Using Wikipedia and Google data to estimate near real-time influenza incidence in Germany

Paul Schneider, *Maastricht University, Netherlands Institute of Health Service Research*
John Paget, *Netherlands Institute of Health Service Research*
Peter Spreeuwenberg, *Netherlands Institute of Health Service Research*
David Barnett, *Maastricht University*
Christel van Gool, *Maastricht University*

Contact: schneider.paulpeter@gmail.com

Introduction

Traditional influenza surveillance systems are costly and involve considerable delay between disease onset and reporting. Previous studies have demonstrated that it is possible to predict the incidence of influenza from relevant Google search queries and Wikipedia page view statistics: The traces of people seeking online health information can be harnessed to monitor influenza activity in near real-time ('Nowcasting'). Until now, research has focused almost exclusively on the US. Here we present our early success in developing a Nowcast–model for influenza in Germany.

Methods

Weekly influenza incidence for Germany was collected from the Robert Koch Institute; and data on 535 potential predictors were identified and retrieved from Wikipedia and Google Correlate/Trends. We split the dataset into a model training set (2010/11 –2015/16), and a validation set (2016/17) to which the algorithms are naïve. Machine-learning algorithms (Elastic net, support vector machine, tree-based methods and others), in combinations with time-series cross-validation, were used for parameter selection and model training. We compared prediction accuracy within the validation dataset against a forecast model.

Findings

The lasso regression model showed the highest prediction accuracy within the training dataset (CV root-mean-squared-error (RMSE) = 1.107; R2 =0.70). Within the validation dataset, the model performed better than the forecast model (RMSE = 3.29 vs. 4.02; R2 = 0.80 vs. 0.58).

Discussion

Our model would have been able to accurately predict the 2016/2017 influenza season in Germany in near real-time. In the future, Nowcasting could provide inexpensive and rapid influenza surveillance, affordable for developing countries and useful for near-real-time epidemic tracking.