Azienda ULSS n. 8 – Asolo
Sistema Informativo Aziendale
P.O. Castelfranco Veneto

Ministero della Salute
Direzione Generale della Ricerca
Sanitaria e della Vigilanza sugli Enti

Regione del Veneto
Assessorato alle Politiche Sanitarie
Direzione per la Prevenzione

Sistema

Informativo

Prevenzione

Assistenza

RELAZIONE CONCLUSIVA

Programma Speciale ex art. 12, comma 2, lett. b), D.Lgs. 502/92

Verso un sistema di sorveglianza e monitoraggio dei bisogni di salute collettiva. Definizione e sperimentazione di un sistema integrato di gestione dei flussi informativi dell'area della prevenzione e dell'area dell'assistenza incentrato sull'utenza.

Responsabile Tecnico Scientifico

Dott. Paolo Spolaore

RINGRAZIAMENTI

Si ringraziano quanti si sono prodigati in questi anni per l'esecuzione del Programma e per il raggiungimento dei suoi scopi prefissati; in particolare:

- i Direttori Generali delle Aziende ULSS n. 3 (Bassano del Grappa), 4 (Thiene), n. 6 (Vicenza), n. 8 (Asolo), n. 9 (Treviso), n. 15 (Alta Padovana), n. 20 (Verona), per la disponibilità dimostrata con l'adesione e la partecipazione al Programma;
- i Referenti Aziendali delle Aziende ULSS coinvolte ed i loro stretti collaboratori, nello specifico i Responsabili dei Sistemi Informativi aziendali ed i Direttori dei Dipartimenti di Prevenzione;
- il personale tecnico e amministrativo delle Aziende che, con dedizione, ha prestato la sua preziosa collaborazione durante le varie fasi della ricerca;
- i collaboratori che a vario titolo hanno prestato nel tempo le proprie capacità e conoscenze ed, in particolare, il dott. Sergio Minello, il dott. Francesco Bisetto, il dott. Giuseppe Battistella, il dott. Roberto Turra, il dott. Francesco Avossa, la dott.ssa Maria Zanandrea;
- il Consorzio "Mario Negri Sud" di S. Maria Imbaro (CH) che ha collaborato attivamente alla elaborazione dei dati e alla definizione degli outcome;
- coloro che, seppur non nominati, hanno contribuito con le proprie conoscenze alla riuscita del Programma.

Hanno contribuito alla elaborazione della presente relazione, e si ringraziano, la dott.ssa Elena Schievano (statistica), il dott. Francesco Avossa (statistico) ed il dott. Roberto Turra per la stesura e l'elaborazione dei testi e dei dati conclusivi, il dott. Antonio D'Ettorre (CMNS) per la parte relativa al prodotto informatico, il dott. Vito Lepore (CMNS) per la parte relativa ai risultati – esempi di applicazione, il dottor Savino Lastella (amministrativo) per la correzione delle bozze e l'editing.

Dott. Paolo Spolaore

INDICE

1. Lo scenario e le criticità

Premessa

"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.

La frase, ripresa da un recente editoriale del Bollettino Epidemiologico Nazionale del ISS (1), introduce in modo efficace l'obiettivo del progetto.

Il tema è ampiamente noto e risponde ai seguenti quesiti:

- I sempre più grandi Data Base Amministrativi (DBA), creati dal crescente aumento dei flussi informativi sanitari in ambito locale e regionale, possono essere utilizzati per scopi e in ambiti diversi da quello meramente amministrativo e/o gestionale?
- Tali fonti di dati sono oggi sufficientemente affidabili da permettere letture ed usi diversi?
- È possibile o almeno proponibile restituire ai tanti dati oggi disponibili il significato ed il contenuto strutturalmente originario e cioè di essere il riflesso di conoscenze, di pratiche e di cultura applicata a specifici contesti assistenziali?
- Ed infine, gli strumenti, la logica, il rigore e la metodologia sviluppata in ambito epidemiologico possono essere applicati – trasferiti su archivi di tale natura, dimensione e qualità?

Certamente i dati amministrativi, per la loro stessa natura di debito informativo, affiancano alla completezza e stabilità della rilevazione, la parzialità del fenomeno rilevato. Le banche-dati relative alle prescrizioni farmaceutiche o ai ricoveri ospedalieri, di cui qui più in dettaglio si parla, illustrano molto meglio il "peso economico" delle prestazioni rispetto alle esigenze di salute per le quali le stesse sono erogate. L'interesse prioritario sugli aspetti quantitativi ha consentito lo sviluppo di analisi approfondite, attualmente disponibili in molte o in tutte le ASL, su:

- prescrizioni farmaceutiche,
- ricoveri ospedalieri,
- mortalità,
- esenzione ticket,
- invalidità,
- altro.

Ciò che per lo più accomuna l'utilizzazione di queste fonti è il livello di dettaglio, talvolta molto sofisticato, ma parziale e frammentario. Infatti l'esplorazione isolata di ciascun archivio-database illustra in modo efficiente e completo solo e soltanto l'aspetto sanitario per il quale sono costruiti risultando, in definitiva, poco informativi sulle interazioni reali tra salute della popolazione assistita, offerta di prestazioni e di assistenza sanitaria ai vari livelli.

La possibilità ormai concreta e praticabile, in diverse realtà locali e/o regionali, di collegamento di archivi a differente contenuto informativo (in particolare: dimissioni ospedaliere, prescrizioni farmaceutiche e mortalità) apre nuove opportunità di studio e approfondimento, in particolare di studi osservazionali di popolazione. (1)

L'ipotesi di lavoro che con questo progetto si vuole verificare parte dalla semplice considerazione ed interesse dell'epidemiologia e della clinica che vuole il paziente, e la sua storia di malattia, al centro dell'osservazione e che tutti gli interventi sono tappe informative di un percorso assistenziale che si deve e si può, se si vuole, ricostruire.

Le "cautele" necessarie per questo lavoro, in parte riprese dal citato editoriale (1) e dal successivo articolo (2), e l'originalità specifica rispetto a quanto già esistente in alcune esperienze a livello regionale, possono essere così sintetizzate:

a) Il rischio di dispersione tra i tanti dati disponibili, raccolti per scopi e con modalità diverse, è alto.
b) La definizione di precisi obiettivi ed il supporto di una metodologia praticabile e condivisa (=epidemiologia) sono indispensabili.
c) Il livello "locale" è la dimensione ideale per testare la continuità tra descrizione e conoscenze epidemiologiche ed interventi, feedback, misure di impatto, valutazioni comparative multicentriche.
d) Osservazione non marginale: siamo all'inizio di un percorso; i concetti, i modelli di analisi, gli orientamenti di ricerca qui sviluppati si prestano in modo importante a creare situazioni di dialogo e collaborazione con i più diversi interlocutori che si incrociano nel lavoro istituzionale.

Il proposito del progetto SIPA è di superare la frammentarietà e disomogeneità dei flussi informativi, per creare un sistema integrato di gestione dei dati, che renda possibile l'analisi e l'interpretazione delle dinamiche epidemiologiche e dell'utilizzo dei servizi sanitari, utili, sia a livello centrale che periferico, per formulare strategie mirate di controllo delle patologie prioritarie e di gestione dell'offerta di risorse sanitarie. La riorganizzazione dei flussi informativi viene di conseguenza effettuata in vista di un impiego continuativo dei dati per conoscere la distribuzione dei fenomeni di mortalità e morbosità e le modalità di erogazione delle prestazioni sanitarie.

Gli *output* del sistema integrato esplicitamente attesi dal progetto sono pertanto due: la conoscenza delle dinamiche epidemiologiche nella popolazione, necessaria per formulare strategie di controllo, e la conoscenza della *performance* dei servizi sanitari, necessaria per la gestione dei servizi.

Nella prima parte di questo documento vengono descritte concisamente le caratteristiche del progetto SIPA, il contesto dentro il quale collocare le attività di epidemiologia e di valutazione dei servizi sanitari, che costituiscono il fine della riorganizzazione dei flussi informativi, definendo, al fine di evitare confusione e fraintendimenti, alcune nozioni fondamentali della Sanità Pubblica, approfondendo i concetti di Sorveglianza, Monitoraggio e Record-Linkage ed infine suggerendo applicazioni pratiche per la Sorveglianza ed il Monitoraggio dei flussi informativi considerati nel progetto SIPA.

Nella seconda parte vengono presentate le proposte di indicatori ottenibili dalle variabili rilevate nei flussi informativi consolidati.

Il progetto

Il progetto si propone di sperimentare modalità di integrazione dei grandi *data base* "amministrativi" creati a partire dai numerosi flussi informativi sanitari al fine di ottenere una più organica descrizione sia dei fenomeni e delle dinamiche epidemiologiche sia dell'utilizzo delle strutture sanitarie. Migliorare la conoscenza della distribuzione dei fenomeni di mortalità e morbosità nella popolazione e la conoscenza delle modalità di erogazione delle prestazioni sanitarie, nella Regione Veneto ed in ciascuna ASL, ha lo scopo di sostenere la definizione di strategie di sanità pubblica e di interventi preventivi specifici e mirati e di fornire un supporto alle decisioni, sia centrali che periferiche, delle politiche di gestione sanitaria.

La struttura operativa del progetto, sottoposta al controllo e con gli indirizzi della Direzione per la Prevenzione della Regione Veneto, è stata guidata da un Nucleo di Coordinamento Tecnico Scientifico di cui facevano parte i rappresentanti dei Dipartimenti di Prevenzione e dei Sistemi Informativi Aziendali delle ASL coinvolte, ritenute le sedi più idonee per la sperimentazione, vale a dire le Aziende nn. 3 – Bassano, 4 – Thiene, 6 – Vicenza, 8 – Asolo, 9 – Treviso, 15 – Cittadella e 20 – Verona.

La gestione amministrativa del budget è stata assegnata alla ASL n. 8 in quanto sede di coordinamento e di raccolta ed elaborazione dei dati.

1.2.1 Scenario

Dal punto di vista internazionale, l'obiettivo n. 35 del documento "Health for All 2000" dell'Organizzazione Mondiale della Sanità pone la priorità dell'"Health Information Support" ovvero dello sviluppo di un sistema informativo sulla salute per la formulazione, implementazione, monitoraggio e valutazione delle politiche per la salute per tutti.

Sul piano della salute in Italia, i principi generali della ristrutturazione del Sistema Informativo Sanitario indicati nel Piano Sanitario Nazionale, per il triennio 1998/2000, rispondevano alla necessità di soddisfare

il più ampliamente e soddisfacentemente possibile la domanda informativa degli operatori ed utilizzatori del Sistema:

- ✓ definizione dei bisogni informativi dei diversi utilizzatori;
- ✓ sviluppo di sistemi orientati al risultato finale del servizio sanitario, in termini di stato di salute, qualità della vita, soddisfazione dei pazienti;
- ✓ integrazione tra i diversi sistemi informativi sanitari e fra questi e gli altri sistemi informativi della pubblica amministrazione;
- ✓ potenziamento dei sistemi informativi a livello locale e sviluppo di connessioni in rete;
- ✓ adozione di protocolli di raccolta ed elaborazione dei dati che soddisfino le esigenze locali e siano compatibili con le necessità informative centrali;
- ✓ valorizzazione e diffusione del patrimonio informativo del SIS.

Il sistema sanitario regionale del Veneto si articola in 21 ASL e 2 Aziende Ospedaliere, convenzionate con le Università, che fanno capo all'Assessorato alle Politiche Sanitarie, nel quale le strutture principali responsabili della programmazione regionale, in dipendenza gerarchica dalla Segreteria del Settore Socio Sanitario, sono quattro: la Direzione per la Prevenzione, la Direzione Piani e Programmi Socio-Sanitari, la Direzione Risorse Socio Sanitarie e la Direzione dei Servizi Sanitari.

Inoltre il sistema regionale sta procedendo alla riorganizzazione delle proprie attività di osservazione epidemiologica e di sorveglianza di sanità pubblica con, da un lato, l'istituzione del Sistema Epidemiologico Regionale proprio con il fine di progredire, tra l'altro, con lo studio dei flussi, dall'altro definendo i livelli Essenziali di Assistenza integrativi e/o aggiuntivi e provvedendo alla stesura dei Piani Sanitari Regionali.

1.2.2 Criticità

Nella Regione Veneto le ASL rilevano routinariamente, in genere per esigenze amministrative, una grande quantità di informazioni. La maggior parte dei flussi informativi è disciplinata da normative regionali, a loro volta in applicazione di normative nazionali o rispondenti a richieste di carattere statistico, un'altra consistente parte non risulta esserlo. Ciò che caratterizza negativamente questi flussi informativi è la loro frammentarietà, incompletezza ed i tempi lunghi di trasmissione, trattamento e analisi.

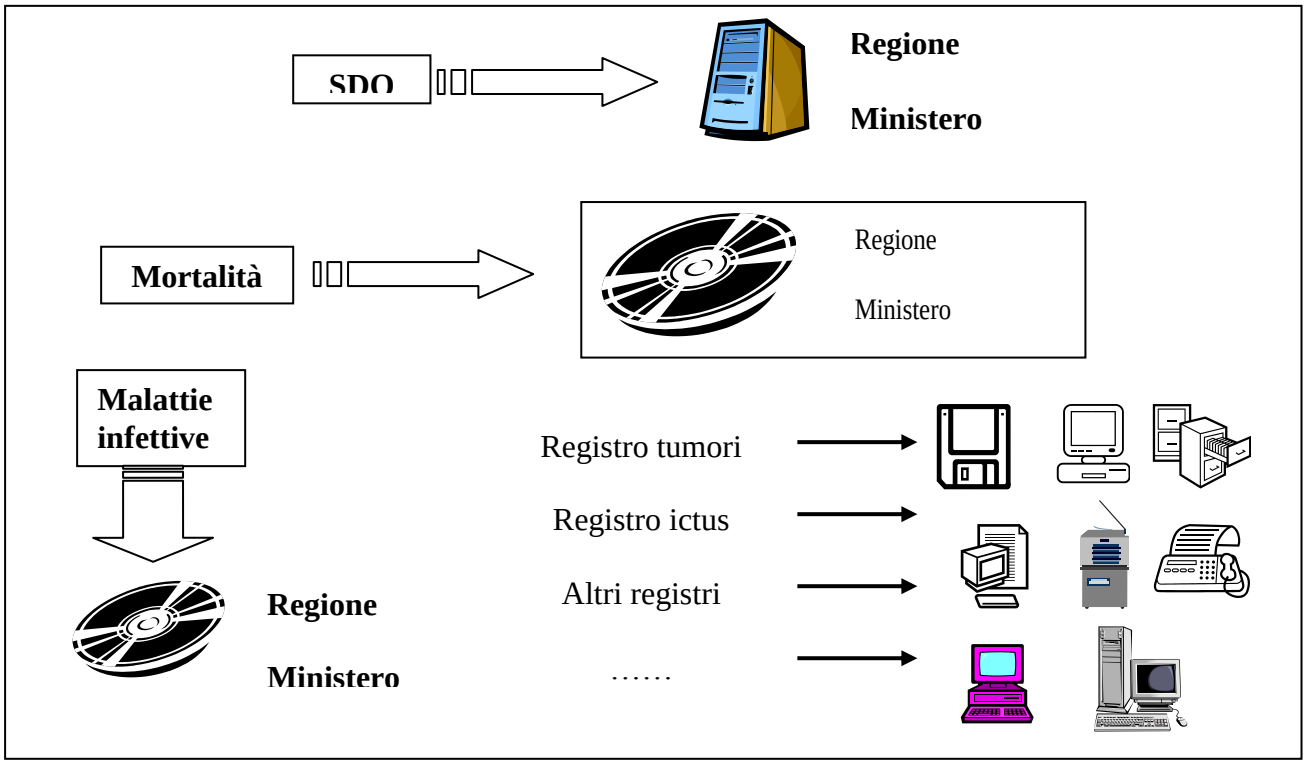


Fig. 1 – Criticità: Frammentarietà, incompletezza, tempi lunghi, unidirezionalità dei flussi.

Oltre alla difformità operativa, un'altra caratteristica è la unidirezionalità dei flussi: i dati trasmessi dalle ASL alle Direzioni Regionali in genere non trovano sbocco in una reportistica, sotto forma di documento stampato o di punto di accesso informatico, che collochi il profilo della singola ASL nel contesto regionale e permetta il confronto con indicatori obiettivi. Inoltre non viene fatto un uso adeguato delle informazioni riguardanti la distribuzione degli eventi di salute nella popolazione: i dati rilevati non vengono impiegati se non occasionalmente per descrivere e analizzare le dinamiche della mortalità e della morbosità.

2. Obiettivi, finalità e razionale dello studio

Obiettivi principali

- ✓ *Censire i numerosi flussi informativi di tipo amministrativo esistenti* nell'area sanitaria, evidenziandone le criticità.
- ✓ *Rendere compatibili i diversi sistemi di dati*, individuando le soluzioni tecniche per il linkage (prodotto informatico specifico) tra diversi data base e per la comunicazione tra ASL e referenti regionali.
- ✓ *Costituire una banca dati integrata*, cioè un archivio degli archivi accessibile con tecnologie informatiche differenti.
- ✓ *Sperimentare una griglia di indicatori utilizzabile* a fine di valutazione dei servizi sanitari e di sviluppo delle conoscenze sulle dinamiche epidemiologiche delle patologie o condizioni morbose ad alto impatto sulla qualità della vita e sui costi del sistema sanitario regionale.
- ✓ *Sviluppare le conoscenze sui processi e gli outcome dell'assistenza sanitaria.*

Finalità e razionale del progetto

- ✓ *Rispondere al fabbisogno informativo generato dalle nuove strategie di sanità pubblica* attraverso il miglioramento dei flussi e la loro integrazione *e essere di supporto nella scelta degli indirizzi della pianificazione sanitaria regionale.*

3. Aspetti generali

Il contesto della Sanità Pubblica

Il termine "Sanità Pubblica" è usato con due significati. Il primo sta a designare l'insieme dei servizi sanitari gestiti dalla Pubblica Amministrazione e finanziati tramite la tassazione ed i ticket. Il suo opposto è "Sanità Privata", cioè l'insieme dei servizi sanitari gestiti da Enti Privati e finanziati tramite il pagamento diretto, integrato da un eventuale sistema di rimborsi, provenienti dalla Pubblica Amministrazione stessa o dalle Assicurazioni Private. Il secondo significato sta a designare l'insieme delle attività dirette a livello comunitario, di popolazione. Il suo opposto è "Assistenza Sanitaria", o l'insieme delle risorse e attività, diagnostico-curative, praticate su base individuale.

Per illustrare come i concetti di azione individuale e azione comunitaria, pubblico e privato si possono intersecare vengono portati alcuni esempi. Le vaccinazioni di massa sono prestazioni erogate a singoli individui, ma il fine non è tanto la protezione individuale (uno scopo che rientra nella Assistenza Sanitaria), quanto la determinazione di una immunità collettiva (*herd immunity*) che impedisce la diffusione epidemica della malattia trasmissibile (uno scopo che rientra nella Sanità Pubblica). La fluorazione dell'acqua degli acquedotti o la fortificazione degli alimenti (es.: iodio nel sale) o il mantenimento di aria non inquinata sono esempi di prestazioni erogate su base comunitaria (sono prestazioni non suddivisibili in unità individuali di erogazione) e con un obiettivo comunitario (ridurre la frequenza di carie nella popolazione, prevenire l'insorgenza di gozzo ipotiroideo, prevenire l'insorgenza o l'aggravamento di malattie respiratorie).

Considerando la Sanità Pubblica nel suo secondo significato, le sue funzioni essenziali sono:

1. diagnosticare lo stato di salute della Comunità descrivendo l'esperienza di mortalità e la diffusione delle malattie all'interno della popolazione (tramite l'epidemiologia descrittiva), e correlare il quadro d'insieme ai determinanti prossimi e remoti (tramite l'epidemiologia analitica, la demografia, la sociologia, l'antropologia, l'ecologia);
2. individuare strategie di controllo ed obiettivi (1. ridurre la diffusione dell'agente e/o dei veicoli e dei vettori; 2. ridurre l'esposizione dei suscettibili all'agente o ai veicoli o ai vettori; 3. ridurre il numero dei suscettibili) agendo su stili di vita e/o ambiente e/o Servizi Sanitari, tenendo conto di evidenza epidemiologica, fattibilità tecnica, economica e sociale;
3. valutare i risultati in termini di stato di salute comunitario (mortalità, morbosità, disabilità).

Per contro, le funzioni importanti nell'Assistenza Sanitaria sono:

1. diagnosticare lo stato di salute individuale (tramite esami clinici, strumentali e di laboratorio);
2. individuare obiettivi (1. Curare; 2. Prevenire una ricaduta; 3. Limitare un danno strutturale o funzionale; 4. Prevenire una complicanza successiva; 5. Alleviare il disturbo presente; 6. Rassicurare;

7. Accompagnare il decesso nel conforto e nella dignità) e strategie di intervento (trattamenti farmacologici, chirurgici, fisici) individuali;
3. valutare i risultati in termini di stato di salute individuale (sintomi, segni clinici, laboratoristici, strumentali).

Sia Sanità Pubblica che Assistenza Sanitaria percorrono le tappe in cui può essere suddiviso il processo di "soluzione dei problemi" (individuazione del problema, individuazione degli obiettivi, considerazioni teorico-pratiche sulla raggiungibilità degli obiettivi, azione, valutazione dei risultati), perché questa è la struttura stessa dell'agire razionale, ma l'una lo applica su scala comunitaria, l'altra su scala individuale. La Sanità Pubblica non è pertanto riducibile al risultato dato dalla sommatoria di gesti clinici individuali, ma è un modo a sé stante di considerare il fenomeno della malattia nelle popolazioni umane: è il punto di vista necessario alla Pubblica Amministrazione, quando mira a tutelare e promuovere il bene comune sanitario allungando la speranza di vita e migliorando la qualità dell'esistenza *della popolazione nel suo insieme*.

Gli studi epidemiologici forniscono evidenze sullo stato di salute della Comunità e sui nessi causali tra esposizione ed esito, contribuendo in modo sostanziale all'individuazione dei problemi di salute e delle strategie di controllo a questi correlate.

Tali studi costituiscono pertanto un valido supporto alle strategie regionali in ambito di salute pubblica e assistenza sanitaria. Oltre alle conoscenze prodotte da studi epidemiologici anche le attività di sorveglianza e monitoraggio possono consentire, attraverso lo sviluppo sistematico della "valutazione", di supportare adeguatamente la programmazione sanitaria regionale. La definizione e la sperimentazione di strumenti informativi/informatici a supporto delle funzioni di programmazione sono fornite attraverso sistemi di raccolta, analisi e diffusione dei dati di Sorveglianza, che il progetto SIPA intende presentare.

Sorveglianza

Sviluppo storico

La parola "Sorveglianza" in ambito sanitario fino agli anni '50 si riferiva al monitoraggio stretto dei contatti di persone con malattie infettive trasmissibili gravi per la rilevazione precoce della comparsa di sintomi, al fine di prendere misure immediate di isolamento. Questo sistema di misure di Sanità Pubblica era entrato in vigore in Germania dal 1766. Nel periodo napoleonico la parola sorveglianza significava "tenere d'occhio da vicino un individuo od un gruppo di individui al fine di cogliere tendenze sovversive". Le radici del termine moderno originano quindi nel controllo sanitario e nella sorveglianza personale. Le origini dell'analisi numerica dei dati sulla Sorveglianza si possono far risalire al 17° secolo, a von Leibnitz e John Graunt, ma i fondatori dei concetti moderni di Sorveglianza sono considerati Lemuel Shattuck e William Farr, vissuti nel 19° secolo.

In Europa la prima forma di rilevazione sistematica dei casi di malattia è stata la notifica obbligatoria dei casi di malattie infettive, iniziata negli Stati Uniti nel 1874, nello Stato del Massachusetts, ed in Italia qualche anno dopo, nel 1881.

L'OMS stabilì nel 1965 l'Unità di Sorveglianza Epidemiologica nella Divisione di Malattie Trasmissibili e nel 1973 adottò la seguente definizione di Sorveglianza:

a) misurazione sistematica dei parametri di salute ed ambientali, registrazione e trasmissione di dati;
b) confronto ed interpretazione di dati al fine di individuare possibili cambiamenti nello stato di salute delle popolazioni e nell'ambiente.

La definizione di sorveglianza presente nel Dizionario di Epidemiologia del Last è la seguente:

L'esame attento di tutti gli aspetti relativi all'insorgenza e alla diffusione delle malattie che sono pertinenti ad un controllo efficace.

Obiettivi della Sorveglianza

Per evitare di attribuire alla Sorveglianza obiettivi che non le sono propri e confonderla con altre attività che pure hanno a che fare con l'elaborazione di variabili relative agli eventi di salute, ma non sono Sorveglianza, è necessario precisare che la Sorveglianza è un'attività diversa dagli studi epidemiologici analitici (osservazionali o sperimentali) e dagli studi di valutazione dei servizi sanitari, che sono attività correlate, ma indipendenti, nell'ambito della Sanità Pubblica, e non comprende l'attuazione di programmi sanitari.

La sorveglianza può rispondere ai seguenti interrogativi epidemiologici: quali sono le malattie che colpiscono la popolazione? Quanti sono i casi di malattia? Come sono distribuiti i casi di malattia a seconda delle caratteristiche personali? Come sono distribuiti i casi di malattia nello spazio e nel tempo? Quanto significative sono le differenze nella frequenza delle malattie tra vari gruppi di popolazione, tra vari aggregati territoriali e tra vari intervalli di tempo? Come interpretare le differenze di frequenza di casi di malattia tra gruppi di popolazione, tra aree geografiche e tra periodi storici?

La sorveglianza ha pertanto compiti descrittivi. Esempio: quanti sono i nuovi casi di carcinoma polmonare diagnosticati in un anno nella popolazione residente nella Regione Veneto? Come si distribuiscono i casi di *K* a seconda del sesso, dell'età, dell'occupazione? Qual è la distribuzione territoriale dei casi di *K* nella regione? Vi sono delle differenze nella frequenza di casi di *K* da ASL ad ASL e, se sì, sono importanti? In che modo è variato nel tempo il numero dei casi di *K* nella regione? C'è stato aumento, diminuzione o stazionarietà? Qual è la distribuzione dell'abitudine del fumo di sigaretta? Le differenze vanno interpretate come differenze reali o come artefatti?

Gli studi epidemiologici analitici invece misurano il grado di associazione tra un'esposizione (una terapia, uno stile di vita, un agente fisico o chimico o biologico) ed un esito (una patologia specifica, così come identificata nella Classificazione Internazionale, o il tempo che trascorre prima della comparsa di un evento relativo alla salute) con misure appropriate di associazione (Rischio attribuibile, Rischio relativo, Coefficiente di Regressione), con un disegno appropriato (osservazionali: caso-controllo, coorte; sperimentali: Studi Clinici Controllati Randomizzati, Interventi di comunità).

Essi rispondono a queste domande: come varia il rischio dell'evento di salute *y* a seconda del variare dell'evento *x*? E' plausibile che l'evento *x* aumenti (o diminuisca) il rischio dell'evento *y*?

Sorveglianza e studi analitici hanno in comune il cuore del metodo epidemiologico, cioè il chiedersi in modo sistematico se ciò che si osserva con i dati riflette la realtà, i fatti, o se è un artefatto, una distorsione (*mantra* dell'epidemiologo: distorsione da selezione? distorsione da informazione? confondimento? casualità?).

Non è fattibile un sistema di rilevazione routinaria di dati sanitari per effettuare studi epidemiologici analitici, perché non sono anticipabili tutte le associazioni testabili tra esposizioni (fattori di rischio) ed esiti (eventi di salute). Gli studi epidemiologici analitici richiedono metodi specifici di misurazione dell'esposizione e dell'esito che non sono inseribili tra le attività correnti dei servizi sanitari, sia per motivi teorici (non è pensabile di misurare un numero illimitato e non prevedibile di variabili), che per ragioni logistiche ed economiche (l'organizzazione ed i costi della rilevazione determinano più svantaggi che vantaggi oltre una certa soglia).

Alcuni studi speciali di tipo analitico sono tuttavia fattibili in caso di *record linkage* (cioè quando gli episodi riguardanti la stessa persona registrati in sistemi separati sono collegati tra loro per ricostituire la storia individuale), sfruttando l'informazione addizionale che un sistema integrato può offrire. Ad esempio è stato possibile indagare sull'associazione tra trattamento con un farmaco (o combinazione di farmaci) e successiva morbosità ospedaliera, sull'associazione tra methyldopa e anemia emolitica e sull'associazione tra amitriptilina e patologia cardiaca utilizzando il *linkage* tra episodi ospedalieri e consumo di farmaci.

Gli interventi di sanità pubblica per definizione non hanno il fine di *conoscere*, ma di generare cambiamenti, di *controllare*. Ad esempio, la normativa che prescrive il divieto di fumare nei locali pubblici non mira a conoscere il numero dei nuovi fumatori o il numero dei fumatori presenti ad una certa data (Sorveglianza), ma a ridurre l'esposizione a quel fattore di rischio. Adoperando la metafora della fisiologia del sistema nervoso, nella Sanità Pubblica, l'epidemiologia (Sorveglianza e studi analitici) costituisce il braccio afferente, gli interventi di controllo (legislazione, vaccinazioni, isolamento, promozione) costituiscono il braccio efferente.

Il legame degli studi analitici con gli interventi di controllo è molto mediato (dagli studi di Peto e Doll su fumo e carcinoma polmonare alle normative antifumo di sigaretta sono passati decenni) mentre il legame della sorveglianza con le decisioni di politica sanitaria e le azioni di controllo e prevenzione è in generale più stretto (l'esempio estremo è la circostanza di un'epidemia di una malattia infettiva: l'individuazione della variazione del numero di casi insorti in un breve arco di tempo dà l'avvio immediato ad indagini epidemiologiche analitiche e a misure di controllo da portare a termine nell'arco di giorni o al massimo di settimane).

Non costituiscono attività di Sorveglianza né gli screening né i controlli periodici. Gli screening sono un intervento di diagnosi precoce offerto a gruppi mirati di persone asintomatiche nel presupposto che il trattamento attuato in fase preclinica modifichi sostanzialmente la prognosi. I programmi di screening possono essere una delle fonti di dati utilizzate da programmi di Sorveglianza, ma non coincidono con l'attività di sorveglianza. Vi possono essere Sistemi di sorveglianza dei tumori anche in assenza di programmi di screening delle neoplasie. Un Registro tumori rileva tutti i casi di neoplasia in una popolazione, ma gli screening possono riguardare solo i tumori della mammella e del collo dell'utero.

I controlli periodici (follow-up) sono prestazioni diagnostiche e terapeutiche offerte a pazienti (es: ipertesi, operati di neoplasia), non ad asintomatici, per assicurare la *compliance*, valutare trattamenti prolungati nel tempo, diagnosticare ricadute trattabili. I programmi di *follow-up* non sono *né screening né tanto meno sorveglianza*. Per sottolineare lo stretto legame tra Sorveglianza e attività di Controllo, la Sorveglianza è stata anche definita "Informazione per l'azione".

Prevenzione e promozione per essere mirati si alimentano anche delle evidenze fornite dall' epidemiologia. Senza sbocchi negli interventi di Sanità Pubblica l'epidemiologia non ha finalità applicate, senza evidenze dall'epidemiologia le strategie di controllo sanitario mancano di giustificazioni. Le due attività sono complementari, ma non coincidono.

Gli obiettivi della Sorveglianza (ciò che ci si può aspettare dalla Sorveglianza, il tipo di informazioni che può fornire) sono pertanto i seguenti:

a) Misurare la frequenza (incidenza e prevalenza) e la distribuzione spazio-temporale (*pattern*) delle malattie nella popolazione (analisi dei dati tramite la **triade epidemiologica** persona-luogo-tempo, per rispondere ai quesiti base: chi si è ammalato, di che cosa, quando e dove?);
b) Misurare i fattori determinanti la frequenza e la distribuzione delle malattie;
c) Individuare i sottogruppi di popolazione a rischio di patologie specifiche.

Le variabili essenziali che permettono di caratterizzare gli eventi sanitari nei termini della triade epidemiologica sono le seguenti:

A. *Persona*: 1) età
2) sesso

3) etnicità
4) stato civile
5) educazione
6) occupazione

L'insorgenza della malattia può essere analizzata in sottogruppi di popolazione definiti da queste variabili. Il numero delle categorie può essere numeroso, ma quelle critiche con cui si analizzano sempre i dati sono le prime due.

B. *Luogo*: a seconda dell'evento oggetto di analisi ed intervento vanno considerati:

1) il luogo di insorgenza della malattia (ad esempio nelle indagini per accertare, descrivere e analizzare surti epidemici o cluster)
2) il luogo di residenza (le dimensioni dell'aggregato geografico e la sua identificazione variano a seconda dello scopo dell'analisi)
3) il luogo di lavoro
4) il luogo di nascita
5) il luogo del decesso.

I luoghi possono essere considerati in categorie come rurale/urbano, domestico/straniero, istituzionale/non-istituzionale.

C. *Tempo*: a seconda dell'evento oggetto di analisi ed intervento di sanità Pubblica e della praticità della rilevazione vanno considerati:

1) la data di insorgenza della malattia (ad esempio nelle indagini per accertare, descrivere e analizzare surti epidemici o cluster)
2) la data della diagnosi della malattia
3) la data del contatto con i servizi sanitari (visita, dimissione ospedaliera, effettuazione di esame di laboratorio o strumentale)
4) la data del decesso.

L'insorgenza della malattia può essere raggruppata e analizzata per settimana, mese, stagione, anno, giorno della settimana, ora, etc.

La descrizione ed il confronto della frequenza delle malattie in sottogruppi di popolazione, fra luoghi differenti e fra tempi differenti è possibile se viene impiegata una definizione standard di caso di malattia, assicurando che ogni caso sia diagnosticato nella stessa maniera. Diversamente le differenze di insorgenza riflettono le differenti modalità di diagnosi, non le differenze nella frequenza delle malattie. La definizione di caso è un insieme di criteri standard per decidere se una persona ha una particolare malattia o altra condizione di salute correlata. Consiste di criteri clinici e, talora, di limitazioni di tempo,

luogo, persona. I casi possono essere classificati in confermati, probabili o sospetti a seconda dell'aderenza ai criteri.

Elementi costitutivi della Sorveglianza

Un sistema di raccolta dati può essere chiamato "Sorveglianza" se sono presenti sia gli obiettivi sia le seguenti caratteristiche di funzionamento:

1. *Continuità* di raccolta, trasmissione, analisi e diffusione dei dati: è un'attività permanente, routinaria, strutturale nell'organizzazione dei servizi sanitari, non episodica o saltuaria.
2. *Uniformità* della raccolta e trasmissione dei dati: chi rileva deve essere stato formato o aver almeno ricevuto linee guida chiare e rigorose su definizione di caso, rilevazione e trasmissione.
3. *Sistematicità* della raccolta e della trasmissione: è impiegata modulistica standard e la periodicità di trasmissione dei dati è regolare.
4. *Regolarità, pertinenza, flessibilità* del ritorno informativo: non ci può non essere una diffusione dei risultati della Sorveglianza, perché essa è informazione per l'azione; gli intervalli della diffusione dei risultati devono essere *regolari*, con la cadenza dipendente dalla natura dell'evento sanitario sotto Sorveglianza (ad esempio settimanale per le malattie trasmissibili, annuale per le malattie non trasmissibili); i dati devono essere inviati ai destinatari *appropriati* a seconda degli utilizzi successivi e presentati in modo da facilitare gli utilizzi successivi; le opzioni di diffusione non devono restringersi alle tradizionali pubblicazioni scritte, ma considerare anche canali informatici, i media, forum pubblici e prevedere circostanze speciali, come la diffusione urgente delle informazioni tramite lettere, telefonate, fax, o televisione.

Utilizzi della Sorveglianza

Le informazioni messe a disposizione dalle attività di Sorveglianza sono utili in Sanità Pubblica quando vengono impiegate per i seguenti usi:

1) documentare la distribuzione degli eventi patologici (tassi di incidenza e/o prevalenza di mortalità e morbosità, dimensione del problema)
2) documentare i cambiamenti nella distribuzione dei fattori di rischio, nella esposizione, nella suscettibilità (ambiente, stili di vita; nuovi sierotipi di virus e batteri; sieroepidemiologia)
3) individuare epidemie e cluster (confronto fra casi attesi e casi osservati)
4) definire le priorità di sanità pubblica in termini quantitativi (le cause principali di morte, malattia, disabilità, determinanti)
5) contribuire alla pianificazione di servizi sanitari (numero di casi attesi di patologia in una popolazione; benefici espressi in unità fisiche: casi di morte o di malattia o di disabilità evitati)

6) documentare variazioni nell'erogazione di prestazioni sanitarie (frequenza di prestazioni chirurgiche; farmacoepidemiologia)
7) valutare l'efficacia delle misure di prevenzione, controllo, promozione (variazione nei tassi di incidenza attribuibili alle attività di Controllo; efficacia delle vaccinazioni).

Pertanto vi può essere un sistema di raccolta dati che funziona con continuità, uniformità, sistematicità e produce report periodici e che non è un sistema di Sorveglianza, ma un sistema sofisticato di gestione informatizzata delle cartelle cliniche; se il sistema non è in grado di produrre misure della frequenza e della distribuzione delle malattie: misurare il numero di visite e di ricoveri non è la misura dei nuovi casi di patologia insorti in sottogruppi specifici di popolazione in un arco di tempo determinato, e ritrovare tempestivamente l'informazione sanitaria nominale per erogare prestazioni al singolo paziente è un utilizzo dell'informazione per l'Assistenza Sanitaria, non per la Sanità Pubblica.

Fonti dei dati raccolti routinariamente

I sistemi di informazione sanitaria per mezzo dei quali i dati sono raccolti routinariamente e resi disponibili per l'analisi sono classificabili in 6 tipi:

1) notifiche obbligatorie delle malattie infettive
2) statistiche vitali (nascite e morti)
3) sistema sentinella (un gruppo selezionato di strutture sanitarie o di medici di base contribuisce alla raccolta dei dati)
4) registri (l'informazione proveniente da più fonti è collegata per ogni individuo nel corso del tempo: ogni nuovo caso è identificato e i casi non sono contati più di una volta; i registri di buona qualità richiedono un uso intensivo di risorse per prolungati periodi di tempo)
5) inchieste sanitarie (un campione rappresentativo di popolazione fornisce informazioni su esposizioni e/o esiti di specifico interesse, periodicamente o in una singola occasione)
6) sistemi amministrativi di raccolta dati (sistemi che raccolgono routinariamente dati per ragioni prevalentemente contabili o legali, ma che contengono informazioni sanitarie utilizzabili per la sorveglianza: ricoveri ospedalieri, visite specialistiche, esami strumentali o di laboratorio).

Nel Progetto SIPA vengono presi in considerazione fonti di dati riconducibili alle statistiche vitali, alle notifiche obbligatorie, e soprattutto ai sistemi amministrativi di raccolta dati; tra questi ultimi non sono inclusi flussi di dati provenienti dai Laboratori.

Criteri di selezione degli eventi da porre sotto Sorveglianza

Se la Sorveglianza è informazione per l'azione, allora il sistema deve raccogliere solo l'informazione prioritaria per l'analisi della situazione, la definizione di obiettivi e la formulazione di strategie e programmi.

Poiché la Sorveglianza è intrinsecamente una disciplina applicata, devono essere riportati solo i dati che possono essere trasformati in indicatori utili alla presa di decisioni di Sanità Pubblica.

Perciò la selezione dei dati raccolti, per ogni aspetto posto sotto Sorveglianza, è cruciale per il funzionamento del sistema.

Un **criterio generale** per decidere quali siano gli eventi sanitari da porre sotto Sorveglianza è che siano eventi per i quali vi è una ragionevole aspettativa che verranno intraprese misure di controllo appropriate.Vi deve essere una risposta affermativa alla seguente domanda: l'informazione raccolta condurrà ad una significativa azione di Sanità Pubblica? Se gli eventi per i quali vi è una aspettativa di azione di controllo sono gli obiettivi di un Piano sanitario, allora è possibile legare Obiettivi del Piano sanitario e Sorveglianza in una griglia che mostra quale componente della sorveglianza misura quale obiettivo.

	Inchieste di Popolazione	Strutture Sanitarie extraospedaliere		Strutture Sanitarie ospedaliere		Laboratori		Statistiche vitali	Altre fonti
Obiettivo di Piano		Tutte le strutture	Siti sentinella	Tutte le strutture	Siti sentinella	Tutte le strutture	Siti sentinella		
Ridurre tot % esposizione *x*, e/o esito *y* entro tempo *t*									

Tab. 1 – Fonti dei dati. Prospetto di misurazione degli obiettivi collegati ai componenti della sorveglianza.

Un insieme di **criteri più specifici** di selezione prende in considerazione frequenza (incidenza, prevalenza, mortalità, anni potenziali di vita perduti), severità (letalità, ospedalizzazioni, disabilità), costi (diretti e indiretti), prevenibilità, trasmissibilità, interesse pubblico.

Un altro **criterio** può essere quello di ispirarsi alle scelte effettuate da istituzioni nazionali o internazionali, per collocare la Sorveglianza in un contesto geografico più vasto e, successivamente, optare per ciò che ha più realistiche probabilità di successo. Nella tabella sottostante sono riportate le patologie non trasmissibili sottoposte a Sorveglianza, per le quali sono disponibili dati su pubblicazioni ufficiali, in Inghilterra, in Italia, negli Stati Uniti, in Canada, e presso l'OMS e le patologie non trasmissibili sulla cui rilevazione indaga il progetto europeo ISARE (Indicateurs de la santè dans les règions en Europe).

Patologia	ITA	UK	USA	CANADA	OMS	ISARE	TOT
Cancro	x	x	x	x	x	x	6
Malattie del cuore e infarto	x		x			x	3
Patologia Cardiovascolare		x		x	x	x	4
Ipertensione	x					x	2
Ipercolesterolemia	x					x	2
Patologia cerebrovascolare						x	1
Comportamento violento			x				1
Bambini maltrattati				x			1
Diabete	x			x	x	x	4
Diabete e condizioni croniche causa di disabilità			x				1
Incidenti		x	x	x		x	4
Salute occupazionale			x	x			2
Patologia perinatale				x			1
Malattie respiratorie (asma)	x	x		x	x	x	5
Malattie reumatiche croniche	x				x	x	3
Osteoporosi	x						1
Malattie muscoloscheletriche		x					1
Salute orale			x		x		2
Genetica umana					x		1
Salute mentale		x	x				2
Schizofrenia						x	1
Disturbi nervosi	x						1
Patologia Neurologica		x				x	2
Parkinson, Sclerosi Multipla						x	1
Salute materno infantile			x				1
Malattie allergiche	x						1
Ulcera gastrica e duodenale	x						1
Malattie del digerente		x					1
Malattie renali		x					1
Insufficienza renale cronica						x	1
Mal di schiena		x					1
Stato nutrizionale		x				x	2

Tab. 2 – Patologie non trasmissibili sottoposte a sorveglianza. Fonti: ITA: rapporto ISTAT, Indagine Multiscopo; UK: National Statistics, Catalogue 2000 ; The health of adult Britain,1997; USA: Healthy People 2000: Midcourse Review; CANADA: Population and Public Health Branch, Laboratory Centre for Disease Control; WHO: Non communicable diseases; ISARE: Regional Health Indicators in Europe Survey.

La valutazione dei sistemi di sorveglianza.

La valutazione dei sistemi di Sorveglianza promuove il miglior impiego delle risorse di Sanità Pubblica assicurando che solo i problemi importanti siano sotto Sorveglianza e che i sistemi di sorveglianza operino in maniera efficiente. Il proposito della valutazione è di migliorare l'informazione fornita e di conseguenza migliorare l'offerta e l'erogazione di servizi. I risultati di una valutazione ben condotta sono:

1. la documentazione del sistema di sorveglianza
2. l'identificazione dei punti deboli del sistema
3. le raccomandazioni per il miglioramento dell'operatività del sistema
4. l'aiuto per definire le necessità di formazione del personale e la giustificazione per la allocazione di risorse.

Lo schema dei compiti da svolgere per effettuare la valutazione di un sistema di sorveglianza è il seguente:

A. Descrivere l'importanza di sanità pubblica di un evento di salute.
B. Descrivere il sistema che deve essere valutato.
C. Indicare il livello di utilità descrivendo le azioni intraprese come risultato dei dati diffusi dalla Sorveglianza.
D. Valutare gli attributi importanti del sistema.
E. Descrivere le risorse impiegate per il funzionamento del sistema.
F. Elencare le conclusioni e le raccomandazioni.

Monitoraggio e valutazione

"Monitoraggio" è la parola indicata per denotare l'insieme delle attività routinarie di rilevazione degli aspetti importanti riguardanti le risorse impiegate, le attività svolte, le prestazioni erogate nei servizi sanitari. Il Monitoraggio è un importante strumento gestionale che consiste di osservazioni frequenti e con cadenza regolare ed implica un costante aggiustamento dell'operatività in relazione ai risultati. Il Monitoraggio va tenuto distinto da "Valutazione" e "Analisi di politica sanitaria".

"Valutazione " è uno studio, un progetto di ricerca approfondita, che riguarda sia gli aspetti considerati nel monitoraggio, sia i risultati conseguiti con le attività dei servizi. La valutazione si prefigge di stabilire in che misura un programma ha realizzato gli obiettivi che si proponeva; non è un'attività routinaria e preferibilmente viene affidata ad un valutatore esterno.

L'"Analisi di politica sanitaria" ha l'obiettivo di definire i problemi e di confrontare alternative di politica sanitaria mirate alla loro soluzione: individua vantaggi e svantaggi di differenti alternative finalizzate a risolvere un problema prioritario.

Nei capitoli successivi verranno richiamati

- i concetti chiave necessari alla comprensione degli obiettivi e dei contenuti del monitoraggio dei servizi sanitari;
- alcuni esempi di valutazione dei servizi sanitari, con richiami alla normativa italiana;
- il possibile utilizzo dei flussi integrati SIPA ai fini della ricerca applicata ai servizi sanitari.

Dimensione sistemica: *input, process, output, outcome, impact*

La cornice concettuale che ormai da più di 30 anni è impiegata per descrivere ed analizzare i servizi sanitari è quella che li descrive in termini di *input*, *process*, *output*, *outcome* ed *impact*.

Input è sia l'insieme delle risorse, umane e finanziarie, disponibili sia l'organizzazione delle risorse. Le risorse possono essere espresse numericamente sia in termini di frequenze assolute che in termini di rapporti di densità o di medie per residente. Esempi: numero di medici di medicina generale per 10.000 residenti; numero di professionista-ore per le vaccinazioni per 1000 residenti della popolazione bersaglio. L'organizzazione è l'insieme delle unità operative gerarchicamente e funzionalmente collegate, caratterizzate da responsabilità e autorità. L'organizzazione non può essere espressa in termini numerici, ma rappresentata in forma diagrammatica (Organigrammi, Matrici organizzative, Diagrammi di flusso) con una simbologia codificata che rappresenta le relazioni gerarchiche e funzionali.

Process è la sequenza di azioni effettuate da ciascuna unità operativa. La sequenza delle azioni può essere rappresentata con Diagrammi di flusso e alcuni aspetti delle sequenze possono essere espressi numericamente con indicatori. Esempio: proporzione di ambulatori di medicina generale con sistema computerizzato di gestione delle cartelle cliniche; proporzione di frigoriferi con termometri esterni per monitorare la temperatura di conservazione dei vaccini.

Output è il volume delle prestazioni erogate o insieme dei contatti tra erogatori di prestazioni ed assistiti; può essere espresso numericamente, sia in termini assoluti che in termini di rapporti di densità o di medie per abitante. L'output dei servizi curativi rapportato alla popolazione prende il nome di Utilizzo. L'output dei servizi preventivi rapportato alla popolazione bersaglio prende il nome di Copertura. Esempi: numero di visite presso i medici di medicina generale per 1000 abitanti in un anno; proporzione della popolazione bersaglio completamente vaccinata.

Gli ***outcome*** rappresentano i risultati delle prestazioni erogate, in termini di miglioramento dello stato di salute. I risultati possono essere espressi numericamente: numero di esposti al fattore di rischio, numero di suscettibili al fattore di rischio, numero di casi di decesso, di disabilità, di malattia evitati, numero di PYLL (anni-persona di dita perduti) evitati, numero di DALYS (anni di vita disabilità-aggiustati) guadagnati. Esempi: proporzione di ipertesi correttamente trattati; proporzione di bambini vaccinati che sono sieroconvertiti.

L'***impact*** è l'insieme degli effetti non sanitari nella comunità attribuibili agli outcome. Esempi: ore di lavoro perdute per patologie fumo-attribuibili.

Di tutte le dimensioni dei servizi sanitari è possibile descrivere le variazioni nel tempo e tra aggregati geografici ed analizzare le differenze. I precedenti esempi possono essere riassunti nella seguente tabella:

	Serie storica	Serie spaziale
input	Variazione nel tempo del numero di medici di medicina generale per 10000 residenti	Variazione tra aggregati geografici (Nazioni, Regioni, ASL) del numero di medici di medicina generale per 10000 residenti
	Variazioni nel tempo del numero di professionista-ore per le vaccinazioni per 1000 residenti	Variazione tra aggregati geografici del numero di professionista-ore per le vaccinazioni per 1000 residenti
process	Variazioni nel tempo della proporzione di ambulatori di medicina generale con sistema computerizzato di gestione delle cartelle cliniche	Variazione tra aggregati geografici della proporzione di ambulatori di medicina generale con sistema computerizzato di gestione delle cartelle cliniche
	Variazioni nel tempo della proporzione di frigoriferi con termometri esterni per monitorare la temperatura di conservazione dei vaccini.	Variazione tra aggregati geografici della proporzione di frigoriferi con termometri esterni per monitorare la temperatura di conservazione dei vaccini.
output	Variazioni nel tempo del numero di visite presso i medici di medicina generale per 1000 abitanti in un anno	Variazioni tra aggregati geografici del numero di visite presso i medici di medicina generale per 1000 abitanti in un anno
	Variazioni nel tempo della proporzione della popolazione bersaglio completamente vaccinata	Variazioni tra aggregati geografici della proporzione della popolazione bersaglio completamente vaccinata
outcome	Variazioni nel tempo della proporzione di ipertesi correttamente trattati	Variazioni tra aggregati geografici della proporzione di ipertesi correttamente trattati
	Variazioni nel tempo della proporzione di bambini vaccinati che sono sieroconvertiti	Variazioni tra aggregati geografici della proporzione di bambini vaccinati che sono sieroconvertiti
impact	Variazioni nel tempo delle ore di lavoro perdute per patologie fumo-attribuibili.	Variazioni tra aggregati geografici delle ore di lavoro perdute per patologie fumo-attribuibili.

Tab. 3 – Variazioni nel tempo e fra aggregati geografici delle dimensioni dei servizi sanitari.

I principali metodi di studio della relazione tra risorse e risultati o tra prestazioni e risultati sono le Analisi delle serie temporali (Disegno di studio Prima-Dopo) e gli Studi ecologici (Studi epidemiologici osservazionali analitici con dati riferiti a gruppi, non ad individui).

	Relazione input-outcome	Relazione output-outcome
Serie storica (disegno prima-dopo)	Variazione della letalità da ictus cerebrale *dopo* l'introduzione della TAC	Variazione del tasso di morbosità di malattia vaccino-prevenibile *dopo* l'attuazione della campagna vaccinale
Studio ecologico	Correlazione tra spesa media sanitaria in servizi materno-infantili e tasso di mortalità infantile	Correlazione tra tasso di copertura vaccinale e tasso di incidenza della malattia vaccino-prevenibile

Tab. 4 – Metodi di studio della relazione tra input e outcome.

Si può anche descrivere e testare il parallelismo nell'andamento delle serie storiche di input, output ed outcome. Esempio: Variazione nel tempo del numero di posti letto ospedalieri per 1000 abitanti e corrispondente variazione nel tempo del tasso grezzo di mortalità.

Efficienza

L'efficienza è una dimensione prettamente economica. Viene abitualmente distinta in efficienza allocativa ed efficienza produttiva.

- Efficienza allocativa: come ottenere un determinato livello di outcome al minimo costo; la relazione è input/outcome; l'input è espresso in termini monetari,
    - outcome espresso in termini di unità fisiche: Analisi costo-efficacia ;
    - outcome espresso in termini di QALY o DALY: Analisi costo-utilità;
    - outcome espresso in termini monetari: Analisi costo-benefici.

La decisione allocativa di più alto livello stabilisce quante risorse allocare ai servizi sanitari e quante, ad esempio, ad educazione, sviluppo comunitario, formazione lavorativa per massimizzare il benessere sociale. La successiva decisione allocativa definisce, entro i limiti delle risorse assegnate ai servizi sanitari, quante risorse allocare fra differenti trattamenti efficaci disponibili.

- Efficienza produttiva: come produrre un determinato livello di output al minimo costo; la relazione è input/output; l'input è espresso in termini monetari, l'output è espresso in termini di unità fisiche: Analisi costo-efficienza (o Costo-minimizzazione); esempio: le stesse prestazioni (output) possono essere erogate sia da medici che da infermieri (input), viene confrontato il costo per prestazione erogata da medici con il costo per prestazione erogata da infermieri.

Qualità

Usualmente vengono considerati sei aspetti costitutivi della qualità:

1) Accessibilità, fisica ed economica - la accessibilità fisica riguarda le componenti spaziali (distanza) e temporali (tempi di spostamento e di attesa) che facilitano od ostacolano l'accesso ai servizi di coloro che potrebbero trarne beneficio; la accessibilità economica riflette la capacità dell'individuo di far fronte ai costi delle prestazioni sanitarie.
2) Accettabilità sociale - riguarda le componenti culturali, psicologiche e motivazionali che caratterizzano le interazioni tra erogatori di prestazioni sanitarie ed assistiti e che facilitano od ostacolano l'accesso.
3) Rilevanza o appropriatezza o qualità clinico-tecnica - la congruenza con gli standard professionali dei profili di cura prestati per specifici problemi clinici.
4) Efficacia - i risultati ottenuti, in una prospettiva clinica o di popolazione.
5) Efficienza - il rapporto tra risorse impiegate e prestazioni erogate (efficienza produttiva) o risultati ottenuti (efficienza allocativa).
6) Equità.

La promozione della qualità dovrebbe basarsi su evidenze empiriche che associano causalmente caratteristiche strutturali con appropriati profili di cura a loro volta associati ad esiti finali favorevoli.

Il contesto delle attività di monitoraggio e valutazione

Nel Servizio Sanitario Inglese l'insieme degli indicatori utilizzati per il monitoraggio della *performance* dei servizi è stato individuato dalla Commissione Corner (istituita nel 1980) che prevede la rilevazione di un set di indicatori rappresentati con profili multi-indicatore, multi-specialità e multi-anno.

Negli Stati Uniti la Joint Commission on Accreditation of Healthcare Organization fin dal 1951 ha sviluppato standard e valutato la performance delle organizzazioni sanitarie, definendo con cadenza annuale un set di indicatori di performance delle strutture ospedaliere.

L'OMS ha pubblicato il rapporto "Health for All 2000" nel cui ambito l'obiettivo 35 (Health information Support) pone la priorità dello sviluppo di un sistema informativo sulla salute per la formulazione, implementazione, monitoraggio e valutazione delle politiche per la salute(21).

In Italia il Piano Sanitario Nazionale 1998-2000 pone tra i suoi obiettivi strumentali l'utilizzo di percorsi diagnostici e terapeutici previsti dalla legge 662/1996 e l'estensione del sistema di indicatori definiti nei DMS del 1995 e del 1996. Tali obiettivi sono stati ripresi dalla successiva programmazione nazionale e regionale; in particolare il DPCM 29/11/01, all. 2/c, introduce il problema della valutazione dell'appropriatezza delle prestazioni sanitarie come condizione necessaria alla corretta implementazione dei Livelli Essenziali di Assistenza. Inoltre la Regione Veneto con DGR 2429/01 istituisce un nuovo flusso informativo socio-sanitario con obiettivi di monitoraggio e valutazione delle attività sanitarie territoriali.

Infine l'istituzione dell'Agenzia per i Servizi Sanitari Regionali con funzioni di accreditamento delle strutture socio-sanitarie rende ancor più evidente l'esistenza di un fabbisogno informativo parzialmente soddisfatto.

Il *record linkage*

Definizione

E' il collegamento di informazioni, che si ritengono riferite allo stesso individuo o alla stessa famiglia, provenienti da fonti indipendenti di documentazione. I documenti si dicono collegati (*linked*) e possono essere trattati come un singolo documento per un individuo o famiglia. Con *linkage* successivi l'informazione può assumere le caratteristiche di una raccolta di storie personali o familiari.

I collegamenti tra vari documenti clinici che si riferiscono ad un singolo paziente sono importanti per gli operatori sanitari che li utilizzano, perché questi ultimi spesso hanno necessità di un documento longitudinale del paziente, cioè dell'insieme di documenti di differenti momenti, erogatori di prestazioni,

strutture sanitarie, collegati per formare una visione delle esperienze di fruizione dell'assistenza sanitaria del paziente lungo l'arco della sua esistenza.

Principali aspetti critici

Dal lato tecnico vi sono tre principali difficoltà nel "*record linkage*":

1. Utilizzare identificatori personali per discriminare tra la persona cui i documenti si riferiscono e tutte le altre persone nella popolazione.
2. Decidere se le discrepanze tra identificatori sono dovute a errori nel riportare i dati del singolo individuo o alla presenza di altri individui.
3. Elaborare la grande massa di dati necessari per il *record linkage* ed allo stesso tempo impiegare tempi ragionevoli per la computerizzazione.

Approcci al *linkage*

La scelta della strategia più appropriata di linkage è sostenuta dal calcolo della quantità di informazione presente in ciascuno dei due archivi di dati che devono essere collegati. Infatti gli identificatori più potenti (numeri unici di identificazione, come codice fiscale o codice sanitario, nome) possono essere assenti. Le strategie possibili sono due: appaiamento deterministico e appaiamento probabilistico.

L'approccio deterministico genera collegamenti basati sul numero di concordanze tra identificatori individuali nei due archivi. Le coppie di documenti (uno da ciascun archivio) sono selezionate sulla base del numero di tali variabili che concordano. Questo approccio è valido con archivi di dati con pochi errori di codifica o con poche variabili di appaiamento. Quando vi sono molti errori di codifica o molte variabili di appaiamento l'approccio probabilistico è più efficiente.

Una certa quantità minima di informazione è comunque necessaria per identificare ciascuna persona come unica. Può essere stimato il numero minimo di variabili che devono concordare per permettere il linkage; ogni individuo in ciascuno degli archivi deve potersi collocare in una "scatola" definita da una combinazione unica di identificatori.

Clustering - Linkage – Validazione

Spesso le banche dati amministrative non si riferiscono al soggetto, ma all'evento sanitario. Una banca dati amministrativa, qual che sia la sua natura (schede di dimissione ospedaliera, prescrizioni farmaceutiche, prestazioni specialistiche, schede di morte, esenzioni ticket, registri di patologie, ecc.), si configura generalmente come una raccolta di informazioni dove l'unità

informativa elementare (record - osservazione) non è quasi mai il soggetto (paziente - assistito), ma la griglia di dati che riguardano il suo ricovero, la sua ricetta, la sua visita specialistica.

Poiché, come è facile immaginare, l'evento associato allo stesso paziente non è unico e irripetibile, c'è da aspettarsi che nel medesimo archivio vi sia più di un record riconducibile allo stesso soggetto o, ancora più verosimilmente, che in ciascuna delle diverse banche dati figuri almeno un'osservazione attribuibile al medesimo assistito. Quindi, risalire dall'informazione legata all'evento a quella legata al soggetto equivale, di fatto, sulla base di dati di identificazione comuni a raccogliere, raggruppare, accomunare i diversi record assegnandoli in *cluster* di appartenenza "anagraficamente" omogenei.

Si tratta, a tutti gli effetti, di una vera e propria operazione di partizionamento della banca dati, sulla base di una relazione di equivalenza applicata ai record e definita dalla "similarità informativa" dei campi anagrafici: le classi di equivalenza generate costituiscono i *cluster* di pazienti.

Analogamente, si può accostare l'operazione di *clustering* a quella di analisi fattoriale, applicata però non sulle colonne del database (variabili), ma sulle righe (*record*) e in cui i fattori individuati rappresentano non gli assi del nuovo spazio a dimensione ridotte, ma l'insieme (meno denso) dei nuovi punti (*cluster* - pazienti) in esso collocati.

Il concetto di **clusterizzazione** resta valido anche quando le banche dati su cui si intende operare sono più d'una, anche se, in generale, la loro diversa natura impone una sua applicazione "sincronizzata", "parallela" e "coerente", in quanto, in questo caso, l'obiettivo non è più semplicemente solo quello di individuare in ciascuno di essi i *cluster*, ma di identificarli in modo da poterne riconoscere la diversità o la coincidenza nei vari flussi informativi.

In altri termini, l'integrazione di basi di dati di diversa natura in un sistema centralizzato basato sul paziente richiede un processo di *data-linkage* che va oltre quello di semplice *data-clustering*, a cui accanto alla fase di individuazione del soggetto va affiancata quella, fondamentale, di identificazione. Quest'ultimo aspetto introduce una complicazione non trascurabile nel percorso di risalita dall'evento al paziente. Infatti sia che si proceda al *clustering* che al *linkage* è necessario definire delle chiavi su cui imperniare il processo di aggregazione. Ciò comporta la scelta di quei campi anagrafici che più si prestano, nella situazione contingente e nell'ambito della problematica legata alla qualità del dato, al raggiungimento dell'obiettivo.

A questo proposito è abbastanza superfluo ricordare come i diversi flussi informativi, proprio per la loro diversa natura e il loro diverso percorso di raccolta, risultano a volta disomogenei e, a volta, addirittura incompatibili in termini di tracciato record e di struttura e formato dell'informazione. Questo determina, nell'ambito del processo di *linkage*, una consistente riduzione del numero di campi anagrafici candidabili ad essere selezionati come chiavi di *join*: infatti, se da una parte la procedura di *clustering* opera in maniera autonoma e asincrona sulle variabili anagrafiche di cui si dispone, o in ogni caso elette a chiavi,

nell'ambito dell'unica banca dati in questione, la procedura di *linkage* non può che operare, in sincronia, su quel gruppo di campi (o parte di essi) comuni ai due o più basi di dati da integrare.

La griglia di dati su cui attuare il *linkage* è dunque di norma più ristretta che nel *clustering* e, nei casi peggiori, può capitare che essa si riduca proprio a quelle variabili qualitativamente meno affidabili e potenzialmente più affette da errori, rendendo di fatto inapplicabile il *join* e di conseguenza inattuabile l'integrazione dell'informazione.

Risulta indispensabile in questi casi la disponibilità di un archivio di riferimento, quale può essere l'Anagrafe Assistiti, o di qualunque altro possa svolgere le funzioni di supporto informativo anagrafico. In questa evenienza, auspicabile anche nel caso del solo *clustering*, si può arrivare ad un miglioramento qualitativo e di completezza dell'informazione, grazie al contributo di recupero, correzione ed integrazione del dato che un tale archivio può fornire.

Questo, consentendo il "riconoscimento" del paziente, permette di correggere il dato laddove risulta non conforme a quanto riscontrato nell'archivio Anagrafe, di recuperarlo in corrispondenza di quei campi nei quali risulta mancante, ma soprattutto di integrare i flussi informativi originari con quelle variabili anagrafiche che risultano assenti e che possono tornare utili nella costruzione delle chiavi di *linkage*.

Di fatto, il passaggio per un archivio di riferimento garantisce una sorta di validazione dell'informazione e, anche dal punto di vista metodologico e procedurale, rende equivalente il *linkage* al *clustering*, nel senso che l'identificazione del paziente consente di operare la clusterizzazione in maniera asincrona sui singoli archivi per poi effettuare, anche in tempi diversi, il *linkage* come semplice *join* tra i pazienti riconosciuti identici.

ESEMPI DI *RECORD LINKAGE* TRA FLUSSI REGIONALI: - REGISTRO TUMORI
- PROGETTO MONICA

- Registro Tumori

  nel Registro tumori della regione Veneto le fonti "linkate" sono tre: le schede di dimissione Ospedaliera (SDO), dette fonte H, le Schede di morte, dette fonte M, e le Schede dell'Anagrafe Assistiti. Con il linkage è possibile determinare i casi prevalenti, i casi incidenti e la durata della sopravvivenza.

- Progetto MONICA

  nel Progetto MONICA le fonti "linkate" sono due: le schede di dimissione ospedaliera (SDO) e le Schede di morte. Con il linkage è possibile determinare l'incidenza di eventi coronarici ed eventi cerebrovascolari, fatali e non.

Possibili utilizzi del *linkage* tra file

Poiché il linkage crea un documento longitudinale, con informazioni su eventi distribuiti in un intervallo di tempo che può essere anche quello dell'intera esistenza (dalla nascita alla morte), è possibile effettuare due tipi di studi epidemiologici:

- Coorte
- Analisi di sopravvivenza

Nel primo tipo gli individui oggetto di studio vengono classificati secondo l'esposizione e si determina la frequenza di comparsa dell'esito di interesse od il valore di una variabile quantitativa; nel secondo tipo l'esito di interesse è il tempo trascorso tra le date di due eventi. Esempio:

- Coorte descrittivi:

esposizione: patologia diagnosticata nella SDO o nell'ESENZIONE TICKET o nella SCHEDA INAIL

esito: ospedalizzazioni successive alla prima
prescrizioni farmacologiche successive al primo ricovero.

- Analisi di sopravvivenza:

inizio: data del ricovero
data dell'esenzione ticket
data del riconoscimento di malattia professionale
data di inizio di terapia farmacologia
data di intervento chirurgico

fine: data del decesso.

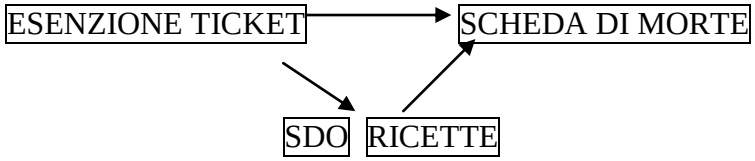
ESEMPI DI UTILIZZO DEGLI ARCHIVI DI DATI REGIONALI COLLEGATI CON *LINKAGE*

Gli archivi sono rappresentati con un rettangolo, il linkage con una freccia, orientata nella direzione del tempo.

1) SDO ⟶ SCHEDA DI MORTE

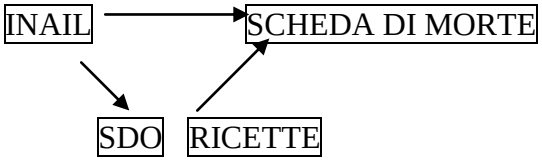
Esempi: sopravvivenza dalla data del primo ricovero o dalla data di un intervento chirurgico o trattamento farmacologico ospedaliero alla data dell'exitus, per le seguenti patologie:

- Diabete
- AIDS
- Malformazioni congenite
- Demenza

2) 

Esempi: - sopravvivenza dalla data di esenzione ticket alla data dell'exitus;
- numero medio di ricoveri ospedalieri o numero medio di giornate di ospedalizzazione o numero medio di ricette in un arco di tempo definito, per le seguenti patologie esenti ticket:

- Diabete
- Ipertensione

3) 

Esempi: - sopravvivenza dalla data di riconoscimento della patologia professionale alla data dell'exitus;
- numero medio di ricoveri ospedalieri o numero medio di giornate di ospedalizzazione o numero medio di ricette, in un arco di tempo definito, per patologie professionali selezionate.

4. Indagine sui flussi informativi delle ASL SIPA

4.1 Censimento, disponibilità dei dati

Durante la prima fase del progetto è stato elaborato e predisposto un questionario informativo per censire i flussi informativi presenti e informatizzati presso le sette ASL coinvolte nella ricerca.
Il questionario, inviato ai responsabili dei Sistemi Informativi Aziendali, presenta le seguenti domande:

- ✓ L'anagrafe assistiti della vostra ASL è informatizzata?
- ✓ Alcuni dei flussi sanitari (archivi della mortalità, SDO, vaccinazioni, specialistica, farmaceutica territoriale e ospedaliera) sono collegati all'anagrafe?
- ✓ Quale priorità date ai flussi (archivi della mortalità, SDO, vaccinazioni, specialistica, farmaceutica territoriale e ospedaliera) relativamente al loro collegamento all'anagrafe sanitaria?
- ✓ Quale utilità date ai flussi (archivi della mortalità, SDO, vaccinazioni, specialistica, farmaceutica territoriale e ospedaliera) relativamente al loro collegamento all'anagrafe sanitaria?
- ✓ Il flusso è digitalizzato nella vostra ASL? Viene mantenuto su supporto informatico? Viene gestito in proprio o viene appaltato a terzi?
- ✓ Come giudicate la formazione specifica del personale che inserisce i dati nel computer?
- ✓ I dati vengono prima gestiti su supporto cartaceo e quindi inseriti al computer? Se sì, viene attuato un controllo degli errori di digitazione ed inserimento? Vi è un controllo sulla codifica da definizione verbale a codice numerico?
- ✓ Vi è la seguente reportistica a disposizione dei Dipartimenti di Prevenzione della vostra ASL?

Flusso	**Report**	**Sì, soddisfacente**	**Sì, insoddisfacente**	**No**
Mortalità	Report almeno annuale			
Scheda Dimissione Ospedaliera	Report strutturati con valore di riferimento			
Vaccinazioni obbligatorie dell'infanzia	Report di copertura vaccinale			
Vaccinazioni facoltative dell'infanzia	Report di copertura vaccinale			
Vaccinazioni dell'adulto	Report di copertura vaccinale			
Farmaceutica Territoriale	Report di consumo			
Farmaceutica Ospedaliera	Report di consumo			

Hanno fatto pervenire i questionari compilati i referenti di cinque Aziende su sette, tutte le anagrafi assistiti delle ASL sono risultate informatizzate.

4.2 Risultati del Censimento Flussi

Il Censimento proposto alle ASL partecipanti ha dato risultati in qualche modo contraddittori nel senso che dal punto di vista della digitalizzazione dei dati e dei flussi tutte le ASL rispondenti registrano ed utilizzano l'informatizzazione di almeno un flusso informativo, ma dal punto di vista delle procedure adottate, della trasmissione dei dati, dell'intervento correttivo e del giudizio sulla preparazione degli addetti al data entry si evidenzia un certo grado di disomogeneità.

Analizzando le risposte fornite da cinque Aziende su sette, le contraddizioni risaltano evidenti in riferimento alle criticità riscontrate:

- ✓ Come precedentemente detto, tutte le ASL coinvolte confermano l'informatizzazione dell'anagrafe assistiti.
- ✓ Tutte le ASL hanno integrato almeno un flusso informativo all'anagrafe assistiti, come descritto nella tabella che segue:

ASL	SDO	Mortalità	Vaccinazioni infanzia	Vaccinazioni adulti	Specialistica	Farmaceutica ospedaliera	Farmaceutica territoriale
3 – Bassano	X	Non integrata	Non integrata	Non integrata		Non integrata	Prog. ARGO
4 – Thiene	X	X	X	Gruppi a rischio	X	Non integrata	X
6 – Vicenza	X		X	X	X	Non integrata	Prog. ARGO
8 – Asolo	X	X	X	X	X	Non integrata	X
9 – Treviso	X	X	X	X	X	Non integrata	Non integrata
20 – Verona	X	X	X	X		Non integrata	Prog. ARGO

Tab. 5 – Integrazione tra flussi informativi e anagrafe assistiti.

Come si può dedurre dalla tabella, quasi tutti i flussi gestiti in proprio o, come nel caso dei flussi della farmaceutica territoriale, nell'ambito del progetto ARGO del CINECA, sono integrati o collegati alle anagrafi degli assistiti. La sola ASL n. 3 risulta carente e in ritardo nell'avviamento delle procedure di integrazione. Va rilevato che la ASL n. 4 ha integrato già dal 1997 il flusso relativo ai Medici di Medicina Generale (ICPC) all'anagrafe e che, nonostante l'assenza del quesito specifico, tutte le ASL stanno integrando il flusso delle malattie infettive all'anagrafe stessa.

- ✓ Per ciò che riguarda la priorità assoluta assegnata dai Dipartimenti di Prevenzione ai diversi flussi informativi relativamente al collegamento con l'anagrafe sanitaria (hanno risposto quattro ASL su sette) risultano essere di primaria importanza i flussi delle SDO e delle vaccinazioni obbligatorie all'infanzia, seguiti a ruota da quelli delle vaccinazioni agli adulti e facoltative all'infanzia. La tabella che segue rende più esplicite le priorità assolute (con le mediane in grassetto) rilevate.

ASL	Mortalità	SDO	Vacc. obbl.	Vacc. facolt.	Vacc. adulti	Special. psich.	Altra special.	Farmac. territ.	Farmac. osped.	ICPC	Mal. infett.
A	5	4	5	5	5	3		4	2		
B	4	5	2	2	1	3	3	3	3	4	
C	4		5	5	5						3
D	3	5	5	4	4			4	0		
Med.	**4**	**5**	**5**	**4,5**	**4,5**	**3**	**3**	**4**	**2**	**4**	**3**
Risp.	*4*	*3*	*4*	*4*	*4*	*2*	*1*	*3*	*3*	*1*	*1*

Tab. 6 – Priorità assoluta dei flussi informativi. (In lettere si riportano le Aziende rispondenti, in forma anonima)

La tabella che segue esprime invece la graduatoria dei flussi informativi, per livello di importanza della integrazione con l'anagrafe sanitaria, come valutata dai DiP (tre rispondenti su sette) e le mediane relative. Mettendo a confronto le mediane delle priorità e della graduatoria si riscontrano notevoli coincidenze ma anche delle difformità (es. Vaccinazioni agli adulti: priorità alta, ultimo flusso in graduatoria).

Flusso	A	B	C	**Mediana**
Mortalità	4	2	1	**2**
SDO	2	1	4	**2**
Vaccinazioni obbligatorie all'infanzia	1	8	2	**2**
Vaccinazioni facoltative all'infanzia	3	9	2	**3**
IPCP		3		**3**
Farmaceutica territoriale	5	5	3	**5**
Farmaceutica ospedaliera	7	4	5	**5**
Specialistica psichiatrica territoriale	6	6	4	**6**
Altra specialistica	5	7		**6**
Vaccinazioni agli adulti	7	10	2	**7**

Tab. 7 – Graduatoria dei flussi informativi. (In lettere si riportano le Aziende rispondenti, in forma anonima)

- ✓ Per ciò che concerne l'utilità dei flussi (tre rispondenti su sette) vengono ritenuti di estrema importanza i flussi informativi sulla mortalità, sulle vaccinazioni (tutte), sui Medici di Medicina Generale, seguiti da quelli sulle SDO e sulla farmaceutica territoriale e in decrescendo quelli sulla specialistica psichiatrica territoriale, sulle malattie infettive, sulla specialistica altra ed infine sulla farmaceutica ospedaliera.
- ✓ I flussi vengono normalmente digitalizzati in tutte le ASL e mantenuti in archivi informatici, fa eccezione la farmaceutica ospedaliera che è informatizzata solo nella ASL n. 20. Tutte le ASL gestiscono in modo informatizzato i dati della specialistica ed altri flussi locali, la farmaceutica territoriale è in parte appaltata a terzi (CINECA) nell'ambito del progetto ARGO a cui partecipano per la Regione Veneto le ASL nn. 1, 5, 6, 12, 14, 15, 16, 17, 18, 19, 20 e 22.
- ✓ Viene evidenziata una grande difficoltà nel conoscere il grado di preparazione specifica degli operatori addetti all'inserimento dei dati in quanto afferenti a varie strutture. Tale difficoltà sottolinea la criticità del processo ed i giudizi espressi (ved. tabella che segue; cinque ASL su sette) pongono in evidenza un elevato grado di disomogeneità.

ASL	Mortalità	SDO	Vacc. obbl.	Vacc. facolt.	Vacc. adulti	Special. psich.	Altra special.	Farmac. territ.	Mal. infett.
A	5		5	5	5				
B	5	3	4	4		4	3	2	
C	0	4	4	4	4				
D	4		4	4	4				5
E	1	1	1		1				
Med.	**4,5**	**3,5**	**4**	**4**	**4**	**4**	**3**	**2**	**5**

Tab 8 – Conoscenza del grado di preparazione degli operatori. (In lettere si riportano le Aziende rispondenti, in forma anonima)

- ✓ Il supporto cartaceo viene utilizzato per i flussi della mortalità, delle SDO e delle vaccinazioni nelle ASL nn. 3, 6 e 9. La ASL n. 20 lo utilizza anche per la farmaceutica. I controlli della codifica e degli errori vengono attuati sistematicamente per alcuni flussi nelle ASL nn. 3, 4, 6 e 9; negli altri casi spesso l'informazione non è nota o mancano i sistemi di controllo.
- ✓ Per quel che riguarda il quesito sulla reportistica ed il suo giudizio, la seguente tabella presenta i dati (ad ogni simbolo corrisponde una ASL – cinque rispondenti su sette).

Flusso	**Report**	**Sì, soddisfacente**	**Sì, insoddisfacente**	**No**
Mortalità	Report almeno annuale	*-/	^	+
Scheda Dimissione Ospedalirera	Report strutturati con valore di riferimento	*+	/	^
Vaccinazioni obbligatorie dell'infanzia	Report di copertura vaccinale	*-^/		+
Vaccinazioni facoltative dell'infanzia	Report di copertura vaccinale	*-^/		+
Vaccinazioni dell'adulto	Report di copertura vaccinale	-^		*+/
Farmaceutica Territoriale	Report di consumo	*/+		
Farmaceutica Ospedaliera	Report di consumo	*+		/

Tab. 9 – Reportistica dei flussi informativi. (I simboli rappresentano le Aziende rispondenti, in forma anonima)

5. La sperimentazione SIPA - Materiali e metodi

Introduzione

L'Obiettivo di costruire una banca dati integrata, a partire da flussi informativi amministrativo-sanitari di diversa natura, che abbia nel paziente/assistito l'unità elementare di riferimento presuppone la possibilità di individuare nei dati a disposizione quegli elementi informativi che consentano di ricondurre al medesimo soggetto osservazioni raccolte in tempi distinti e provenienti da flussi diversi.

E' chiaro che una tale operazione risulta fortemente condizionata dalla completezza e qualità del dato: questo, per sua natura, è affetto da errori vari, dovuti per lo più ad inappropriatezza ed approssimazione, che in fasi successive e per cause molteplici finiscono per deteriorarne la qualità e comprometterne l'attendibilità, fino a renderlo nei casi estremi totalmente inutilizzabile.

Quando ciò avviene nella componente anagrafica dell'informazione, ossia in quei campi quali il nome e cognome, il codice sanitario, quello fiscale, la data di nascita, il sesso e il comune di nascita, è facilmente comprensibile come lo sforzo di individuare il soggetto sulla base di elementi di identificazione comuni può risultare, se non vano, comunque di difficile attuazione.

Se si considera, poi, che proprio perché trattasi di informazioni che riguardano l'identità del soggetto e quindi sottoposte ai vincoli e alle restrizioni previste da quella parte della legislazione che mira a garantire la tutela della privacy, molto spesso queste informazioni sono volutamente, in *toto* o in parte, non riportate e dunque non disponibili.

Lo scenario informativo con cui ci si deve confrontare è quindi, da questo punto di vista, molto spesso non proprio favorevole: la griglia dei riferimenti anagrafici è composta da molti campi vuoti, da altri parzialmente compilati, da altri ancora riempiti in maniera approssimativa; solo poche variabili sono attendibili e utilizzabili ai fini del processo di identificazione del soggetto di riferimento.

È chiaro che quella descritta è una situazione che può variare da caso a caso: a volte realtà particolari, grazie magari anche allo scrupolo dell'operatore che archivia il dato, offrono scenari migliori e più favorevoli; in linea di massima, però, si può affermare che, soprattutto andando a ritroso negli anni, quando il supporto dei sistemi di gestione informatizzata del dato era ancora limitato se non inesistente, la precarietà dell'informazione di cui si dispone fa sì che il meccanismo di attribuzione di una osservazione ad un soggetto non sia mai banale né tanto meno immediato.

Da quanto detto finora si evince che la procedura di clustering e, per estensione, quella di linkage, richiedono oltre alla individuazione e alla selezione dei campi anagrafici su cui incentrare il meccanismo di aggregazione dei record, la definizione precisa e rigorosa dei "criteri di appartenenza" di una

osservazione ad un cluster, ossia la scelta delle regole e delle specifiche in base alla quali sia possibile assegnare in maniera deterministica e univoca un record nel proprio cluster di competenza.

Si tratta, per usare le analogie algebriche e statistiche citate prima, di definire la relazione di equivalenza da applicare sulla banca dati in esame in modo da partizionarne i record in classi di equivalenza "anagraficamente" omogenee, oppure, equivalentemente, di fissare l'algoritmo di fattorizzazione che generi i fattori più idonei a raccogliere e sintetizzare le caratteristiche anagrafiche comuni ai vari record.

Mentre la fase di scelta delle chiavi di clustering è diversa da caso a caso e legata alla disponibilità e alla qualità del dato, quella di definizione dei criteri di appartenenza va standardizzata e automatizzata in modo da garantire, a parità delle condizioni iniziali, la riproducibilità dei risultati.

È quanto ci si è proposto di realizzare ideando e sviluppando la routine ReClust, un algoritmo di clusterizzazione implementato in linguaggio SAS e che, basato su una logica iterativa e a passi successivi, mira a raggruppare i record secondo una logica di prevalenza nelle corrispondenze e nel *matching* tra le varie chiavi di aggregazione.

- Il contributo più specifico del presente lavoro è costituito dalla/e procedure di record linkage che, come richiamato dalla citazione di apertura, non è un concetto né "nuovo" né "originale" e certamente esiste da prima della massiccia diffusione della tecnologia informatica.
- E' altrettanto evidente che in assenza di ipotesi ragionevoli, di domande valide, di protocolli sensati il collegamento di differenti Data Base non potrà che ulteriormente accentuare il rischio di dispersione tra i tanti dati.
- I percorsi di analisi e gli scenari modello qui presentati costituiscono già una sintesi delle tante possibilità esplorate. La selezione è stata guidata principalmente dal buon senso, dalla chiarezza degli obiettivi, dalla robustezza delle conclusioni.
- La metodologia adottata richiama continuamente procedure di gestione-dati altamente sofisticate ma ormai sufficientemente sviluppate e testate in modo adeguato e che, quindi, possono essere rese disponibili allo sviluppo di concrete applicazioni.
- Gli archivi utilizzati sono stati selezionati per la loro immediata possibilità di implementazione nelle diverse realtà e particolare cura è stata posta nella preparazione di procedure utilizzabili in concreti contesti assistenziali (=trasferibilità).
- Archivio regionale delle Dimissioni Ospedaliere **(SDO)**; archivio regionale delle schede Istat di morte **(RECAM)**; Anagrafe Assistiti **(AA)** con aggiornamenti e cancellazioni per decesso; archivio prescrizioni farmaceutiche **(PF)** con individuazione dell'assistito; Anagrafe dei Medici di Medicina Generale e Pediatri di Libera Scelta **(AMMG)** sono stati preliminarmente esplorati per qualità e

affidabilità dei dati. Il livello raggiunto può oggi ritenersi sufficiente anche se delle riserve appaiono necessarie.

- Il miglioramento del dato amministrativo è facilmente documentabile nelle diverse realtà locali ma con significative differenze geografiche. Tale processo si è inoltre sviluppato differentemente nel corso degli ultimi anni e perciò particolare attenzione va posta nell'uso delle serie storiche e nel collegamento di archivi provenienti da diverse realtà locali. Quando questa attenzione è mantenuta la riproducibilità dei dati è buona.
- Sulla base di tali premesse è evidente che il percorso metodologico e gli strumenti debbono essere letti in stretta connessione con i due scenari modello di seguito presentati e rispettivamente sviluppati a livello regionale, il primo, a livello di aussl, il secondo:
    1. Interventi di rivascolarizzazione cardiaca (cod 36.xx) della regione Veneto anno 1999. Mortalità e riospedalizzazione precoce (a 30 e 60 gg)
    2. Mortalità e assorbimento di risorse (ospedalizzazione e prescrizioni farmaceutiche) della popolazione di una ASL con particolare riferimento alla popolazione anziana (>= 65aa.). Periodo considerato (Gennaio 2001 Giugno 2002).

5.1 Indicatori per la sorveglianza e il monitoraggio

Sorveglianza - Applicazioni pratiche ai flussi informativi considerati nel progetto SIPA

I paragrafi che seguono intendono illustrare quali sono i flussi informativi oggetto del Progetto SIPA, quali sono le informazioni di interesse sanitario che si possono trarre da essi, in che misura possiedono le caratteristiche che li rendono idonei all'utilizzo ai fini della Sorveglianza, per quali eventi sanitari è proponibile l'utilizzo dei flussi ai fini di Sorveglianza.

Flussi considerati nel progetto SIPA

I flussi informativi regionali presi in considerazione nell'ambito del progetto SIPA sono elencati nella tabella seguente, assieme all'informazione di interesse sanitario da essi ricavabile.

Fonti di dati	*Informazione*
1) Registro anagrafe assistiti	Denominatori dei tassi
2) Certificati di morte	Mortalità di popolazione
3) Schede di Dimissione Ospedaliera (SDO)	Morbosità ospedaliera Frequenza di interventi chirurgici
4) Registro vaccinazioni pediatriche obbligatorie	Copertura vaccinazioni obbligatorie pediatriche
5) Registro vaccinazioni pediatriche facoltative	Copertura vaccinazioni facoltative adulti

Fonti di dati	*Informazione*
6) Registro delle vaccinazioni di adulti	Copertura vaccinazioni adulti
7) Prestazioni specialistiche territoriali	Morbosità extraospedaliera
8) Prestazioni psichiatriche territoriali	Morbosità psichiatrica extraospedaliera
9) Consumo farmaci ospedalieri	Consumo di farmaci ospedalieri
10)Consumo farmaci territoriali	Consumo di farmaci territoriali
11)Notifiche Obbligatorie malattie infettive	Morbosità da malattie infettive
12)Pronto Soccorso	Morbosità extraospedaliera Incidenti
13)Esenzioni ticket	Morbosità da patologie non trasmissibili
14)Infortuni INAIL	Incidenti sul lavoro

Tab. 10 – I flussi considerati dal SIPA e le relative informazioni ricavabili.

Esempi di integrazione dati	*Informazioni rilevabili*
SDO del reparto di malattie infettive + notifiche obbligatorie + consumo antibiotici + schede mortalità + vaccinazioni	Morbosità , consumo farmaci Letalità Efficacia vaccini
SDO + specialistica territoriale + consumo farmaci + schede di morte	Morbosità , consumo farmaci, interventi chirurgici Letalità

Tab. 11 – Esempi di integrazione dati di flusso e relative informazioni rilevabili.

Caratteristiche dei flussi

Tra i risultati raggiunti dal progetto vi è la ricognizione e analisi della situazione esistente riguardante i flussi informativi.

Nella tabella seguente vengono riassunte in modo semplificato, per ogni flusso, le caratteristiche, emerse nella ricognizione effettuata nell'ambito del progetto SIPA, che è necessario considerare per valutarne l'idoneità alla Sorveglianza.

Caratteristica	1 anagrafe assistiti	2 certif. morte	3 SDO	4 vacc. pediat. obbl.	5 vacc. pediat. facolt.	6 vacc. adulti	7 prest. special. terr.	8 prest. psich. terr.	9 farmaci osped.	10 farmaci terr.	11 notifiche obblig.	12 P.S.	13 esen. ticket	14 infor. inail
a) *Continuità*	si	si	si	si	si	si	si	si	si	si	si	si	si	si
b) *Uniformità*														
- *disciplina regionale*	si	si	si	no	no	no	no	si	si	no	si	no	no	si
- *formaz. personale*	no	si	no	no	no	no	no	no	no	no	no	no	no	si
c) *Sistematicità*:														
<u>standard</u>														
- modulo cartaceo	no	si	si	no	no	no	no	no	no	no	no	no	no	si
- software locale:														
architettura	no	no	no	no	no	no	no	no	no	no	no	no	no	si
struttura data-base	si	si	si	no	no	no	no	no	no	no	no	no	no	si
- controlli qualità	no	no	no	no	no	no	no	no	no	no	no	no	no	
- *set di dati per referente regionale*	si	si	si	no	no	no	no	si	si	no	si	no	no	si
- *modalità di trasmissione dei dati*	si	si	si	no	no	no	no	si	si	no	si	no	no	si
- *regolarità di trasmissione dei dati*	si	si	si	no	no	no	no	si	si	no	si	no	no	si

Tab. 10 – Caratteristiche dei flussi informativi.

Esempi di lettura della tabella

Il "si" in corrispondenza della *Continuità* significa che per quel flusso tutte le ASL del progetto SIPA effettuano una rilevazione routinaria e non saltuaria. Il "si" in corrispondenza di *disciplina regionale* significa che per quel flusso vi sono disposizioni prescrittive regionali riguardo alla rilevazione e trasmissione dei dati. Il "no" in corrispondenza di **software locale-architettura** significa che, per la gestione informatizzata locale del flusso, le ASL impiegano software differenti (il software non è standard); questo non significa che la **struttura del data-base** (cioè nome, tipo e lunghezza delle variabili rilevate) sia differente; in *corsivo* sono le caratteristiche importanti, quelle la cui standardizzazione è cruciale perché da quel flusso si possa ricavare una reportistica che abbia un valore di Sorveglianza delle dinamiche epidemiologiche o di Monitoraggio dell'utilizzo dei servizi sanitari.

Quando i dati inviati al referente regionale sono dati aggregati, non analizzabili successivamente nei termini della triade epidemiologica persona-luogo-tempo (non sono utili per determinare frequenza e distribuzione spazio-temporale degli eventi di salute nei sottogruppi di popolazione), anche se omogenei (standard) tra le varie ASL (come il Rapporto Annuale) in corrispondenza di *set di dati per il referente regionale* viene scritto "no" (non c'è standardizzazione).

Proposte di eventi da porre sotto Sorveglianza

Le patologie non trasmissibili da porre sotto Sorveglianza possono essere selezionate sulla base di un insieme di criteri espliciti, tra i quali vanno elencati:

a) gli obiettivi del Piano sanitario Regionale
b) il livello raggiunto di continuità, uniformità, sistematicità dei flussi
c) la costituzione ed accessibilità di *warehouse* (archivi di archivi) di dati
d) gli esempi internazionali e nazionali di sistemi di sorveglianza
e) quanto si conosce in termini di severità, prevedibilità, interesse pubblico degli eventi sanitari candidabili

Una proposta di patologie da porre sotto sorveglianza che tiene conto dei criteri sopraelencati è la seguente:

PATOLOGIA	*FONTE*
1) Diabete mellito e sue complicanze*	SDO; Scheda di morte
2) Malformazioni congenite*	Registro malformazioni congenite
3) Epatopatie croniche*	SDO; Scheda di morte
4) Fratture di femore*	SDO; Scheda di morte
5) Incidenti da traffico*	Pronto soccorso; Scheda di morte
6) Patologie cardiovascolari	SDO; Esenzione ticket; Scheda di morte
7) Patologie cerebrovascolari*	SDO; Esenzione ticket; Scheda di morte
8) Malattie respiratorie croniche	SDO; Esenzione ticket; scheda di morte
9) Demenza*	SDO; Esenzione ticket;

Tab. 11 – Proposta di patologie da sottoporre a sorveglianza. * - Progetti di sorveglianza già avviati in Regione Veneto.

Nello sviluppo di un sistema computerizzato la documentazione è un elemento critico. Per ognuno degli eventi sottoposti a Sorveglianza i referenti devono avere a disposizione un manuale con la seguente documentazione:

- una copia del modulo cartaceo, se c'è, su cui vengono rilevate le informazioni;
- una descrizione del software utilizzato per la gestione informatizzata locale;
- una descrizione della struttura del data-base elettronico (nome, tipo e lunghezza di ogni variabile);
- una copia delle videate utilizzate per il data entry;
- la struttura del report finale, con l'elenco di tavole, grafici ed analisi statistiche;
- l'elenco dei destinatari del report finale;
- un diagramma con la rappresentazione del flusso di dati, dal compilatore iniziale all'archivio finale;
- una descrizione della gestione della documentazione cartacea (dove e come avviene l'archiviazione finale) e delle copie dei data base elettronici (dove e come vengono tenute, protette e aggiornate le copie dei data base);
- una descrizione della gestione dei dati :
  - ✓ chi e come compila il modulo cartaceo (profilo professionale, anzianità di lavoro, training formale);
  - ✓ chi e come effettua il data entry (profilo professionale, anzianità di lavoro, training formale);
  - ✓ come avviene l'individuazione degli errori;
  - ✓ eventuali manipolazioni di variabili e creazione di variabili secondarie;
  - ✓ come viene garantita la sicurezza dei dati e la confidenzialità.

Per i progetti con caratteristiche di Sistema di sorveglianza già avviati in Regione Veneto è raccomandabile inoltre predisporre un piano di valutazione per:

- valutare e dare priorità agli eventi posti sotto sorveglianza;
- valutare la qualità dell'informazione epidemiologica prodotta;
- valutare come i risultati della Sorveglianza influenzino la formulazione di strategie di controllo;
- identificare gli elementi del sistema che possono essere valorizzati per migliorare la qualità dell'informazione.

Monitoraggio - Applicazioni pratiche ai flussi informativi considerati nel processo SIPA

Nella tabella sottostante sono presentate, per ogni flusso informativo, le informazioni classificabili come output e come outcome. Quest'ultime possono essere utilizzate ai fini delle Sorveglianza e per una eventuale valutazione dei risultati in una prospettiva di popolazione.

Flussi	*Informazione di output*	*Informazione di outcome*
1 – Anagrafe assistiti		
2 – Certificati di morte		Numero morti Per causa
3 – SDO	Numero ricoveri per DRG Numero giornate di ospedalizzazione Numero interventi chirurgici	Morbosità ospedaliera (n. casi di patologia Ospedalizzata)
4 – Vaccinazioni pediatriche obbligatorie	Numero vaccinati	
5 – Vaccinazioni pediatriche facoltative	Numero vaccinati	
6 – Vaccinazioni adulti	Numero vaccinati	
7 – Prestazioni special. territoriali	Numero di visite per specialità	Morbosità extraospedaliera (n. casi di patologia extraospedaliera)
8 – Prestazioni psich. territoriali	Numero di visite	Morbosità extraospedaliera (n. casi di patologia extraospedaliera)
9 – Farmaci ospedalieri	DDD per farmaco	
10 – Farmaci territoriali	Numero ricette Numero pezzi DDD per farmaco	Morbosità extraospedaliera (n. casi di patologia extraospedaliera)
11 – Notifiche obbligatorie		Morbosità (n. casi Malattie Vaccino-Prevenibili, HIV/AIDS, STD, etc.)
12 – Pronto Soccorso	Numero visite	Morbosità (n. casi di patologia extraospedaliera
13 – Esenzione ticket		Morbosità (n. casi di patologia extraospedaliera malattie non trasmissibili)
14 – Infortuni INAIL		Morbosità

Tab. 12 – Informazioni di output e di outcome ai fini della Sorveglianza.

Le informazioni relative all'output vanno collegate con quelle relative alla popolazione (Anagrafe degli assistiti) e con quelle relative ai bilanci (Ufficio Gestione e controllo) per costruire alcuni indicatori relativi all'utilizzo, alla copertura e all'efficienza produttiva.

L'analisi dell'utilizzo e dell'efficienza produttiva avviene tramite il confronto fra ASL e fra periodi di tempo differenti, **valutando la posizione relativa e/o la posizione rispetto a standard**. La tabella riassume indicatori e loro utilizzo.

Flusso	*Utilizzo*		*Efficienza produttiva*	
	Serie storiche	Serie spaziali	Serie storiche	Serie spaziali
Anagrafe assistiti	Variazione nel tempo struttura per sesso ed età della popolazione	Variazione fra ASL della struttura per sesso ed età della popolazione	Variazioni nel tempo della Spesa media per abitante - grezza - specifica per servizio	Variazioni fra ASL della Spesa media per abitante - grezza - specifica per servizio
Certificato di morte	Variazione nel tempo dei tassi di mortalità specifici per causa	Variazione tra ASL dei tassi di mortalità specifici per causa		
SDO	Variazioni nel tempo dei tassi di ricovero DRG specifici o di procedure chirurgiche selezionate	Variazioni fra ASL dei tassi di ricovero DRG specifici o di procedure chirurgiche selezionate		
Vaccinazioni pediatriche obbligatorie	Variazioni nel tempo del tasso di copertura Specifico per vaccino	Variazioni tra ASL del tasso di copertura Specifico per vaccino	Variazioni nel tempo del Costo per bambino vaccinato	Variazioni fra ASL del Costo per bambino vaccinato
Vaccinazioni pediatriche facoltative	Variazioni nel tempo del tasso di copertura specifico per vaccino	Variazioni fra ASL del tasso di copertura specifico per vaccino	Variazioni nel tempo del Costo per bambino vaccinato	
Vaccinazioni adulti	Variazioni nel tempo del tasso di copertura specifico per vaccino	Variazioni fra ASL del tasso di copertura specifico per vaccino	Variazioni nel tempo del Costo per adulto vaccinato	
Prestazioni specialistiche territoriali	Variazioni nel tempo del tasso di - visite ambulatoriali (prime visite nell'anno e visite totali) specifiche per specialità - procedure diagnostiche(laboratoristiche e strumentali) - procedure terapeutiche	Variazioni fra ASL del tasso di - visite ambulatoriali (prime visite nell'anno e visite totali) specifiche per specialità - procedure diagnostiche(laboratoristiche e strumentali) - procedure terapeutiche	Variazioni nel tempo del Costo medio - per assistito - per residente	Variazioni fra ASL del Costo medio - per assistito - per residente
Prestazioni psichiatriche territoriali	Variazioni nel tempo del tasso di - visite ambulatoriali (prime visite nell'anno e visite totali) - procedure terapeutiche	Variazioni fra ASL del tasso di - visite ambulatoriali (prime visite nell'anno e visite totali) - procedure terapeutiche	Variazioni nel tempo del Costo per - visita ambulatoriale - procedura terapeutica	Variazioni fra ASL del Costo per - visita ambulatoriale - procedura terapeutica
Farmaci ospedalieri	Variazioni nel tempo di DDD /per 10.000 ab farmaco specifiche	Variazioni fra ASL di DDD /per 10.000 ab farmaco specifiche	Variazione nel tempo del Costo medio per farmaci per ricovero	Variazione fra ASL del Costo medio per farmaci per ricovero
Farmaci territoriali	Variazioni nel tempo di - numero di ricette per 1000 ab - di DDD /per 10.000 ab farmaco specifiche	Variazioni fra ASL di - numero di ricette per 1000 ab - DDD /per 10.000 ab farmaco specifiche	Variazione nel tempo del Costo medio per farmaci per abitante	Variazione fra ASL del Costo medio per farmaci per ricovero
Notifiche obbligatorie	Variazioni nel tempo di - numero di notifiche /1000ab	Variazioni fra ASL di - numero di notifiche /1000ab		
Pronto soccorso	Variazioni nel tempo di - numero di visite/1000ab - numero di prestazioni/1000ab	Variazioni fra ASL di - numero di visite/1000ab - numero di prestazioni/1000ab	Variazione nel tempo del costo medio per visita	Variazione fra ASL del costo medio per visita
Esenzioni ticket	Variazioni nel tempo di - numero di esenti/1000ab	Variazioni fra ASL di - numero di esenti/1000ab		
Infortuni INAIL	Variazioni nel tempo di - numero di inf/1000ab	Variazioni fra ASL di - numero di inf/1000ab		

Tab. 13 – Analisi dell'utilizzo e dell'efficienza produttiva dei flussi informativi.

5.2 *Linkage* tra Data Base amministrativi

La routine *ReClust.*

Le poche varianti di funzionamento previste sono legate da una parte alla disponibilità o meno di banche dati anagrafiche di riferimento e, in quest'ultima ipotesi, alla scelta e/o alla necessità di effettuare in maniera asincrona o meno la clusterizzazione dei diversi archivi (nell'eventualità del data linkage), dall'altra all'esigenza di intervenire nel flusso naturale del programma, per alterare il peso e la priorità assegnata alle chiavi di clustering onde favorire la generazione di un numero inferiore di classi, ma più "ampie" dal punto di vista della numerosità dei propri elementi. In ogni caso, ReClust opera simulando o, più propriamente, riconducendo i vari scenari a quello classico della clusterizzazione: anche quando i flussi informativi sono più di uno e, pur disponendo di un supporto informativo anagrafico, si preferisce rendere sincrona la procedura di aggregazione nei diversi archivi, l'algoritmo fonde le varie banche dati facendo perno sulla griglia di variabili anagrafiche comuni selezionate come chiavi e, rendendo sequenziali i record di diversa origine, agisce in una sola fase, fatta di più passi, ma comunque unica nei suoi tempi di espletazione, sull'unico mega-archivio così costruito.

Con questa logica di esecuzione ReClust adempie alla duplice necessità, propria del linkage, di individuare il cluster, ma, nello stesso tempo, di identificarlo: si tratta di una identificazione relativa e non assoluta, ma più che sufficiente per riconoscere in fase di output, quando gli archivi originari vengono restituiti nella loro forma separata, i cluster facenti capo allo stesso soggetto (e dunque concettualmente linkati grazie allo stesso codice di identificazione) da quelli che figurano in maniera mutuamente esclusiva in uno solo dei flussi informativi.

La presenza di un archivio "validante", invece consente di distinguere e differire le fasi di clusterizzazioni, ossia di forzare la procedura di clustering (perché sempre di essa si tratta) a procedere in maniera asincrona: si tratta di una scelta obbligata, se si selezionano chiavi di aggregazioni non coincidenti per i vari archivi, facoltativa, se il set di variabili anagrafiche responsabili del clustering sono comuni ai diversi database.

In ogni caso, l'algoritmo di clustering opera tanti cicli di clusterizzazione quanti sono i flussi informativi da integrare, ma il linkage tra le classi è comunque assicurato dall'identificazione "assoluta" che l'archivio anagrafico consente di ottenere grazie all'attribuzione univoca del codice paziente all'atto del riconoscimento dello stesso in seno all'archivio validante. Ricapitolando:

- Assenza di Archivio Anagrafe:
  - Chiavi di clustering comuni ai flussi informativi coinvolti.
    - ✓ Procedura di clustering unica e sincrona - identificazione relativa del paziente.

- Presenza di Archivio Anagrafe:
  - Chiavi di clustering comuni ai flussi informativi coinvolti.
    - ✓ Procedura di clustering unica e sincrona - identificazione assoluta del paziente.
    - ✓ Procedura di clustering ripetuta e asincrona - identificazione assoluta del paziente.
  - Chiavi di clustering distinti per i diversi flussi informativi coinvolti.
    - ✓ Procedura di clustering ripetuta e asincrona - identificazione assoluta del paziente.

A sua volta, la procedura di clustering, a prescindere dallo scenario in cui è chiamata ad operare, consta di passi ben definiti, alcuni dei quali si ripetono ciclicamente, altri condizionabili in misura più o meno forte dall'esterno, altri ancora eseguiti solo in presenza o assenza di un archivio di riferimento. Di fatto si tratta di operazioni sostanzialmente riconducibili ai seguenti 5 step:

- Scelta dei campi di clustering;
- Definizione delle chiavi di clustering;
- Applicazione della procedura di Soundex;
- Assegnazione dei codici di chiave;
- Individuazione e identificazione del cluster.

Di seguito verranno descritte nel dettaglio ciascuna di queste fasi e saranno riportati i riferimenti ai risultati ottenuti nell'uso pratico della procedura ReClust, applicata ai due contesti presi in esame nell'ambito del progetto SIPA:

- ASL n. 8 di Asolo.

  Flussi amministrativi presi in esame:
  - Schede di dimissione ospedaliera dei residenti (periodo: Gennaio 2001 - Giugno 2002);
  - Prescrizioni farmaceutiche territoriali (periodo: Gennaio 2001 - Giugno 2002);
  - Anagrafe Assistibili (storica) con indicazione dell'evento morte aggiornata al 2002.
- Regione Veneto.

  Flussi amministrativi presi in esame:
  - Schede di dimissione ospedaliera (Anno 1999);
  - Schede di morte (Anno 1999).

Scelta dei campi di clustering.

La selezione delle variabili anagrafiche su cui applicare la procedura di clusterizzazione è ovviamente un'operazione che non compete all'algoritmo ReClust, ma precede cronologicamente il lancio della routine.

Si tratta di uno step che è affidato all'utente-analista e alla sua conoscenza della situazione reale circa la

disponibilità del dato e la qualità del contenuto informativo di ciascun campo anagrafico.

Come si è avuto modo di anticipare più volte in precedenza, la precarietà e l'incompletezza dell'informazione non lasciano molti margini alla fantasia dell'operatore, nel senso che tutto ciò che concerne la "descrizione anagrafica" del soggetto va preso in considerazione e, magari con i dovuti adeguamenti e raffinamenti, reso disponibile per le fasi successive del clustering.

I campi che solitamente vanno esaminati per valutarne l'idoneità all'uso possono essere quelli amministrativi come codice fiscale e codice di tessera sanitaria, oppure quelli prettamente anagrafici come nome e cognome, data di nascita, sesso e comune di nascita, e, quando si rivela necessario per ristrettezza di informazioni, anche quelli che descrivono la situazione sociale dell'individuo come comune di residenza, stato civile, professione, titolo di studio, ecc. (l'uso di questi campi, ovviamente, va ponderato con estrema attenzione, perché essi, non definendo, in generale, una condizione "costante" e duratura nel tempo del soggetto, possono risultare tanto più inappropriati per la sua individuazione quanto più è ampio l'arco temporale a cui fanno riferimento i flussi informativi esaminati).

Come già anticipato, l'unico vincolo imposto dalla procedura di clusterizzazione alla scelta dei campi deriva dalla necessità di dover lavorare su una griglia di variabili comuni agli archivi da linkare, quando non è disponibile una banca dati anagrafica che consenta di rendere asincrono il processo, e dunque, in questa ipotesi, vanno esclusi dalla selezione quei campi non sempre presenti in tutti gli archivi, anche se ottimali, laddove figurano, dal punto di vista della qualità del dato.

Inoltre, può capitare a volte che nell'ambito delle variabili scelte possano comparire valori che non rientrano nel *range* di quelli attesi o consentiti: si tratta molto spesso di valori convenzionali che vengono malauguratamente inseriti come indicativi di "dato mancante".

È buona norma, in questi casi, procedere alla rimozione di questi valori, cancellandoli e lasciando vuoto il contenuto del campo, in modo da evitare che essi possano condizionare negativamente l'andamento dell'operazione di clustering con l'individuazione di classi solo in apparenza "anagraficamente" omogenee.

Un'altra regola fondamentale a cui attenersi in fase di preparazione e ripulitura del dato è quella di codificare in maniera opportuna e coerente (adottando o un proprio sistema di codifica o uno proposto a scopo amministrativo) quei campi che per loro natura non lo sono (ad esempio sesso, stato civile, professione, ecc.) o per i quali è stata preferita una descrizione per esteso (comune di nascita, data di nascita, ecc.).

Ciò evidentemente serve a rendere uniforme il contenuto informativo delle variabili e ad evitare che l'originaria disomogeneità delle stesse possa minare e compromettere in partenza il buon esito della procedura di clusterizzazione.

Allo stesso scopo, è consigliabile per quelle variabili per le quali non è possibile introdurre una

rappresentazione codificata, quali ad esempio nome e cognome, adottare un formato omogeneo (tutto maiuscolo o minuscolo) evitando l'uso di apostrofi, lettere accentate, caratteri separatori, ecc.

Infine, l'intervento dell'operatore può risultare auspicabile e lungimirante, in previsione degli step successivi della routine ReClust, se è finalizzato da una parte alla conversione in formato alfanumerico di variabili (come la data di nascita) che si intende sottoporre alla procedura di Soundex e dall'altro alla creazione di nuove variabili, risultanti dalla composizione di campi originariamente distinti (quali ad esempio cognome e nome) che contribuirebbero, come si vedrà, a rendere maggiormente "stabile" ed efficace la procedura di Soundex.

Ricapitolando, la fase di selezione dei campi anagrafici è riconducibile ai seguenti punti:

- Scelta di variabili comuni agli archivi in esame;
- Cancellazione di valori convenzionali;
- Codifica dei campi descrittivi;
- Omogeneizzazione delle stringhe alfanumeriche;
- Conversione in stringhe delle variabili da sottoporre alla procedura di Soundex;
- Creazione di campi composti da sottoporre alla procedura di Soundex.

Nel caso pratico dei flussi informativi della ASL di Asolo e della Regione Veneto sono state prese in considerazione le seguenti variabili anagrafiche, di cui, per ciascuna, viene riportato il dato di incompletezza, ossia il numero di valori mancanti (o resi tali dopo il lavoro di ripulitura) e il corrispondente tasso percentuale:

- ASL n. 8 di Asolo
  - Schede di dimissione ospedaliera dei residenti (n. record: 44.688):
    - ✓ Codice sanitario (n. missing: 0 - 0 %);
    - ✓ Codice fiscale (n. missing: 12.347 - 27,6 %);
    - ✓ Cognome e nome (n. missing: 0 - 0 %);
    - ✓ Data di nascita (n. missing: 0 - 0 %);
    - ✓ Sesso (n. missing: 0 - 0 %).
  - Prescrizioni farmaceutiche territoriali (n. record: 1.769.095):
    - ✓ Codice sanitario (n. missing: 7.888 - 0,4 %).
- Regione Veneto
  - Schede di dimissione ospedaliera (n. record: 1.022.108):
    - ✓ Codice sanitario (n. missing: 117.316 - 11,5 %);
    - ✓ Codice fiscale (n. missing: 442.766 - 43,3 %);
    - ✓ Cognome e nome (n. missing: 114.394 - 11,2 %);

- ✓ Data di nascita (n. missing: 0 - 0 %);
- ✓ Sesso (n. missing: 6 - 0 %).

- Schede di morte (n. record: 43.546):
  - ✓ Codice sanitario (n. missing: 34.583 - 79,4 %);
  - ✓ Codice fiscale (n. missing: 1 - 0 %);
  - ✓ Cognome e nome (n. missing: 0 - 0 %);
  - ✓ Data di nascita (n. missing: 0 - 0 %);
  - ✓ Sesso (n. missing: 0 - 0 %).

Come si può notare, nel caso del *linkage* tra le prescrizioni farmaceutiche e le S.D.O. per la ASL di Asolo, la disponibilità dell'Archivio Anagrafe ha reso possibile di selezionare un set di variabili diverso per i due archivi, senza dover rinunciare, come sarebbe stato necessario in caso contrario, al contributo informativo che campi come cognome e nome, data di nascita e sesso, in virtù della loro completezza, sono in grado di fornire alla procedura di clusterizzazione. Per entrambi i contesti, infine, sono state attuate le procedure di ripulitura, raffinamento, ricodifica e riconversione sopra esposte.

Definizione delle chiavi di clustering.

Dopo la scelta dei campi, il passo più delicato dell'intera fase di preparazione, predisposizione e configurazione dei parametri che regolano il funzionamento dell'algoritmo ReClust è probabilmente quello di definizione delle chiavi di *clustering*, ossia della combinazione delle singole variabili selezionate in set di uno o più campi che individuano la griglia informativa su cui poi la procedura di *clustering* opererà l'aggregazione dei record in classi omogenee.

Da questa scelta dipende la composizione dei *cluster* finali e soprattutto il loro livello di omogeneità: è dunque fondamentale una definizione delle chiavi appropriata e coerente, in funzione di quello che è l'obiettivo della clusterizzazione, ossia l'individuazione di classi di record che abbiano la caratteristica comune di essere attribuibili al medesimo soggetto-paziente.

Il criterio base che deve guidare nella definizione delle chiavi è dettato dalla esigenza di tendere il più possibile a individuare chiavi primarie o univoche, come lo sono per loro natura, ad esempio, il codice fiscale e il codice sanitario: ciò significa combinare in maniera opportuna i singoli campi (utilizzando, se necessario, più volte lo stesso campo in più pattern) in modo da costruire chiavi multicampo che somiglino o si comportino come chiavi primarie.

Mettere insieme, ad esempio, nome, cognome, sesso, data di nascita e comune di nascita significa costruire una chiave a 5 dimensioni che, almeno nel contesto in cui si opera, può essere ritenuta di fatto equivalente al codice fiscale e, quindi, a tutti gli effetti, univoca.

A tal proposito, va tenuto presente che un pattern di variabili può costituire in pratica una chiave primaria,

anche se non lo è in teoria, se applicato in uno scenario informativo particolare, quale può essere, ad esempio, quello delle schede di dimissione ospedaliere nell'ambito di una patologia rara.

In questo contesto una chiave tridimensionale composta da cognome, nome e data di nascita ha in effetti valenza di univocità, se non altro perché si può ritenere che i casi di omonimi coetanei siano da escludersi a priori proprio grazie all'ambito particolare in cui si opera.

Dall'altro canto, nella logica di funzionamento di ReClust, una variabile, coinvolta nella costruzione di una chiave di *clustering*, perde la sua identità singola nel momento in cui confluisce nel pattern multicampo e trasferisce ad esso il suo contributo informativo; analogamente la chiave multidimensionale eredita da ciascuno dei campi che la compongono la qualità del dato, compresa, e questo è da non sottovalutare, l'eventuale incompletezza informativa.

Ciò, in altri termini, significa che in corrispondenza di un record il valore della chiave è considerato mancante se almeno uno dei suoi campi non riporta il dato: pertanto la fase di definizione delle chiavi di *clustering* porta con sé, come effetto collaterale, la problematica legata all'estensione, alla propagazione e all'amplificazione della incompletezza informativa.

Di questo effetto si deve tener conto in fase di costruzione della chiave: se è vero che maggiore è la sua dimensione (ossia il numero di campi che la compongono) più la sua incidenza è assimilabile a quella di una chiave primaria, è altrettanto vero che più campi si coinvolgono nella sua definizione, maggiore è la probabilità di ingenerare perdita di informazione.

Dunque, una chiave di *clustering* deve essere il risultato di un giusto mix tra l'esigenza di individuare un pattern di variabili capace di assicurare alla fine *cluster* omogenei e la necessità di salvaguardare la quantità del contenuto informativo, elemento primario per il conseguimento di tale risultato.

Altro aspetto determinante per il raggiungimento dell'obiettivo finale è l'attribuzione del livello di priorità a ciascuna chiave di *clustering*. Si tratta sostanzialmente di stabilire l'ordine con cui ciascuna chiave e ciascuna combinazione di chiavi dovranno intervenire nella fase ciclica dell'algoritmo ReClust per assegnare i record nei diversi *cluster* che vengono individuati nel corso del processo iterativo.

Senza entrare nel dettaglio tecnico del funzionamento della routine, va detto che la procedura esegue un numero di cicli legato esponenzialmente al numero di chiavi: in ciascuna iterazione viene stabilita l'appartenenza dei record alle varie classi in base al *matching* o, più semplicemente, all'equivalenza informativa degli stessi in corrispondenza di un set di chiavi, il cui numero dal valore massimo scende progressivamente con l'eliminazione sequenziale e alternata delle chiavi con più basso livello di priorità.

In particolare, se si utilizzano 3 chiavi K_1, K_2 e K_3 con priorità rispettivamente alta, media e bassa, il numero di cicli sarà pari a $2^3 - 1 = 7$ e il pattern su cui iterativamente verrà eseguito il confronto informativo sarà composto nell'ordine dalle seguenti chiavi:

- Ciclo con 3 chiavi:
  - K_1, K_2, K_3;
- Cicli con 2 chiavi:
  - K_1, K_2;
  - K_1, K_3;
  - K_2, K_3;
- Cicli con 1 chiave:
  - K_1;
  - K_2;
  - K_3;

La regola da seguire per l'attribuzione dei livelli di priorità è diretta conseguenza di quanto detto in precedenza riguardo al grado di univocità e completezza delle chiavi: è buona norma privilegiare quelle che danno maggiori garanzie in termini sia di capacità di individuazione di *cluster* omogenei che di qualità e presenza del dato.

Molto spesso può risultare utile per diversi motivi, come ad esempio l'inaffidabilità o l'elevata incompletezza del dato in corrispondenza di chiavi primarie a cui si è attribuito priorità massima, innalzare in qualche misura il peso delle chiavi secondarie senza però modificare l'ordine di intervento nella procedura di clusterizzazione.

Mediante una opportuna configurazione dei parametri di input della routine ReClust è possibile influenzare la costruzione dei *cluster*, ritardando l'attribuzione di alcuni record in determinate classi fino all'iterazione che coinvolge singolarmente la chiave a cui si è deciso di attribuire peso maggiore.

Agendo in maniera simultanea sul peso di tutte le chiavi di *clustering*, l'effetto finale non è quello di ripristinare l'iniziale rapporto di priorità, bensì di accrescere l'effetto nella clusterizzazione dei cicli con pattern di una sola chiave: il risultato è l'individuazione di classi più "voluminose" in termini di numero di record raggruppati e di conseguenza meno numerose come effetto della partizione.

Ricapitolando, la fase di definizione delle chiavi di *clustering* deve tener conto dei seguenti aspetti:

- Scelta di chiavi il più possibile univoche;
- Limitazione della perdita della completezza informativa;
- Assegnazione del livello di priorità delle chiavi;
- Eventuale innalzamento del peso delle singole chiavi.

In riferimento all'applicazione pratica della procedura ReClust alle due realtà locali della ASL di Asolo e della Regione Veneto, sono state costruite le seguenti chiavi di *clustering* elencate di seguito in ordine decrescente di priorità e per le quali è riportato il dato assoluto e percentuale di completezza:

- ASL n. 8 di Asolo
  - Schede di dimissione ospedaliera dei residenti (n. record: 44.688):
    - ✓ K_1: Codice sanitario (n. missing: 0 - %);
    - ✓ K_2: Codice fiscale (n. missing: 12.347 - 27,6 %);
    - ✓ K_3: Cognome e nome – Data di nascita - Sesso (n. missing: 0 - 0 %).

- Prescrizioni farmaceutiche territoriali (n. record: 1.769.095):
  - ✓ K_1: Codice sanitario (n. missing: 7.888 - 0,4 %).

➢ Regione Veneto.

- Schede di dimissione ospedaliera (n. record: 1.022.108):
  - ✓ K_1: Codice sanitario (n. missing: 117.316 - 11,5 %);
  - ✓ K_2: Codice fiscale (n. missing: 442.766 - 43,3 %);
  - ✓ K_3: Cognome e nome – Data di nascita - Sesso (n. missing: 114.400 - 11,2 %).
- Schede di morte (n. record: 43.546):
  - ✓ K_1: Codice sanitario (n. missing: 34.583 - 79,4 %);
  - ✓ K_2: Codice fiscale (n. missing: 1 - 0 %);
  - ✓ K_3: Cognome e nome – Data di nascita - Sesso (n. missing: 0 - 0 %).

Come emerge da questi dati la scelta delle chiavi ha tenuto conto in entrambi i contesti della disponibilità di chiavi primarie, anche se una certa incompletezza del dato ha reso necessario rafforzare il peso delle chiavi secondarie. Infine, si noti come l'opportuna combinazione dei campi ha limitato ai minimi termini la perdita del contenuto informativo.

Applicazione della procedura di Soundex.

Scelti i campi anagrafici, definite le chiavi di *clustering*, stabilito il loro livello di priorità, per l'avvio della clusterizzione vera e propria rimane da fissare i criteri e le specifiche con cui si debba operare il confronto informativo tra i record e per mezzo dei quali sia possibile non solo cogliere la correlazione tra i dati, ma misurarne in qualche maniera l'equivalenza informativa.

Accomunare dei record e assegnarli ad una classe di equivalenza significa, per prima cosa, rilevare il *matching* e la coincidenza dei dati in corrispondenza del pattern di chiavi su cui si opera. Ciò può sembrare addirittura banale quando il processo di aggregazione si impernia su campi nei quali il contenuto informativo è rappresentato da un codice.

È abbastanza evidente che quando le variabili in gioco sono chiavi, per di più primarie, come il codice fiscale o il codice sanitario, due o più record si possono attribuire allo stesso soggetto solo e quando c'è perfetta coincidenza tra i valori delle due stringhe: anche un solo carattere diverso, per la natura e il significato stesso dei codici, deve far supporre che si tratti di individui distinti.

Ma quando si lavora con campi il cui contenuto informativo ha un formato "descrittivo", come è nei casi di nome e cognome, è necessario ribaltare i termini del discorso e chiedersi: fino a che punto la discrepanza o la non "sovrapponibilità" delle stringhe è da imputarsi ad una effettiva e reale distinzione dei soggetti a cui si riferiscono? Oppure, viceversa, in che misura si possono tollerare "divergenze"

nell'informazione anagrafica e ritenere due osservazioni riconducibili allo stesso paziente, senza incorrere nell'errore di identificare due soggetti di fatto distinti?

È necessario, allora, a questo scopo introdurre il concetto di "metrica" e di "distanza fonetica" tra le stringhe e definire una funzione che sia in grado di misurare la verosimiglianza di due parole.

Nella procedura di Soundex è stato appunto implementato un algoritmo che fa uso della funzione ***spedis*** (SAS®/BASE), capace di valutare la *spelling distance* come un costo normalizzato risultante da una serie di operazione sulle lettere (cancellazione, inserimento, sostituzione, raddoppio, ecc.) da effettuare in sequenza per convertire uno dei due termini di confronto (*keyword*) nell'altro (*query word*).

In dettaglio, per determinate coppie di record da confrontare (può trattarsi di due osservazioni dell'unico mega-archivio aggregato da clusterizzare, in caso di *clustering* sincrono, o di un osservazione proveniente da una delle banche dati da linkare e di un'altra presente nell'archivio validante nella quale si tenta il "riconoscimento" del paziente, in caso di *clustering* asincrono), la procedura di Soundex, tenendo fissa una delle due voci, genera per l'altra una sequenza di voci derivate, ottenute permutando in tutti i modi possibili le sottostringhe di cui è composta.

Quindi valuta le distanze di spelling tra la prima voce e ciascuna di quelle derivate dalla seconda: la distanza minima viene considerata la misura di verosimiglianza tra i due record. L'esempio seguente illustra il concetto appena esposto:

Record n. 1 (R1)	**Record n. 2** (R2)
Cognome e Nome	*Cognome e Nome*
Rossini Maria Grazia	Rossini Graziella Maria

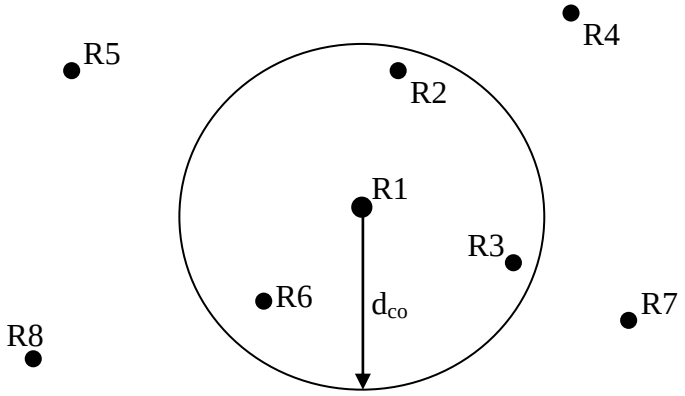
	Voci derivate	*Distanza fonetica*
$R2_1$:	Maria Graziella Rossini	$d(R1,R2_1) = 65$
$R2_2$:	Maria Rossini Graziella	$d(R1,R2_2) = 43$
$R2_3$:	Graziella Maria Rossini	$d(R1,R2_3) = 57$
$R2_4$:	Graziella Rossini Maria	$d(R1,R2_4) = 50$
$R2_5$:	Rossini Maria Graziella	$d(R1,R2_5) = 7$
$R2_6$:	Rossini Graziella Maria	$d(R1,R2_6) = 30$

Misura di verosimiglianza fonetica tra R1 e R2 = distanza minima = $d(R1,R2_5) = 7$

A questo punto, una volta definito il concetto di distanza tra due record in corrispondenza di uno o più campi, è possibile introdurre un valore di *cut-off* con cui discriminare i record "vicini" da quelli "lontani".

Così facendo, in pratica, per ogni record viene definito un intorno circolare di raggio pari al valore di *cut-off*: tutti i record che da esso hanno una distanza inferiore a questo valore soglia e dunque cadono

all'interno del suddetto cerchio vengono considerati dal punto di vista informativo, correlati ed equivalenti al record di riferimento e, quindi, sono candidati a finire nel suo stesso *cluster* di appartenenza; tutti gli altri che ne rimangono fuori, invece, sono da ritenere ad esso non omogenei e, quindi, potenzialmente, riconducibili a soggetti distinti.



Con questa logica, di fatto, il concetto di *matching* viene esteso ed assume un connotato meno rigido: il *join* tra due voci avviene non più solo quando c'è perfetta corrispondenza, ma anche quando la discrepanza tra di esse si mantiene entro limiti accettabili.

Dalla logica di funzionamento della procedura di Soundex appena descritta, risulta chiaro che i campi a cui essa è applicabile devono avere un formato alfanumerico, come si è già avuto modo di precisare in precedenza.

La funzione SAS ***spedis*** esegue le sue manipolazioni sulle lettere o, più in generale, sui caratteri delle stringhe e, dunque, non è pensabile che possa agire su variabili numeriche: se si ritiene utile e vantaggioso applicare il Soundex a campi come la data di nascita, è necessario accertarsi che questa figuri come una stringa e, se così non fosse, è indispensabile provvedere alla necessaria riconversione di formato.

Come detto, inoltre, la distanza fonetica è una grandezza normalizzata e il termine di standardizzazione è rappresentato dalla lunghezza della stringa: ciò appare anche abbastanza naturale dal momento che è plausibile che cambiamenti di lettere in parole brevi abbiano un peso superiore rispetto alle medesime variazioni in stringhe più lunghe.

Proprio per far fronte a ciò, onde poter accettare anche minime discrepanze tra stringhe corte, senza per questo dover aumentare eccessivamente il valore della distanza di *cut-off* (con il rischio, che ciò comporterebbe, di "appiattimento" dell'informazione), può risultare consigliabile, quando ciò non accada già in origine, comporre campi singoli, come nome e cognome, in uno combinato ed unico e riferirsi a questo come variabile da utilizzare, in generale, nell'ambito dell'intera procedura di clusterizzazione ed, in particolare, in seno alla procedura di Soundex.

Tornando all'esempio precedente, se il nome fosse stato considerato separatamente, la distanza minima tra i due record, ossia la misura di verosimiglianza fonetica, sarebbe quasi raddoppiata:

Record n. 1	**Record n. 2**	
Nome	*Nome*	*Verosimiglianza fonetica*
Maria Grazia	Graziella Maria	distanza minima = 12

Non si tratta ancora di una distanza particolarmente rilevante, ma di fronte a livelli di *cut-off* minimi, questi due record sarebbero valutati come potenzialmente non correlati. In generale l'uso di campi combinati e, dunque, la presenza di stringhe medio-lunghe conferiscono alla procedura di Soundex caratteristiche di maggiore "stabilità".

Un ultimo aspetto importante di cui bisogna tenere conto perché questo algoritmo possa operare con efficacia senza produrre effetti spuri indesiderati riguarda l'individuazione del set di record su cui operare il confronto mediante la procedura di Soundex.

È chiaro che confrontare tutte le possibili coppie di record, oltre che essere improponibile dal punto di vista della complessità computazionale e dei tempi di calcolo difficilmente gestibili soprattutto con archivi di enormi dimensioni, è alquanto pericoloso dal momento che porterebbe ad "incrociare", tra l'altro, stringhe molto vicine tra di loro anche se relative, senza ombra di dubbio, a soggetti diversi, come illustrato dal seguente esempio:

Record n. 1	**Record n. 2**	
Cognome e Nome	*Cognome e Nome*	*Verosimiglianza fonetica*
Rossini Maria Grazia	Rossini Mario	distanza minima = 17

Allora la procedura di Soundex limita la ricerca dei record su cui valutare le reciproche distanze in quel set di osservazioni che le chiavi con più alto livello di priorità (rispetto a quella in cui figura il campo sul quale si sta applicando l'algoritmo) hanno già segnalato come potenzialmente correlate dal loro contenuto informativo e la cui eventuale vicinanza fonetica, quindi, andrebbe interpretata legittimamente come un'ulteriore conferma del loro grado di apparentamento.

Questo evidenzia, allora, l'importanza di limitare l'uso della procedura di Soundex ai campi di quelle chiavi che non solo abbiano le caratteristiche finora descritte, ma che, per loro natura, non denotino un carattere di univocità, e, per ruolo svolto nell'ambito della procedura di clusterizzazione, abbiano un livello di priorità di secondo piano.

Il Soundex, quindi, si rivela uno strumento estremamente potente, ma, allo stesso tempo, delicato, che va usato col dovuto equilibrio, bilanciando in maniera opportuna i parametri di input messi a disposizione dalla routine ReClust (scelta della variabile da sottoporre a Soundex, attribuzione del livello di priorità della corrispondente chiave, definizione della distanza di *cut-off*).

Ricapitolando, i concetti basilari su cui poggia la procedura di Soundex sono i seguenti:

- Nozione di metrica e distanza di spelling;

- Misura di verosimiglianza fonetica;
- Distanza di cut-off;
- Scelta di campi combinati, alfanumerici e non univoci da sottoporre a Soundex;
- Attribuzione di un basso livello di priorità alle corrispondenti chiavi.

Nell'ambito dell'applicazione della routine ReClust ai dati relativi alla ASL n. 8 di Asolo e della Regione Veneto, si sono scelte, in entrambe le simulazioni, le variabili data di nascita, cognome e nome come campi a cui applicare l'algoritmo di Soundex.

Gli ultimi due, in alcuni dei flussi informativi in esame, erano presenti come campi distinti e, quindi per i motivi sopra esposti, si è provveduto a combinarli in un'unica variabile. Questa e la data di nascita, insieme con il sesso, sono state utilizzate nell'ambito di un'unica chiave, a cui, come visto in precedenza, è stato assegnato il livello di priorità più basso. Infine si è scelto come livello di *cut-off* della distanza fonetica un valore di soglia pari a 15.

Assegnazione dei codici di chiave.

Una volta selezionate le chiavi e, soprattutto, definite le modalità con cui esse intervengono nell'individuazione del legame informativo tra i record, per raggiungere lo scopo finale di partizionamento delle osservazioni in classi di equivalenza, è necessario attribuire loro dei codici numerici che, in qualche maniera, rappresentino l'effetto, momentaneo e parziale, dell'azione clusterizzante operata singolarmente e autonomamente da ciascuna chiave.

Tali codici, con il loro valore, costituiscono, in altri termini, il marchio che ogni chiave lascia sul record come risultato dell'operazione di *matching* che la routine ReClust affida a ciascun pattern multicampo: sulla base di questo responso, ogni singolo record, di fatto viene potenzialmente assegnato a dei raggruppamenti temporanei che rappresentano il punto di partenza per l'attribuzione ai *cluster* finali, che avverrà solo attraverso l'azione combinata e alternata delle chiavi, regolata dai loro livelli di priorità, descritta nel paragrafo 3.2.

L'assegnazione dei codici di chiave ai record non è, come può sembrare, un'operazione superflua o ridondante: infatti, essa non consiste semplicemente nell'attribuire lo stesso valore del codice di chiave esclusivamente alle sole osservazioni che presentano il medesimo responso informativo in corrispondenza di quel pattern di variabili; al contrario, coerentemente alla novità interpretativa (introdotta dalla procedura di Soundex) del concetto di *matching*, inteso non più solo come coincidenza tra voci, ma come equivalenza tra stringhe, il medesimo valore del codice di chiave viene esteso anche a tutti quei record che, seppur discordanti, si rivelano non eccessivamente distanti in termini di verosimiglianza fonetica e, quindi, secondo i criteri dell'algoritmo di Soundex, annoverabili nello stesso raggruppamento.

Nello schema seguente viene illustrato con un esempio l'idea appena descritta:

	Cognome e Nome	*Data di nascita*	*Sesso*	*Codice di chiave*
R_1	Rossini Maria Grazia	01/01/1960	2	348615
R_2	Rossini Maria Grazia	01/11/1960	2	348615
R_3	Rossini Maria Graziella	01/01/1960	2	348615
R_4	Rossini Graziella Maria	01/01/1960	2	348615

L'assegnazione del codice di chiave viene ripetuta, come è logico, per ciascuna delle chiavi che sono state definite, a partire da quella con più alto livello di priorità, per finire con quella a cui è stata assegnata priorità inferiore.

Alla fine, quindi, ciascun record è contraddistinto da tanti codici quante sono le chiavi ed è prevedibile che i raggruppamenti che ciascuno di essi individua non si sovrappongano mai in maniera coerente, cosa che avverrebbe solamente nella situazione puramente teorica ed ideale di banche dati dal contenuto informativo completo e qualitativamente perfetto. In ogni caso, una volta generati i codici di chiave, la routine ReClust abbandona definitivamente l'uso dei campi anagrafici su cui aveva finora lavorato e prosegue da questo momento nell'operazione di clusterizzazione affidandosi solo ed esclusivamente a questi codici, i quali racchiudono in sé, dal punto di vista informativo, tutto ciò che è necessario e indispensabile per poter raggiungere l'obiettivo finale.

L'assegnazione dei valori dei codici di chiave viene realizzata, inoltre, in modo da discriminare i record non solo in base al raggruppamento in cui vengono inseriti, ma anche in funzione del contributo e dell'attendibilità informativa che essi possono garantire nella fase di generazione dei *cluster* finali.

A questo scopo, viene utilizzato il valore nullo per contraddistinguere le osservazioni che presentano valore mancante nella chiave multicampo a cui il codice si riferisce: queste, dunque, sono raccolte in un'unica classe (contrassegnata dal valore 0) nella quale sono accomunate non dalla omogeneità informativa, ma piuttosto dall'incompletezza del dato e, quindi, dall'impossibilità di influenzare l'attribuzione ad un determinato *cluster* almeno attraverso la chiave in questione.

Col valore negativo, nel solo caso in cui si utilizzi una banca dati con funzione di supporto informativo anagrafico a cui si affida la funzione di riconoscimento e validazione del dato, viene, invece, contraddistinto il record che attraverso la chiave in questione non trova riscontro in seno a questo archivio, né in termini di verosimiglianza fonetica né tanto meno mediante un *matching* completo.

Questa volta, però, assegnando un diverso valore (< 0) del codice di chiave per i diversi raggruppamenti di record, accomunati solo dal fatto di non essere riconosciuti nell'archivio validante, ma comunque distinti dal punto di vista del contenuto informativo, si continua a salvaguardare la distinzione tra le varie classi di appartenenza: il segno negativo del codice di chiave "segnala" semplicemente che quel raggruppamento con quella chiave non è stato identificato dalla banca dati anagrafica, ma il suo valore assoluto ne consente in ogni caso l'individuazione.

Quest'ultimo aspetto evidenzia, inoltre, come le modalità di assegnazione dei valori dei codici di chiave siano diverse a seconda che si disponga o meno di un archivio validante ed è proprio in ragione di ciò che poi, all'atto della costruzione dei *cluster* finali, si avrà, a seconda dei casi, una identificazione assoluta o relativa del paziente.

Nel caso di indisponibilità di una banca dati anagrafica di riferimento, la routine ReClust, come si è detto, agisce in generale sul mega-archivio ottenuto disponendo in sequenza i flussi informativi da linkare ed è sull'insieme di questi record che provvede alla generazione dei valori dei codici in corrispondenza di ciascuna chiave.

Molto semplicemente, per ogni pattern multicampo un indice numerico positivo viene progressivamente incrementato ogni qual volta risulta individuato, secondo la logica ampiamente descritta in precedenza, un raggruppamento omogeneo di osservazioni e il valore ottenuto viene così assegnato ad ogni singolo record: fanno eccezione come anticipato le sole osservazioni che presentano valore mancante in corrispondenza della chiave, per le quali il codice assume sistematicamente valore nullo.

Alla fine del processo iterativo, composto da tanti cicli quante sono le chiavi, ogni record ha associato un numero corrispondente di codici, ciascuno di valore diverso, che determinano la sua appartenenza ai vari raggruppamenti individuati dalle singole chiavi.

Riprendendo l'esempio precedente e assumendo di lavorare su due archivi (DB=1 e 2) e di avere scelto le tre chiavi Codice sanitario (K_1), Codice fiscale (K_2) e Cognome e nome - Data di nascita - Sesso (K_3), una possibile situazione finale non si discosterebbe di molto dal seguente schema:

	DB	*Codice San*	*Cod K_1*	*Codice fiscale*	*Cod K_2*	*Cognome e Nome*	*Data di nasc.*	*S*	*Cod K_3*
R_1	1	123456789	5218	RSSMGR60A41X111X	74623	Rossini Maria Grazia	01/01/1960	2	348615
R_2	2	123456789	5218	RSSMGR60A41X111X	74623	Rossini Maria Grazia	01/01/1960		0
R_3	1		0	RSSMGR60A41X111X	74623	Rossini Maria Grazia	01/11/1960	2	348615
R_4	1	123456789	5218		0	Rossini Maria Graziella	01/01/1960	2	348615
R_5	2	123456789	5218	RSSMGR60A41X111X	74623	Rossini Maria Graziella		2	0
R_6	1	123456789	5218	RSSMGR60A41X111X	74623	Rossini Graziella Maria	01/01/1960	2	348615
R_7	2	123456789	5218		0	Rossini Maria Grazia	31/12/1959	2	348616

Nel caso, invece, in cui si disponga di una banca dati anagrafica utilizzata per la validazione dei flussi informativi in esame, il processo di assegnazione dei codici di chiave, a prescindere dalla modalità sincrona o meno della procedura di clusterizzazione, avviene all'atto del confronto tra il singolo record e l'archivio validante.

Nell'ipotesi in cui esso venga "riconosciuto" o perché foneticamente simile alla voce presente nell'anagrafe di riferimento o addirittura attraverso un *matching* completo con essa, l'osservazione

"eredita" in corrispondenza della chiave, in quel momento in azione, il codice paziente che preventivamente, in fase di inizializzazione, la routine ReClust ha provveduto ad assegnare in maniera univoca ad ogni record dell'archivio validante. Solo nell'eventualità contraria in cui il record non trovi riscontro nell'anagrafe, allora, all'osservazione viene assegnato il valore corrente di un indice numerico, questa volta negativo, che, nel frattempo, si procede opportunamente a decrementare per ogni nuovo raggruppamento omogeneo non identificato nell'archivio validante.

Ciò avviene per tutti quei record nei quali il dato è presente in corrispondenza della chiave in questione; per tutti gli altri, invece, caratterizzati da incompletezza informativa, il valore del codice di chiave, così come accadeva in precedenza, viene impostato automaticamente a 0.

Anche adesso l'operazione è ripetuta ciclicamente tante volte quante sono le chiavi, così da attribuire a ciascuna osservazione un corrispondente numero di codici di chiave: ora, però, ed è questa la novità rispetto al caso in cui non si disponeva di supporto informativo anagrafico, se ciascuna chiave ha assicurato attraverso l'archivio validante un riconoscimento coerente ed univoco, i valori dei codici di chiave coincideranno tra di loro, creando così le premesse per l'identificazione assoluta del *cluster*.

Se nell'esempio precedente si fosse utilizzato un archivio validante applicato in maniera asincrona ai due database, si sarebbe ottenuto un risultato molto simile a quanto illustrato nello schema seguente:

	DB	*Codice San*	*Cod K_1*	*Codice fiscale*	*Cod K_2*	*Cognome e Nome*	*Data di nasc.*	*S*	*Cod K_3*
R_1	1	123456789	193468	RSSMGR60A41X111X	193468	Rossini Maria Grazia	01/01/1960	2	193468
R_2	1		0	RSSMGR60A41X111X	193468	Rossini Maria Grazia	01/11/1960	2	193468
R_3	1	123456789	193468		0	Rossini Maria Graziella	01/01/1960	2	193468
R_4	1	123456789	193468	RSSMGR60A41X111X	193468	Rossini Graziella Maria	01/01/1960	2	193468
R_1	2	123456789	193468	RSSMGR60A41X111X	193468	Rossini Maria Grazia	01/01/1960		0
R_2	2	123456789	193468	RSSMGR60A41X111X	193468	Rossini Maria Graziella		2	0
R_3	2	123456789	193468		0	Rossini Maria Grazia	31/12/1959	2	-173

In entrambi i casi, come già anticipato, i codici di chiave così ottenuti condensano tutto il contenuto informativo necessario per procedere alla individuazione (relativa o assoluta) dei cluster finali e permettono in tal modo alla routine ReClust di abbandonare l'uso dei campi anagrafici che da questo momento in poi non intervengono più nella procedura di clusterizzazione.

Ricapitolando, i punti salienti che caratterizzano la logica di generazione e assegnazione dei codici di chiave sono i seguenti:

- Attribuzione di valori in base alla omogeneità informativa dei record;

- Generazione di valori nulli o negativi per record incompleti o non riconosciuti;
- Creazione di tanti codici quante sono le chiavi in uso;
- Assegnazione di codici assoluti o relativi in presenza o meno di archivio validante;
- Abbandono dell'uso dei campi anagrafici nel prosieguo della clusterizzazione.

Nella dimostrazione pratica dell'uso di ReClust sui dati locali (ASL n. 8 di Asolo) e regionali (Veneto) si è avuta l'occasione di testare i due meccanismi di generazione dei codici di chiave, dal momento che nel primo caso si disponeva dell'Anagrafe Assistibili, a cui è stato affidato il ruolo di archivio validante, che ha consentito per ogni chiave un'aggregazione dei record asincrona e assoluta, mentre nel secondo caso si è proceduto ad un'assegnazione sincrona dei codici sul mega-archivio costituito dalle schede di dimissione ospedaliera e dalle schede di morte, pervenendo ad una identificazione relativa dei raggruppamenti.

Di seguito vengono riportati i dati riassuntivi sulla composizione, qualità informativa e numero delle classi attraverso i valori dei codici di chiave generati nei due contesti:

- ASL n. 8 di Asolo
  - Schede di dimissione ospedaliera dei residenti (n. record: 44.688)
    - ✓ K_1: Codice sanitario:
      - n. record con codice nullo (dato mancante): 0 - 0 %;
      - n. raggruppamenti con codice negativo (non riconosciuti): 228;
      - n. raggruppamenti con codice positivo: 30.443.
    - ✓ K_2: Codice fiscale:
      - n. record con codice nullo (dato mancante): 12.347 - 27,6 %;
      - n. raggruppamenti con codice negativo (non riconosciuti): 1.991;
      - n. raggruppamenti con codice positivo: 20.058.
    - ✓ K_3: Cognome e nome – Data di nascita - Sesso:
      - n. record con codice nullo (dato mancante): 0 - 0 %;
      - n. raggruppamenti con codice negativo (non riconosciuti): 224;
      - n. raggruppamenti con codice positivo: 30.443.
  - Prescrizioni farmaceutiche territoriali (n. record: 1.769.095)
    - ✓ K_1: Codice sanitario:
      - n. record con codice nullo (dato mancante): 7.888 - 0,4 %;
      - n. raggruppamenti con codice negativo (non riconosciuti): 19.051;
      - n. raggruppamenti con codice positivo: 160.708.

- Regione Veneto
  - Schede di dimissione ospedaliera + Schede di morte (n. record: 1.065.654)
    - ✓ K_1: Codice sanitario:
      - n. record con codice nullo (dato mancante): 151.899 - 14,3 %;
      - n. raggruppamenti con codice positivo: 623.563 .
    - ✓ K_2: Codice fiscale:
      - n. record con codice nullo (dato mancante): 442.767 - 41,5 %;
      - n. raggruppamenti con codice positivo: 444.715.
    - ✓ K_3: Cognome e nome – Data di nascita - Sesso:
      - n. record con codice nullo (dato mancante): 114.400 - 10,7 %;
      - n. raggruppamenti con codice positivo: 653.627.

Individuazione e identificazione del cluster.

Una volta valutato l'effetto clusterizzante attribuibile all'azione autonoma e separata di ogni singola chiave, la routine ReClust può finalmente procedere alla costruzione dei *cluster* finali in cui raccogliere i vari record sulla base del responso derivante dall'azione combinata, concomitante e alternata, delle chiavi. Secondo quanto già descritto nel paragrafo 3.2, in base al livello di priorità assegnato loro, le chiavi, attraverso i loro codici, da cui, a questo punto, sono contraddistinte, sono chiamate a fornire l'indicazione definitiva su come ripartire le osservazioni in classi, operando, questa volta, in maniera simultanea e secondo delle regole e priorità che ne disciplinano l'intervento.

Si è detto in precedenza che l'attribuzione dei record ai *cluster* avviene in modo iterativo nel corso di cicli ricorsivi il cui numero complessivo è legato esponenzialmente a quello delle chiavi. In ciascuno di essi è chiamato ad operare un pattern diverso sia per dimensione (numero di chiavi) che per costituzione (identità delle chiavi coinvolte): in particolare se N indica il numero delle chiavi, in corrispondenza del generico ciclo la struttura del pattern è la seguente:

$$K_{i_1}, K_{i_2}, K_{i_3}, ..., K_{i_m} \qquad \text{ciclo i-esimo}$$

dove m varia regressivamente da N a 1 e il generico indice può assumere valore compreso tra 1 e N. Considerando che per ogni valore di *m* (iterazione principale) è possibile individuare $\binom{N}{m}$ pattern *m*-dimensionali (iterazioni secondarie), alla fine si hanno 2^N- 1 cicli in cui si susseguono tutte le possibili combinazioni a gruppi di *m* delle N chiavi. In ciascuno di essi la routine ReClust prende in esame i raggruppamenti di record individuati nella fase precedente da ognuna delle *m* chiavi coinvolte e ne valuta

la rispettiva congruità e corrispondenza: quelle classi che risultano perfettamente "allineate" sulle *m* chiavi, nelle quali, cioè, ciascun codice di chiave (con valore positivo) si mantiene costante per tutte e sole le osservazioni in corrispondenza delle quali lo sono tutti gli altri codici, vengono elette automaticamente a *cluster* in quanto rappresentano chiaramente aggregazioni omogenee di record che tutte le *m* chiavi in gioco hanno saputo coerentemente riconoscere (individuazione del *cluster*).

La sua identificazione, invece, cioè l'attribuzione di un codice d'identità, in base alla stessa logica seguita a suo tempo per i codici di chiave, avviene con modalità diverse a seconda che si disponga o meno del supporto informativo di un archivio validante: in caso negativo, così come allora, viene assegnato come codice di *cluster* il valore di un indice positivo, appositamente incrementato ogni volta che viene identificata una nuova classe, con il quale è possibile ottenere l'identificazione relativa del *cluster*; nell'eventualità in cui, invece, si utilizzi una banca dati di riferimento anagrafico, allora, per quanto è stato detto in precedenza, i vari codici di chiave, che hanno ereditato dall'archivio validante un valore univoco rappresentativo del paziente, non solo risultano costanti sul *cluster* in questione, ma presentano un valore comune perfettamente identico che, quindi, automaticamente viene trasferito alla classe appena individuata (identificazione assoluta del *cluster*).

Nello schema seguente, viene illustrata con un esempio questa duplice possibilità (si suppone di operare in corrispondenza dell'iterazione che coinvolge un pattern bidimensionale costituito dalle chiavi K_1 e K_3):

Indisponibilità di Archivio Validante

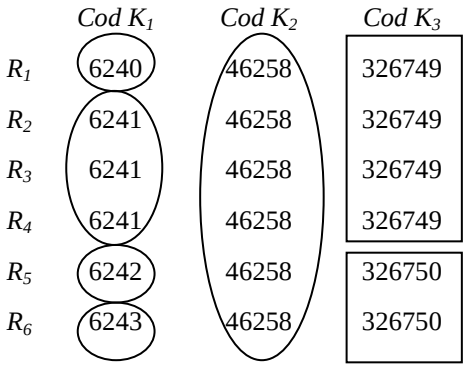
	Cod K_1	*Cod* K_3	*Cod Cluster*
R_1	4523	57924	23456
R_2	4523	57924	23456
R_3	4523	57924	23456
R_4	4523	57924	23456

Disponibilità di Archivio Validante

	Cod K_1	*Cod* K_3	*Cod Cluster*
R_1	10452	10452	10452
R_2	10452	10452	10452
R_3	10452	10452	10452
R_4	10452	10452	10452

Una possibile variante al meccanismo di identificazione del *cluster* appena descritto può essere introdotta innalzando il peso di alcune chiavi a cui si ritiene utile affidare una "forza clusterizzante" superiore a quella che si riuscirebbe a garantire loro attraverso il solo livello di priorità: in questo modo, senza alterare l'ordine di intervento delle chiavi nel processo iterativo, si ammette che una chiave "valorizzata" possa intervenire, seppure con un contributo passivo ed esterno, anche in tutti i passi del processo ricorsivo in cui essa rimane esclusa dal pattern m-dimensionale a cui è affidato il ruolo attivo di definizione delle classi. In particolare, se il *cluster* potenzialmente individuato dal gruppo delle m chiavi chiamate ad operare raccoglie un gruppo di osservazioni di numerosità inferiore rispetto al raggruppamento in cui la chiave valorizzata è riuscita singolarmente ed autonomamente ad aggregarle, allora essa, appellandosi ad una sorta di "diritto di veto" conferitole con l'innalzamento del suo peso

relativo, ha la facoltà di bloccare l'identificazione di quel *cluster* (e quindi di salvaguardare l'integrità del suo raggruppamento) differendolo, nella migliore delle ipotesi, al ciclo successivo:



Nello schema precedente viene illustrata con un semplice esempio una tipica situazione che comunemente si presenta nel processo di clusterizzazione. Si supponga di disporre di 3 chiavi K_1, K_2 e K_3, con priorità rispettivamente alta media e bassa, di trovarsi nel quinto dei 7 cicli previsti, quando è il turno della chiave K_1 (pattern monodimensionale) ad intervenire nella definizione dei *cluster* e di aver attribuito alla sola chiave K_3 un peso relativo superiore; allora, il pattern costituito dalla chiave K_1, agendo dal canto suo secondo i criteri di sua competenza, sarebbe indotto a costruire 4 *cluster*, il primo comprendente il record R_1, il secondo i record R_2, R_3 e R_4, il terzo il record R_5 e il quarto il record R_6: si tratterebbe, però, di tutti *cluster* di numerosità inferiore ai due raggruppamenti che la chiave valorizzata K_3 saprebbe garantire in corrispondenza dei sei record. Allora in virtù del potere conferitole, la chiave K_3 non dà il via libera all'individuazione dei 4 *cluster*, ma, e questo è importante, non ha neppure la facoltà di imporre i suoi due raggruppamenti come *cluster* finali, perché ciò significherebbe un ribaltamento del livello di priorità delle chiavi e, dunque, del loro ordine di intervento nel processo di clusterizzazione (che, in generale, è inviolabile e, in particolare, non modificabile da un peso relativo superiore): pertanto, l'aggregazione dei sei record è rimandata al ciclo successivo, quando è chiamata ad intervenire la chiave K_2, la quale, riuscendo a prospettare una classe unica capace di raccogliere addirittura tutte e sei le osservazioni, ottiene semaforo verde dalla chiave valorizzata K_3 e, dunque, procede all'individuazione del *cluster*; solo nell'ipotesi in cui neppure in questo passo si fossero configurate le condizioni necessarie per l'aggregazione dei record, sarebbe stato necessario rimandare tutto all'iterazione finale, quando, essendo il suo turno, la chiave K_3 risulterebbe, solo allora, pienamente legittimata a proporre ed ottenere la promozione a *cluster* dei suoi due raggruppamenti.

Terminata la fase iterativa costituita dai 2^N - 1 cicli, la maggior parte dei record risultano raggruppati nei loro *cluster* di competenza. È possibile, però, che una minoranza di osservazioni non risulti attribuita a nessuna classe, in quanto, presumibilmente, durante il processo ricorsivo vincoli stringenti e veti incrociati non hanno creato i presupposti indispensabili per definire in maniera netta ed incontrovertibile la loro appartenenza.

Queste osservazioni, dette record residuali, a seconda del segno dei valori dei codici di chiave, possono essere di varia natura e corrispondentemente, avranno un trattamento diverso nel momento in cui andrà stabilito il loro destino. Si possono avere tre diverse categorie di record residuali:

a) record con valore positivo in almeno uno dei codici di chiave;
b) record con valore nullo in tutti i codici di chiave;
c) record con valori negativi o nulli dei codici di chiave.

Nel caso a) si tratta di osservazioni che sono rimaste escluse da tutti i *cluster* finora creati perché evidentemente in nessuno dei cicli della fase iterativa hanno saputo soddisfare i criteri e le specifiche prima descritte che regolano la ripartizione dei record in classi: molto probabilmente sono quelle osservazioni, parzialmente incomplete dal punto di vista informativo, che derivano dalla "frammentazione" dei raggruppamenti inizialmente individuati dalle chiavi a più basso livello di priorità, che, per effetto dell'azione combinata e preminente di pattern di chiavi con priorità superiore, si sono "frantumati" liberando record spuri che successivamente nessuna altra chiave o pattern di chiavi è stato in grado di raccogliere e aggregare:

	Cod K_1	*Cod* K_2	*Cod* K_3
R_1	3712	61239	296729
R_2	3712	61239	296730
R_3	0	0	296730

In quest'esempio il pattern bidimensionale di chiavi K_1, K_2 aggrega in un unico *cluster* i record R_1 e R_2 lasciando spaiato il record R_3, che la chiave K_3 aveva, invece, inserito in un unico raggruppamento temporaneo insieme con R_2: quindi, R_3 non può più essere recuperato, almeno durante questa fase iterativa, né dalle chiavi K_1 e K_2 che, nei passi successivi del processo ricorsivo, chiamati ad agire come pattern monodimensionali, non sono in grado di assegnarlo ad un nuovo *cluster* a causa della incompletezza informativa dell'osservazione, né dalla chiave K_3 che si ritrova il suo raggruppamento, originariamente individuato, ormai frammentato (R_2 è stato già assegnato) e quindi non più proponibile ad essere eletto come *cluster* finale. Allora, record residuali come questo, alla fine della fase iterativa, vengono ridistribuiti nei *cluster* già costruiti in base alla comunanza informativa che mostrano di avere con quelle osservazioni che, al contrario, vi hanno già trovato collocazione: la ricerca dei record "apparentati" e, quindi, indirettamente, delle classi più idonee ad accogliere i residuali, avviene, ancora una volta, attraverso una procedura ciclica costituita da un numero variabile di passi, non superiore, in ogni caso, al numero delle chiavi, durante la quale, procedendo secondo l'ordine stabilito dal livello di priorità delle chiavi, si prova ad individuare il record o i record che in corrispondenza del codice di chiave in esame abbiano lo stesso valore (positivo) del residuale.

Una volta trovata la corrispondenza, il processo ricorsivo si interrompe e l'osservazione residuale, come

in una specie di effetto domino, viene assegnata al *cluster* di appartenenza del record a cui è accomunato, acquisendo, a pieno titolo, il diritto di cittadinanza in quella classe, di cui eredita il codice identificativo. Nell'esempio precedente, dopo i due tentativi affidati alle chiavi K_1, K_2 e risultati vani per l'incompletezza informativa su quei campi del record residuale R_3, è K_3 che, rilevandone il legame con R_2 in corrispondenza del suo codice di chiave, estende i confini del *cluster* precedentemente definito dal pattern bidimensionale K_1, K_2 e comprendente già R_1 e R_2 e ne crea uno più ampio capace di abbracciare anche il nuovo elemento R_3.

Nel caso b), invece, si ha a che fare con tutte quelle osservazioni che sono incomplete, dal punto di vista informativo, in corrispondenza di tutti i gruppi di campi scelti per la procedura di clusterizzazione, e che, quindi, essendo prive dell'elemento basilare per il conseguimento di qualunque obiettivo di aggregazione, ossia il dato, risultano irrimediabilmente escluse da ogni tentativo di raggruppamento.

Pertanto, ad esse non viene assegnato nessun valore del codice di *cluster* che, dunque, resta indefinito e, di conseguenza, non consente alcun coinvolgimento di questi record nella successiva eventuale fase di *linkage* e di integrazione dell'informazione.

Il caso c) si può verificare, per quanto detto in precedenza, solo nell'eventualità in cui nella procedura di *clustering* intervenga una banca dati anagrafica alla quale si assegni funzioni di archivio validante.

In questa ipotesi, quando un record non viene riconosciuto in seno a tale base di dati di riferimento attraverso nessuna delle chiavi di *clustering*, tutti i corrispondenti codici di chiave assumono valori negativi o, tutt'al più, nulli in caso di assenza del dato.

Le osservazioni con queste caratteristiche non vengono prese in considerazione nel processo iterativo di individuazione e identificazione dei *cluster* affidato all'azione combinata e alternata delle chiavi, nel quale ad essere esaminati sono i soli record contraddistinti da almeno un valore positivo nei corrispondenti codici: infatti, solo in questo caso, tale valore, essendo rappresentativo del soggetto ed essendo stato ereditato dall'archivio validante, può essere così trasferito al *cluster* al momento della sua individuazione e utilizzato come codice di identificazione assoluto.

I record residuali, caratterizzati, invece, da codici di chiave negativi, non possono garantire un meccanismo di aggregazione e soprattutto di identificazione basato su questo principio; essi, segno negativo a parte, possono semmai essere regolati in fase di clusterizzazione da criteri simili a quelli che disciplinano e governano la ripartizione in classi in assenza di banca dati validante, quando le osservazioni sono contraddistinte da valori dei codici di chiave ottenuti attraverso l'incremento di un indice progressivo e, quindi, non derivanti da una fonte oggettiva e assoluta e, in quanto tali, utilizzabili solo per un'identificazione relativa del *cluster*.

Allora, per l'aggregazione di questi record residuali si procede con una nuova applicazione del processo iterativo affidato ancora ai diversi pattern di chiavi e regolato da modalità del tutto analoghe a quelle che

non prevedono, appunto, l'uso di un archivio di riferimento: il risultato finale consiste, diversamente dal punto a), non nella semplice ridistribuzione delle osservazioni residuali in classi già costruite, ma nella produzione di nuovi *cluster* che, però, rispetto a quelle di prima generazione, risultano non validate: è per questo motivo che, proprio per rimarcarne la diversa origine e differente valenza, si preferisce non assegnare loro un codice identificativo, ma ci si limita a segnalarne l'identificazione.

Va detto, infine, che i record residuali che anche dopo questa fase risultino non assegnati ad alcuna classe, vengono, questa volta, ridistribuiti, secondo le modalità descritte nel punto a), tra i *cluster* appena creati o, su preciso input esterno, anche nell'ambito di quelli di prima generazione.

Terminata finalmente la fase vera e propria di clusterizzazione, la routine ReClust provvede a fornire anche uno strumento per poter valutare il livello di affidabilità, qualità e completezza dei *cluster* così ottenuti. Anche in questo caso, lo scenario è diverso a seconda che si disponga o meno di un archivio validante.

Quando non si utilizza un supporto informativo anagrafico di riferimento, l'algoritmo si limita a qualificare la composizione di ogni singolo *cluster* individuato e, in particolare, per ciascuna chiave che ha contribuito al *clustering* viene fornita l'indicazione di omogeneità, disomogeneità, completezza o incompletezza informativa della classe; in dettaglio:

- col simbolo 1 è indicata perfetta omogeneità del *cluster*, nel senso che il valore del codice di chiave si mantiene costante su tutta la classe;
- col simbolo 0 è indicata disomogeneità più o meno spiccata, nel senso che il valore del codice di chiave varia nella classe, assumendo valore diverso in corrispondenza di due o più record;
- col simbolo X è indicata assenza totale d'informazione (su quella chiave) e il corrispondente codice assume valore nullo;
- col simbolo Y è indicata incompletezza informativa parziale, nel senso che in alcuni record del *cluster* il dato è mancante (valore del codice nullo), mentre nelle rimanenti osservazioni è omogeneo (valore del codice costante).

In presenza, invece, di una banca dati di riferimento, l'informazione fornita non è solo relativa alla composizione del *cluster*, bensì è indicativa anche del livello di appropriatezza dell'appartenenza del generico record a quella classe e, quindi, misurando, oltre che la completezza informativa, anche il grado di congruità dell'appartenenza dell'osservazione in corrispondenza di ciascuna chiave, si configura più come una grandezza specifica del record che del *cluster*; in dettaglio:

- col simbolo 1 è contrassegnato il record che legittimamente è inserito in quella classe, nel senso che esso, attraverso la chiave di riferimento, è stato riconosciuto nell'archivio validante proprio nel soggetto con cui poi viene identificata la classe (valore del codice di chiave uguale al valore del codice di *cluster*);

- col simbolo 0 è contraddistinto il record che, attraverso la chiave di riferimento, non ha avuto riscontro in seno all'archivio validante e, dunque, è inserito in quella classe solo per l'effetto dell'azione delle altre chiavi (valore del codice di chiave negativo e, quindi, necessariamente diverso dal valore del codice di cluster);
- col simbolo Ø è indicato il record che appartiene illegittimamente a quella classe, nel senso che nell'archivio validante è stato riconosciuto in un soggetto diverso da quello con cui, per effetto delle altre chiavi, è stata identificata la classe (valore del codice di chiave positivo, ma, comunque, diverso dal valore del codice di *cluster*);
- col simbolo X è contrassegnato il record che è caratterizzato da incompletezza del dato in corrispondenza della chiave di riferimento e la cui appartenenza alla classe è stata decretata dal contributo informativo delle altre chiavi (valore del codice di chiave nullo).

A questo punto la routine ReClust è in grado di restituire in output, nella loro forma originaria e in maniera del tutto trasparente, gli archivi che sono stati sottoposti alla sua azione clusterizzante. L'informazione aggiuntiva rappresentata dal codice di *cluster* consente di individuare all'interno di ogni banca dati il soggetto-paziente a cui si può far risalire ciascun record-evento e, in più, grazie alla sua natura di codice identificativo, esso assume il ruolo di chiave di *linkage* per l'integrazione dei diversi flussi informativi.

Non solo: proprio per effetto della sua caratteristica di univocità, tale codice consente di eliminare nei sistemi integrati ottenuti qualunque riferimento anagrafico al paziente e, nel pieno rispetto delle norme che mirano alla salvaguardia della privacy, assicura il totale anonimato del soggetto di riferimento.

Ciò è esattamente quanto è stato fatto nell'applicazione pratica della routine ReClust ai flussi informativi della ASL n. 8 di Asolo e della Regione Veneto.

Di seguito vengono riportati i principali risultati ottenuti dalla clusterizzazione nei due contesti, con particolare riferimento al numero di pazienti individuati nei diversi archivi e al livello qualitativo e di omogeneità riscontrato nella composizione dei *cluster*:

- ASL n. 8 di Asolo (clusterizzazione asincrona con archivio validante).
  - Schede di dimissione ospedaliera dei residenti (n. record: 44.688)
    - ✓ Campi anagrafici conservati:
      - Data di nascita;
      - Sesso;
      - Evento morte (ereditato dall'anagrafe);
      - Data di morte (ereditata dall'anagrafe);
      - Comune di residenza (ereditato dall'anagrafe).
    - ✓ Numero cluster-pazienti individuati (totale: 30.667, di cui 30.493 identificati [99,4 %])

- N. pazienti presenti nel solo archivio S.D.O.: 4.428 (14,5 %);
- N. pazienti comuni all'archivio prescrizioni: 26.065 (85,5 %).

✓ Composizione dei *cluster*:

Qualità K_1	*Qualità K_2*	*Qualità K_3*	*N. Record*	*%*
1	1	1	29.513	66
1	X	1	12.114	27,1
1	0	1	2.757	6,2
0	X	0	173	0,4
0	X	1	49	0,1
1	1	0	32	0,1
1	Ø	1	14	0
1	0	0	12	0
1	X	0	11	0
0	1	1	8	0
0	0	0	4	0
0	0	1	1	0

▪ Prescrizioni farmaceutiche territoriali (n. record: 1.769.095)

✓ Campi anagrafici conservati:

- Data di nascita (ereditata dall'anagrafe);
- Sesso (ereditata dall'anagrafe);
- Evento morte (ereditato dall'anagrafe);
- Data di morte (ereditata dall'anagrafe);
- Comune di residenza (ereditato dall'anagrafe).

✓ Numero cluster-pazienti individuati (totale: 187.647, di cui 160.708 identificati (85,6 %))

- N. pazienti presenti nel solo archivio prescrizioni: 134.643 (83,8 %);
- N. pazienti comuni all'archivio S.D.O.: 26.065 (16,2 %).

✓ Composizione dei *cluster*:

Qualità K_1	*N. Record*	*%*
1	1.735.733	98,1
0	25.474	1,4
X	7.888	0,4

➢ Regione Veneto (clusterizzazione sincrona)

▪ Schede di dimissione ospedaliera + Schede di morte (n. record: 1.065.654)

✓ Campi anagrafici conservati:

- Data di nascita;
- Sesso;
- Data di Morte.

✓ Numero cluster-pazienti individuati (totale: 720.656, di cui 720.635 identificati [99,9 %])

- N. pazienti presenti nel solo archivio S.D.O.: 677.137 (94 %);
- N. pazienti presenti nel solo archivio Schede di morte: 13.027 (1,8 %);

- N. pazienti comuni ai due archivi: 30.471 (4,2 %).

✓ Composizione dei *cluster*:

Qualità K_1	*Qualità K_2*	*Qualità K_3*	*N. Cluster*	*%*
1	X	1	254.542	35,3
1	1	1	241.513	33,5
X	1	1	66.496	9,2
1	1	X	61.020	8,5
X	X	1	26.905	3,7
1	Y	1	18.039	2,5
Y	Y	1	13.087	1,8
Y	1	1	7.281	1
1	Y	Y	4.942	0,7
X	1	X	4.350	0,6
1	1	Y	4.070	0,6
Y	1	Y	3.148	0,4
Y	0	1	3.064	0,4
1	0	1	2.866	0,4
X	Y	1	1.073	0,1
1	X	X	1.055	0,1
Y	X	1	847	0,1
1	0	Y	799	0,1
0	1	Y	673	0,1
Y	Y	Y	665	0,1
0	Y	1	643	0,1
Y	1	X	596	0,1
X	0	1	472	0,1
0	X	1	331	0
1	0	X	298	0
0	1	1	259	0
1	X	0	199	0
1	Y	0	192	0
0	1	X	175	0
0	0	1	149	0
Y	0	Y	121	0
1	1	0	120	0
1	0	0	119	0
Y	Y	0	105	0
1	X	Y	104	0
X	1	Y	75	0
0	Y	Y	44	0
Y	0	0	40	0
Y	1	0	36	0
1	Y	X	34	0
X	1	0	21	0
X	X	X	21	0
0	0	Y	17	0
Y	0	X	15	0
0	0	0	10	0
0	Y	0	8	0
Y	X	0	7	0
0	1	0	4	0
0	0	X	2	0
X	Y	0	2	0
X	Y	Y	1	0
Y	Y	X	1	0

5.3 Definizione degli indicatori costruiti a partire da un archivio integrato ed esempi di applicazione

Primo modello: livello regionale

Interventi di rivascolarizzazione cardiaca (cod 36.xx) della regione Veneto anno 1999. Mortalità e ri-ospedalizzazione precoce (a 30 e 60 gg).

Il titolo è sufficientemente esplicativo degli obiettivi del primo scenario – modello.

Alcune possibili domande a cui l'analisi richiesta potrebbe fornire delle risposte sono:

- Quanti sono gli interventi di rivascolarizzazione cardiaca in un anno?
- Quanti sono i pazienti trattati per rivascolarizzazione cardiaca in un anno?
- Qual è il carico ospedaliero per ciascun paziente trattato?
- Quali indicatori di esito sono ragionevolmente individuabili e concretamente valutabili?

L'apparente ovvietà e ragionevolezza dei quesiti appena formulati deve, ora, confrontarsi con la disponibilità dei dati, la capacità degli strumenti e delle metodologie da utilizzare. Certamente non sfugge che la scelta degli esiti da valutare, mortalità e ri-ospedalizzazione precoce, qualifica in senso clinico-epidemiologico la nostra indagine. E' opportuno sottolineare la rilevanza di questo aspetto che da solo rappresenta la sostanziale differenza tra ciò che "normalmente" avviene in uno studio epidemiologico e che altrettanto "normalmente" è escluso dalle valutazioni descrittive delle Banche Dati Amministrative.

L'altro passaggio, altrettanto rilevante, indicato dalla serie di domande proposte, è quello di misurare-descrivere non solo le prestazioni eseguite ma, soprattutto, i pazienti trattati.

Se, a questo punto, il razionale dello studio è sufficientemente definito, la successiva fase di analisi può essere illustrata nei seguenti passi.

Pur disponendo dell'intero archivio informatizzato delle dimissioni ospedaliere dal 1995 al 2001 abbiamo dovuto limitare la nostra indagine al solo anno 1999 per la necessaria concomitante disponibilità dell'archivio Schede di Morte 1999 (completo, informatizzato e validato). E' seguita la preparazione degli archivi amministrativi:

- SDO Regione Veneto (Anno 1999);
- Schede di morte Regione Veneto (Anno 1999);

su cui è stata applicata la procedura di clusterizzazione e *linkage* (**ReClust**) con la definitiva individuazione, attraverso una serie di chiavi identificative, dei soggetti contemporaneamente presenti nei due archivi.

Si è quindi proceduto alla selezione di tutte le dimissioni ospedaliere e schede di morte dei soggetti che nel corso dell'anno prescelto avevano subito almeno una procedura di rivascolarizzazione cardiaca. Detta prestazione è stata identificata dalla presenza dei caratteri "36" (=Interventi sui vasi del cuore) nelle prime due posizioni dei codici di intervento e/o procedure. Finalmente è stato generato il file dati contenente N° 13.825 record, denominato **"CUORE"**, il cui contenuto informativo è strettamente correlato agli obiettivi della nostra indagine.

L'allegato 1 riporta tutte le procedure classificate dall'ICD 9 CM con codice 36 e successive sottocategorie che sono considerate nella presente analisi (Interventi sui vasi del cuore).

Risultati

I risultati di seguito presentati sono stati selezionati allo scopo di illustrare le potenzialità della metodologia adottata e come ulteriore chiarimento della procedura **ReClust**, pertanto non esplorano l'intero contenuto informativo dai data base esaminati (SDO e Schede di Morte).

Una ulteriore precisazione: la procedura **ReClust** costituisce la tappa fondamentale dell'intero processo in quanto provvede alla identificazione del soggetto sia per ciascuno dei due Data Base esplorati (SDO e Schede di Morte). All'identificazione univoca dei soggetti segue una fase altrettanto importante che porta alla costruzione del vero e proprio **file di dati integrati** (denominato **"CUORE"** nell'esempio considerato) il cui contenuto informativo è effettivamente il risultato della integrazione e del *linkage* tra i due Data Base considerati.

Le successive tabelle illustrano il tracciato record del file **"CUORE"** considerando 4 casi concreti, esemplificativi di 4 diversi percorsi assistenziali e di differenti esiti:

Tabella 14. Caso n° 1; Sesso M, 66 aa, un solo ricovero nell'anno 1999 con procedura di rivascolarizzazione (36.01=PTCA); vivo alla dimissione, vivo al 31/12/99; giornate di degenza utilizzate: 6 gg.; DRG prodotto:112, INTERVENTI SUL SISTEMA CARDIOVASCOLARE PER VIA PERCUTANEA (tariffa Lit. 13.136.710).

Tabella 15. Caso n° 2; Sesso M, 72 aa; tre ricoveri ordinari nell'anno 1999; 2° ricovero con procedura di rivascolarizzazione (36.13=Bypass aortocoronarico di tre arterie coronariche); vivo alle dimissione, deceduto al 31/12/99 - oltre i 60 gg.; riospedalizzato entro 30 gg; totale giornate di degenza utilizzate: 56 gg.; DRG prodotti: 122, 107, 316 (Totale tariffe Lit. 36.340.080).

Tabella 16. Caso n° 3; Sesso M, 85 aa; tre ricoveri ordinari nell'anno 1999; 2° ricovero con procedura di rivascolarizzazione (36.04=Infusione trombolitica nell'arteria intracoronaria); deceduto in ospedale - entro 60 gg; riospedalizzato - entro 60 gg; totale giornate di degenza utilizzate: 60 gg.;

DRG prodotti: 88, 475, 99 (Totale tariffe Lit. 26.727.380)

Tabella 17. Caso n° 4; Sesso M, 74 aa; dieci ricoveri (di cui 3 in regime di DH) nell'anno 1999; 5° ricovero con procedura di rivascolarizzazione (36.15= Bypass singolo mammaria interna-arteria coronaria; 36.12=Bypass aortocoronarico di due arterie coronariche); vivo alla dimissione, vivo al 31/12/99; ripetutamente riospedalizzato - entro 30 e 60 gg -; totale giornate di degenza utilizzate: 119 gg.; DRG prodotti: 133, 122, 122, 127, 107, 138, 132, 86, 85, 86 (Totale tariffe Lit. 47.424.948)

Appare evidente come informazioni rilevate dai due archivi interrogati si completino scambievolmente. In Tabella 14, insieme alla illustrazione del primo caso clinico, è riportato anche la descrizione del tracciato record del file **"CUORE".** In questo caso il singolo record coincide con il singolo paziente e con il contenuto informativo della singola SDO.

Diversa la situazione per i successivi tre casi caratterizzati da ricoveri multipli nel corso dell'anno 1999 e con decesso sia alla dimissione (Caso N° 3) che dopo la dimissione (Caso N° 2). Il singolo paziente, e la sua storia ospedaliera, viene descritta tra 3 record (Casi N° 2 e 3) o 10 record (Caso N° 4). L'avvenuto decesso (Casi N° 2 e 3) viene rilevato dall'incrocio con l'archivio schede di morte ed attribuito al corrispondente soggetto insieme all'informazione se detto decesso è avvenuto entro 30 o 60 gg dalla data di dimissione della SDO con procedura di rivascolarizzazione.

Tabella 14. Caso n° 1; Sesso M, 66 aa, un solo ricovero nell'anno 1999 con procedura di rivascolarizzazione (36.01=PTCA); vivo alla dimissione, vivo al 31/12/99; giornate di degenza utilizzate: 6 gg.; DRG prodotto: 112, INTERVENTI SUL SISTEMA CARDIOVASCOLARE PER VIA PERCUTANEA (tariffa Lit. 13.136.710)

Variabile	Ricovero	Descrizione
cod_rec1	360328	Codice di record
cod_clu	33564	Codice di clusterizzazione; identifica il soggetto
n1	1	Numero elementi cluster=Numero Sdo per soggetto
intcuore	1	Identifica Sdo con procedura 36.xx (1=si; 2=no)
deceduto	0	Sogg. Deceduto nell'anno 1999 (1=si; 2=no)
dec_osp	0	Sogg. Deceduto in ospedale (1=si; 2=no)
dec30gio	0	identifica il paziente deceduto entro 30 giorni dalla data di dimissione della SDO con procedura 36.xx
dec60gio	0	identifica il paziente deceduto entro 60 giorni dalla data di dimissione della SDO con procedura 36.xx;
paz30gio	0	identifica il paziente che si riospedalizza almeno una volta nell'arco dei 30 giorni successivi al ricovero con procedura 36.xx;
paz60gio	0	identifica il paziente che si riospedalizza almeno una volta nell'arco dei 60 giorni successivi al ricovero con procedura 36.xx
sdo30gio	0	identifica il ricovero avvenuto nell'arco dei 30 giorni successivi a quello con procedura 36.xx
sdo60gio	0	identifica il ricovero avvenuto nell'arco dei 60 giorni successivi a quello con procedura 36.xx
codice	XXXXXX	Codice presidio ospedaliero
subcod	00	_flag_
anno	1999	Anno dimissione
codreg	050	Codice regione
ulssres	XXX	Codice aussl di residenza
numsched	99xxxxxx	Numero identificativo Sdo
sesso	1	1= M; 2=F
nascita	25/12/1933	Data di nascita
ric_reg	1	Regime ricovero (1=Ordinario; 2=DH)
datarico	20/01/1999	Data di ricovero
giornidh	0	N° accesso in DH
datadimi	26/01/1999	Data di dimissione
dim_rep	0801	Codice reparto di dimissione (08=CARDIOLOGIA)
dim_mod	2	Modalità di dimissione
data_dec		Data decesso
dim_dia	413	Diagnosi principale alla dimissione; 413 = Angina pectoris
pat_co1	411	I° Diag. concomitante; 411=Altre forme acute e subacute di cardiopatia ischemica
pat_co2	5355	II° Diag. concomitante; 5355=Gastrite e gastroduodenite non specificate
pat_co3		III° Diag. concomitante
pat_co4		IV° Diag. concomitante
pat_co5		V° Diag. concomitante
int_chi	8856	Intervento; 88.56=Arteriografia coronarica con catetere doppio
int_al1	3601	36.01=Angioplastica percutanea transluminale coronarica di vaso singolo [PTCA] senza menzione di agente trombolico
int_al2		
los	6	Durata della degenza in gg
Drg	112	112=INTERVENTI SUL SISTEMA CARDIOVASCOLARE PER VIA PERCUTANEA
importo	**13.136.710**	Tariffa DRG in Lit.

Tabella 15. Caso n° 2; Sesso M, 72 aa; tre ricoveri ordinari nell'anno 1999; 2° ricovero con procedura di rivascolarizzazione (36.13= Bypass aortocoronarico di tre arterie coronariche); vivo alle dimissione, deceduto al 31/12/99 - oltre i 60 gg.; riospedalizzato entro 30 gg; totale giornate di degenza utilizzate: 56 gg.; DRG prodotti: 122, 107, 316 (Totale tariffe Lit. 36.340.080)

Variabile	1° ricovero	2° ricovero	3° ricovero
cod_rec1	112270	112270	112270
cod_clu	676096	676096	676096
n1	3	3	3
intcuore	0	1	0
deceduto	1	1	1
dec_osp	0	0	0
dec30gio	0	0	0
dec60gio	0	0	0
paz30gio	1	1	1
paz60gio	1	1	1
sdo30gio	0	0	1
sdo60gio	0	0	1
codice	XXXXXX	XXXXXX	XXXXXX
subcod	00	00	00
anno	1999	1999	1999
codreg	050	050	050
ulssres	XXX	XXX	XXX
numsched	99xxxxxx	99xxxxxx	99xxxxxx
sesso	1	1	1
nascita	22/04/1927	22/04/1927	22/04/1927
ric_reg	1	1	1
datarico	10/06/1999	30/06/1999	06/08/1999
giornidh	0	0	0
datadimi	24/06/1999	31/07/1999	17/08/1999
dim_rep	0801	4801	4801
dim_mod	2	2	2
data_dec	23/10/1999	23/10/1999	23/10/1999
dim_dia	4100(1)	4140	585(2)
pat_co1	4110	410	4259
pat_co2	5850(2)	8872	2113
pat_co3	4140	V458	
pat_co4		4259	
pat_co5			
int_chi	8954	3613(7)	5498
int_al1	8952	3961(8)	
int_al2	8856(3)	9670	
int_al3	3722(4)	5498	
int_al4	8853(5)		
int_al5	8872		
los	14	31	11
drg	122(6)	107(9)	316(10)
importo	**5.593.020**	**24.492.500**	**6.254.560**

Descrizione dei Codici per le Patologie, le Procedure e i DRG più rilevanti

(1) 410.0 Infarto miocardico acuto della parete anterolaterale
(2) 585 Insufficienza renale cronica
(3) 88.56 Arteriografia coronaria con catetere doppio
(4) 37.22 Cateterismo cardiaco del cuore sinistro
(5) 88.53 Angiocardiografia del cuore sinistro
(6) DRG 122, MAL CARDIOVASCOLARI CON INFARTO MIOCARDICO ACUTO SENZA COMPLICANZE CARDIOVASCOLARI DIMESSI VIVI;
(7) 36.13 Bypass aortocoronarico di tre arterie coronariche
(8) 39.61 Circolazione extracorporea ausiliaria per chirurgia a cuore aperto
(9) DRG 107 BYPASS CORONARICO SENZA CATETERISMO CARDIACO
(10) DRG 316 INSUFFICIENZA RENALE

Tabella 16. Caso n° 3; Sesso M, 85 aa; tre ricoveri ordinari nell'anno 1999; 2° ricovero con procedura di rivascolarizzazione (36.04= Infusione trombolitica nell'arteria intracoronaria); deceduto in ospedale - entro 60 gg - ; riospedalizzato - entro 60 gg -; totale giornate di degenza utilizzate: 60 gg.; DRG prodotti: 88, 475, 99 (Totale tariffe Lit. 26.727.380)

Variabile	**1° ricovero**	**2° ricovero**	**3° ricovero**
cod_rec1	43275	43275	43275
cod_clu	539499	539499	539499
n1	3	3	3
intcuore	0	1	0
deceduto	1	1	1
dec_osp	1	1	1
dec30gio	0	0	0
dec60gio	1	1	1
paz30gio	0	0	0
paz60gio	1	1	1
sdo30gio	0	0	0
sdo60gio	0	0	1
codice	XXXXXX	XXXXXX	XXXXXX
subcod	00	00	00
anno	1999	1999	1999
codreg	050	050	050
ulssres	XXX	XXX	XXX
numsched	99xxxxxx	99xxxxxx	99xxxxxx
sesso	1	1	1
nascita	03/06/1914	03/06/1914	03/06/1914
ric_reg	1	1	1
datarico	08/04/1999	03/05/1999	26/05/1999
giornidh	0	0	0
datadimi	23/04/1999	20/05/1999	23/06/1999
dim_rep	2101	2101	2101
dim_mod	2	2	1
data_dec	23/06/1999	23/06/1999	23/06/1999
dim_dia	4912	7860	7860
pat_co1	4280	4912	4912
pat_co2	7860		4140
pat_co3			
pat_co4			
pat_co5			
int_chi		3604	
int_al1		9672	
int_al2		3891	
int_al3		3893	
int_al4			
int_al5			
los	15	17	28
drg	88	475	99
importo	**5587200**	**17443025**	**3697155**

Tabella 17. Caso n° 4; Sesso M, 74 aa; dieci ricoveri (di cui 3 in regime di DH) nell'anno 1999; 5° ricovero con procedura di rivascolarizzazione (36.15= Bypass singolo mammaria interna-arteria coronaria; 36.12=Bypass aortocoronarico di due arterie coronariche); vivo alla dimissione, vivo al 31/12/99; ripetutamente riospedalizzato - entro 30 e 60 gg -; totale giornate di degenza utilizzate: 119 gg.; DRG prodotti: 133, 122, 122, 127, 107, 138, 132, 86, 85, 86 (Totale tariffe Lit. 47.424.948)

Variabile	**1° rivovero**	**2° rivovero**	**3° rivovero**	**4° rivovero**	**5° rivovero**	**6° rivovero**	**7° rivovero**	**8° rivovero**	**9° rivovero**	**10° rivovero**
cod_rec1	207512	207512	207512	207512	207512	207512	207512	207512	207512	207512
cod_clu	77532	77532	77532	77532	77532	77532	77532	77532	77532	77532
n1	10	10	10	10	10	10	10	10	10	10
intcuore	0	0	0	0	1	0	0	0	0	0
deceduto	0	0	0	0	0	0	0	0	0	0
dec_osp	0	0	0	0	0	0	0	0	0	0
dec30gio	0	0	0	0	0	0	0	0	0	0
dec60gio	0	0	0	0	0	0	0	0	0	0
paz30gio	1	1	1	1	1	1	1	1	1	1
paz60gio	1	1	1	1	1	1	1	1	1	1
sdo30gio	0	0	0	0	0	1	0	0	0	0
sdo60gio	0	0	0	0	0	1	1	0	0	0
codice	xxxxxx	xxxxxx	xxxxxx	xxxxxx	xxxxxx	xxxxxx	xxxxxx	xxxxxx	xxxxxx	xxxxxx
subcod	00	00	00	00	00	00	00	00	00	00
anno	1999	1999	1999	1999	1999	1999	1999	1999	1999	1999
codreg	050	050	050	050	050	050	050	050	050	050
ulssres	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
numsch	99xxxxxx	99xxxxxx	99xxxxxx	99xxxxxx	99xxxxxx	99xxxxxx	99xxxxxx	99xxxxxx	99xxxxxx	99xxxxxx
sesso	1	1	1	1	1	1	1	1	1	1
nascita	10/02/1925	10/02/1925	10/02/1925	10/02/1925	10/02/1925	10/02/1925	10/02/1925	10/02/1925	10/02/1925	10/02/1925
ric_reg	2	1	1	1	1	1	1	2	1	2
datarico	21/01/1999	21/03/1999	21/03/1999	23/04/1999	10/05/1999	22/05/1999	17/06/1999	14/07/1999	28/08/1999	05/10/1999
giornidh	1	0	0	0	0	0	0	2	0	3
datadimi	21/01/1999	21/03/1999	13/04/1999	10/05/1999	20/05/1999	05/06/1999	07/07/1999	04/08/1999	25/09/1999	17/12/1999
dim_rep	0801	2601	0801	2601	0701	2601	2601	2601	2601	2601
dim_mod	1	6	2	2	2	2	2	1	2	1
data_dec										
dim_dia	4140	410	410	4281	4140	4268	4148	5119	5119	5119
pat_co1	4148	414	4148	4148		5119	V458		4148	
pat_co2	4409	4409	4140	4409		2858	5119		2428	
pat_co3			4409			4409	4239		4414	
int_chi	8907	8872	8856	8965	3961	3491	8965	3491	905	3491
int_al1	9301	9919		8954	3615	8954	3491		8741	
int_al2	8871	8965		9396	3612	8879	8906		8801	
int_al3		8952			8907	8744	8744		8744	
los	1	1	23	17	10	14	20	2	28	3
drg	133	122	122	127	107	138	132	86	85	86
importo	**257632**	**1398255**	**5593020**	**4522625**	**21582500**	**3933350**	**3681635**	**325483**	**5804965**	**325483**

Dall'esame dei quattro casi proposti e del contenuto informativo del file dati **"CUORE"** ne ricaviamo alcune osservazioni:

- ciascun record del file dati integrati **"CUORE"**, in definitiva, corrisponde a ciascuna SDO dei soggetti individuati per aver eseguito nel corso dell'anno 1999 almeno una procedura di rivascolarizzazione cardiaca. A ciascun record sono state aggiunte le variabili necessarie alla nostra indagine e ricavabili sia all'interno dello stesso archivio SDO (Riospedalizzazione) che dell'archivio correlato (Mortalità);
- ciò è stato possibile per la condizione, relativamente favorevole, di collegamento di due soli archivi di cui uno (Schede di morte) senza possibilità di duplicati per lo stesso soggetto;
- condizioni meno favorevoli, quali osservazioni multiple per lo stesso soggetto e più di due archivi correlati, dovranno necessariamente essere affrontati con diversa strategia come meglio illustrato nel modello successivo;
- dopo l'applicazione della procedura di ReClust, con l'attribuzione di un codice univoco e anonimo per ciascun soggetto, tutte le successive analisi risultano completamente sganciate dai dati anagrafici del soggetto.

A questo punto possiamo facilmente e finalmente analizzare il nostro file dati integrati "CUORE" ed affrontare i quesiti posti dalla nostra indagine.

Sintesi risultati (vedi anche Figura 2):

- Nel corso dell'anno 1999 nella regione Veneto sono stati eseguiti N° 6.677 interventi di rivascolarizzazione cardiaca cosi come emerge dalla semplice analisi descrittiva delle dimissioni ospedaliere dell'anno esaminato (vedi Tabella 18). La tipologia di intervento e la posizione occupata nella SDO (intervento chirurgico e/o procedura 1, 2, 3, 4, 5) è riportata in Tabella 18.
- L'analisi appena più approfondita rivela che tali interventi sono riportati in N° 5421 SDO; pertanto N° 1256 (=6677 meno 5421) sono interventi di rivascolarizzazione cardiaca concomitanti o contemporanei allo stesso episodio di ricovero.
- La successiva clusterizzazione ha consentito di individuare il numero dei pazienti sottoposti a detti interventi (N° 5099 soggetti).
- Di questi 5099 pazienti, N° 292 hanno eseguito due interventi e N° 30 più di due interventi nel corso di differenti episodi di ricovero nell'anno 1999.
- Il numero complessivo degli episodi di ricovero, sia in regime ordinario che di DH (con e senza procedure di rivascolarizzazione cardiaca), attribuito ai 5099 soggetti nel corso di tutto l'anno 1999 è di N°13.825 ricoveri; con una spesa complessiva, DRG pesata, di Lit. 133.784.272.527 e complessive N°126.487 giornate di degenza.

- I soli 5421 ricoveri con procedure di rivascolarizzazione cardiaca comporta invece una spesa totale di Lit 98.426.017.477 (73% della spesa complessiva) ed un numero di giornate di degenza di 53.698 gg.(42% delle giornate complessive).
- Si sono individuati rispettivamente N° 2034 e N° 1502 riospedalizzazioni entro 60 e 30 gg dalla data di dimissione della SDO con procedura di rivascolarizzazione, mentre il numero dei pazienti con una o più riospedalizzazioni entro 60 e 30 gg sono rispettivamente 1664 e 1374.
- Il *linkage* con l'archivio schede di morte ha permesso di individuare, per la nostra popolazione di N° 5099 soggetti, N° 184 decessi avvenuti entro l'anno 1999 di cui ben 160 in ospedale alla dimissione dell'ultimo ricovero. Il numero dei pazienti deceduti, per qualsiasi causa, a 60 e 30 gg dalla data di dimissione della SDO con procedura di rivascolarizzazione era rispettivamente di 146 e 135 soggetti.

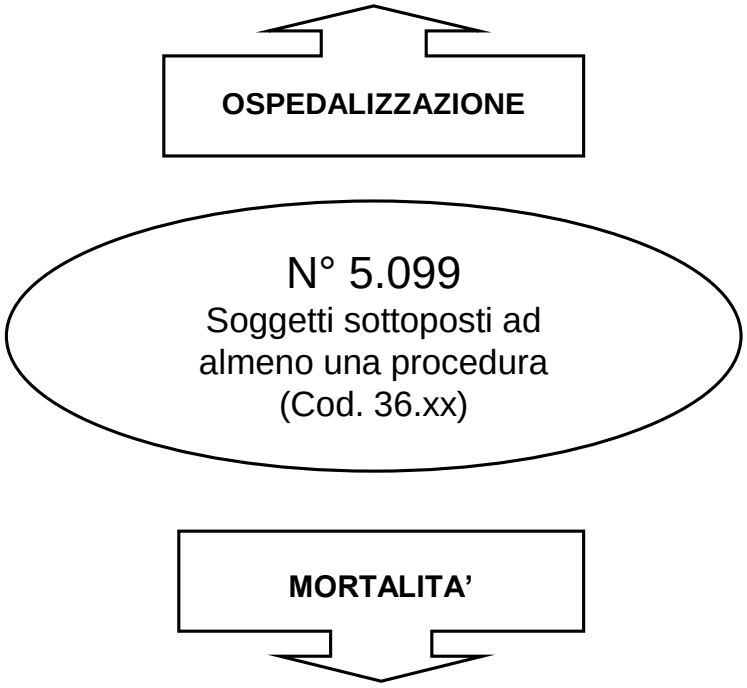
L'analisi per causa di decesso e/o riospedalizzazione non era inclusa nella presente indagine è, pertanto sarà oggetto di una specifica e successiva analisi.

Tabella 18. Numero complessivo procedure di rivascolarizzazione cardiaca eseguiti nell'anno 1999 nella regione Veneto

ICD 9 CM	int_chi	int_al1	int_al2	int_al3	int_al4	int_al5	Totale	Descrizione
36.01	1418	766	204	141	32	13	2574	Angioplastica percutanea transluminale coronarica di vaso singolo [PTCA] senza menzione di agente trombolico
36.02	9	5	3	1	1		19	Angioplastica percutanea transluminale coronarica su vaso singolo con agente trombolitico
36.03	3	1			1		5	Angioplastica dell'arteria coronarica a torace aperto
36.04	7	2	1	1	1		12	Infusione trombolitica nell'arteria intracoronaria
36.05	303	145	61	18	5	6	538	Angiopl. perc. transluminale coronarica vaso mult. [PTCA] eseguita durante stesso int., con o senza menzione agente tromb.
36.09	384	311	89	52	7	12	855	Altra rimozione di ostruzione dell'arteria coronarica
36.10	3		1				4	Bypass aortocoronarico per rivascolarizzazione cardiaca, NAS
36.11	120	163	64	11	2		360	Bypass aortocoronarico di una arteria coronarica
36.12	398	245	84	12	5		744	Bypass aortocoronarico di due arterie coronariche
36.13	495	116	45	10	8	1	675	Bypass aortocoronarico di tre arterie coronariche
36.14	137	28	20	3	3		191	Bypass aortocoronarico di quattro o più arterie coronariche
36.15	412	172	58	5	4	2	653	Bypass singolo mammaria interna-arteria coronarica
36.16	15	2					17	Bypass doppio mammaria interna-arteria coronarica
36.19	3						3	Altro bypass per rivascolarizzazione cardiaca
36.2							0	Rivascolarizzazione cardiaca mediante innesto arterioso
36.3	3	2	1	1	1		8	Altra rivascolarizzazione cardiaca
36.91	1	8	4				13	Riparazione di aneurisma dei vasi coronarici
36.99	2	2		2			6	Altri interventi sui vasi del cuore
Totale	3713	1968	635	257	70	34	6677	

Figura **2.** Anno 1999 – Regione Veneto
Interventi di rivascolarizzazione cardiaca (Codici ICD 9 CM 36.xx)

	Complessiva	Solo procedure 36.xx	Entro 30 gg.*	Entro 60 gg.*
N° SDO	13.825	5.421	1.502	2.034
N° gg di degenza	126.487	53.698	14.641	20.552
Tariffa DRG in Lit	133.784.272.527	98.426.017.477	7.383.579.572	10.175.511.595
N° pazienti	5.099	5.099	1.374	1.664



Mortalità	N° Decessi	Tasso grezzo x 1000
Mortalità per tutte le cause, anno 1999	184	36.1
In Ospedale	160	31.4
Mortalità entro 30 gg.*	135	26.5
Mortalità entro 60 gg.*	146	28.6

**** dalla data di dimissione della SDO con proc. 36.xx***

Secondo modello: livello ASL

Mortalità e assorbimento di risorse (ospedalizzazione e prescrizioni farmaceutiche) della popolazione di una ASL con particolare riferimento alla popolazione anziana (>= 65aa.). Periodo considerato (Gennaio 2001 Giugno 2002).

Lo scenario qui considerato si caratterizza immediatamente, rispetto al precedente, per il livello locale della ASL e, di conseguenza, per la più immediata e tempestiva disponibilità degli archivi necessari e, infine, per l'utilizzo, nel modello proposto, del rilevante ed imponente flusso informativo costituito dalle prescrizioni farmaceutiche.

Obiettivo principale è quello di valutare, per ciascun assistito, i costi diretti (per ospedalizzazione e prescrizioni farmaceutiche **(PF)**) prodotti nel periodo considerato.

Obiettivi secondari:

- individuare la popolazione di assistiti con maggiore assorbimento di risorse;
- valutare se a tali maggiori costi corrisponde una maggiore gravità clinica (=appropriatezza);
- individuare eventuali sottogruppi di popolazione per analisi e/o interventi mirati.

Come nel modello precedente sono stati individuati gli archivi disponibili e valutati per completezza e validità interna:

- SDO ASL n. 8 (Gennaio 2001-Giugno 2002);
- Prescrizioni farmaceutiche ASL n. 8 (Gennaio 2001-Giugno 2002);
- Registro Mortalità ASL n. 8, da Anagrafe Assistiti (Anni 2001-2002).

Su detti archivi è stata applicata la procedura di clusterizzazione e *linkage* (**ReClust**) (descritta in dettaglio in "Materiali e Metodi") con la definitiva individuazione di tutti i soggetti e la loro contemporanea e multipla presenza negli archivi considerati.

Sottolineiamo, ancora una volta, che l'applicazione della procedura consente, alla fine, l'attribuzione di un codice di clusterizzazione univoco per ciascun paziente con rispetto del completo anonimato e permette di sganciare, tutte le successive analisi, dai dati anagrafici dell'assistito.

Tutti i soggetti che hanno ricevuto almeno una **PF** e/o un ricovero ospedaliero nel periodo in esame sono stati inclusi nella presente analisi.

Si è quindi proceduto alla costruzione del **file dati integrati** (denominato **ASOLO**) il cui contenuto informativo riassume ed integra tutti i DBA considerati.

Il tracciato record è di seguito mostrato e, come appare evidente, la sua struttura è completamente diversa da quella dei singoli DBA di partenza (Tabella 19).

Infatti ciascun record corrisponde ad uno ed uno solo assistito e riassume la sua "storia assistenziale" nel periodo esaminato.

In considerazione della natura esemplificativa del presente modello si è scelto di privilegiare alcune "variabili" di tipo generale e di significato complessivo. Infatti il contenuto informativo dei singoli DBA analizzati è enorme; basti pensare a: diagnosi, procedure, DRG, farmaci. Ciò che ci preme, in questa sede, è dimostrare come, con le stesse modalità esecutive, sia immediatamente possibile costruire singoli **file dati integrati** che esplorano, di volta in volta, singoli problemi ed ipotesi di lavoro. Nello sviluppo del modello vengono ulteriormente forniti alcuni esempi.

La variabile "costo", di cui qui si discute in modo particolareggiato, è certamente la variabile più immediata, più sintetica, più maneggevole, di immediato utilizzo quando si collegano DBA differenti (quali SDO e PF) in cui proprio il costo o la tariffe rappresenta, spesso, l'elemento comune, "unificante", di più facile interpretazione e forse, proprio per ciò, di più facile manipolazione.

Perciò si è scelto, nel modello presentato, di associare sempre, alla variabile costo un'altra variabile, altrettanto sintetica e maneggevole, quale la mortalità o stato in vita. La forte correlazione che emerge tra le due variabili e/o fenomeni (aumento dei costi, incremento della mortalità) ha suggerito l'opportunità di selezionare un sottocampione di assistiti di ASOLO per i quali è stato possibile ricostruire "la storia-il carico assistenziale" nell'anno che precede il decesso; il file corrispondente è denominato **ASOLODEC** e mantiene la stessa struttura record del file di origine ASOLO illustrata in Tabella 19.

Tabella 19. Tracciato record file ASOLO (N° 165.128 record) e ASOLODEC (N° 890 record), descrizione variabili.

Variable	Type	Len	Label
cod_paz	Num	8	Codice Paziente
sesso	Num	3	Sesso
nascita	Char	10	Data di Nascita
asoclust	Char	4	Cluster di appartenenza
n_presc	Num	8	N. Prescrizioni
n_st_far	Num	8	N. max Prescr. dello stesso farmaco
n_dv_far	Num	8	N. max Prescr. farmaci diversi al mese
n_ga	Num	8	N. diversi Gruppi Anatomici
patt_gt	Char	15	Pattern dei Gruppi Terapeutici Assunti
lab_gt	Char	75	Label dei Gruppi Terapeutici Assunti
sp_farm	Num	8	Spesa Farmaci
n_ric_t	Num	8	N. Totale Ricoveri
n_ric_o	Num	8	N. Ricoveri Ordinari
n_drg	Num	8	N. diversi DRG
patt_dim	Char	23	Pattern dei Reparti di Dimissione
lab_dim	Char	75	Label dei Reparti di Dimissione
patt_drg	Char	75	Pattern dei DRG
long_los	Num	8	N. Giorni Degenza in Lungodeg. e Riabil.
tot_los	Num	8	N. Giorni Degenza complessiva
deceduto	Num	3	Deceduto
dec_osp	Num	8	Deceduto in Ospedale
data_dec	Char	10	Data Decesso
giopremo	Num	8	Giorni intercorsi tra ultima SDO e morte
sp_ric_t	Num	8	Spesa Totale Ricoveri
sp_ric_o	Num	8	Spesa Ricoveri Ordinari

L'accurata scelta di "variabili" significative e sintetiche per il file dati ASOLO fa si che ciascuna delle variabili riportate in tabella 6 possa essere riguardata come indicatore di:

- costo;
- assorbimento di risorse;
- cronicità
- complessità assistenziale
- presenza di patologie o di polipatologie
- gravità clinica

Infatti la rilettura di ciascuna riga della tabella 6 ci permette di ri-classificare le variabili presentate come indicatori (Tabella 19 bis).

Tabella 19 bis. Tracciato record file ASOLO e ASOLODEC, descrizione indicatori.

Variable	Label	INDICATORE di	DESCRIZIONE
cod_paz	Codice Paziente		Identificativo anonimo del paziente
sesso	Sesso		Dati anagrafici
nascita	Data di Nascita		Dati anagrafici
Asoclust	Cluster di appartenenza	Appartenenza a definiti sottogruppi	Nell'esempio sono identificati pazienti con costi crescenti
n_presc	N. Prescrizioni	Assorbimento di risorse	N° totale di prescriz. Farmaceu. nel periodo esaminato
n_st_far	N. max Prescr. dello stesso farmaco	Cronicità,	più prescrizioni dello stesso principio attivo – ATC7
n_dv_far	N. max Prescr. farmaci diversi al mese	Complessità,	più patologie farmaco trattate con più principi attivi – ATC7
n_ga	N. diversi Gruppi Anatomici	Complessità,	più farmaci di diversi gruppi anatomici – ATC1
patt_gt	Pattern dei Gruppi Terapeutici Assunti	Complessità	Tipi di farmaci assunti per sottogruppo terap. ATC3
lab_gt	Label dei Gruppi Terapeutici Assunti		Specifica dei farmaci prescritti per sottogruppo terap. ATC3
sp_farm	Spesa Farmaci	Costo	Spesa complessiva per prescrizioni farmaceutiche
n_ric_t	N. Totale Ricoveri	Assorbimento di risorse	N° complessivo SDO per ricoveri ordinari e DH
n_ric_o	N. Ricoveri Ordinari	Assorbimento di risorse	N° complessivo SDO solo per ricoveri ordinari
n_drg	N. diversi DRG	Complessità	N° diversi DRG prodotti nel periodo esaminato
patt_dim	Pattern dei Reparti di Dimissione	Complessità	Percorsi assistenziali intraospedalieri
lab_dim	Label dei Reparti di Dimissione	Complessità	Specifica le Discipline dei reparti di dimissione
patt_drg	Pattern dei DRG	Poli patologie	Specifica il/i DRG prodotti
long_los	N. Giorni Degenza in Lungodeg. e Riabil.	Assorbimento di risorse	N° complessivo giornate di degenza in reparti ospedalieri di Lungodegenza e Riabilitazione
tot_los	N. Giorni Degenza complessiva	Assorbimento di risorse	N° complessivo giornate di degenza; tutti i reparti
deceduto	Deceduto	Gravità clinica	
dec_osp	Deceduto in Ospedale	Gravità clinica	
data_dec	Data Decesso		
giopremo	Giorni intercorsi tra ultima SDO e morte	Gravità clinica	
sp_ric_t	Spesa Totale Ricoveri	Costo	
sp_ric_o	Spesa Ricoveri Ordinari	Costo	

Nello sviluppo del modello qui presentato solo alcuni indicatori sono stati utilizzati a titolo esemplificativo.

Su una popolazione di 223.749 (dato Istat – di cui 35.197>=65 aa) della Aussl 8 di Asolo N° 165.128 soggetti hanno ricevuto almeno una PF e/o un ricovero ospedaliero nel periodo considerato e pertanto sono riconosciuti come assistiti ed inclusi nell'analisi. Su questa stessa popolazione negli anni 2001 e

2002 sono stati registrati, dalla Anagrafe Assistiti, complessivamente N° 3.621 decessi con la relativa data di morte.

Ripetute analisi esplorative hanno permesso di individuare i seguenti criteri di stratificazione per gravosità assistenziale e gravità clinica e qui proposti per la loro semplicità e praticità:

1. età >= 65aa;
2. N° prescrizioni dello stesso farmaco (principio attivo, ATC7) >=3 (indicatore di cronicità);
3. N° ricoveri ordinari >=2 (indicatore di forte assorbimento di risorse).

Sulla base di questi semplici criteri la nostra popolazione di assistiti può essere successivamente stratificati nei seguenti quattro gruppi:

Tabella 20. Stratificazione della popolazione assistita della ASL 8 di Asolo per gravosità assistenziale e gravità clinica

DESCRIZIONE	N. Soggetti
Assistiti, tutte le età; almeno 1 prescrizione e/o 1 ricovero	**165.128**
Assistiti, età ≥ 65 aa; almeno 1 prescrizione e/o 1 ricovero	**33.614**
Assistiti, età ≥ 65 aa; almeno 1 prescrizione + almeno 1 ricovero	**7.870**
Assistiti, età ≥ 65 aa; ≥ 3 prescrizione + ≥ 2 ricoveri	**2.729**

Dove il 1° gruppo comprende tutti gli assisti della nostra ASL campione, il 2° gruppo la sola popolazione anziana, il 3° gruppo la popolazione anziana con almeno una prescrizione e un ricovero ordinario, il 4° gruppo la popolazione anziana più gravosa per cronicità (più di 2 PF dello stesso principio attivo) e carico assistenziale (più di un ricovero ordinario).

Come atteso tali sottogruppi di popolazione si caratterizzano per un progressivo incremento nell'assorbimento di risorse con il seguente gradiente di costo medio per assistito per le due componenti di spesa considerate (ricoveri ospedalieri e PF) (Tabella 21).

Tabella 21 - Gradiente di spesa per i diversi gruppi di popolazione considerati.

DESCRIZIONE	N. Totale	Costo medio per Paziente (ricoveri) Lit x 1000	Costo medio per Paziente (ric.ord.) Lit x 1000	Costo medio per Paziente (farmaci) Lit x 1000	Costo medio per Paziente Lit x 1000
Assistiti, tutte le età; almeno 1 prescrizione e/o 1 ricovero	**165.128**	**1.081**	**906**	**552**	**1.633**
Assistiti, età ≥ 65 aa; almeno 1 prescrizione e/o 1 ricovero	**33.614**	**2.625**	**2.298**	**1.295**	**3.920**
Assistiti, età ≥ 65 aa; almeno 1 prescrizione + almeno 1 ricovero	**7.870**	**9.889**	**9.297**	**1.747**	**11.635**
Assistiti, età ≥ 65 aa; ≥ 3 prescrizione + ≥ 2 ricoveri	**2.729**	**16.190**	**15.417**	**2.173**	**18.363**

Il costo medio complessivo per paziente cresce da Lit 1.633.000 del primo gruppo a Lit 18.363.000 per il quarto gruppo con un incremento di 11,2 volte. Il 4° gruppo, composto solo di 2.729 soggetti, ovvero solo 1,6% della popolazione assistita, assorbe oltre il 18% della spesa complessiva per farmaci e ricoveri ospedalieri.

Un quinto gruppo costituito dal quartile superiore degli assistiti più gravosi della tabella precedente realizzano un costo medio di Lit 34.059.000. Solo 682 assistiti (0,4% della popolazione) assorbono l'8,6% della spesa complessiva (Tabella 21 bis).

Tabella 21 bis. Costi medi nel sottogruppo di pazienti con maggior carico assistenziale

DESCRIZIONE	N. Totale	Costo medio per Paziente (ricoveri) Lit x 1000	Costo medio per Paziente (ric.ord.) Lit x 1000	Costo medio per Paziente (farmaci) Lit x 1000	Costo medio per Paziente Lit x 1000
Quartile superiore Assistiti, età ≥ 65 aa; ≥ 3 prescrizione + ≥ 2 ricoveri	**682**	**31.206**	**30.129**	**2.853**	**34.059**

Come già anticipato al dato di costo, espressione di maggiore carico assistenziale abbiamo voluto associare il dato di mortalità come espressione di gravità clinica (Tabella 22).

Tabella 22. Gravosità assistenziale e gravità clinica. Costo medio per paziente e tasso grezzo di mortalità nei diversi sottogruppi di popolazione.

DESCRIZIONE	N. Totale	Costo medio per Paziente Lit x 1000	Tot Decessi Anni 2001 2002	N° Decessi in Ospedale	Tasso Mortalità x1000 xAnno	% Decessi in Ospedale
Assistiti, tutte le età; almeno 1 prescrizione e/o 1 ricovero	**165.128**	**1.633**	**3621**	**1372**	**11**	**38**
Assistiti, età ≥ 65 aa; almeno 1 prescrizione e/o 1 ricovero	**33.614**	**3.920**	**3059**	**1163**	**46**	**38**
Assistiti, età ≥ 65 aa; almeno 1 prescrizione + almeno 1 ricovero	**7.870**	**11.635**	**1776**	**933**	**113**	**53**
Assistiti, età ≥ 65 aa; ≥ 3 prescrizione + ≥ 2 ricoveri	**2.729**	**18.363**	**831**	**433**	**152**	**52**
Quartile superiore Assistiti, età ≥ 65 aa; ≥ 3 prescrizione + ≥ 2 ricoveri	**6.82**	**34.059**	**246**	**147**	**180**	**60**

Parallelamente all'incremento del costo medio, i cinque gruppi di popolazione considerati presentano un altrettanto evidente incremento del tasso grezzo di mortalità dall'11 al 180 per mille.

Da rilevare, inoltre, l'alto tasso di mortalità intra-ospedaliera, dal 38 fino al 60% dei decessi.

E' ovviamente possibile, utilizzando tutti gli altri indicatori-variabili, del file dati integrati ASOLO, descrivere il nostro campione e i diversi gruppi di popolazione per pattern di prescrizione, oppure per DRG, per disciplina dei reparti di dimissione, per giornate di degenza, per ricorso al ricovero in riabilitazione ecc. La completezza dell'analisi ci appare, però, poco atta ad illustrare ulteriormente le potenzialità del modello. Più utile, come sempre, esplorare, a titolo esemplificativo, alcune ipotesi. Abbiamo perciò selezionato, su richiesta di alcuni gruppi di clinici, gruppi di paziente per alcune patologie di interesse.

Il caso DIABETE.

La popolazione diabetica farmaco trattata è stata individuata, utilizzando il DBA delle prescrizioni farmaceutiche, per la presenza di almeno 3 prescrizione di antidiabetici (A10 della classificazione ATC). Utilizzando lo stesso tipo di analisi fin qui illustrata è stato possibile calcolare la spesa media del paziente diabetico nella ASL esaminata (Tabella 23)

Tabella 23. Confronto tra costo medio per assistito nella popolazione generale e nella popolazione diabetica farmaco-trattata della ASL n. 8 di Asolo.

DESCRIZIONE	N. Totale	Costo medio per Paziente Lit x 1000	N. Paz. con A10	% Paz. con A10	Costo medio per Diabetico Lit x 1000	% Costo
Assistiti, tutte le età; almeno 1 prescrizione e/o 1 ricovero	**165.128**	**1.633**	**6.188**	**3,7**	**4.313**	**264,1**
Assistiti, età ≥ 65 aa; almeno 1 prescrizione e/o 1 ricovero	**33.614**	**3.920**	**3.600**	**10,7**	**5.284**	**134,8**
Assistiti, età ≥ 65 aa; almeno 1 prescrizione + almeno 1 ricovero	**7.870**	**11.635**	**1.186**	**15,1**	**12.444**	**107,0**
Assistiti, età ≥ 65 aa; ≥ 3 prescrizione + ≥ 2 ricoveri	**2.729**	**18.363**	**534**	**19,6**	**18.309**	**99,7**
Quartile superiore Assistiti, età ≥ 65 aa; ≥ 3 prescrizione + ≥ 2 ricoveri	**682**	**34.059**	**146**	**21,4**	**32.167**	**94,4**

N° 6188 sono i soggetti cronicamente trattati con antidiabetici tra gli assistiti della ASL esaminata (3,7%) il loro costo medio è di Lit 4.313.000, ovvero 264 % in più rispetto al costo medio per assistito dell'intera popolazione. Tale netto incremento tende a ridursi fino a scomparire nei gruppi di popolazione più gravosi. Pazienti diabetici sono comunque sempre presenti nei diversi sottogruppi di popolazione esaminati anzi la loro percentuale tende progressivamente a crescere nei gruppi di pazienti più gravosi.

Come atteso il diabete si conferma come una delle patologie più rilevanti in termini di carico economico e di assorbimento di risorse. Una sottoanalisi ha confermato la sottostima del diabete rilevato attraverso le SDO in cui il diabete non viene riportato in oltre un terzo dei casi.

Il caso dell'ONCOEMATOLOGIA

Nel secondo esempio le patologie selezionate sono, per definizione, rare, ma , per la loro gravità, facilmente rilevabile dalle SDO. Sono stati selezionati sia pazienti con almeno una diagnosi di Mieloma Multiplo che pazienti con almeno una diagnosi di Leucemia Acuta.

Tabella 24. Costo medio per assistito affetto da Leucemia Acuta e Mieloma Multiplo

DESCRIZIONE	N. Totale	Costo medio per Paziente (ricoveri) Lit x 1000	Costo medio per Paziente (ric.ord.) Lit x 1000	Costo medio per Paziente (farmaci) Lit x 1000	Costo medio per Paziente Lit x 1000
Assistiti, tutte le età; almeno 1 ricovero con diagn. di LEUCEMIA	29	25.971	22.573	2.050	28.021
Deceduti (01/01/02 - 30/06/02), tutte le età; almeno 1 ricovero con diagn. di LEUCEMIA Anno precedente la morte	2	30.350	30.350	7.777	38.127
Assistiti, tutte le età; almeno 1 ricovero con diagn. di MIELOMA	44	12.661	11.002	2.546	15.207
Deceduti (01/01/02 - 30/06/02), tutte le età; almeno 1 ricovero con diagn. di MIELOMA Anno precedente la morte	5	19.079	18.308	4.072	23.151

Tabella 24 bis. Assistiti affetti da Leucemia Acuta e Mieloma Multiplo. Mortalità e distribuzione per età e sesso

DESCRIZIONE	N. Totale	Tot Decessi Anni 2001 2002	N° Decessi in Ospedale	N° Maschi	N° Femmine	Età Media Maschi	Età Media Femmine
Assistiti, tutte le età; almeno 1 ricovero con diagn. di LEUCEMIA	29	14	8	14	15	50.0	54.5
Deceduti (01/01/02 - 30/06/02), tutte le età; almeno 1 ricovero con diagn. di LEUCEMIA Anno precedente la morte	2			2	0	65.7	
Assistiti, tutte le età; almeno 1 ricovero con diagn. di MIELOMA	44	21	9	19	25	72.5	71.8
Deceduti (01/01/02 - 30/06/02), tutte le età; almeno 1 ricovero con diagn. di MIELOMA Anno precedente la morte	5			3	2	73.4	75.5

Come atteso, i numeri esaminati sono molto piccoli ma identificano delle popolazioni di pazienti estremamente gravi e gravosi. Tutti i dati riportati di costo, di prevalenza ed anche quelli relativi all'età dei soggetti sono coerenti con il tipo di patologia esaminata e confermano la validità del modello e della analisi anche quando piccoli gruppi vengono esaminati.

Questa ultima analisi introduce l'ultimo scenario sviluppato a titolo esemplificativo. Nel periodo che precede il decesso (un anno, nell'esempio qui riportato) i costi sanitari diretti, sempre, aumentano.

L'entità del fenomeno è illustrato nella tabella 25 dove mostriamo il semplice confronto tra alcune variabile estratte dal file ASOLO con le 165.128 osservazioni relative a tutta la popolazione assistita e le corrispondenti variabili estratte dal file ASOLODEC con le sole 890 osservazioni relative ai soli assistiti per i quali è stato possibile ricostruire i 12 mesi che precedevano il decesso.

Tabella 25. Confronto di alcune variabili, utilizzate come indicatori di gravità, tra tutti gli assistiti della ASL esaminata e la popolazione deceduta dopo 12 mesi di osservazione.

	Assistiti tutte le età	**Deceduti anni 2001-2002**
Periodo considerato	**Gennaio 2001-Giugno 2002**	**12 mesi precedenti il decesso**
N° soggetti	**165.128**	**890**
Costo medio per Paziente (ricoveri) Lit x 1000	**1.081**	**6.225**
Costo medio per Paziente (farmaci) Lit x 1000	**552**	**711**
Costo medio per Paziente (Totale) Lit x 1000	**1.633**	**6.936.**
Media N. Totale Ricoveri	**0,27**	**2,08**
Media N. Giorni Degenza complessiva	**1,78**	**24,72**
Media giornate Degenza in Lungodeg. e Riabil	**0,19**	**4,39**
Media N. diversi DRG	**0,25**	**1,68**
Media N. totale Prescrizioni	**10,48**	**25,62**
Media N. max Prescr. dello stesso farmaco	**4,49**	**7,65**
Media N. max Prescr. farmaci diversi al mese	**2,11**	**4,25**

Pur avendo considerato per la popolazione di riferimento (N° 165.128) un periodo di 18 mesi, la popolazione deceduta presenta, per tutte le variabili, valori significativamente e marcatamente più elevati. Al di là dei diversi commenti, che ciascuno dei dati presentati suggerisce, l'aspetto di rilievo che ci preme sottolineare è che tutte le variabili esaminate effettivamente correlano simultaneamente con una maggiore

gravità clinica, maggiore carico assistenziale e maggiore assorbimento di risorse. Il significato e la validazione di dette variabili ne risulta confermata.

6. Conclusioni

Il sistema informativo socio-sanitario si caratterizza per l'esistenza di numerosi flussi "correnti" generati a fini per lo più amministrativi, spesso in risposta a specifiche normative. Il contenuto di tali flussi differisce per tipologia di informazioni rilevate, modalità di classificazione e codifica delle stesse, sistemi di controllo di qualità esistenti, segmenti del processo sociosanitario da cui vengono generati, output-indicatori prodotti, modalità e livelli di archiviazione ed utilizzo degli stessi flussi.

L'enorme sviluppo della tecnologia informatica consente oggi di utilizzare i grandi *data base* amministrativi esistenti non solo a meri fini amministrativi ma anche a fini di sorveglianza di sanità pubblica di monitoraggio e valutazione dei servizi sanitari. La copertura del territorio e la tempestività del ritorno informativo così ottenuti rendono possibile rispondere in modo appropriato, sostenibile ed equo ai fabbisogni informativi generati dalle nuove politiche di sanità pubblica e di razionalizzazione del sistema sanitario.

L'utilizzabilità dei flussi informativi correnti a fini di sorveglianza, pianificazione e valutazione degli interventi di sanità pubblica e di organizzazione dei servizi assistenziali è quindi associata principalmente alla loro integrabilità sull'utente ed alla qualità del loro contenuto.

L'integrabilità dei sistemi informativi è a sua volta associata alla standardizzazione dei flussi ed alla compatibilità dei sistemi informatici esistenti; la qualità del flusso informativo è associata alla accuratezza ed alla completezza dei dati.

Il concetto di "integrazione", a prescindere dal contesto del sistema informativo, consiste nella definizione di una più completa e coordinata "entità" tramite aggiunta o configurazione di diverse parti o elementi, incorporati in una cornice o unità più ampia, creando un tutto funzionante ed eliminando le ridondanze.

Tuttavia, spesso le informazioni sono difficilmente collegabili perché

- sono state raccolte in formati incompatibili, usando differenti definizioni, identificatori personali, sistemi di classificazione o strategie di campionamento;
- non vi è una infrastruttura di comunicazioni attraverso la quale i dati possono essere ottenuti, aggregati, trasferiti;
- politiche, normative e consuetudini organizzative, sebbene necessarie per la protezione della confidenzialità, possono impedire senza motivo l'accesso e la condivisione dell'informazione.

L'integrazione delle informazioni per la sanità pubblica e per la valutazione dei servizi sanitari consiste nell'unire questi frammenti di informazione combinando o collegando assieme i sistemi di dati che

contengono l'informazione. Perché l'integrazione sia possibile è necessario tuttavia che siano verificate alcune condizioni, in particolare:

- l'esistenza di standard uniformi di dati;
- l'esistenza di una rete di comunicazioni;
- la possibilità di accedere ai dati esistenti in un'ottica di condivisione all'interno di un quadro normativo che garantisca la sicurezza e la privacy.

Lo sviluppo di standard di dati richiede l'accordo di fornitori e utilizzatori su:

- comuni definizioni degli elementi e dei termini dei dati;
- comuni sistemi di classificazione;
- protocolli compatibili di telecomunicazione;
- specificazioni tecniche che permettono ai differenti sistemi di essere confrontati e collegati.

In termini generali una infrastruttura elettronica di comunicazioni include equipaggiamento, protocolli e software che consentono agli utilizzatori di connettersi e scambiare dati con altri utilizzatori attraverso reti locali o più ampie (LAN - local area network - o WAN - wide area network). Una componente essenziale di ogni rete di comunicazioni è l'aderenza a un insieme di protocolli e standard che governano come sono fatte le connessioni tra membri delle reti locali e ampie.

Un sistema integrato di informazione e sorveglianza di sanità pubblica si fonda su una serie di accordi tra coloro che hanno i dati e coloro che utilizzano i dati, per superare le barriere normative, organizzative e culturali che incoraggiano l'indipendenza piuttosto che la cooperazione. Questi accordi forniscono la base di un flusso efficiente di dati agli appropriati utilizzatori minimizzando il carico della raccolta, proteggendo la confidenzialità e massimizzando l'utilità analitica.

I dati necessari per lo svolgimento delle funzioni di Sanità Pubblica e di valutazione dei servizi sanitari vengono raggruppati in sette categorie di informazione[*]:

1) report di eventi di salute riguardanti gli individui (per es. Sorveglianza delle malattie trasmissibili soggette a notifica; Sorveglianza delle malattie non-trasmissibili);
2) statistiche vitali della popolazione (nascite e morti);
3) informazione sullo stato di salute, comportamenti a rischio ed esperienze di popolazione (per es. regolari inchieste su dieta, abitudine al fumo, uso della cintura di sicurezza);
4) informazioni sulla potenziale esposizione ad agenti ambientali (per es. inquinanti di aria, acqua, suolo; igiene degli alimenti; diffusione e abitudini dei vettori di malattie);

[*] Il Rapporto Katz dal titolo "Integrare l'informazione di Sanità Pubblica ed i Sistemi di Sorveglianza" del 1995 è il documento fondamentale nel processo di creazione di un tale sistema perseguito dal CDC negli Stati Uniti a partire dal 1993. Dopo tale rapporto venne istituito il Consiglio per l'informazione ed i sistemi di sorveglianza di sanità pubblica (HISSB) di cui il relatore del Rapporto ne è l'attuale presidente.

5) informazione sui programmi di Sanità Pubblica esistenti (per es. Copertura della popolazione, costo);
6) informazioni utili per la sanità pubblica, ma ottenuta da organizzazioni non direttamente coinvolte nell'attività di Sanità Pubblica (per es. Servizi demografici, ARPA, Polizia, Camere di Commercio);
7) informazioni sui servizi sanitari e sull'impatto dei servizi sanitari sulla salute.

La sperimentazione condotta con lo studio SIPA oltre al valore aggiunto della costituzione di un gruppo permanente di tecnici della sanità sia produttori che utilizzatori dei dati, ha consentito di evidenziare la notevole variabilità del contenuto,del livello di informatizzazione e di integrazione dei flussi esistenti nelle Aziende che hanno partecipato al progetto, selezionate tra l'altro in base al criterio di una maggior grado di informatizzazione e di qualità dei flussi informativi sanitari esistenti nelle stesse.

Considerati i possibili livelli di utilizzo dei sistemi informativi integrati a fini di governo del sistema sanitario, la sperimentazione ha riguardato essenzialmente due livelli il livello della Regione Veneto ed il livello dell'Azienda.

Dal punto di vista del livello regionale sono stati utilizzati i due principali flussi informativi sanitari che rispondono a requisiti di completezza e robustezza e cioè la scheda di dimissione ospedaliera (SDO) e il registro delle cause di morte (SKM).

Dal punto di vista aziendale è stato possibile utilizzare, oltre a quelli suddetti, anche altri flussi informativi e cioè l'anagrafe sanitaria, la farmaceutica territoriale e la specialistica ambulatoriale.

Dal punto di vista operativo sono state messe a punto e sperimentate varie procedure di linkage, fra le quali la *ReClust*, descritta nel capitolo 5, è risultata particolarmente efficace.

L'integrazione degli archivi SDO e SKM a livello regionale ha consentito di descrivere – misurare, in un particolare ambito di ricerca (gli interventi di rivascolarizzazione cardiaca), non solo le prestazioni eseguite, come normalmente avviene nelle analisi condotte sui sistemi sanitari, ma anche e soprattutto i casi trattati. Sono stati valutati i carichi assistenziali per gli interventi in questione ed anche gli esiti in termini di mortalità precoce e di riospedalizzazione.

Tale approccio assolutamente innovativo sembra adeguato, se opportunamente sviluppato, ai fabbisogni informativi generati dall'avvio dei processi di accreditamento e di *risk management* su scala regionale. L'analisi della variabilità territoriale e temporale dei costi e degli esiti degli interventi selezionati è indubbiamente utile per migliorare l'appropriatezza degli interventi e l'equità nella allocazione delle risorse.

A livello aziendale l'utilizzo, oltre ai flussi considerati a livello regionale, anche dell'imponente flusso delle prescrizioni farmaceutiche ha consentito di individuare profili di pazienti omogenei per ricorso ai servizi sanitari e per esito e di valutare, per ciascun gruppo, i costi diretti dell'assistenza. Questo risultato

rappresenta un enorme valore aggiunto del progetto: è assolutamente evidente che l'estensione di tale modello di analisi ad altre realtà potrebbe consentire di valutare i costi dell'assistenza in relazione a specifiche tipologie di pazienti (si pensi ai costi associati all'invecchiamento) garantendo una maggior equità nell'allocazione delle risorse finanziarie.

Altro elemento di rilievo dei risultati della sperimentazione a livello aziendale utilizzabile a fini di controllo interno risulta essere quello della relazione osservata tra costi e gravità clinica con evidenti implicazioni sulla possibilità di fornire informazioni adeguate e tempestive in grado di favorire la valutazione della appropriatezza clinica delle prestazioni erogate.

Pur in presenza di forti difficoltà tecnico - organizzative e di scarsa diffusione di conoscenze e competenze sull'utilizzo dei flussi informativi correnti al fine della pianificazione dell'assistenza della prevenzione sanitaria, si può infine affermare che i risultati del progetto SIPA possono rappresentare un importante contributo al miglioramento ed alla diffusione delle conoscenze necessarie a riorientare i clinici e gli amministratori alla gestione dei pazienti, i primi, e del sistema sanitario, i secondi, in base alle prove di efficacia degli interventi da porre in essere.

7. Bibliografia

Bibliografia essenziale

1. Raschetti R.
Editoriale
BEN - Notiziario ISS - Vol.16 - n.1 Gennaio 2003

2. Carlotta Sacerdote1, Marco Dalmasso2, Giovannino Ciccone1, Moreno Demaria3 e Roberto Gnavi
Utilizzo di differenti chiavi identificative di soggetti presenti in diversi archivi
BEN - Notiziario ISS - Vol.16 - n.1 Gennaio 2003

3. Lepore V, D'Ettorre A, Valerio M, Corrado D, De Camillis P, Romero M, Scurti V, Monesi G, Ferrarese A, Monesi L, Mollo F, Tognoni G.
Dalla farmacoepidemiologia all'epidemiologia dell'assistenza.
Giorn Ital Farm Clin 2002; 16: 102-7.

4. Monesi L, Rosso Fernandez C, D'Ettorre A, Corrado D, Lepore V, Sasso E, Tognoni G, Mollo F, Monesi G, Ferrarese A.
I database amministrativi come fonti di ricerca epidemiologica: il percorso clinico- assistenziale del diabete mellito.
Giorn Ital Farm Clin 2002; 16: 158-64.

5. Lepore V, D'Ettorre A, Valerio M, Corrado D, De Camillis P, Pellegrini F, Romero M, Scurti V, Monesi L, Ferrarese A, Mollo F, Monesi G.
La variabilità in Medicina Generale. Il caso dei pazienti «gravi-gravosi»
Giorn Ital Farm Clin 2002: 16, 220-5.

6. Blakely T, Salmond C.
Probabilistic record linkage and a method to calculate the positive predictive value.
Int J Epidemiol. 2002 Dec;31(6):1246-52. Review.

7. Agabiti N, Ancona C, Forastiere F, Arca M, Perucci CA.
Evaluating outcomes of hospital care following coronary artery bypass surgery in Rome, Italy.
Eur J Cardiothorac Surg. 2003 Apr;23(4):599-606.

8. Schuerenberg BK.
Electronic records find long-term use.
Health Data Manag. 2003 Feb;11(2):112-4.

9. Oden A, Fahlen M.
Oral anticoagulation and risk of death: a medical record linkage study.
BMJ. 2002 Nov 9;325(7372):1073-5.

10. Pinnelli A.
Record linkage in the study of infant mortality: some aspects concerning data quality
Statistica. 1984 Oct-Dec;44(4):675-86.

11. Gomatam S, Carter R, Ariet M, Mitchell G.
An empirical comparison of record linkage procedures.
Stat Med. 2002 May 30;21(10):1485-96.

12. Harvey JN, Craney L, Kelly D.
Estimation of the prevalence of diagnosed diabetes from primary care and secondary care source data: comparison of record linkage with capture-recapture analysis.
J Epidemiol Community Health. 2002 Jan;56(1):18-23.

13. Hammar N, Alfredsson L, Rosen M, Spetz CL, Kahan T, Ysberg AS.
A national record linkage to study acute myocardial infarction incidence and case fatality in Sweden.
Int J Epidemiol. 2001 Oct;30 Suppl 1:S30-4.

14. Cook LJ, Olson LM, Dean JM.
Probabilistic record linkage: relationships between file sizes, identifiers and match weights.
Methods Inf Med. 2001 Jul;40(3):196-203.

15. Dal Maso L, Zanetti R, Orengo MA, Tagliabue G, Guzzinati S, Cavallieri F, Serventi L, Mangone L, Ferretti S, Milandri C, Pannelli F, Balzi D, Tonini G, Gafa L, Rezza G, Franceschi S.
Methodological issues and first results of a record linkage between AIDS and Cancer Registries in Italy
Epidemiol Prev. 2000 May-Jun;24(3):109-16.

16. Morgan CL, Currie CJ, Peters JR.
Relationship between diabetes and mortality: a population study using record linkage.
Diabetes Care. 2000 Aug;23(8):1103-7.

17. Kelman C, Smith L.
It's time: record linkage--the vision and the reality.
Aust N Z J Public Health. 2000 Feb;24(1):100-1.

18. Bernillon P, Lievre L, Pillonel J, Laporte A, Costagliola D.
Record-linkage between two anonymous databases for a capture-recapture estimation of underreporting of AIDS cases: France 1990-1993. The Clinical Epidemiology Group from Centres d'Information et de Soins de l'Immunodeficience Humaine.
Int J Epidemiol. 2000 Feb;29(1):168-74.

19. Brameld KJ, Holman CD, Bass AJ, Codde JP, Rouse IL.
Hospitalisation of the elderly during the last year of life: an application of record linkage in Western Australia 1985-1994.
J Epidemiol Community Health. 1998 Nov;52(11):740-4.

20. Evans JM, MacDonald TM.
Record-linkage for pharmacovigilance in Scotland.
Br J Clin Pharmacol. 1999 Jan;47(1):105-10.

21. WHO Regional Office for Europe
Health for all targets: the health policy for Europe
European Health for All Series n. 4, 1993, Copenhagen

8. Allegati

CODICI INCLUSI NELLA ANALISI (36xx in uno qualsiasi dei campi procedure – interventi)

ICD-9-CM
CLASSIFICAZIONE DELLE MALATTIE E DEI TRAUMATISMI E DEGLI INTERVENTI CHIRURGICI E DELLE PROCEDURE DIAGNOSTICHE E TERAPEUTICHE

36 Interventi sui vasi del cuore
Incl.: come approccio chirurgico:
sternotomia (mediana) (trasversa)
toracotomia
Codificare anche bypass cardiopolmonare, eventualmente eseguito [circolazione extracorporea] [macchina cuorepolmone] (39.61)

36.0 Rimozione di ostruzione dell'arteria coronarica ed inserzione di stent

36.01 Angioplastica coronarica percutanea transluminale di vaso singolo [PTCA] o aterectomia coronarica senza menzione di agente trombolitico
Codificare anche eventuale inserzione di stent coronarici (36.06)
Angioplastica mediante palloncino di arteria coronarica
Angioplastica percutanea coronarica SAI
PTCA SAI
Aterectomia coronarica

Escl.: angioplastica coronarica percutanea transluminale su vaso multiplo [PTCA] o aterectomia coronarica effettuata durante lo stesso intervento (36.05)

36.02 Angioplastica coronarica percutanea transluminale di vaso singolo [PTCA] o arterectomia coronarica con menzione di agente trombolitico
Codificare anche eventuale inserzione di stent coronarici (36.06)
Angioplastica mediante palloncino di arteria coronarica con infusione di agente trombolitico [streptochinasi]
Aterectomia coronarica

Escl.: angioplastica coronarica percutanea transluminale su vaso multiplo effettuata durante lo stesso intervento [PTCA] o aterectomia coronarica eseguita durante lo stesso intervento(36.05)
PTCA su singolo vaso o aterectomia senza menzione di agente trombolitico (36.01)

36.03 Angioplastica dell'arteria coronarica a torace aperto
Codificare anche eventuale inserzione di stent coronarici (36.06)
Sull'arteria coronarica:
endoarteriectomia (con innesto a patch)
tromboendoarteriectomia (con innesto a patch)
Chirurgia aperta per attenuazione diretta di ostruzione dell'arteria coronarica

Escl.: angioplastica con bypass di arteria coronarica (36.10-36.19)

36.04 Infusione trombolitica nell'arteria coronarica
Infusione mediante iniezione, infusione o cateterismo diretto nell'arteria coronarica.
Infusione di enzimi
Inibitore piastrinico

Escl.: somministrazione per via intravenosa [IV Infusione] (99.29)
infusione associata con qualunque altra procedura codificata in 36.02, 36.03

36.05 Angioplastica coronarica percutanea transluminale su vaso multiplo [PTCA] o aterectomia coronarica eseguita durante lo stesso intervento, con o senza menzione di agente trombolitico
Angioplastica mediante palloncino di arterie coronariche multiple
Aterectomia coronarica

Codificare anche eventuale infusione di agente trombolitico nell'arteria coronarica (36.04)

Codificare anche eventuale inserzione di stent coronarici (36.06)

Escl.: PTCA su singolo vaso o aterectomia coronarica
senza menzione di agente trombolitico (36.01)
con menzione di agente trombolitico(36.02)

36.06 Inserzione di stent nell'arteria coronarica
Innesto di stent

Codificare anche eventuale angioplastica coronarica a torace aperto (36.03)

Codificare anche eventuale angioplastica coronarica percutanea transluminale [PTCA] o aterectomia coronarica (36.01, 36.02, 36.05)

36.09 Altra rimozione di ostruzione dell'arteria coronarica
Angioplastica coronarica SAI

Escl.: rimozione mediante angioplastica a torace aperto (36.03)
rimozione mediante angioplastica percutanea transluminale coronarica [PTCA] o aterectomia coronarica (36.01-36.02, 36.05)

36.1 Bypass per rivascolarizzazione cardiaca

Codificare anche bypass cardiopolmonare [circolazione extracorporea] [macchina cuore-polmone] (39.61)

36.10 Bypass aortocoronarico per rivascolarizzazione cardiaca, SAI
Interventi con supporto per catetere, protesi o innesto venoso:
rivascolarizzazione diretta:
cardiaca
coronarica
muscolocardiaco
miocardio
rivascolarizzazione del cuore

SAI

36.11 Bypass aortocoronarico di una arteria coronarica

36.12 Bypass aortocoronarico di due arterie coronariche

36.13 Bypass aortocoronarico di tre arterie coronariche

36.14 Bypass aortocoronarico di quattro o più arterie coronariche

36.15 Bypass singolo mammaria interna-arteria coronarica
Anastomosi (singola):
arteria mammaria ad arteria coronaria
arteria toracica ad arteria coronaria

36.16 Bypass doppio mammaria interna-arteria coronarica
Anastomosi doppia:
arteria mammaria ad arteria coronaria
arteria toracica ad arteria coronaria

36.17 Bypass dell'arteria coronaria addominale
Anastomosi:
arteria gastroepiploica e arteria coronaria

36.19 Altro bypass per rivascolarizzazione cardiaca

36.2 Rivascolarizzazione cardiaca mediante innesto arterioso
Impianto di:
ramo dell'aorta [rami aortici ascendenti] nel muscolo cardiaco
vasi sanguigni nel miocardio
arteria mammaria interna [arteria toracica interna] nel:
muscolo cardiaco
miocardio
ventricolo
parete ventricolare
Rivascolarizzazione cardiaca indiretta SAI

36.3 Altra rivascolarizzazione cardiaca
Abrasione dell'epicardio
Cardio-omentopessi
Poudrage intrapericardico
Innesto miocardico: tessuto adiposo mediastinico, omento, muscoli pettorali

36.9 Altri interventi sui vasi del cuore

Codificare anche bypass cardiopolmonare [circolazione extracorporea] [macchina cuore-polmone] (39.61)

36.91 Riparazione di aneurisma dei vasi coronarici

36.99 Altri interventi sui vasi del cuore
Esplorazione, incisione, legatura dell'arteria coronarica
Riparazione di fistola arteriovenosa

Miniglossario. Parole chiave in Sanità Pubblica.

Su alcune parole chiave ricorrenti nell'ambito della salute Pubblico è indispensabile vi sia accordo sul significato dei termini, per evitare che la stessa parola significhi cose diverse a diverse persone, privandola di ogni utilità in un contesto scientifico.

Epidemiologia è lo studio della distribuzione e dei determinanti di stati o eventi correlati alla salute in specifiche popolazioni, e l'applicazione di questo studio al controllo di problemi di salute. Risponde alle domande: chi si ammala di che cosa, dove, quando, come e perché? Un epidemiologo identifica e previene le malattie in una data popolazione, un clinico identifica e tratta le malattie in un individuo. L'epidemiologia si basa su due assunti fondamentali: a. Le malattie non insorgono per caso; b. Le malattie non sono distribuite casualmente nella popolazione; la loro distribuzione indica qualcosa riguardo a come e perché il processo patologico è insorto.

Rischio è la probabilità che un individuo sviluppi una malattia in un periodo di tempo determinato.Il suo valore varia tra 0 (o 0%), l'evento non si verifica mai, e 1 (o 100%), l'evento si verifica sempre.

Progetto è uno sforzo intensivo, limitato nel tempo, con un insieme ben definito di risultati finali, abbastanza complesso da essere suddivisibile in sottocompiti che richiedono coordinamento e controllo in termini di esecuzione, tempo e costo. Non è particolarmente rilevante per la sua definizione che sia su piccola o larga scala o su breve o lungo periodo. Le fasi in cui può essere distinto sono: ideazione, selezione, pianificazione, temporalizzazione, monitoraggio, controllo, valutazione e terminazione. (Ha poco senso l'espressione " Progetto obiettivo", perché l'aggiunta del termine obiettivo è ridondante e perché non esiste un Progetto non-obiettivo; ha senso invece differenziare tra progetti sanitari, sociali e socio-sanitari).

Programma è un gruppo di progetti simili.

Gestione è una serie di tappe sistematiche, sequenziali, o parzialmente sovrapponentisi, dirette al raggiungimento degli scopi e degli obiettivi di un'organizzazione; chi gestisce organizza le attività, assegna ad esse il personale, e accerta che siano effettuate come pianificato.

Medicina preventiva è la branca della medicina che si concentra nel mantenere sana la popolazione, evitando l'insorgenza delle malattie e promovendo lo stato di benessere fisico ed emotivo.