ISDS
INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

# Wikipedia Usage Estimates Prevalence of Influenza-like Illness in Near Real-Time

David McIver*[1, 2] and John S. Brownstein[1, 2]

[1]Boston Children's Hospital, Boston, MA, USA; [2]Harvard Medical School, Boston, MA, USA

## Objective

The purpose of this work was to develop a novel method of estimating the amount of influenza-like illness (ILI) in a population, in near-real time, by using a source of information that is completely open to the public and free to access. We investigated the usefulness of data gathered from Wikipedia to estimate the prevalence of ILI in the United States, using data from the Centers for Disease Control and Prevention (CDC) as well as Google Flu Trends.

## Introduction

Each year, there are an estimated 250,000–500,000 deaths worldwide that are attributed to seasonal influenza, with anywhere between 3,000–50,000 deaths occurring in the United States of America (US). In the US, the Centers for Disease Control and Prevention (CDC) continuously monitors the level of influenza-like illness (ILI) circulating in the population. While the CDC ILI data is considered to be a useful indicator of influenza activity, its availability has a known lag-time of between 7–14 days. To appropriately distribute vaccines, staff, and other healthcare commodities, it is critical to have up-to-date information about the prevalence of ILI in a population.

To this end, we have created a method of estimating current ILI activity in the US by gathering information on the number of times particular Wikipedia articles have been viewed. Not only is the information held within Wikipedia articles very useful on its own, but statistics and trends surrounding the amount of usage of particular articles, frequency of article edits, region specific statistics, and countless other factors make the Wikipedia environment an area of interest for researchers. Furthermore, Wikipedia makes all of this information public and freely available, greatly increasing and expediting any potential research studies that aim to make use of their data.

## Methods

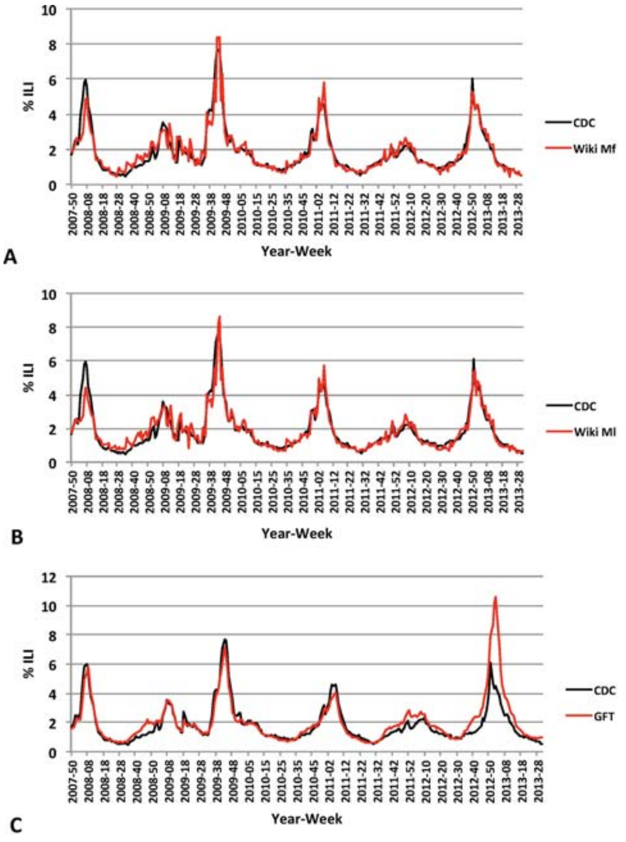
Data was collected on how often, with results up to the hour, selected Wikipedia articles were viewed. This information was compared to both official CDC ILI data and Google Flu Trends GFT data. A Poisson model was created which estimated ILI prevalence using all of the selected Wikipedia articles, and a separate model was produced using Lasso regression that automatically and dynamically selects the most appropriate Wikipedia articles to fit the data. A split-sample analysis was used to test the reliability of the Lasso model, comparing half of the data representing the 2007-2010 flu seasons to the second half representing the 2011-2013 flu seasons.

## Results

Our Wikipedia-based Poisson model accurately estimates the level of ILI activity in the American population, up to two weeks ahead of the CDC, with an absolute average difference between the two estimates of just 0.27% over 294 weeks of data. Wikipedia-derived ILI models performed well through both abnormally high media coverage events (such as during the 2009 H1N1 pandemic) as well as unusually severe influenza seasons (such as the 2012–2013 influenza season). Wikipedia usage accurately estimated the week of peak ILI activity 17% more often than Google Flu Trends data and was often more accurate in its measure of ILI intensity.

## Conclusions

This study is unique in that it is the first scientific investigation into the harnessing of Wikipedia usage data over time to estimate the burden of disease in a population. The application of Wikipedia article view data has been demonstrated to be effective at estimating the level of ILI activity in the US, when compared to CDC data. Wikipedia article view data is available daily (and hourly, if necessary), and can provide a reliable estimate of ILI activity up to 2 weeks in advance of traditional ILI reporting. This study exemplifies how non-traditional data sources may be tapped to provide valuable public health related insights and, with further improvement and validation, could potentially be implemented as an automatic sentinel surveillance system for any number of disease or conditions of interest as a supplement to more traditional surveillance systems.



## Keywords

Surveillance; Influenza; Digital Disease Detection

*David McIver
E-mail: david.mciver@childrens.harvard.edu