Le impronte digitali al servizio dell'epidemiologia

Le potenzialità della digital epidemiology per la salute sono rilevanti

Il lavoro dell'epidemiologo è un po' come quello dell'investigatore: si basa sulla ricerca di tracce da analizzare, e oggi gran parte di queste tracce sono dati digitali lasciati quotidianamente da chi interagisce su internet attraverso i social media e le app. Grazie alle dettagliate informazioni sulla mobilità degli esseri umani su scala globale, l'analisi di questi dati rende possibile predire come epidemie e pandemie si propaghino: fotografando in modo dettagliato come gli individui si muovono con aerei o treni, nel momento in cui arriva una pandemia, si possono determinare predizioni e previsioni su quando e quanti casi verranno importati nei vari paesi. È proprio ciò che è stato fatto nel 2009 durante la pandemia di H1N1 da un team di ricerca della Fondazione Isi – Istituto per l'interscambio scientifico, che, grazie ai dati sulla mobilità degli esseri umani e a un sofisticato framework computazionale in grado di modellizzare la propagazione della pandemia, è riuscito a predire con tre, quattro mesi di anticipo il picco pandemico di H1N1 nell'emisfero nord.

Si può anche predire cosa accadrà in termini di diffusione e impatto sulla popolazione. Utilizzando un modello di simulazione a base individuale l'Istituto superiore di sanità (Iss), sempre nel 2009, ha potuto simulare e prevedere l'andamento della pandemia influenzale in Italia per valutare l'impatto delle possibili misure di contenimento adottate in Italia. Per ottenere il modello di popolazione sono stati utilizzati dati censuali italiani che hanno permesso di costruire una popolazione sintetica, all'interno della quale si è potuto simulare il modello di trasmissione dell'influenza pandemica e le relative misure da adottare per ridurre il contagio (misure di distanziamento sociale, chiusura delle scuole, uso di antivirali, uso dei vaccini, ecc.). Queste previsioni si sono dimostrate un valido supporto per le autorità sanitarie indicando in anticipo con precisione il picco della pandemia in Italia al momento della dichiarazione di emergenza di sanità pubblica internazionale da parte dell'Organizzazione mondiale della sanità.

Ma questi sono soltanto alcuni dei tanti esempi che illustrano le potenzialità della *digital epidemiology*, la disciplina che utilizza il web come luogo di raccolta di dati di interesse epidemiologico rimodellando i confini dell'epidemiologia tradizionale, senza però sostituirla. In pratica, essa rappresenta un sistema di vigilanza della salute collettiva che, con l'intervento di una modellistica computazionale, integra i sistemi di analisi dell'epidemiologia tradizionale con i dati digitali globali per costruire un quadro della diffusione di una malattia in un determinato momento e anche fare previsioni a lungo termine.

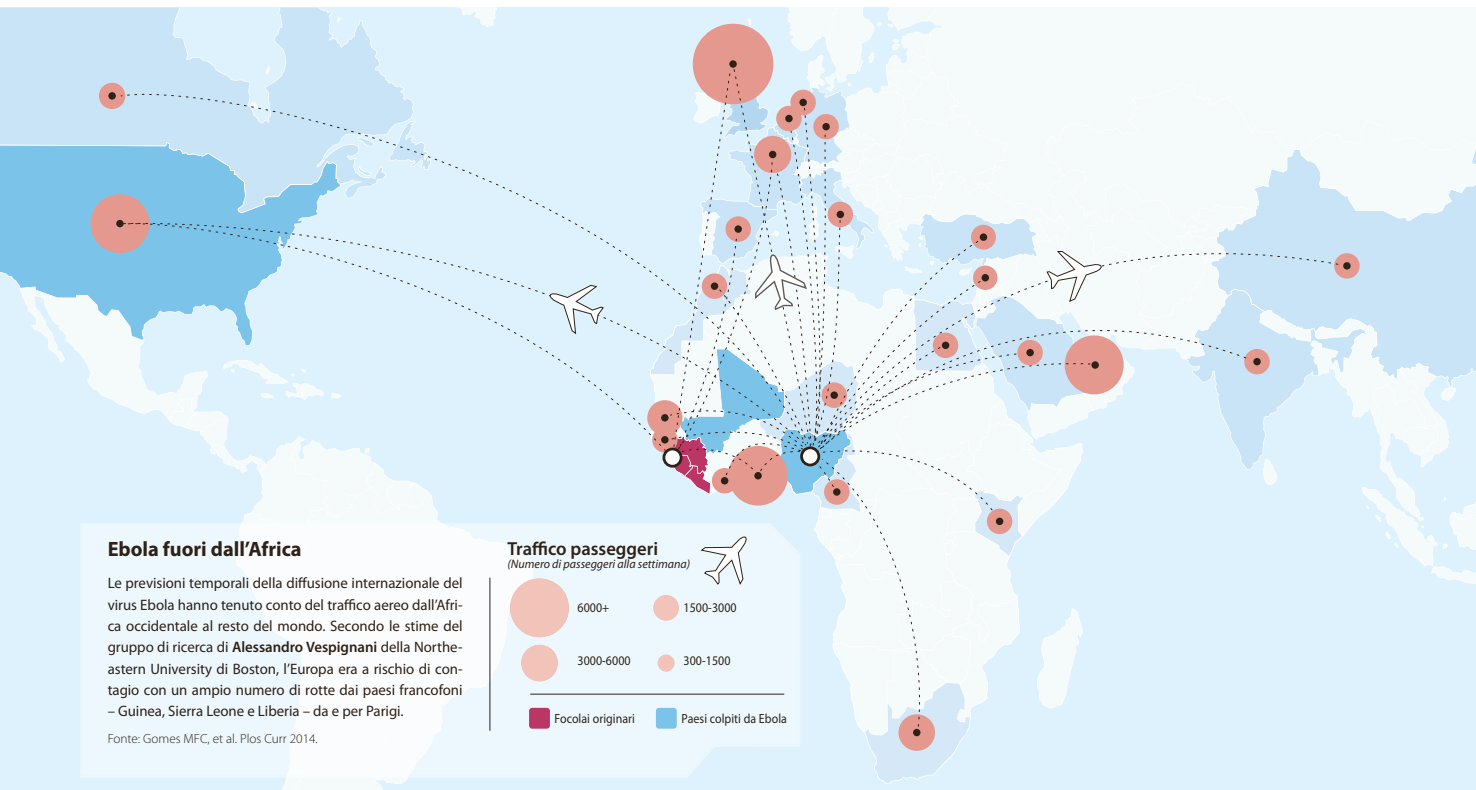
L'epidemiologia digitale ha un record di successi ormai decennale. Basti pensare alla piattaforma Healthmap, fondata 10 anni fa da John Brownstein della Harvard School. Brownstein è stato il primo a utilizzare le tracce digitali che in ogni istante vengono lasciate sui blog, sui siti istituzionali e sui social media per avere informazioni sulle epidemie e sulla diffusione delle malattie a livello globale. Healthmap adotta il classico approccio "passivo" alla rete che si basa, per l'appunto, sull'analisi delle tracce digitali generate per scopi diversi da quelli epidemiologici ma che possono essere usati per studi in questo campo.

Esiste poi un approccio "attivo", che utilizza la rete per reclutare volontari a cui chiedere informazioni circa la loro condizione di salute. Si tratta sempre di dati digitali generati tramite web ma appositamente per scopi epidemiologici. L'approccio "attivo" è quello su cui si fonda, ad esempio, InfluenzaNet, una piattaforma web interattiva volta a raccogliere dati sull'influenza stagionale – con una risoluzione geografica e temporale molto alta – per informare modelli predittivi. La sorveglianza viene condotta su una coorte di volontari che annualmente, all'inizio della stagione influenzale, vengono invitati a ripor-

Daniela Paolotti
Fondazione Isi – Istituto per l'interscambio scientifico
Project manager di Influweb

Caterina Rizzo
Centro nazionale di epidemiologia, sorveglianza e promozione della salute
Istituto superiore di sanità

> *Invece di concentrarci sulla "rivoluzione dei big data" è forse arrivato il momento di focalizzarci sulla "rivoluzione di tutti i dati".*
>
> — David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani



Ebola fuori dall'Africa

Le previsioni temporali della diffusione internazionale del virus Ebola hanno tenuto conto del traffico aereo dall'Africa occidentale al resto del mondo. Secondo le stime del gruppo di ricerca di **Alessandro Vespignani** della Northeastern University di Boston, l'Europa era a rischio di contagio con un ampio numero di rotte dai paesi francofoni – Guinea, Sierra Leone e Liberia – da e per Parigi.

Fonte: Gomes MFC, et al. Plos Curr 2014.

tare la loro condizione di salute sia che stiano bene sia che abbiano dei sintomi respiratori. Con questo approccio non si raggiungono le dimensioni dei big data, poiché il numero degli individui raggiunti con questa modalità non è paragonabile ai milioni di utenti di Facebook o Twitter, ma il numero è tale per cui il segnale epidemiologico che si ottiene è sufficientemente accurato. Inoltre, con la modalità della sorveglianza partecipativa si possono ottenere informazioni da persone che non si recano dal medico in caso di febbre, ma che non hanno problemi a compilare un questionario sul web quando sono a casa da malati. InfluenzaNet è stato sperimentato per la prima volta in Olanda e Belgio nella stagione influenzale 2003/2004. Ora viene utilizzato in 10 paesi europei, tra cui l'Italia con la Fondazione Isi e l'Iss, la Francia con l'Inserm e l'Inghilterra con la Public health England, e ha inoltre ispirato delle piattaforme analoghe negli Stati Uniti e in Australia. Si è quindi creato un sensore digitale globale di volontari, sia dell'emisfero nord che di quello sud, che ogni anno durante la stagione influenzale riportano il proprio stato di salute. Questo è un enorme passo avanti nella sorveglianza globale dell'influenza.

Le applicazioni dei big e small data

L'epidemiologia digitale viene spesso associata ai big data che, secondo la definizione contenuta in un rapporto del Congresso Usa del 2012, rappresentano "grandi volumi di dati ad alta velocità, complessità e variabilità che richiedono tecniche e tecnologie avanzate per la raccolta, l'immagazzinamento, la distribuzione, la gestione e l'analisi dell'informazione". Ma in realtà si può parlare di epidemiologia anche per gli small data. Non necessariamente tutti i dati che si raccolgono dal web, dai social media o dai cellulari devono essere *big*. Possono essere *small* nel senso che sono sempre digitali ma le dimensioni del dataset non sono talmente grandi da richiedere software e tecnologie particolari per poter processare in modo efficiente l'enorme ammontare di dati in tempi ragionevoli. E viceversa dati della genomica raccolti per vie tradizionali possono rientrare nella categoria dei big data. Sarebbe quindi più corretto riferirsi non alla dimensione del volume dei dati bensì alla sorgente parlando quindi di dati digitali.

Questi dati vengono usati non solo per contare quanti casi di una certa malattia vengono osservati in una certa popolazione o per capire se la diffusione di un certo comportamento, come smettere di fumare, può avere un impatto positivo sulla salute della persona. Si impiegano anche per sondare se la popolazione è favorevole o meno alle vaccinazioni (il cosiddetto *sentiment*) oppure per monitorare i fattori di rischio comportamentali per patologie legate all'obesità, al fumo, all'abuso di sostanze alcoliche, alla non attività fisica. Infatti, sono state messe a punto delle tecniche per estrapolare i comportamenti a rischio degli utenti iscritti ai social media ed eventualmente inviare a questi soggetti messaggi specifici che possano influire e determinare un cambiamento o un miglioramento del loro stile di vita.

Un'altra applicazione di notevole impatto è la sorveglianza sugli eventi avversi ai farmaci. L'Agenzia europea dei medicinali ha identificato le esigenze dell'utilizzo di questi dati provenienti dai social media, e in generale dai media, per identificare segnali che possano aiutare la farmacovigilanza routinaria che di solito viene effettuata attraverso le segnalazioni dei medici o della popolazione. È uno strumento potenzialmente molto importante perché consente di catturare tutto ciò che viene avvertito dai cittadini come evento avverso che per la sanità pubblica è molto difficile da identificare.

Questioni aperte

Chiaramente i tipi di dati su cui si basa l'epidemiologia digitale non sono raccolti per scopi epidemiologici, e quindi non si tratta di una coorte di pazienti selezionati in modo rappresentativo, ben controllata. I limiti di questo approccio sono legati, in parte, al cosiddetto *selection bias*, cioè il problema di selezionare adeguatamente il campione da analizzare. Si pone una duplice problematica: da un lato il segnale rilevabile da questi dati non è pulito, dall'altro nei paesi in via di sviluppo l'accesso al web è limitato. È per tale ragione che i dati di Twitter funzionano bene per la sorveglianza dell'influenza in quei paesi in cui questo tipo di social network è molto diffuso, ma non altrettanto bene in paesi come l'Italia dove è ancora sottoutilizzato.

Tuttavia i vantaggi della *digital epidemiology* sono tali che questo tipo di limite è superabile. Diversi articoli scientifici hanno verificato per esempio che il dato sull'incidenza dell'influenza misurato contando il numero di volte che gli utenti menzionano su Twitter parole legate alle sindromi influenzali è strettamente correlato al tipo di segnale rilevato con la sorveglianza nazionale, ovvero dai medici sentinella che riportano settimanalmente il numero di casi di influenza diagnosticati tra i loro pazienti. Quindi è vero che il segnale di Twitter è rumoroso e non rappresentativo, ma l'abbondanza dei dati è tale per cui aggregandolo a livello di paese la curva che si ottiene è strettamente simile a quella ottenuta attraverso la sorveglianza basata sui medici sentinella. E questo ci conforta perché – nonostante tutti i problemi legati al tipo di dati – il segnale che si ottiene è comunque robusto e affidabile. La stessa cosa è stata fatta anche per altri sistemi. Ad esempio, lo stesso Brownstein aveva pubblicato un articolo dove dimostrava che negli Usa il numero di "click" degli utenti di Wikipedia su articoli il cui titolo conteneva parole collegate ai sintomi influenzali generava una incidenza fortemente correlata a quella rilevata con i dataset tradizionali dai Centers for disease control.

Conclusioni

Nonostante le criticità, le potenzialità che la *digital epidemiology* porta all'epidemiologia tradizionale e alla salute pubblica sono numerose e rilevanti. Anzitutto la disponibilità di dati in tempo reale consente di monitorare costantemente l'evolvere per esempio di un'epidemia, o, nell'ambito della farmacovigilanza, di identificare segnali relativamente a eventi avversi a farmaci che possono completare la sorveglianza routinaria e la farmacovigilanza, solitamente effettuate attraverso le segnalazioni da parte degli operatori sanitari. Questo paradigma della sorveglianza partecipativa è stato esteso anche a paesi meno sviluppati in cui il sistema sanitario non è perfettamente funzionante ma in cui la gran parte della popolazione dispone di un telefono cellulare attraverso il quale può accedere al web e compilare un questionario sui propri sintomi. L'idea, oltre che sfida, per l'epidemiologia è di integrare la sorveglianza routinaria con il digitale e con un approccio partecipativo per avere infor- *a p. 10* ⟶

Il flop di Google flu trends

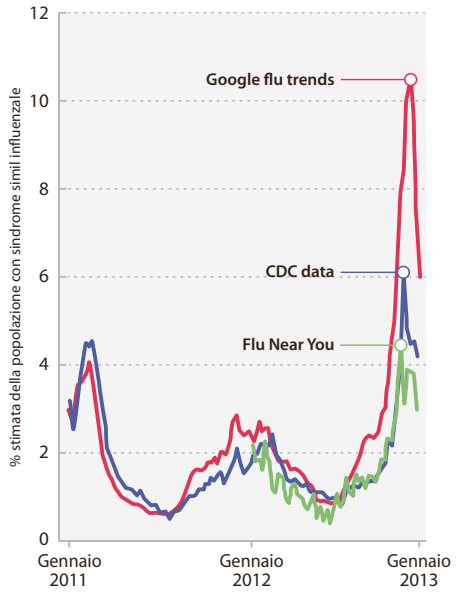
L'attività dei motori di ricerca quali Google che conta centinaia di milioni di utenti attivi è potenzialmente un segnale affidabile ma non sempre preciso. Un esempio di ciò è Google flu trends, considerato un vero e proprio flop. Google flu trends si basava sull'analisi delle ricerche fatte tramite il motore di ricerca di Google di parole collegate ai sintomi influenzali quali febbre, mal di gola, raffreddore. Il numero di volte che gli utenti chiedevano al motore di ricerca queste informazioni veniva utilizzato come specchio del numero di casi di influenza fra la popolazione.

Dopo diversi inverni di mappature perfette delle epidemie influenzali, nel 2013 il sistema aveva fallito clamorosamente sovrastimando i casi di influenza. Il problema è che Google usava un modello statistico impiegato per produrre previsioni da una settimana all'altra e che veniva allenato soltanto sui dati della stagione corrente, quando invece la dinamica dell'influenza è tale per cui si osserva sempre lo stesso andamento stagionale, ma se si analizza nel dettaglio si osserva che ogni stagione è diversa. Inoltre, il fatto che una persona cerchi la parola "influenza" con Google non è indicativo del motivo per cui lo fa: potrebbe eseguire la ricerca perché ha l'influenza o perché ne ha sentito parlare molto dai media.

Il flop di Google flu trends potrebbe però essere imputabile non tanto alla qualità dei dati digitali quanto piuttosto al modello di calcolo impiegato che non è mai stato reso noto alla comunità scientifica. Prima del 2013 e quindi della sua chiusura, Google flu trends veniva usato estensivamente in tanti paesi dove era assente un sistema esteso di medici sentinella e, al di là dei suoi limiti, rappresentava comunque uno strumento importante di sorveglianza per sanità pubblica. Migliorandolo nella modellistica per questi paesi potrebbe rappresentare comunque uno strumento valido per l'epidemiologia e integrare i dati ottenuti attraverso i sistemi di sorveglianza. •

I picchi di influenza

Previsioni della percentuale della popolazione statunitense con sindrome simil influenzale fatte utilizzando l'algoritmo di Google flu trends, i dati dei Centers of disease control della rete sentinella di medici e Flu Near You, la piattaforma sviluppata da HealthMap insieme all'American Public Health Association in cui gli utenti inseriscono settimanalmente i propri sintomi. Nel 2013 Google flu trends aveva sovrastimato il picco influenzale.



da p.9 → mazioni epidemiologiche molto accurate. Con la disponibilità di dataset amministrativi e di dati provenienti dal web e con lo sviluppo di sistemi innovati per il caricamento, lo storage e l'analisi di grandi quantità di dati, buona parte dei quali non strutturati, l'epidemiologia si trova ad affrontare un cambiamento critico: riuscire a integrare dati da sorgenti tradizionali e nuove, e garantire una comprensione più chiara e accurata del mondo. Come hanno sottolineato Alessandro Vespignani e colleghi non dovremmo tanto parlare di "rivoluzione dei big data" ma di "rivoluzione di tutti i dati". F

— Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Surveillance sans frontieres: internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med 2008;5:e151.

— Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data. Nature 2009;457:1012-4.

— Hay SI, George DB, Moyes CL, Brownstein JS. Big Data Opportunities for Global Infectious Disease Surveillance. PLoS Med 2013;10:e1001413.

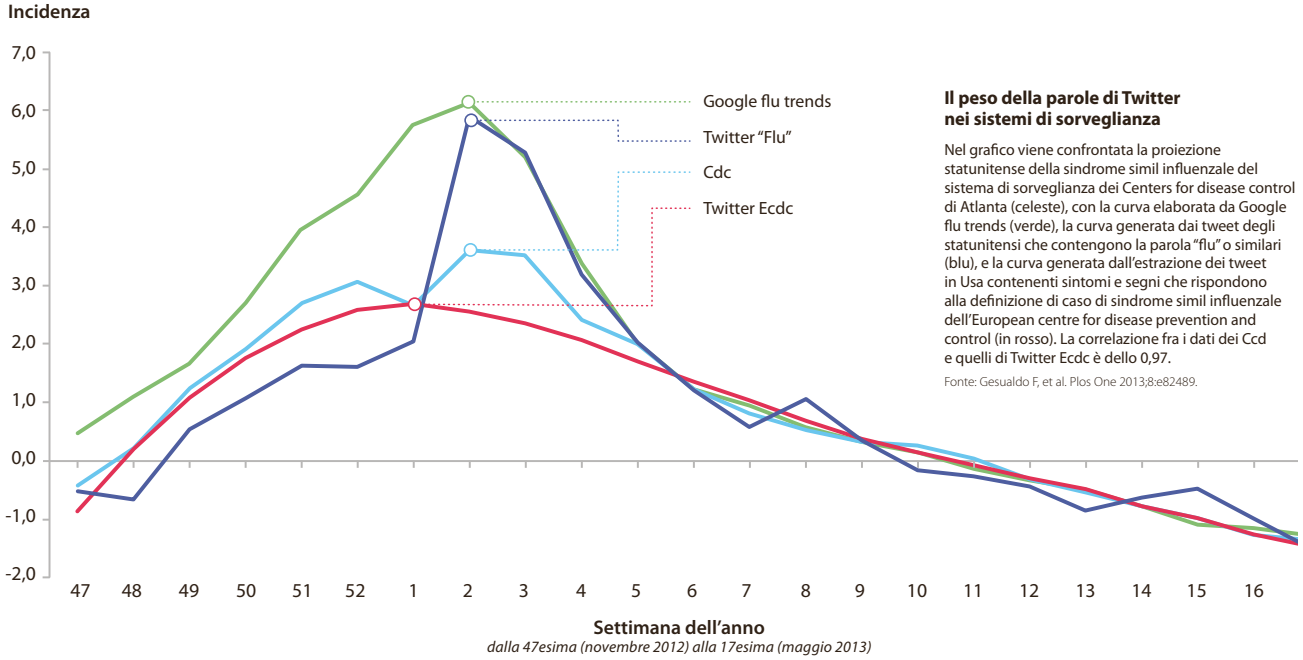
— Salathé M, Bengtsson L, Bodnar TJ, et al. Digital epidemiology. PLoS Comput Biol 2012;8:e1002616.

— Paolotti D, Carnahan A, Colizza V, et al. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. Clinical Microbiology and Infection 2014;20:17-21.

— Zhang Q, Gioannini C, Paolotti D, et al. Social Data Mining and Seasonal Influenza Forecasts: The FluOutlook Platform. Computer Science 2015; 9286:237-40.

— Cantarelli P, Debin M, Turbelin C, et al., The representativeness of a European multicenter network for influenza-like-illness participatory surveillance. BMC Public Health 2014;14:984.

— Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science 2014;343:1203-5.



Il peso della parole di Twitter nei sistemi di sorveglianza

Nel grafico viene confrontata la proiezione statunitense della sindrome simil influenzale del sistema di sorveglianza dei Centers for disease control di Atlanta (celeste), con la curva elaborata da Google flu trends (verde), la curva generata dai tweet degli statunitensi che contengono la parola "flu" o similari (blu), e la curva generata dall'estrazione dei tweet in Usa contenenti sintomi e segni che rispondono alla definizione di caso di sindrome simil influenzale dell'European centre for disease prevention and control (in rosso). La correlazione fra i dati dei Ccd e quelli di Twitter Ecdc è dello 0,97.

Fonte: Gesualdo F, et al. Plos One 2013;8:e82489.

Cercare i piccoli numeri nei grandi numeri

I database amministrativi al servizio dell'epidemiologia

Francesca Dominici

PhD, Harvard T.H. Chan School of public health

In ambito epidemiologico generalmente per big data si intendono i dati sulla salute della popolazione estrapolati dai cosiddetti dati amministrativi, come ad esempio le ricevute rilasciate al paziente al momento di una visita medica, di un esame clinico o di un ricovero. Diversamente dalle informazioni registrate nella cartella clinica del paziente, i dati amministrativi peccano di scarsa accuratezza perché non danno delle indicazioni diagnostiche precise, ragione per cui spesso vengono criticati. Di contro però hanno il vantaggio di coprire popolazioni di grandi dimensioni difficilmente raggiungibili con uno studio epidemiologico tradizionale, disegnato in modo mirato per misurare una serie di fattori di rischio e di malattie su campioni di popolazione numericamente molto inferiori.

"Il valore aggiunto dei big data è di aumentare l'informazione già presente negli studi epidemiologici tradizionali."

Disporre di una grande quantità di dati in poco tempo, anche se non accurati, è un valore aggiunto in diversi ambiti epidemiologici, ad esempio per gli studi sugli effetti negativi dell'inquinamento atmosferico sulla salute umana che, essendo effetti di piccole dimensioni, richiedono ampi campioni per essere rilevati attraverso metodi statistici che tengano conto delle interazione dei vari fattori. Sia negli Usa sia in Europa, si sta quindi sperimentando la strada di incrociare i database amministrativi dell'intera popolazione con le rilevazioni dei principali inquinanti presenti nell'aria ottenute dalle stazioni di monitoraggio e dai satelliti.

Ad esempio alla Harvard T.H. Chan School of public health abbiamo potuto verificare sui grandi numeri che la riduzione del particolato sottile al di sotto degli standard annuali per la qualità dell'aria riduce i decessi e le ospedalizzazioni per tutte le cause e per malattie respiratorie o circolatorie. Abbiamo raccolto i dati registrati dalla Medicare current beneficiary survey di 34.427 iscritti a Medicare con 5355 diversi codici postale e li abbiamo incrociati con i livelli medi annui di pm 2,5 rilevati in ogni codice postale. Alla Johns Hopkins Bloomberg school of public health, della University di Baltimora, è stato condotto uno studio sull'associazione causale tra malattie cardiache e respiratorie ed esposizione alle polveri sottili estrapolando dalla banca dati National claims history di Medicare i dati di 11,5 milioni di iscritti a Medicare, con più di 65 anni, che vivevano in 204 contee urbane statunitensi.

È importante precisare che questi studi vengono effettuati sempre combinando i dati amministrativi con i dati epidemiologici. L'analisi dei dati non può, infatti, ignorare quanto già conosciamo delle relazioni tra inquinanti e condizioni di salute e sui fattori di rischio. Essenzialmente il valore aggiunto dei big data è di aumentare l'informazione già presente negli studi epidemiologici tradizionali. F