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.

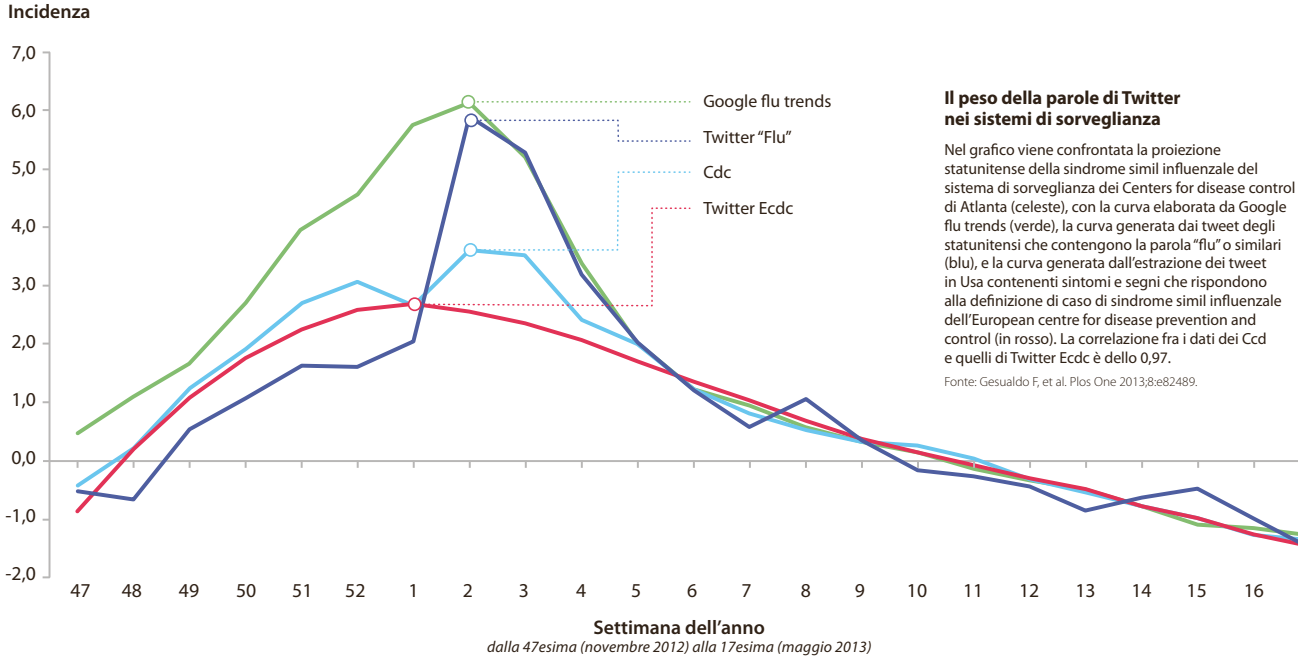
— Salathé M, Bengtsson L, Bodnar TJ, et al. Digital epidemiology. PLoS Comput Biol 2012;8:e1002616.

— 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