# Searching for better flu surveillance? A brief communication arising from Ginsberg *et al. Nature* 457, 1012-1014 (2009)

International Society for Disease Surveillance (ISDS) Distributed Surveillance Taskforce for Real-time Influenza Burden Tracking and Evaluation (DiSTRIBuTE) Working Group

ISDS DiSTRIBuTE Working Group: Donald R. Olson[1], Atar Baer[2], Michael A. Coletta[3], Lana Deyneka[4], Ryan Gentry[5], Amy Ising[6], Erin L. Murray[7], Marc Paladini[7], Justin Pendarvis[8], Karl Soetebier[9], Kevin J. Konty[7], Jill Schulmann[10], Jeffrey Engel[4], Julia Gunn[8], Robert T. Rolfs[11] & Farzad Mostashari[7]

*[1]International Society for Disease Surveillance, New York, NY, USA. [2]Public Health – Seattle & King County, Seattle, WA, USA. [3]Virgina Department of Health, Richmond, VA, USA. [4]North Carolina Division of Public Health, Raleigh, NC, USA. [5]Indiana Department of Health, Indianapolis, IN, USA. [6]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [7]New York City Department of Health and Mental Hygiene, New York, NY, USA. [8]Boston Public Health Commission, Boston, MA, USA. [9]Georgia Division of Public Health, Atlanta, GA, USA. [10]Markle Foundation, New York, NY, USA. [11]Utah Department of Health, Salt Lake City, UT, USA.*

Through retrospectively analyzing billions of internet search queries, Ginsberg *et al.*[1] identified a collection of specific searches that track the course of influenza-like illness (ILI) reported by the US Centers for Disease Control and Prevention (CDC)[2]. Prospective monitoring during 2007-2008 found high correlation between Google estimates and CDC-reported ILI, with next-day timeliness compared to the 1-2 week delay reported in traditional CDC ILI surveillance[1]. The assertion by Ginsberg *et al.*[1], however, that internet search term estimates enable public health officials to respond better to seasonal and pandemic influenza does not take into account the current practice of public health, or the state of the art in electronic disease surveillance.

Local and state health departments in the U.S. have increasingly used electronic syndromic surveillance systems to monitor influenza-related morbidity[3]. The advantages of monitoring febrile, respiratory and ILI syndromes using electronic outpatient and emergency department (ED) data at the local and state level have previously been shown[4-7]. The Distributed Surveillance Taskforce for Real-time Influenza Burden Tracking and Evaluation (DiSTRIBuTE) is a collaborative working group of state and local health departments conducting this type of syndrome-based surveillance as part of daily public health practice[8]. Designed in a framework consistent with the Markle Foundation Connecting for Health guidelines for health-data sharing, security and patient privacy[9], the DiSTRIBuTE network electronically receives data from regional health departments aggregated by day, syndrome, age-group and 3-digit zip code. In order to evaluate the robustness and timeliness of the surveillance system proposed by Ginsberg *et al.* [1] against other indicators of influenza activity, we compared correlations between publicly available query-based trends from Google[1], ILI data from CDC sentinel providers[2], outpatient and ED visit data from DiSTRIBuTE participating sites[8], and influenza viral culture data from CDC-collaborating laboratories.

Trends in Google estimates from 2006-2008 paralleled national influenza isolate reporting and aggregated DiSTRIBuTE data, as well as CDC-ILI (Fig. 1a,b). In addition, Google state-level estimates were highly correlated with corresponding DiSTRIBuTE data for the 5 states and 3 cities in the system (range, 0.90 to 0.97). Cross-correlation during 2007-2008 found the 3 city-level DiSTRIBuTE systems leading corresponding Google state-level estimates (at one week lead, range 0.92 to 0.96, data not shown), and examination of age-specific DiSTRIBuTE data found earlier increases in visits among school-age children and younger working-age adults, compared to infants (<24 months), and older adults (>45 years) (Fig. 1c). While correlations of CDC-ILI with both DiSTRIBuTE and Google data were very high when calculating coefficients for the entire influenza season (0.97 and 0.98, respectively), limiting the data to shorter time periods found higher coefficients with DiSTRIBuTE data during the first half of each influenza season, when influenza detection is most critical (Fig. 1d).

Our findings validate Ginsberg *et al*'s[1] assertion that Google search estimates provide a publicly available and timely view consistent with influenza activity across the US, but suggest that the Google estimates are not as accurate or timely an indicator of influenza as ED visits during the early influenza season, particularly in certain age groups. More concerning, however, is the inability for health officials to characterize and understand increases in this system, collect additional information and obtain specimens from affected individuals for definitive diagnosis, particularly given the tenuous and inferred relationship between search queries and true illness. Without the ability to link detection to investigation and response[10], this proposed surveillance system may offer little utility to public health practitioners.

Future research could illustrate whether the greater timeliness in ED visits at the city level compared to state-wide Google estimates reflects an underlying urban-rural difference, and investigate the utility of internet search queries for other surveillance needs such as epidemic acute gastroenteritis and allergic asthma. But closer cooperation with public health practitioners is required to create tools for investigation of such increases, and to enable more direct communication between public health authorities and the searching public.

1. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. & Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012-1014 (2009). http://dx.doi.org/10.1038/nature07634

2. CDC Flu Activity & Surveillance. http://www.cdc.gov/flu/weekly/fluactivity.htm

3. Buehler, J.W., Sonricker, A., Paladini, M., Soper, P. & Mostashari, F. Syndromic Surveillance Practice in the United States: Findings from a Survey of State, Territorial, and Selected Local Health Departments. *Advances in Disease Surveillance* **6**, 3 (2008). http://www.isdsjournal.org/article/view/2618/2517

4. Olson, D.R., Heffernan, R.T., Paladini, M., Konty, K., Weiss, D. & Mostashari, F. Monitoring the impact of influenza by age: Emergency department fever and respiratory complaint surveillance in New York City. *PLoS Med.* **4,** e247 (2007). http://dx.doi.org/10.1371/journal.pmed.0040247

5. Baer, A., Coberly, J., Hung, L., Burkom, H., Loschen, W., Lombardo, J. & Duchin, J. Classification of Emergency Department Syndromic Data for Seasonal Influenza Surveillance. *Advances in Disease Surveillance* **4**, 233 (2007). http://www.isdsjournal.org/article/view/2132/1708

6. Murray, E.L., Soetebier, K. & Cameron, W. Syndromic Surveillance and Influenza-like Illness in Georgia. *Advances in Disease Surveillance* **4**, 179 (2007). http://www.isdsjournal.org/article/view/2082/1650

7. Pendarvis, J., Gunn, J., Smith, A.K., Donovan, M. & Barry, A. Sneezes vs. Wheezes: Syndrome Definitions for Influenza-like Illness. *Advances in Disease Surveillance* **2**, 115 (2007). http://www.isdsjournal.org/article/view/870/752

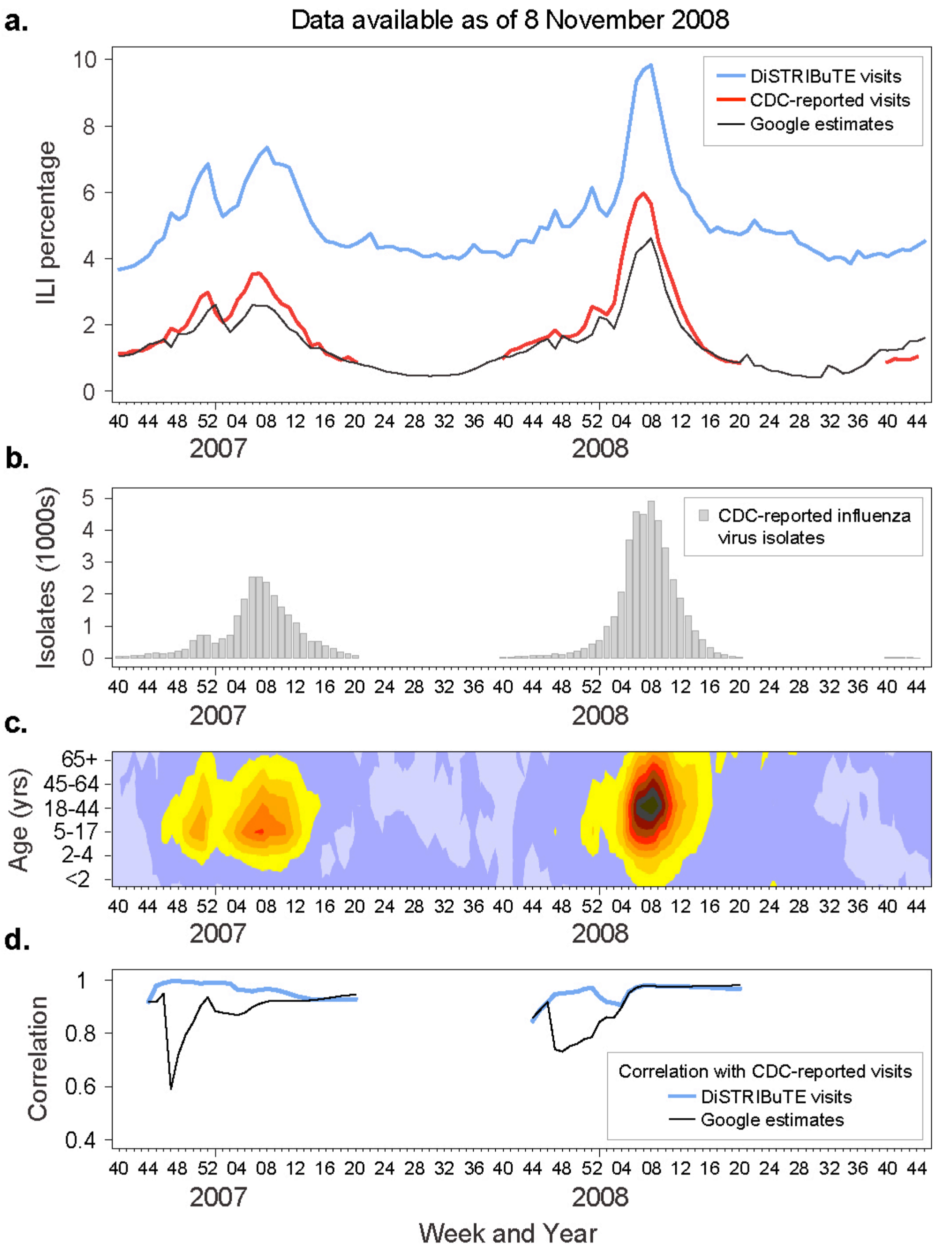
8. Olson, D.R., Paladini, M., Buehler, J.W. & Mostashari, F. Review of the ISDS Distributed Surveillance Taskforce for Real-time Influenza Burden Tracking & Evaluation (DiSTRIBuTE) Project 2007/08 Influenza Season Proof-of-concept Phase. *Advances in Disease Surveillance* **5**, 155 (2008). http://www.isdsjournal.org/article/view/3307/2456

9. Markle Foundation Connecting for Health Common Framework http://www.connectingforhealth.org/commonframework/

10. Teutsch S.M. & Churchill, R.E. *Principles and Practice of Public Health Surveillance*, second edition (Oxford University Press, 2000) pp 17 -29.

Correspondence should be addressed to Donald R. Olson (drolson@gmail.com)

**Figure 1. Comparison of CDC and DiSTRIBuTE surveillance data with Google search engine query based estimates in the US, 1 October 2006 (week 2006-40) through 8 November 2008 (week 45-2008). a,** Observed weekly proportion of fever, respiratory and influenza-like syndrome ED visits reported by DiSTRIBuTE network participant sites (blue), proportion of US CDC-reported sentinel ILI visits (red), and model ILI estimates based on Google search engine query data (black). Weekly DiSTRIBuTE visits were recorded year-round by participating US health department electronic surveillance systems representing 6 of 9 US surveillance regions, from 3 large-city and 5 state systems reporting 31 million total and 281,000 average weekly visits during the period. Weekly CDC-reported visits were recorded during each 33-week influenza season from all 9 US regions, covering 31 million total and 436,000 average weekly visits. **b,** CDC-collaborating laboratory influenza isolates reported in the US are shown by week (grey). **c,** Age-specific temporal epidemic response surface plot[3] shows relative increase in aggregated DiSTRIBuTE visit-proportion over lower-quartile baseline as colour-gradient by week and age group. **d,** Correlations of observed CDC-reported ILI against DiSTRIBuTE data (blue) and Google estimates (black), show coefficients calculated progressively through each influenza season from week 44 (coefficient based on the 5 points from week 40 to 44) through the end of each seasons, week 20 (coefficient based on the 33 points from week 40 to 20). During 2006-2007, DiSTRIBuTE morbidity data were more highly correlated than Google search query estimates for weeks 2006-45 to 2007-13. During 2007-2008, DiSTRIBuTE data were more highly correlated than Google estimates for weeks 2007-47 to 2008-03.



# Supporting materials in response to comments from Authors Ginsberg *et al. Nature* 457, 1012-1014 (2009)

Regarding draft shared with Ginsberg *et al.* on 16 January 2009.

Our Communication Arising questions the assertion by Ginsberg *et al.* that internet search data can enable public health officials to respond better to seasonal and pandemic influenza: We believe their paper does not take into account the current practice of public health electronic disease surveillance.

We present previously unpublished data and analysis from a network of local health department electronic disease surveillance systems. These local systems capture near-time data and analyse it on a daily basis, with equal or greater timeliness as the Google system. The local surveillance data are used routinely to validate and characterize increases and epidemic signals, and to drill down into the data and reach out directly to facilities or clinicians to investigate events. Our working group had many concerns with the paper, they included: first, that the 1-2 week delay reported with US CDC sentinel physician network data was presented as a straw man comparison for evaluation of internet search timeliness; second, the lack of age-specific and regional aggregation, and inability to validate data or investigate epidemic signals in search data were critical shortcomings. Also, at the heart of our concern was a general question: Simply, *When is epidemic flu detected?*

Despite the implicit assertion in the title of the Author’s Letter “Detecting influenza epidemics using search engine query data” Nature 457, 1012-1014 (2009), we were concerned that the internet search data presented were not *detecting* influenza epidemics as much as following their course once well on there way. Detection implies identifying something that is otherwise hidden, and the early seasonal patterns and correlations suggest the emergence and early waves of influenza were not being detected in the search data (Supporting Figure 1a, shaded periods). The monumental (and commendable) data-driven approach Ginsberg *et al.* undertook may have favoured over-fitting their model to non-flu and peak-flu periods, thus missing signs of the emergence and early wave impact from influenza each season – which is precisely when timely detection and characterization of epidemics must occur.

Our direct response to the Author’s comments and suggestions follow.

## *Nature* Author Comment #1

> Your analysis of early-season correlations relies on a "sliding window" of variable length, ranging from 5 weeks to the entire season. In other words, at the beginning of a season, correlations are measured over a small number of weeks, while at the end of a season, correlations are measured over a large number of weeks.
>
> As a result, any momentary discrepancies in correlation are amplified at the beginning of each season and smoothed over at the end of each season. To best understand how correlations vary throughout a flu season, we contend that a fixed-size sliding window would be more appropriate. By measuring each correlation over the same number of consecutive weeks, readers gain a better understanding of how accuracy varies throughout the season.

We agree with the Authors that a fixed-size moving correlation window can be additionally informative. However, we believe the expanding window week-by-week correlations based on cumulative seasonal data, as we present in our Communication Arising, provides an appropriate comparison for two reasons:

(1) public health knowledge of influenza at any point in time is based on the sum of the accumulated data to-date, not simply on the previous 10 weeks. Our wish was to present correlations based on each season’s accumulated context.

(2) early seasonal discrepancies can be important data for public health, even if only based on a small number of data points (ie, by week 45 there are typically hundreds of culture confirmed flu cases reported by US CDC collaborating laboratories, and likely thousands of unrecognised infections).

To address discrepancies due to limitations in early seasonal correlations, we presented our comparisons starting when correlation coefficients reached significance ($p<0.05$), which was the fifth week each season. To capture a correlation measure that mirrors the actual use of surveillance data, the expanding window provided a coefficient with significance based on the accumulated experience that season, with the final coefficient providing a summary measure for the entire season. For comparison, we present our analysis based on the expanding window showing correlation coefficients for the epidemic as elapsed (Supporting Figure 1b), as well as on a 10-week moving window as the Authors suggest (Supporting Figure 1c).

## *Nature* Author Comment #2

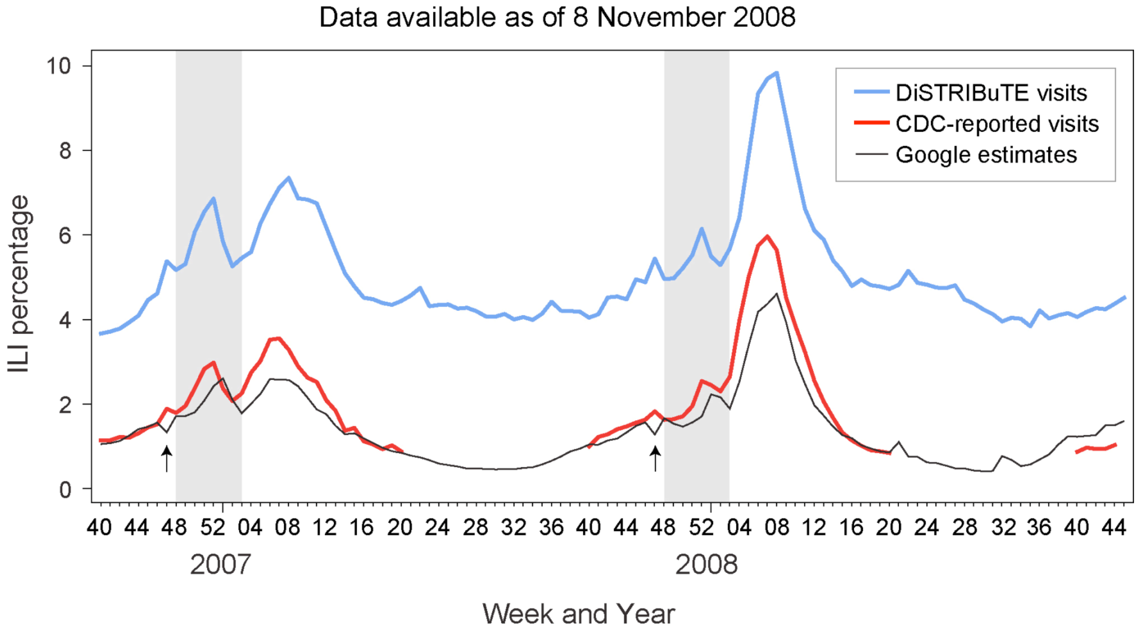
> As you can see in figure "a", our model's ILI percentages are typically underestimated during the week of Thanksgiving (week 47) each year. This makes sense, as search habits (much like physician visit habits) are different during holiday weeks. Our simple univariate model, which does not explicitly consider week-of-year as a variable, is not tuned to detect and adjust for this pattern. If you regenerate figure "d" while excluding week 47, the apparently massive dip in correlation disappears. Thus, it seems that your analysis is overly sensitive to a single week's variation.
>
> We propose that you revise figure "d" using a fixed-size sliding window; preliminary analysis shows that N=10 weeks may be interesting. In this case, your first data point would be the correlation over weeks 40-49. However, as we don't have access to raw DiSTRIBuTe data, we cannot directly perform this analysis on your data.

We agree with the Authors that it makes sense that holiday behaviour can significantly impact internet search patterns. We are concerned, however, with what the implications are of having an underestimation in Google search fractions during the US Thanksgiving holiday that is inversely proportional to the actual observed shift in illness proportions seen in both the CDC and DiSTRIBuTE visit data.

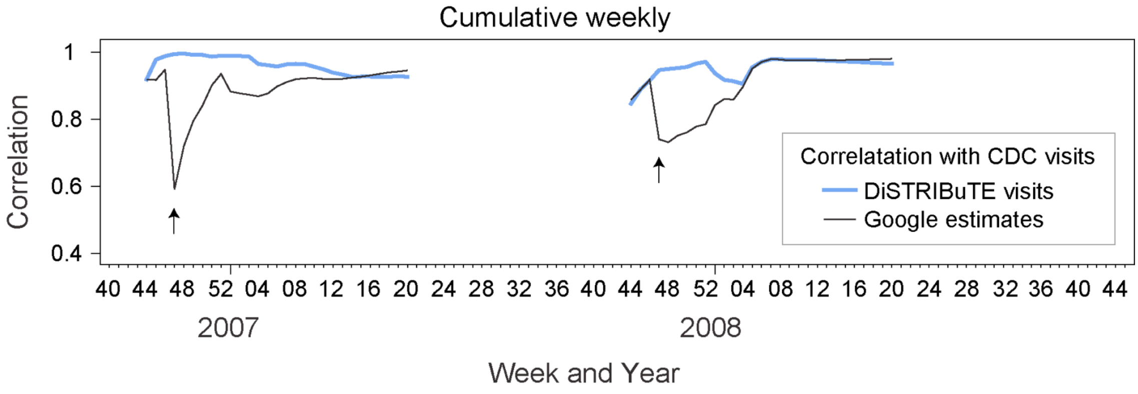
To address the Author’s concerns that Google week 47 data were not suitable for comparison with CDC visits (their gold standard), we followed their recommendation and removed it from the initial (Supporting Figure 1d), and supplementary moving-window correlations (Supporting Figure 1e). As suggested, the removal of this week from the Google data made the “apparently massive dip in correlation disappear”. However, it did not erase lower correlation coefficients that were seen during the following weeks.

The removal of the Google week 47 effect did not show the analysis to be overly sensitive to a single week’s variation. Rather it made the lagged increases and early waves in the Google estimate time-series more apparent during the 8 weeks that followed (shaded periods, Supporting Figure 1a). Comparing the CDC and DiSTRIBuTE visit time-series data with Google estimates during this period found the patterns in the disease surveillance data sources to be consistent, while the increases and early waves seen in the internet search data appeared to be notably lagged.

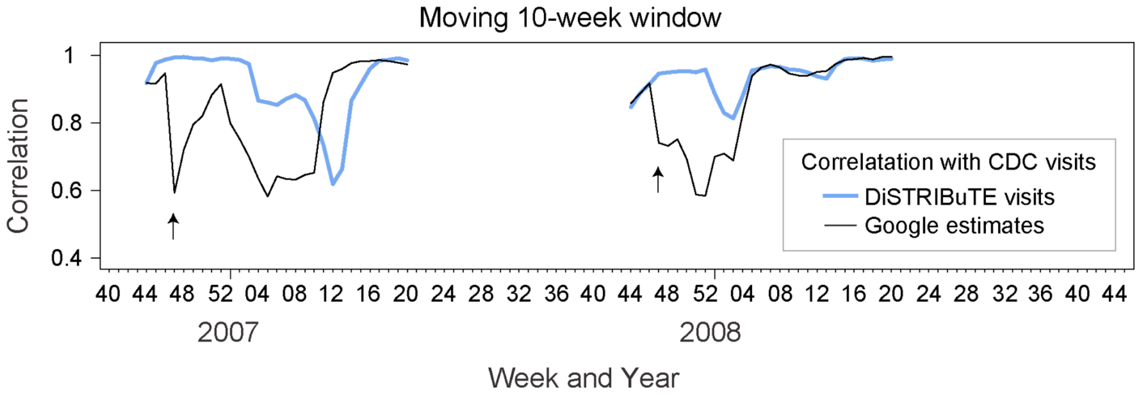
**Supporting Figure 1. Comparison of US influenza-related morbidity surveillance data. Figures (a) and (b) present data and analysis as shown in the accompanying Brief Communication Arising. Figures (c)-(e) present analysis based on suggestions arising in correspondence with the Authors.**



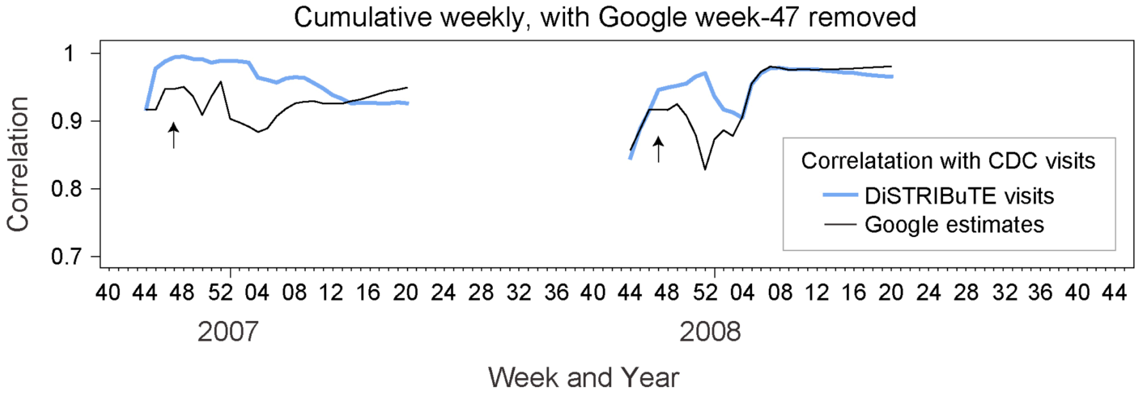
**a, Observed weekly proportion of DiSTRIBuTE fever, respiratory and influenza-like syndrome visits (blue), CDC-reported ILI visits (red) and Google search query based ILI estimates (black):** The US Thanksgiving holiday weeks are indicated (arrows, week 47), and the 8-week period immediately following is highlighted (shaded, weeks 48-03).



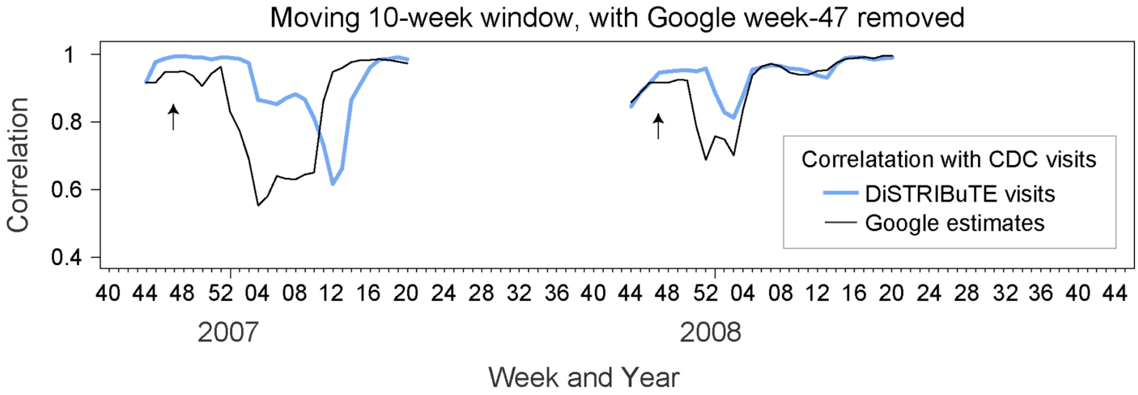
**b, Cumulative weekly correlation with CDC-reported data:** Coefficients calculated against DiSTRIBuTE data (blue) and Google estimates (black), shown progressively through each influenza season from week 44 (coefficient based on 5 points, week 40 to 44) through the end of each season, week 20 (coefficient based on 33 points, week 40 to 20). During both seasons, a dramatic drop in correlations between Google estimates and CDC-reported ILI occurred week 47 (arrows), corresponding with the US Thanksgiving holiday. During 2006-2007, DiSTRIBuTE data were more highly correlated than Google estimates for weeks 2006-45 to 2007-13. During 2007-2008, DiSTRIBuTE data were more highly correlated than Google estimates for weeks 2007-47 to 2008-03.



**c, Moving 10-week window correlations:** Cumulative correlation coefficients were calculated for weeks 44 to 49, and as a moving 10-week window from week 50 (coefficient based on 10 points from week 41 to 50) through the end of each season, week 20 (coefficient based on 10 points from week 11 to 20). During 2006-2007, DiSTRIBuTE data were more highly correlated through week 2007-10, and Google estimates were more highly correlated during weeks 2007-11 to 2007-16. During 2007-2008, DiSTRIBuTE data were more highly correlated for weeks 2007-47 to 2008-05. The Thanksgiving week drop (arrows) were interestingly not the periods with the lowest correlations coefficients in the Google data.



**d, Cumulative weekly correlation with CDC-reported data, with the Google week-47 data removed:** Overall Google correlations were improved by removing the Thanksgiving holiday data points (arrows). The DiSTRIBuTE data, however, remained more highly correlated than the Google estimates during the 2006-2007 season through week 2007-13, and during the 2007-2008 season for the weeks 2004-47 to 2008-03.



**e, Moving 10-week window correlations, with the Google week-47 data removed.** Google correlations early each season were improved by removing the Thanksgiving holiday data points (arrows). The DiSTRIBuTE data, however, remained more highly correlated during 2006-2007, through week 2007-10, and during 2007-2008 for weeks 2007-47 to 2008-05.