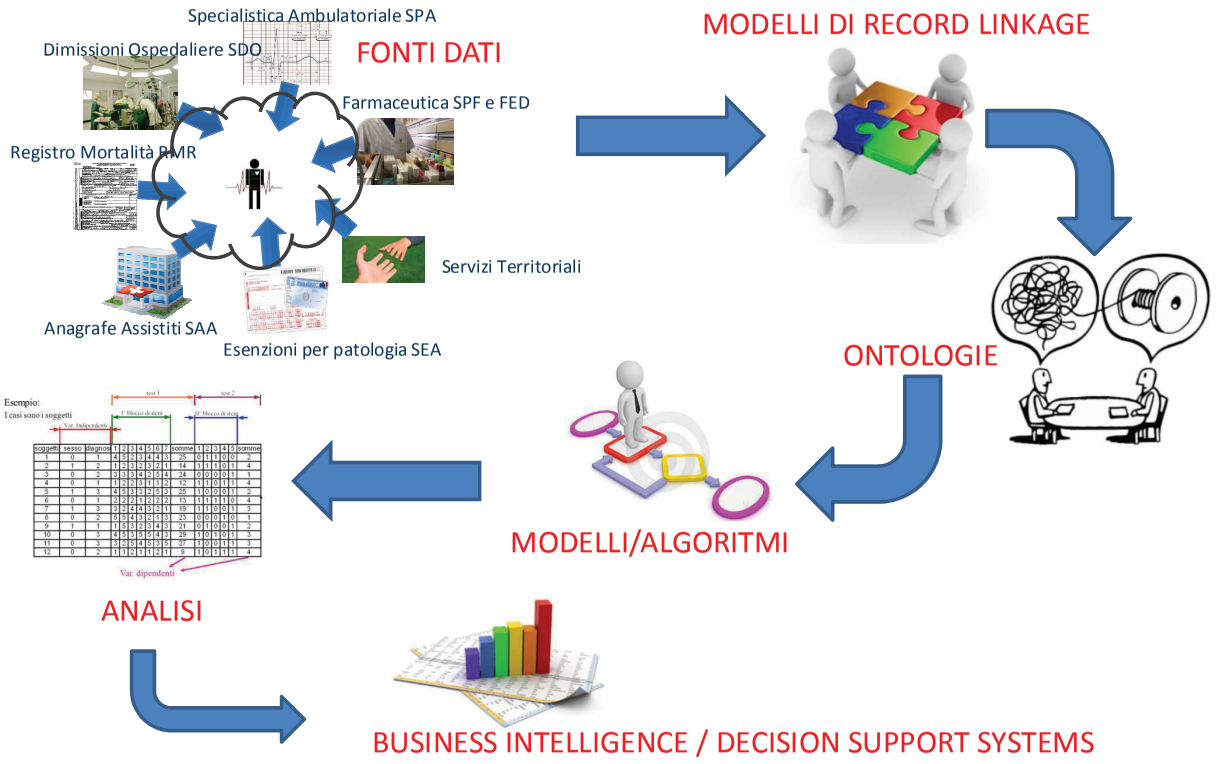
Milano Bicocca, 14 aprile 2016

ARCHES CONSENSUS CONFERENCE

to establish whether it is possible to elaborate a standardized operating protocol for the management and use of Administrative Electronic Databases and Health Databases as a source of reliable epidemiological information for effective health policy

Disponibilità del dato
Autorizzazione all'utilizzo e relative modalità
Metodologie di integrazione
Capacità di sintesi dell'informazione
Validità

IL QUADRO DEGLI "STRUMENTI"
per colmare il gap fra disponibilità informativa e conoscenza

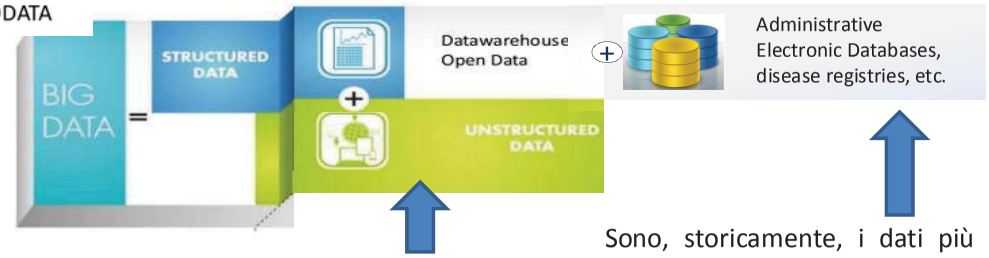


Dal concetto di «fonte dati» a quello di Big Data

Esistono già strumenti capaci di interfacciarsi con portali esistenti (opendata, I.stat, etc) ed importare in modo automatizzato un database (http://www.odinet.it/) :

- CSV
- XLS
- MDB
- DBF
- Shapefile
- RDF
- SDMX
- ODATA

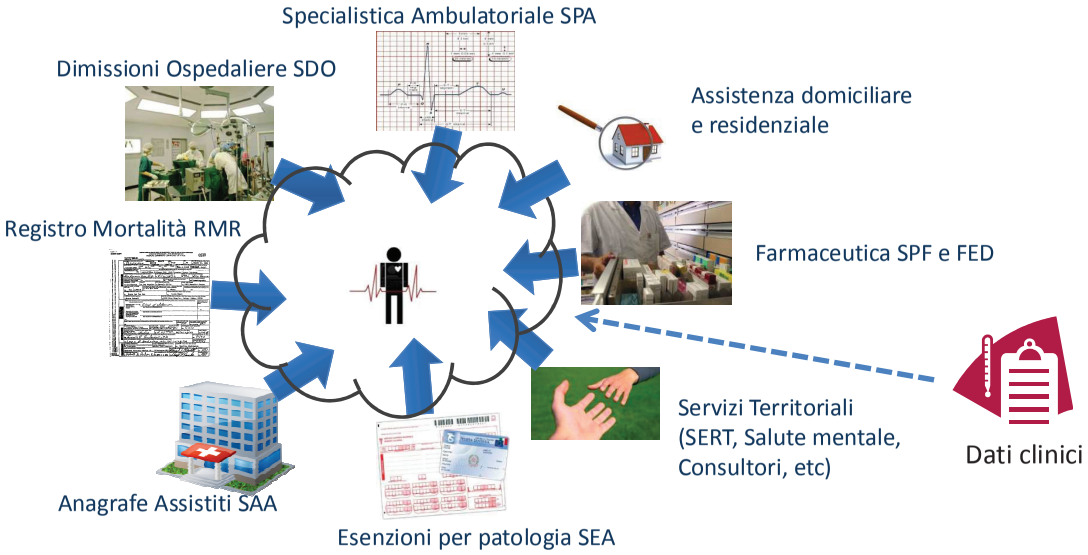
Si tratta, tendenzialmente di informazioni di contesto, georeferenziate, senza alcun riferimento che possa consentire l'identificazione dell'unità statistica



Sono i dati conservati senza alcuno schema. Un esempio possono essere i file contenenti testi a carattere narrativo prodotti per mezzo di uno dei più diffusi software di editing testuale o un file multimediale. In questo caso, i sistemi di gestione di dati utilizzabili sono quelli basati sul modello *dell'information retrieval*. Sono i più complessi da standardizzare ed utilizzare

Sono, storicamente, i dati più «semplici» da gestire, perché strutturati e descritti con precisione nei loro tracciati.
L'unità statistica deve essere opportunamente de-identificata, ma è possibile procedere all'integrazione fra DB attraverso metodiche di record-linkage anche deterministico

Le fonti di dati strutturati in sanità: flussi amministrativi elettronici e dati clinici

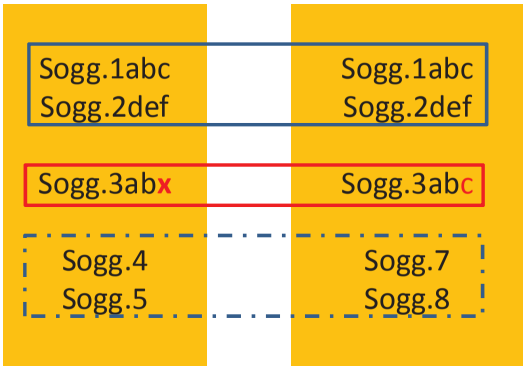


Dal database al datawarehouse

http://www.istat.it/it/files/2013/12/met_norme_03_16_metodi_statistici_record_linkage.pdf

MODELLI DI RECORD LINKAGE

"Each person in the world creates a book of life. This book starts with birth and ends with death. Its pages are made up of the records of the principal events in the life. Record linkage is the name given to the process of assembling the pages of this book into a volume" - HL Dunn, 1946.



- *Data cleaning* e standardizzazione
- Identificazione delle variabili che compongono la chiave
- Scelta della metodologia di linkage

Approcci di linkage

linkage DETERMINISTICO

Archivio A

nome e cognome		genere	data di nascita	comune di nascita
CRL	FRN	F	15/01/1967	010330
GNN	CRR	M	23/10/1953	020001

Archivio B

nome e cognome		genere	data di nascita	comune di nascita
CRL	FRN	F	15/01/1967	010330
GVN	CRR	.	23/10/1953	020001

Regola decisionale: due record appartengono alla stessa unità se, e solo se, tutti i campi della chiave di linkage coincidono.

linkage SEMIDETERMINISTICO

Archivio A

nome e cognome		genere	data di nascita	comune di nascita
	FRN		15/01/1967	010330
	CRR		23/10/1953	020001

Archivio B

nome e cognome		genere	data di nascita	comune di nascita
	FRN		15/01/1967	010330
	CRR		23/10/1953	020001

Regola decisionale: due record appartengono alla stessa unità se, e solo se, tutti i campi della chiave di linkage "ridotta" coincidono.

linkage PROBABILISTICO

$$\mathbf{w} = f(\gamma_1, \gamma_2, \ldots, \gamma_i, \ldots, \gamma_k)$$

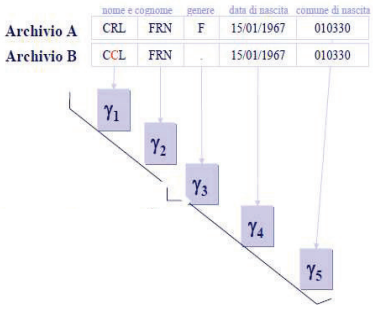
Ad ogni coppia di record viene assegnato un peso complessivo (W) che calcolato come funzione del potere discriminante e dell'attendibilità dei campi identificativi

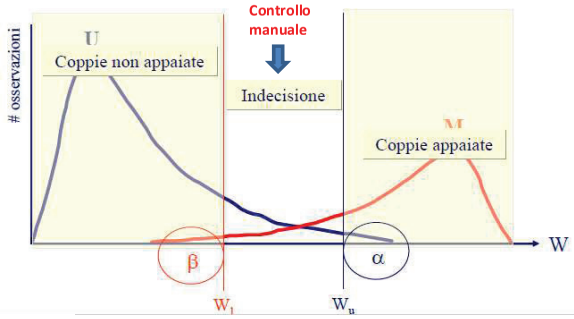
A THEORY FOR RECORD LINKAGE*

IVAN P. FELLEGI AND ALAN B. SUNTER

Dominion Bureau of Statistics

m = Probabilità che il campo concordi, dato che i record confrontati si riferiscono allo stesso individuo (ATTENDIBILITA')
u = Probabilità che il campo concordi, dato che i record confrontati si riferiscono a individui diversi (1-u =POTERE DISCRIMINANTE)





Chiave di linkage	Passo 1
Nome	Nome
Cognome	Cognome
Sesso	Sesso
Giorno nascita	Giorno nascita
Mese di nascita	Mese di nascita
Anno di nascita	
Comune di nascita	
Comune di residenza	Comune di residenza

blocco

Dal confronto fra i risultati dei due passi una coppia si definisce appaiata se il peso in entrambi i passi supera quello della corrispondente soglia (minor peso positivo complessivo osservato? Altra soglia?)

Chiave di linkage	Passo 2
Nome	Nome
Cognome	Cognome
Sesso	Sesso
Giorno nascita	
Mese di nascita	
Anno di nascita	Anno di nascita
Comune di nascita	Comune di nascita
Comune di residenza	

blocco

Valutazione dell'algoritmo di linkage

PPV represents the proportion of matched pairs classified by the algorithm as matches that are true matches.

$$f\text{-}measure = ((\beta^2 + 1.0) * Sensitivity * PPV) / (\beta^2 * Sensitivity + PPV)$$

Sensitivity measures the ability of an algorithm to correctly classify true matches as matches.

β is equal to the user-assigned relative importance of sensitivity over PPV.

While there is no hard rule, a good linkage algorithm will typically have values of sensitivity, PPV, and the f-measure in excess of 95 percent.

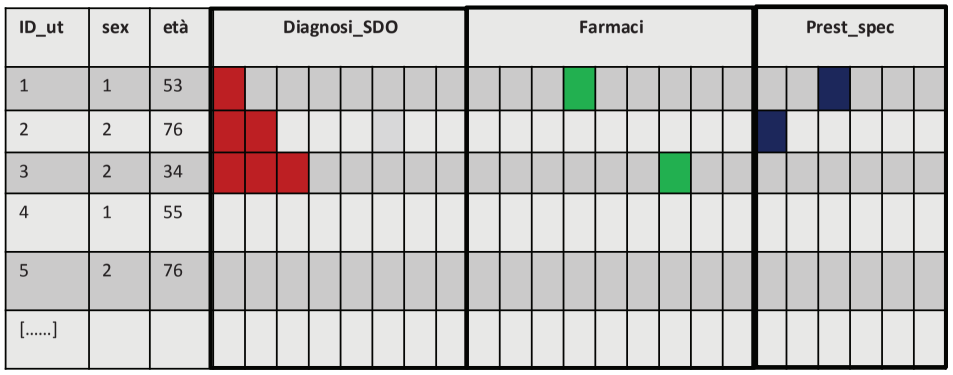
Linking Data for Health Services Research: A Framework and Instructional Guide [Internet].

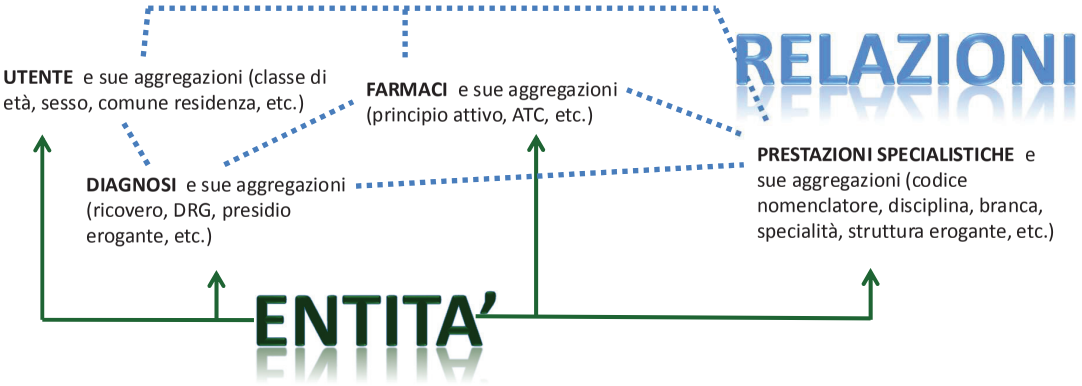
Show details

Dusetzina SB, Tyree S, Meyer AM, et al.
Rockville (MD): Agency for Healthcare Research and Quality (US); 2014 Sep.

http://www.ncbi.nlm.nih.gov/books/NBK253312/

Il risultato del linkage *vulgo:* il «TABELLONE»





Aspetti legali nell'integrazione degli archivi sanitari elettronici

Epidemiology Biostatistics and Public Health - 2013, Volume 10, Number 3
THEME: OBSERVING REAL WORLD CLINICAL PRACTICE

Legal aspects regarding the use and integration of electronic medical records for epidemiological purposes with focus on the Italian situation

Antonietta Stendardo[1,2], Francesca Preite[3,4], Rosaria Gesuita[5], Simona Villani[5], Antonella Zambon[6] and the SISMEC "Observational Studies" working group

Dati sensibili →
Dati semi-sensibili
Dati comuni →
Dati giuridici

Dati PERSONALI

Art.2 – matrice generale di tutela del diritto alla privacy
Artt. 3, 13, 14, 15 e 21 - fanno riferimento a sfere determinate della riservatezza o salvaguardano valori ed interessi che possono essere pregiudicati dall'attività di trattamento dei dati personali

- direttiva 95/46/CE
- L. 675/1996 – Legge sulla Privacy
- Dlvo 196, 30 giugno 2003 -> *Codice di Protezione dei Dati Personali in vigore dal 1 gennaio 2004*
- 25 gennaio 2012 - proposta relativa al nuovo quadro giuridico europeo in materia di protezione dei dati

Ma più interessante, a nostro avviso, è il precedente creato **Autorizzazione generale al trattamento dei dati personali effettuato per scopi di ricerca scientifica - 1° marzo 2012** *(Pubblicato sulla Gazzetta Ufficiale n. 72 del 26 marzo 2012), prorogata fino al 31/12/2013*

Pillole di record linkage

- Le performance delle tecniche di linkage sono strettamente legate alla qualità dei dati disponibili

- Il **RL deterministico esatto** è caratterizzato dai più bassi livelli di sensibilità e il suo utilizzo è limitato alle situazioni in cui sono disponibili codici univoci di identificazione di buona qualità. **All'aumentare del numero di archivi** da integrare, aumenta parallelamente il tempo di elaborazione, ma paradossalmente può risultare più efficace l'accoppiamento (se basato su più campi!!!)

- Rimane aperta e ampiamente dibattuta la questione dell'assenza di **gold standard per le stime di incidenza e prevalenza** ottenute dall'integrazione di più archivi *(vedere dibattito E&P 32 (6) 2008)*.....ma questa non può essere la scusa per non fruire dell'immenso capitale informativo routinariamente prodotto

- Altrettanto dibattuta è la questione dell'ottimizzazione e aggiornamento degli algoritmi di **definizione del «caso» patologico**

Concettualizzare il «caso»: il tipo ontologico

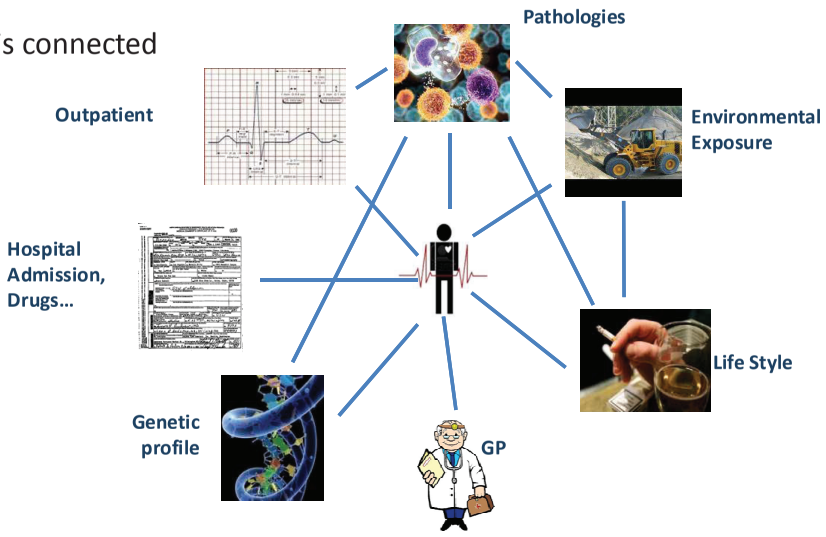
Il malato post-moderno non è riducibile a un sintomo e, meno che mai, è deducibile da un sintomo. Egli richiede sempre più una *conoscenza intera*, non circoscritta al corpo malato, ma estesa alla persona e alla sua esperienza. Per avere un'immagine intera del malato bisogna interconnettere le diverse dimensioni del malato. L'ontologia è la comprensione, attraverso diversi tipi di conoscenza, di una complessità del malato costitutiva dei suoi modi di essere.

Il *tipo* è uno *schema* di caratteristiche, attributi, qualità, proprietà, dunque un gruppo di tratti correlativi del malato, nel quale il *tratto* è, a sua volta, un gruppo di caratteristiche, attributi, ecc. La nozione di *tipo* equivale, in fin dei conti, a un mettere in relazione aspetti diversi del malato e della malattia, anche utilizzando tipologie biologiche, della personalità, sociali, antropologiche. L'idea di *tipo ontologico* può essere ripensata come una sorta di *grammatica*, che tiene insieme segni, sintomi, tratti della personalità, circostanze sociali e malattia.

Ivan Cavicchi, 2002; p. 72;78

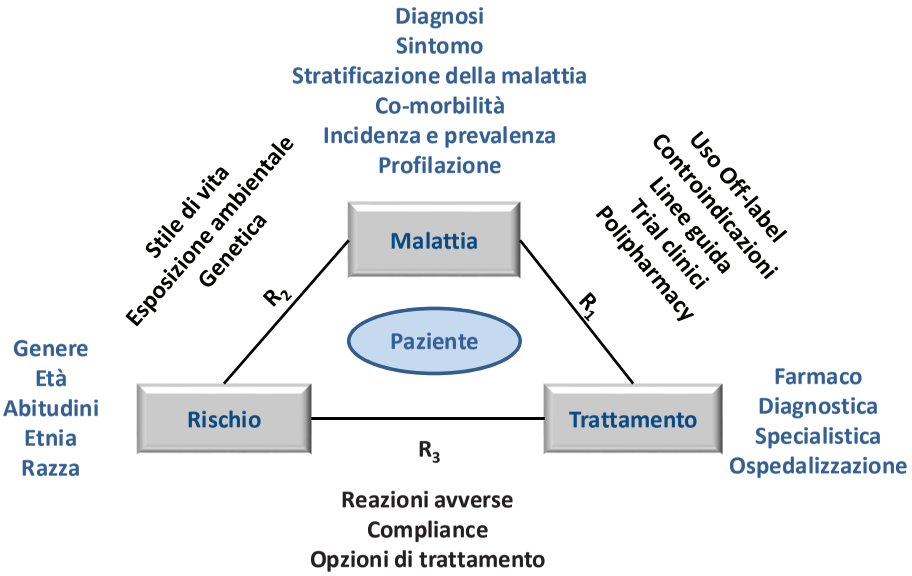
Data modeling: la visione d'insieme

In real world, a patient is connected
to GP
to pathologies
to drugs
to outpatients
to diagnosis
to life style habits
to genetic profile
.....



and continue with 'what is important' in order to build the model

Ontologia di dominio salute: concetti e relazioni

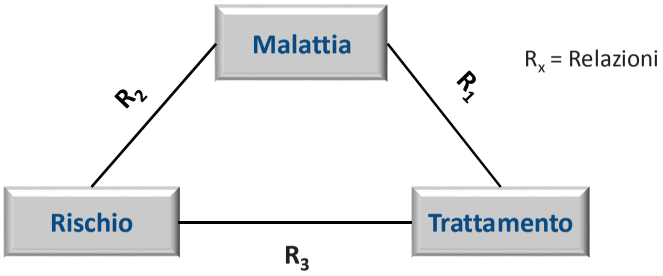


Modificato da M. Liebman – IPQ Analytics

Ontologia di dominio salute

1. Processo di creazione

- Il processo comincia con lo sviluppo dei concetti. L'ontologia si basa su tre (?) prospettive che convergono e si sovrappongono in modo naturale nel dominio di conoscenza.
- Le tre prospettive sono il rischio, la malattia e il trattamento.
- L'ontologia evolve sulla base dei concetti associati alle tre prospettive e alle loro relazioni



Modificato da M. Liebman – IPQ Analytics

Concetti	Relazioni	Concetti
{Malattia}	{R_1}	{Farmaco}
{Malattia}	{R_2}	{Rischio}
{Farmaco}	{R_3}	{Rischio}

La malattia e' un processo in evoluzione

Criteri per la costruzione di Ontologie

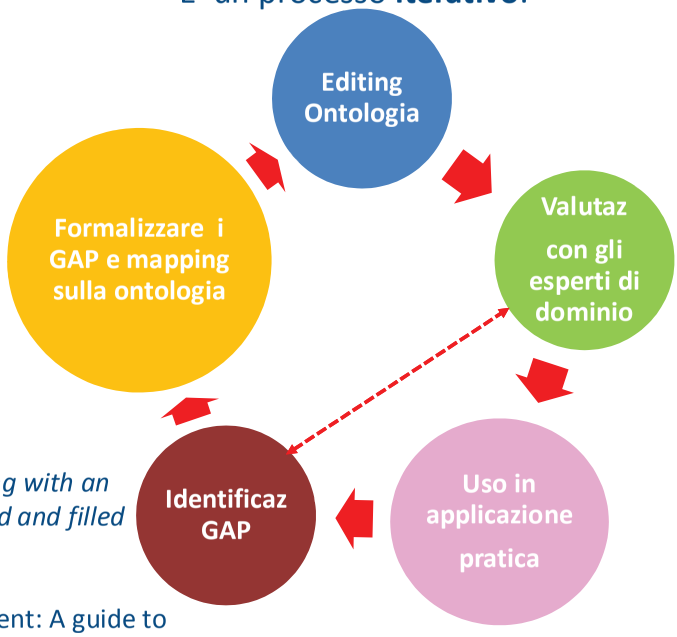
Non esiste un'unica metodologia corretta

la soluzione migliore dipende dall'uso che si deve fare dell'ontologia.

Processo **complesso**:

- DEFINIRE lo SCOPO ontologia
- Ricerca bibliografica
- Riuso dell'esistente
- Fonti dati (OPEN e non)
- Coinvolgimento esperti di dominio
- Multidisciplinarietà (Clinici, Epidemiologi, Informatici, …)

E' un processo **iterativo**:

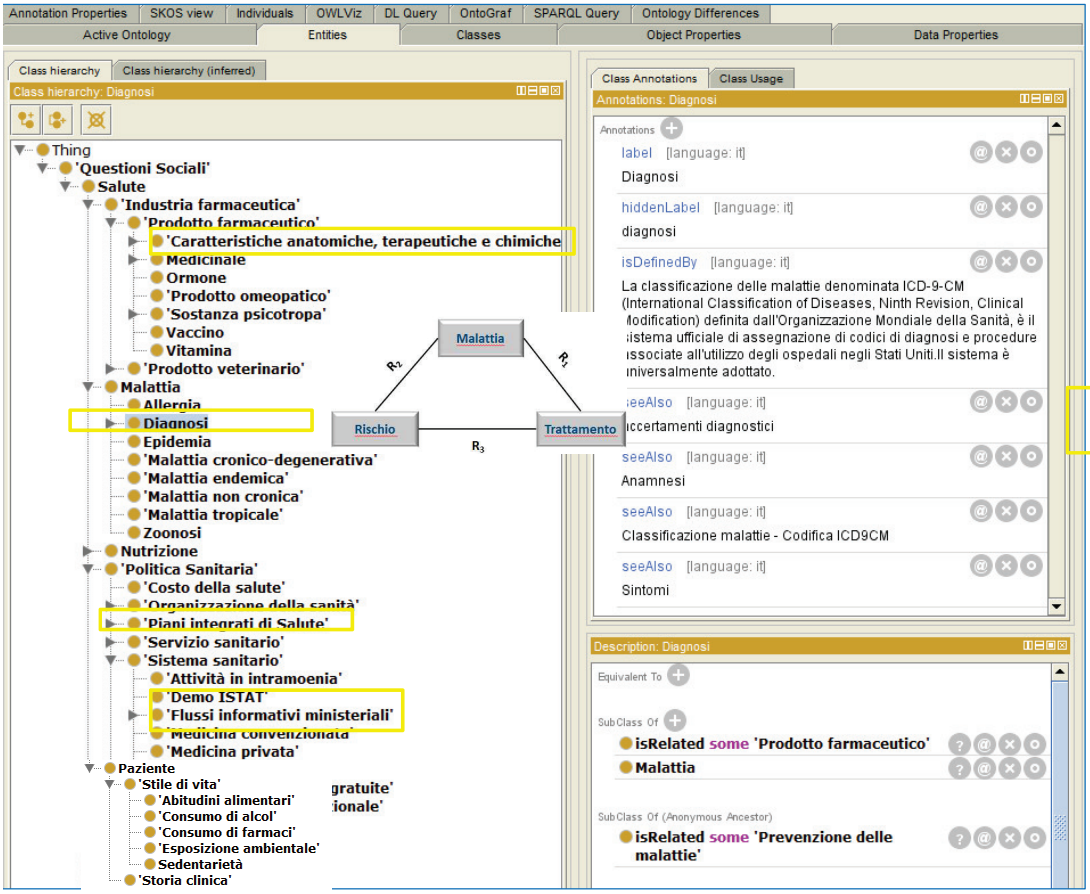


Tutto dipende dallo scopo

Ontology development is a dynamic process starting with an initial rough ontology, which is later revised, refined and filled in the details

[N. F. Noy and D. McGuinness. Ontology development: A guide to creating your first ontology]

L'ontologia di dominio Salute



healthOn.owl

Sub-levels

Estensioni

Dal tipo ontologico all'algoritmo di ricerca

ESEMPIO:

Paziente diabetico

- ✓ Almeno una prescrizione farmaceutica per insulina o analoghi (ATC 3:A10A) con almeno DDD del 10%

OPPURE

- ✓ Almeno una prescizione farmaceutica per agenti ipoglicemici orali (ATC 3: A10B) con almeno DDD del 30%

OPPURE

- ✓ Diagnosi di diabete mellito di tipo 2 (ICD9-CM:250)

OPPURE

- ✓ Esenzione per patologia diabetica (codice esenzione 013).

http://www.aslmi1.mi.it/modulistica/doc_view/5058-guida-alla-bda.html

Il «tabellone» delle co-morbidità

Matrice di dati de-identificati costruita attraverso un processo di record-linkage applicato ai flussi DOC

Ogni assistito viene caratterizzato attraverso il proprio profilo patologico, definito mediante l'applicazione di opportuni algoritmi alle informazioni disponibili, e attraverso il proprio profilo di utilizzo dei servizi/prestazioni/farmaci.

SOGG	SES	ETA	COM	MOR	DIS	PSI	DIP	TRA	INS	HIV	NEC	DIA	CAR	BRO	GAS	NEU	AUT	ENDO	RARE	PAR	ALTR	TRF_DRG	TRF_FAR	TRF_AMB
19	F	77		0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	0	0	1	0	7251	1534
3913960	F	76		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1239973	F	4		0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	42	19
337184	F	13		0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	47	0
597157	M	45		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
665940	F	11		0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	104	0
2977400	M	14		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	109
259862	F	50		0	0	0	0	0	0	0	0	1	1	0	0	1	0	1	0	0	1	0	508	306
1457249	F	85		0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	1	0	1277	116
47971	F	68		0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	0	979	210
3168886	M	9		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	29	0
2319526	F	22		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	20	31
1132310	F	77		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	86
543853	F	51		0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	1	0	92	227
2871564	M	83		0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	665	502
3141964	M	61		0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	349	0
3536034	F	34		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3136	0	79

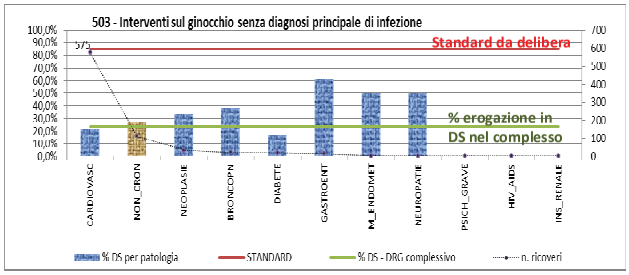
Mortalità
Disabile Grave
Psichiatrico Grave
Dipendenze da sostanze psicoattive
Trapiantato
Insufficienti Renali Cronici
HIV Positivo ed AIDS Conclamato
Neoplastici
Diabetici
Cardiovasculopatici
Broncopneumopatici
Gastroenteropatici
Neuropatici
Malattie Autoimmuni
Malattie Endocrine e Metaboliche
Malattie Rare
Partorienti
Altro

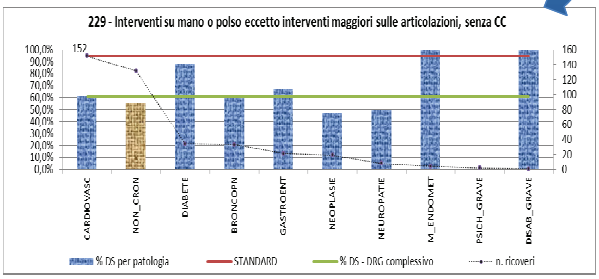
L'utilizzo degli algoritmi a supporto della programmazione sanitaria

DRG CHIRURGICI

Tabella 6.6: Primi 12 DRG chirurgici ordinati in base al risparmio tariffario stimato dalla conversione da regime ordinario a Day Surgery

Codice DRG e descrizione	% Std Patto Salute	Importo medio		Risparmio stimato	% su risparmio stimato
		Ordinario	DS		
359 - Interventi su utero e annessi non per neoplasie maligne senza CC	80%	3.017,08	2.316,65	- 280.734,28	45,3
8 - Interventi su nervi periferici e cranici e altri interventi su sistema nervoso senza CC	75%	3.034,01	2.159,41	- 49.633,49	8,0
538 - Escissione locale e rimozione di mezzi di fissazione interna eccetto anca e femore senza CC	85%	1.485,37	1.275,78	- 33.869,62	5,5
227 - Interventi sui tessuti molli senza CC	80%	1.752,78	1.498,08	- 31.838,10	5,1
503 - Interventi sul ginocchio senza diagnosi principale di infezione	85%	2.144,02	2.077,99	- 31.497,46	5,1
315 - Altri interventi sul rene e sulle vie urinarie	83%	6.606,06	5.377,50	- 26.340,33	4,2
229 - Interventi su m ano o polso eccetto interventi maggiori sulle articolazioni, senza CC	95%	1.390,25	1.218,04	- 23.722,41	3,8
36 - Interventi sulla retina	90%	2.452,80	2.135,05	- 23.513,63	3,8
360 - Interventi su vagina, cervice e vulva	90%	2.311,12	1.753,85	- 21.399,02	3,5
266 - Trapianti di pelle e/o sbrigliamenti eccetto per ulcere della pelle/cellulite senza CC	95%	2.352,46	2.022,20	- 14.977,63	2,4
55 - Miscellanea di interventi su orecchio, naso, bocca e gola	70%	1.832,58	1.770,69	- 14.388,11	2,3
339 - Interventi sul testicolo non per neoplasie maligne, età > 17 anni	95%	1.476,19	1.301,49	- 10.892,47	1,8
TOTALE				**- 562.806,56**	**91,0**





La validazione continua……

CARPEDIEM *– Comorbidity And Risk Profiles Evaluation in Diabetes and hEart Morbidities*

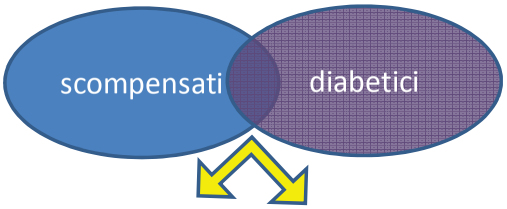
Residenti AUSL 5 e AUSL 11

Lo schema di progetta

ANAG.
SDO
SPA
SPF+FED
ESENZ.

Algoritmi identificazione SCOMPENSO e DIABETE + altre comorbidità

2011-2012



Registri CCM scompenso e diabete
(giugno 2010- marzo 2011)

	In carico CCM	NON in carico CCM
Solo compenso		
Solo diabete		
Scompenso + diabete		

2013-2014

VALIDAZIONE CRITERI

PDTA soggetti in carico VS **PDTA soggetti NON in carico**

I risultati preliminari su una porzione dei dati

I livelli di sensibilità e specificità dell'algoritmo sono buoni, in particolare per quanto riguarda il diabete

scompensati algoritmo	Scompensati CCM +	Scompensati CCM -	
+	645	1687	2332
-	46	1388	1434
	691	3075	3766

SENSIBILITA' = 93.3% [91.5% – 95.2%]
SPECIFICITA' = 45.1% [43.4% – 47.0%]

diabetici algoritmo	diabetici CCM +	diabetici CCM -	
+	2819	45	2864
-	442	460	902
	3261	505	3766

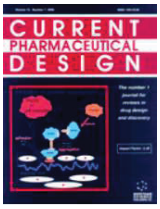
SENSIBILITA' = 86.4% [85.3% – 87.6%]
SPECIFICITA' = 91.1% [88.6% – 93.6%]

La capacità predittiva si riduce quando viene valutata la comorbidità di scompenso e diabete

algoritmo	Comorbidità scompenso + diabete CCM +	Comorbidità scompenso + diabete CCM -	
+	148	1329	1477
-	38	2251	2289
	186	3580	3766

SENSIBILITA' = 79.6% [73.8% – 85.4%]
SPECIFICITA' = 62.9% [61.3% – 64.5%]

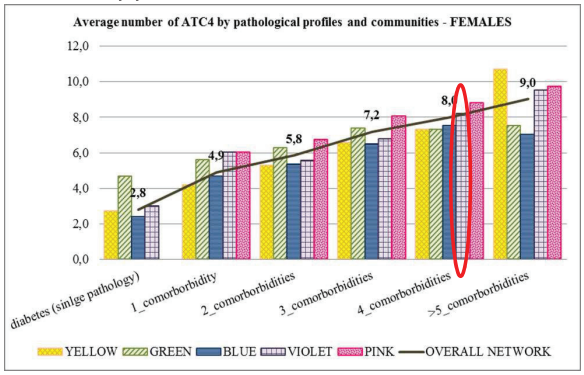
La versatilità degli algoritmi

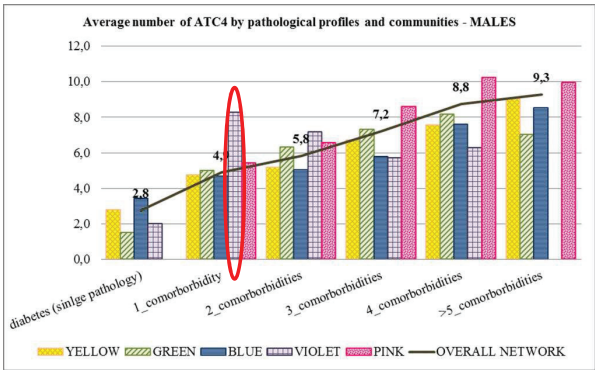


Poly-pharmacy among the Elderly: Analyzing the Co-morbidity of Hypertension and Diabetes

Michela Franchini*, Stefania Pieroni*, Loredana Fortunato*, Sabrina Molinaro* and Michael Liebman†

The aim of this study was to identify the patterns of drug use among diabetes patients aged 75 years and older, using administrative healthcare data and examining the complexity of all available data sources by means of a network-based approach.





The average number of ATC4 classes included in the patients pharmacological treatments increases from 2.8 to 9.0 and over, accordingly to the pathological profile complexity

Male patients belonging to the violet community and characterized by a pathological pattern including only diabetes and neoplasm, show an average number of ATC4 drug classes equal to female patients belonging to the same community, but having a more complex pathological profile

Algoritmi di ricerca
e
integrazione tra basi di dati

Michela Franchini
Stefania Pieroni
Loredana Fortunato
Sabrina Molinaro

www.epid.ifc.cnr.it