Article: Copeland, Patrick, et. al. “Google Disease Trends: An Update”. International Society of Neglected Tropical Diseases 2013. Page 3.

Google Inc. ran a project called Google Flu Trends (GFT) which used search data to predict a daily estimate of the occurrence of flu two weeks prior to an official publication/survey. The primary purpose of the Google Flu Trends was to forecast the start and peak of the flu season. The secondary purpose is to determine magnitude of flu-like illnesses. Google Flu Trends has been surprisingly accurate from 2008 to 2013. Its predicted flu occurrence was similar to the actual occurrence of flu as surveyed by the Center for Disease Control. However, in January 2013, Google Flu Trends overestimated over twice the CDC reported incidence. The research problem and goal of this article was to determine why the last season’s predictions were so high and what improvements could be made to the model used.

The hypothesis that was formed was that Google Flu Trend search queries were highly susceptible to media bias. In other words, when the media reported incidents of flu, people began to use search keywords associated with flu.

Google Flu Trends sought to compensate for these sudden changes in search volume. They believed that media driven flu searches were short term – period of 3 to 7 days – and these “inorganic” results could be removed from the model. The system must use times series data and validate if the latest results are within expected variance from historical data.

Google Flu Trends did not re-evaluate its model after every flu season because the results were better than expected for subsequent years after 2008. It admitted that if they reviewed their model they would have improved prediction results for 2013.

The proposed approach is to use an ElasticNet model and to dampen media spikes. There was post processing validation in that these models and additional factors did bring the expected results close to actual CDC results. However, Google Flu Trends ended soon after the 2013 season so the proposed solution could not be actively tested against future data.

The study determined that the proposed solution would not have been very effective against prolonged media bias over an entire season as it would be one “continuous spike” instead of a 3 to 7 day spike. By both dampening the media spikes and using a different regression model such as the ElasticNet, Google Flu Trends would have still over predicted the 2013 flu levels.

*Critique the study. What was done well and what could be improved? What is the future of the research?*