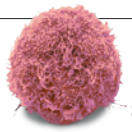
NEWS IN FOCUS

GENETIC DIAGNOSIS Data barriers hamper search for meaning in mutations **p.156**

FUNDING US science agencies gird themselves for the budget axe **p.158**

MALARIA Plant source of key drug faces lab-made competition **p.160**



BIOMEDICINE A Texas-style showdown over stem-cell therapy **p.166**

JOHN ANGELILLO/UPI/NEWSCOM

The latest US influenza season is more severe and has caused more deaths than usual.

EPIDEMIOLOGY

When Google got flu wrong

US outbreak foxes a leading web-based method for tracking seasonal flu.

BY DECLAN BUTLER

When influenza hit early and hard in the United States this year, it quietly claimed an unacknowledged victim: one of the cutting-edge techniques being used to monitor the outbreak. A comparison with traditional surveillance data showed that Google Flu Trends, which estimates prevalence from flu-related Internet searches, had drastically overestimated peak flu levels. The glitch is no more than a temporary setback for a promising strategy, experts say, and Google is sure to refine its algorithms. But as flu-tracking techniques based on mining of web data and on social media proliferate, the episode is a reminder that they will complement, but not substitute for, traditional epidemiological surveillance networks.

"It is hard to think today that one can provide disease surveillance without existing systems," says Alain-Jacques Valleron, an epidemiologist at the Pierre and Marie Curie University in Paris, and founder of France's Sentinelles monitoring network. "The new systems depend too much on old existing ones to be able to live without them," he adds.

This year's US flu season started around November and seems to have peaked just after Christmas, making it the earliest flu season since 2003. It is also causing more serious illness and deaths than usual, particularly among the elderly, because, just as in 2003, the predominant strain this year is H3N2 — the most virulent of the three main seasonal flu strains.

Traditional flu monitoring depends in part on national networks of physicians who report cases of patients with influenza-like illness (ILI) — a diffuse set of symptoms, including high fever, that is used as a proxy for flu. That estimate is then refined by testing a subset of people with these symptoms to determine how many have flu and not some other infection.

With its creation of the Sentinelles network in 1984, France was the first country to computerize its surveillance. Many countries have since developed similar networks — the US system, overseen by the Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, includes some 2,700 health-care centres that record about 30 million patient visits annually.

But the near-global coverage of the Internet and burgeoning social-media platforms such as Twitter have raised hopes that these technologies could open the way to easier, faster estimates of ILI, spanning larger populations.

The mother of these new systems is Google's, launched in 2008. Based on research by Google and the CDC, it relies on data mining records of flu-related search terms entered in Google's search engine, combined with computer modelling. Its estimates have almost exactly matched the CDC's own surveillance data over time — and it delivers them several days faster than the CDC can. The system has since been rolled out to 29 countries worldwide, and has been extended to include surveillance for a second disease, dengue.

Google Flu Trends has continued to perform remarkably well, and researchers in many countries have confirmed that its ILI estimates are accurate. But the latest US flu season seems to have confounded its algorithms. Its estimate for the Christmas national peak of flu is almost double the CDC's (see 'Fever peaks'), and some of its state data show even larger discrepancies.

It is not the first time that a flu season has tripped Google up. In 2009, Flu Trends had to tweak its algorithms after its models badly underestimated ILI in the United States at the start of the H1N1 (swine flu) pandemic — a glitch attributed to changes in people's search behaviour as a result of the exceptional nature of the pandemic (see http://doi.org/djw73f).

Google would not comment on this year's ▶

NATURE.COM
See maps showing reports of flu-like symptoms in France: go.nature.com/w954hn

difficulties. But several researchers suggest that the problems may be due to widespread media coverage of this year's severe US flu season, including the declaration of a public-health emergency by New York state last month. The press reports may have triggered many flu-related searches by people who were not ill. Few doubt that Google Flu will bounce back after its models are refined, however.

"You need to be constantly adapting these models, they don't work in a vacuum," says John Brownstein, an epidemiologist at Harvard Medical School in Boston, Massachusetts. "You need to recalibrate them every year."

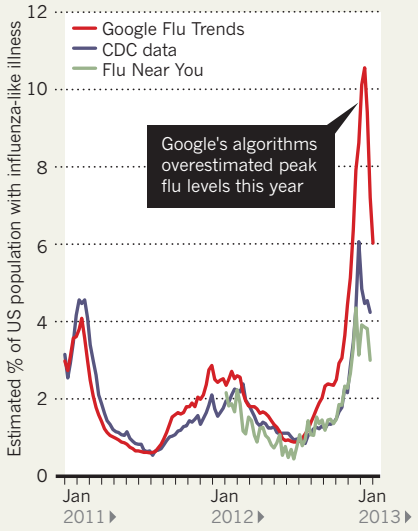
Brownstein is one of many researchers trying to harness the power of the web to establish sentinel networks made up not of physicians, but of ordinary citizens who volunteer to report when they or someone in their family are experiencing symptoms of ILI. 'Flu Near You', a system run by the HealthMap initiative co-founded by Brownstein at Boston Children's Hospital, was launched in 2011 and now has 46,000 participants, covering 70,000 people.

Similar systems are springing up in Europe. For example, GrippeNet.fr, run by French researchers in collaboration with national health authorities, has attracted more than 5,500 participants since its creation a year ago, with 60–90 people joining each week.

Lyn Finelli, head of the CDC's Influenza Surveillance and Outbreak Response Team, feels that such crowdsourcing techniques hold great promise, especially because the questionnaires are based on clinical definitions of ILI and so yield very clean data. And both Flu Near You and GrippeNet.fr have a representative age distribution of participants. The CDC has worked with Flu Near You on its development, and Finelli herself has signed up: "I submit my family's data every week," she says.

FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.



SOURCES: GOOGLE FLU TRENDS (WWW.GOOGLE.ORG/FLUTRENDS); CDC; FLU NEAR YOU

Other researchers are turning to what is probably the largest publicly accessible alternative trove of social-media data: Twitter. Several groups have published work suggesting that models of flu-related tweets can be closely fitted to past official ILI data, and various services, such as MappyHealth and Sickweather, are testing whether real-time analyses of tweets can reliably assess levels of flu.

But Finelli is sceptical. "The Twitter analyses have much less promise" than Google Flu or Flu Near You, she says, arguing that Twitter's signal-to-noise ratio is very low, and that the most active Twitter users are young adults and so are not representative of the general public.

Michael Paul, a computer scientist at Johns Hopkins University in Baltimore, Maryland, disagrees. He is part of a team that is developing Twitter-based disease monitoring, and says that Google search-term data probably have just as much noise. And although Internet-based surveys may boast less noise, their smaller size means that they may be prone to sampling errors. "I suspect that passive monitoring of social media will always yield more data than systems that rely on people to actively respond to surveys, like Flu Near You," Paul says.

To reduce the noise, the Johns Hopkins team has recently analysed a subset of a few thousand flu-related tweets, looking for patterns indicating which tweets showed that the tweeter was actually ill rather than simply, say, pointing to news articles about flu. They then used this information to retrain their models to weed out irrelevant flu-related tweets. Paul says that a paper in press will show that this greatly improves their results.

Already, web data mining and crowdsourced tracking systems are becoming a part of the flu-surveillance landscape. "I'm in charge of flu surveillance in the United States and I look at Google Flu Trends and Flu Near You all the time, in addition to looking at US-supported surveillance systems," says Finelli. "I want to see what's happening and if there is something that we are missing, or whether there is a signal represented somewhat differently in one of these other systems that I could learn from." ■

MEDICINE

Data barriers limit genetic diagnosis

Tools for data-sharing promise to improve chances of connecting mutations with symptoms of rare diseases.

BY ERIKA CHECK HAYDEN

For the first five months of Harrison Harkins' life, doctors had little idea about what was causing his spinal malformation and inability to gain weight. But in November 2011, Matthew Bainbridge, a computational biologist at Baylor College of Medicine in Houston, Texas, found a clue. After analysing genetic data from Harrison and his parents, Bainbridge discovered that the child had an abnormal version of a gene called *ASXL3*.

But Bainbridge had no easy access to records of other children with *ASXL3* mutations, and could not be sure that this mutation was the culprit. So he did what many scientists do: he networked. A Dutch team put Bainbridge in touch with German researchers who were treating another boy with an *ASXL3* mutation — and symptoms similar to Harrison's. After finding two further cases in an internal Baylor database, Bainbridge felt that the connection was concrete. He describes the syndrome seen in all four children, and probably caused by *ASXL3* mutations, in a paper published on 5 February (M. N. Bainbridge *et al. Genome Med.* **5,** 11; 2013).

Researchers are using new tools to increase the pace of discoveries such as Bainbridge's. Efforts to connect sequences with symptoms — or in genetic parlance, genotype with phenotype — have taken on increased urgency as clinical sequencing gains traction and funders put more money towards rare diseases. Researchers are planning to address the barriers to data sharing at a workshop in April, after the first International Rare Diseases Research Consortium Conference in Dublin. "There is a very positive feeling in the community that things are changing for the better," says Peter Robinson, a computational biologist at the Charity University Hospital in Berlin.

Thousands of people have had their genomes sequenced, but a reluctance to surrender ownership of the valuable data, along with the privacy concerns of researchers and families (see 'Families find solace in