# *Quando i big data possono diventare "scomodi"*

**Dalla produzione e gestione dei dati alle scelte politiche**

**Il termine "big data" è diventato di gran moda. Sulla base dell'esperienza di chi ha visto già molto tempo fa nei dati amministrativi la potenzialità di trarre da questi nuova conoscenza, la nuova attenzione al tema è frutto di qualche novità particolare o siamo di fronte alla riproposizione di qualcosa che ben si conosce?**

Dati amministrativi? Non so cosa siano i dati amministrativi. Sistemi informativi con vari contenuti possono essere utilizzati per scopi diversi, anche di carattere amministrativo, ma queste sono caratteristiche dell'utilizzazione, non dei dati per sé. Tuttavia le finalità per le quali viene disegnato e gestito un sistema informativo, e il contesto sociale, culturale e istituzionale determinano in modo rilevante la riproducibilità e la validità dei dati. Ancor di più: la variabilità temporale e geografica degli utilizzi di un sistema informativo si associano sempre a variabilità della riproducibilità e della validità dei dati. Le cosiddette schede di dimissione ospedaliera (sdo), oggi molto utilizzate per la remunerazione prospettica delle prestazioni di assistenza ospedaliera, erano certamente già presenti in Italia negli anni Settanta, prima in modalità campionaria e poi sistematica, ben prima che negli anni Novanta iniziasse il loro uso a fini "amministrativi". Ma la remunerazione dei soggetti erogatori di assistenza ospedaliera è stata introdotta in modo eterogeneo nel tempo, tra diverse regioni, per diverse tipologie di aziende, pubbliche e private; ancor oggi molti ospedali, soprattutto pubblici e nel meridione, non sono finanziati sulla base del valore della loro produzione che viene calcolato attenendosi alle sdo. Questa forte eterogeneità geografica, istituzionale e temporale dell'uso di questi dati determina una grande, ma non sempre attentamente considerata, eterogeneità della riproducibilità e della validità dei dati che riguardano i ricoveri ospedalieri, non solo in Italia peraltro. Quando il Programma nazionale esiti (Pne) ha iniziato la propria attività, l'estrema variabilità della "qualità" dei dati delle sdo ha proposto una sfida metodologica e operativa molto impegnativa.

**Alcuni esempi?**

Negli anni Novanta in Italia non solo non era possibile ma non era nemmeno progettata la interconnesione sistematica a livello nazionale tra le sdo e le schede di morte; solo alcune regioni avevano anticipato i tempi con propri sistemi informativi ospedalieri e con i cosiddetti registri nominativi delle cause di morte, interconnetendoli. Quindi a livello nazionale era possibile stimare esclusivamente la mortalità intraospedaliera. All'inizio del nuovo secolo le stime di mortalità intraospedaliera dopo un episodio di infarto miocardico acuto (ima) davano valori relativamente omogenei nel nord e centro Italia, attorno a circa il 10%, con un rapido progressivo decremento delle stime nel sud, per giungere a straordinari valori inferiori al 4% in Sicilia. Ricordo, oggi con un sorriso, le fantasiose interpretazioni di questi risultati da parte di alcuni, anche illustri, non solo clinici ma anche epidemiologi: la dimostrazione del potente ruolo protettivo della dieta mediterranea (ipotesi "etiologica": fattore protettivo per letalità dopo ima). Oppure la clamorosa sconfessione dell'opinione molto diffusa della bassa qualità dell'assistenza ospedaliera nel meridione. I "dati" empirici mostravano chiaramente il ruolo protettivo della dieta mediterranea e/o l'ottima efficacia del trattamento degli episodi di ima negli ospedali del sud. Ma il controllo dei dati sdo consentì di tener conto di un'altra ipotesi valutativa: per fattori culturali e sociali in molte aree del meridione le famiglie ritenevano, e in molti casi oggi ancora ritengono, disonorevole la morte di un congiunto in ospedale. Quindi, anche per cinici fenomeni speculativi altresì connessi a organizzazioni mafiose, le persone decedute in ospedale venivano dimesse come vive e contro il parere dei sanitari, e trasportate a casa dove veniva certificato il decesso. L'interconnessione tra sdo e registri di morte permi-

*Intervista a*

**Carlo Alberto Perucci**

*già direttore del Programma nazionale esiti*

se quindi di svelare il fenomeno e migliorare la validità delle stime della mortalità intraospedaliera dopo infarto che nel sud risultava in media simile a quella del centro e del nord. Ma questo problema di validità dei dati, non è uno dei soliti difetti italiani: credo che ancora oggi l'Organizzazione per la cooperazione e lo sviluppo economico (Ocse) stimi la mortalità intraospedaliera post-infarto per tutti i paesi europei, poiché in alcuni non sarebbe possibile l'interconnessione sistematica tra sistemi informativi ospedalieri e i registri di morte. Gli stili di cura inpatient e outpatient sono eterogenei nel tempo e da un paese a un altro, e il confronto della mortalità intraospedaliera è affetto da forti distorsioni; tuttavia molti esperti e molti politici usano talora in modo acritico le stime dell'Ocse senza tener conto di questi rilevanti limiti di validità.

Ancor oggi un'analisi "esplorativa", così cara a certi maneggioni di big data, consentirebbe di osservare alcuni singolari fenomeni. Ad esempio, negli ospedali italiani, ipertensione, diabete, broncopneumopatia cronica ostruttiva e altre patologie croniche sembrano essere fattori protettivi nella mortalità a 30 giorni dopo ima, ancor più protettivo sembrerebbe il fumo di tabacco. Artefatti… In realtà questi risultati dipendono esclusivamente da un noto fenomeno di codifica "competitiva", quando patologie croniche, meno rilevanti in un contesto di misurazione per "intensità assistenziale", vengono registrate e codificate su sdo solo nei casi di minore gravità.

**Nulla di nuovo, dunque…**

Certamente oggi si presentano maggiori opportunità per disponibilità di tecnologie informatiche, sia hard sia soft, che consentono in tempi brevi il trattamento di grandi volumi di dati e lo sviluppo di metodi statistici avanzati che ne possono permettere analisi molto "potenti". Tuttavia, questa grande disponibilità di dati e tecniche di analisi aumenta i problemi di riproducibilità e di validità non solo dei dati ma, soprattutto, delle stime e delle loro interpretazioni.

Probabilmente torna il grande dilemma metodologico tra approcci induttivi e metodi ipotetico-deduttivi. Personalmente ritengo importante formulare ipotesi, etiologiche e valutative, basate sulle conoscenze disponibili, da sottoporre a processi di falsificazione utilizzando metodologie rigorose e trasparenti. Attenzione, basarsi sulle conoscenze disponibili non significa assolutamente considerare solo ipotesi plausibili, semmai avere il coraggio di sottoporre a valutazione anche o soprattutto ipotesi altamente improbabili, contrarie al senso comune dominante e sgradite alla politica e cultura del momento. Tanto più sarà big la quantità dei dati disponibili, tanto maggiori saranno i problemi metodologici nella loro produzione, analisi e utilizzazione, e tanto maggiori dovranno essere il rigore nella formulazione a priori delle ipotesi e la trasparenza nella discussione e interpretazione delle informazioni derivate, dei loro limiti, dei potenziali errori casuali e sistematici.

> *"L'aumento delle conoscenze disponibili, anche attraverso i big data, non è destinato a ridurre l'incertezza, semmai a renderci maggiormente consapevoli di tutto quello che non conosciamo."*

Nessun disegno di studio, su nessuna ipotesi etiologica o valutativa, è privo di errori sistematici e casuali. I risultati di grandi trial clinici randomizzati (rct) sono affetti da distorsioni (soprattutto, ma non solo, per selezione e modificazione delle misure di effetto), come lo sono quelli dei grandi studi osservazionali (soprattutto per confondimento). Forse, tuttavia, qualcuno pensa, per esempio nel campo della valutazione di efficacia dei trattamenti sanitari, di superare il rigore (apparente) metodologico dei trial, per affermare una pratica corrente di studi osservazionali usando soprattutto big data, molto "quick and dirty", perché ritenuti più manipolabili e adattabili a specifiche finalità commerciali. La risposta tuttavia non è la difesa a oltranza degli studi sperimentali e del loro (apparente) rigore ma l'affermazione, negli studi osservazionali che usano big data, di metodi scientifici rigorosi, altrettanto o più complessi e costosi di quelli degli rct.

Anni fa, nella fase sperimentale di Pne, gli esperti di una grande società di consulenza, che fornisce al Ministero della salute servizi di sistemi informativi, chiesero di produrre e fornire per ciascun indicatore di esito modelli di risk adjustment e di controllo del confondimento da inserire nella base di dati del Nuovo sistema informativo sanitario (Nsis), in modo da produrre "automaticamente" stime di esito aggiustate che tenessero conto dei fattori confondenti di ciascun indicatore. Sulle prime parve straordinario come i consulenti avessero (finalmente) colto l'idea del confondimento nei confronti – concetto spesso assai ostico da capire negli ambienti professionali sanitari, per non dire in quelli politici. Ma fu difficile spiegare come non fosse possibile produrre procedure di risk adjustment standard, ma fosse necessario sviluppare modelli di risk adjustment specifici di ciascun confronto, di ciascun periodo temporale, valutandone la validità e le potenziali distorsioni legate, soprattutto, alla eterogeneità temporale e geografica della validità dei dati (big) utilizzati, e agli effetti sulla precisione delle stime aggiustate.

A proposito di dati "aggiustati", mi si permetta un episodio. A una Commissione del Senato, tra pochi senatori, alcuni annoiati, altri interessati, vi è una rapida presentazione di Pne; con diapositive sintetiche che riportano stime comparative tra ospedali, si sottolinea che i risultati sono aggiustati. Una senatrice, furibonda, chiede la parola: "Come vi permettete di presentare al Senato delle Repubblica dati aggiustati!". Molti altri aneddoti potrebbero essere raccontati sulla grande difficoltà, da parte dei politici e soprattutto dei giornalisti, a interpretare correttamente gli errori casuali, le "magiche" p e gli effetti del caso.

Emergono atteggiamenti culturali e politici talora contrastanti: sottovalutazione degli aspetti di metodo; rifiuto di considerare errori casuali e sistematici o, viceversa, enfatizzazione opportunistica di errori; sopravalutazione degli aspetti informatici e gestionali dei dati. In generale si riproduce il noto paradosso nei sistemi sanitari: quando mancano informazioni per decidere, i decisori sostengono di essere costretti a decidere in assenza di informazioni utili; quando invece le informazioni sono disponibili, e talora abbastanza "forti", i decisori preferiscono non considerarle nemmeno, per potere decidere come vogliono. Quindi istituzioni, politici e decisori sono pronti, a parole, a creare e finanziare grandi sistemi informativi, basi di dati sempre più grandi, infrastrutture informatiche costosissime e faraoniche, salvo poi non sapere o volere utilizzare le "evidenze" prodotte dalla loro utilizzazione e analisi, soprattutto quando limitano i gradi di libertà delle decisioni. Di questo processo sono spesso complici esperti, anche epidemiologi, pronti a produrre risultati, analisi, informazioni funzionali, a giustificare decisioni (politiche) già prese. E guai a quelli che formulano ipotesi e producono conoscenze considerate contrarie, non compatibili, comunque fastidiose, critiche con gli orientamenti decisionali dei potenti di turno.

L'aumento delle conoscenze disponibili, anche attraverso l'uso di big data, non è destinato a ridurre l'incertezza, semmai a renderci maggiormente consapevoli di tutto quello che non conosciamo. In altre parole la produzione di enormi volumi di dati e di informazioni non semplifica gli scenari decisionali, anzi costringe i decisori ad aver a che fare con livelli più complessi di incertezza e a essere esposti a maggiori e più documentate critiche.

Mi permetto di dire che il problema non è quanto big siano i dati, ma quanto sono grandi, forti e oneste l'autonomia, l'indipendenza e l'integrità di coloro che li progettano, gestiscono, analizzano e interpretano. *a p.28* →

*da p.27* →

**Con i big data si fa spesso riferimento (anche) a flussi di dati non strutturati generati dai social. I limiti legati alla qualità del dato, alla definizione di protocolli affidabili rispetto a quesiti clinici precisi, alla difficoltà di gestire la privacy sono simili a quanto già presente negli attuali flussi amministrativi?**

Personalmente rifiuto la definizione di "flussi amministrativi": non ho nessun pregiudizio nell'uso di questi dati, a condizione che siano utilizzati, analizzati e interpretati tenendo conto, con metodi rigorosi, delle loro caratteristiche di riproducibilità e validità.

In quanto alla questione della cosiddetta privacy, mi si consenta una premessa. Non solo in Italia certo, ma soprattutto nella degradata e arretrata condizione delle nostre burocrazie, nella sua attuale modalità di funzionamento, la cosiddetta Autorità garante della privacy potrebbe essere da alcuni considerata, alla stregua del fumo di tabacco, dell'inquinamento ambientale e di alcune epidemie di malattie trasmissibili, uno dei principali fattori di rischio per la salute della popolazione con un potente fenomeno di modificazione/moltiplicazione di effetto con le altre burocrazie. Ce ne sarebbero molti di buoni motivi. Un esempio: attorno al 2010, anche sulla base delle proposte di Agenas, allora diretta da un innovatore intelligente come Fulvio Moirano, il Ministero della salute, un po' controvoglia in alcune sue burocrazie, diede nuova spinta a un processo di revisione dei contenuti informativi delle sdo e a un progetto di integrazione delle basi di dati del Sistema sanitario nazionale (Ssn), peraltro già avviato all'inizio degli anni Duemila con il cosiddetto progetto Mattoni, con una esplicita interconnessione tra i diversi cosiddetti flussi informativi correnti. Era un obiettivo ambizioso l'interconnessione a livello nazionale delle informazioni individuali tra sdo, farmaceutica, specialistica, emergenza, pronto soccorso, riabilitazione e tutti gli altri sistemi informativi su base individuale del Ssn con l'anagrafe tributaria, allora, e ancor oggi, unica anagrafe di popolazione attiva a livello nazionale. Questa importante azione di riorganizzazione del sistema informativo del Ssn viene esplicitamente sancita per iniziativa di Renato Balduzzi, divenuto Ministro della salute e grande sostenitore dei sistemi di valutazione, e con il contributo decisivo di alcuni senatori, tra i quali mi piace ricordare Lionello Cosentino, dalla Legge numero 135, del 7 agosto 2012. Stiamo parlando delle cosidetta spending review del governo Monti che all'articolo 15 comma 25 bis, così recita: *"Ai fini della attivazione dei programmi nazionali di valutazione sull'applicazione delle norme di cui al presente articolo, il Ministero della salute provvede alla modifica e integrazione di tutti i sistemi informativi del Servizio sanitario nazionale, anche quando gestiti da diverse amministrazioni dello Stato, e alla interconnessione a livello nazionale di tutti i flussi informativi su base individuale. Il complesso delle informazioni e dei dati individuali così ottenuti è reso disponibile per le attività di valutazione esclusivamente in forma anonima ai sensi dell'articolo 35 del decreto legislativo 23 giugno 2011, n.118. Il Ministero della salute si avvale dell'Agenas per lo svolgimento delle funzioni di valutazione degli esiti delle prestazioni assistenziali e delle procedure medico-chirurgiche nell'ambito del Servizio sanitario nazionale. A tal fine, Agenas accede, in tutte le fasi della loro gestione, ai sistemi informativi interconnessi del Servizio sanitario nazionale di cui al presente comma in modalità anonima".*

Più chiaro di così il legislatore non poteva essere. Ma l'attuazione di questa legge, avanzatissima per quei tempi e certamente nel senso della creazione di big data per la valutazione e il governo del Ssn, avrebbe poi comportato: la acquisizione dei pareri della cabina di regia del Nisis, della Conferenza permanente per i rapporti tra lo Stato, le regioni e le province autonome, e dell'Autorità garante della privacy, del Consiglio di stato; poi il nulla osta della Presidenza del Consiglio dei ministri, il visto del Ministro guardasigilli, la registrazione della Corte dei conti, infine la pubblicazione sulla *Gazzetta ufficiale* e l'attuazione da parte delle regioni, prevista allora per l'inizio del 2015. Il 2015 è passato, il 2016 sta finendo e i decreti sulla interconnessione e quello sulle sdo non sono ancora pubblicati sulla *Gazzetta ufficiale*…

Anni e anni per realizzare una riorganizzazione dei sistemi informativi che, in altri paesi e in altre culture, avrebbe richiesto solo pochi atti amministrativi. In questo ritardo spaventoso ha giocato un inammissibile ruolo ostativo, dilatorio l'Autorità garante della privacy con continui rinvii e obiezioni, sempre alla scadenza dei termini, in un tiramolla paralizzante e defatigante.

**Quali potrebbero essere le ragioni di questo ritardo?**

A taluni è venuto il sospetto che in questo incredibile ritardo abbiano giocato fattori non dichiarati, come i contrasti tra gestori commerciali dei sistemi informativi in diversi ministeri o l'ostilità di qualche settore professionale. Ad esempio la nuova sdo dovrebbe contenere l'identificazione dei chirurghi per ciascuna procedura chirurgica, informazione che consentirebbe di stimare il volume di attività dei professionisti e di valutarne gli effetti sugli esiti, ma anche di controllare possibili distorsioni nello svolgimento delle attività professionali. Magari, ad esempio, si potrebbe anche scoprire che alcuni illustri cattedratici o alcune scuole di specializzazione non hanno sufficienti volumi di attività.

Di fatto le burocrazie hanno fino ad oggi bloccato questi importanti cambiamenti nelle informazioni disponibili per il Ssn, impedendo tante possibili analisi etiologiche e valutative importanti per la tutela della salute della popolazione. Quanti studi su fattori di rischio ambientale e occupazionale sarebbero stati possibili se i dati interconnessi del Ssn fossero stati tempestivamente resi disponibili alla agenzie competenti e alle strutture di ricerca qualificate? Quanti farmaci avrebbero avuto una più rapida e valida valutazione comparativa di efficacia (o di inefficacia)? Quanta inappropriatezza in strutture e servizi sanitari avrebbe potuto essere meglio identificata? Quante tecnologie sanitarie avrebbero potuto essere meglio valutate? Quanta della tanto osannata "eccellenza" avrebbe potuto essere validamente certificata, consentendo ai cittadini scelte più informate nei luoghi di cura?

Viene spontanea una domanda: sono corpi e settori burocratici arretrati e conservatori a ostacolare lo sviluppo di una più valida base informativa sulla quale fondare valide analisi etiologiche o valutative nel Ssn? Oppure esiste una diffusa, nemmeno tanto nascosta, volontà politica che non vuole informazioni valide e tempestive sulla salute della popolazione e sull'efficacia dei servizi sanitari per poter continuare a decidere arbitrariamente senza spiegare le ragioni delle decisioni e senza dover rispondere dei loro effetti?

Ovvero, anche nel Ssn è la burocrazia il male oscuro di questo paese, oppure è una certa politica che usa la burocrazia per non dover rispondere delle proprie scelte?

**La velocità di produzione e messa a disposizione dei nuovi big data sembra un vantaggio rispetto ai dati generati dalle rilevazioni fatte fino ad oggi con i sistemi correnti. Si tratta di un vero avanzamento? E riguardo alla varietà e al volume?**

Credo di aver risposto a questa domanda. Grandi ed efficienti basi di dati di buona qualità sarebbero molto utili alla ricerca etiologica e valutativa e consentirebbero di fornire importanti e tempestive informazioni per le decisioni di gestione e di governo dei sistemi sanitari. Ben vengano quindi big, anche very, extremely big data, a condizione tuttavia che siano utilizzati con trasparenza e rigore metodologico, con la consapevolezza che la moltiplicazione delle fonti, della quantità e tipologia dei dati e dei loro trattamenti moltiplica le fonti di errore. Ancora una volta, guardando una fotografia è bene ricordare che le caratteristiche della macchina fotografica sono altrettanto importanti della (teorica) realtà che il fotografo intende riprendere. Ogni misura è un esercizio di errore, ma nulla possiamo conoscere di quello che non misuriamo.

In quanto alla velocità mi pongo una domanda: come è possibile pensare a velocissimi big data quando per ottenere una semplice modifica dei contenuti informativi della sdo ci sono voluti cinque anni? Avremmo già disponibili grandi e relativamente veloci sistemi informativi che, se interconnessi, costituirebbero una base di dati grande e potente, ma le burocrazie che paralizzano il paese ne impediscono la realizzazione e l'uso.

Siamo veramente convinti che la politica, nelle sue diverse articolazioni, voglia veramente i big data che, se utilizzati con metodi scientifici rigorosi e in modo trasparente, potrebbero produrre informazioni capaci di condizionare e criticare le scelte politiche, limitandone comunque i gradi di libertà, l'arbitrio e, soprattutto, costringendo tutti a rendere esplicita l'incertezza e a rendere ragione delle proprie scelte?

> *"Ogni misura è un esercizio di errore, ma nulla possiamo conoscere di quello che non misuriamo."*