Quantifying the relationship between diseases and symptoms using big data

Shih Han Lin¹,a, Hsin Hui Shao²,b

¹No.3 Keelung Road Sec.4 Taipei Taiwan

² No. 92, Sec. 2, Zhongshan N. Rd., Taipei City 10449, Taiwan

¹bonbonshan@gmail.com, bchocosophi@gmail.com
Abstract

Background/Problem/Objective
Crises in endemic transmitted diseases affect humans worldwide, and the symptoms these diseases cause may provide firsthand information about these disorders. We suggest that massive new data sources resulting from human interaction with the Internet may offer a unique perspective on the relationship between illness and symptoms.

Methods/Intervention
By analyzing changes in Google query volumes for search terms related to disease, we find a pattern that may define the relationship between symptoms and disorders. We first retrieved pattern data from Google Trend using the common cold as the primary disease, and sore throat, stuffy nose, sneeze, fever, cough, and headache as symptoms. Pearson's correlation coefficient was calculated using SPSS to determine the relationship between the symptoms and the disease.

Results (of evaluation)
Data created since 2013/1/13 was retrieved from Google Trend on a weekly basis. A total of 261 sets of data were calculated to create a high correlation coefficient of 0.925 between the common cold and the stuffy nose symptom. The cough symptom has the second highest correlation coefficient of 0.925, sore throat has a correlation coefficient of 0.853, and fever has a correlation coefficient of 0.626, which was significant at the 0.01 level in a two-tailed test.

Conclusions/Lessons learned.
Data on the relationship between diseases and symptoms often comes from facilities such as government, hospitals, and clinics, where the data is collected through the documentation of physicians and nurses. A conventional study can be limited by the region, the number of patients and the interpretation of the specialist. However, with access to Google Trend's big data, millions or even billions of data points are accumulated directly from the patient. Another contribution of this study is that the quantified relationship between symptoms and diseases can be used to educate future physicians or even artificial intelligence.

Keywords: big data, symptoms, diseases, relationship, GFT
Introduction

The cumulative sizes of big data indicates various facets of our daily activities point out a fundamental opening for experts to tackle basic questions about the complicated world we live in. (King, 2011) [1].

Every day, large numbers of people around the world use search engines, which create a whole new data source (Hulth, Rydevik, & Linde, 2009) [2]. Google Trends (GT) [3] is a search engine provided by Google for Internet users to follow the volume of search queries. GT provides information on how often a particular search term is entered relative to the total search volume in the world, and this information is available in different languages. Using GT, it is possible to demonstrate the dynamic situation of Internet health-seeking behaviors. (Carneiro & Mylonakis, 2009) [4].

Google Flu Trends is an algorithm assessing billions of Internet search queries from Google users from numerous regions. Having been found to correlate well with reported influenza, it can improve influenza prediction models (Klembczyk et al., 2016) [5].

It is possible to view medical decisions and diagnoses along a spectrum, with categorical reasoning at one end and probability reasoning at the other (Szolovits & Pauker, 1978) [6]. In this process, the relationship between symptoms and diseases plays an important role at every step. Complex symptoms scale such as depression and anxiety scale is often used to screen and diagnose patients (Zigmond & Snaith, 1983) [7]. However, not much research has sought to quantify the relationship between a single symptom and disease. Our goal is to use big data to picture the actual relationship between symptoms and diseases.

Data on the relationship between diseases and symptoms often come from facilities like governments, hospitals, and clinics, where they are documented by physicians and nurses. These data are usually limited by region, number of patients, and the interpretation of the specialists. “Observed relationships between symptoms and diseases can be biased by consultation, disease verification, and referral patterns” (Knottnerus, 1987) [8].

Method
By analyzing changes in Google query volumes for search terms related to diseases, we found a pattern that may define the relationship between symptoms and diseases.

We first pulled out pattern data from GT using the common cold as the disease, and sore throat, stuffy nose, sneezing, fever, cough, and headache as its symptoms. Data from January 13, 2013 was pulled on a weekly basis. Pearson’s correlation coefficient was calculated using SPSS to determine the relationship between symptoms and disease.

Results
A total of 261 sets of data were collected. The results showed a high correlation coefficient of 0.925 between the common cold and stuffy nose. Cough had the second highest correlation coefficient of 0.925 with the common cold. Sore throat had a correlation coefficient of 0.853, and fever 0.626. These correlations were significant at the 0.01 level in the two-tailed test.

Fig 1. Scatterplot of common cold vs cough
Fig 2. GT interest in the common cold, nasal congestion, cough, sore throat, and fever since 2013.

Another set of data from January 13, 2013 was pulled from GT on a weekly basis. A total of 260 sets of data were calculated to indicate a correlation coefficient of 0.765 between influenza and stuffy nose. Cough had the highest correlation coefficient of 0.775. Sore throat had a correlation coefficient of 0.650, and fever 0.618. These correlations were significant at the 0.01 level in the two-tailed test.

Fig 3. GT interest in influenza, nasal congestion, cough, sore throat, and fever since 2013.

GT data started since 2004, and if we look at Fig 4, we can see that the amount of data for the first five years was relatively little when compared to last year. Thus, the correlation might be different when comparing different time segments.
Fig 4. GT interest in the common cold, nasal congestion, cough, sore throat, and fever since 2004.

Due to the H1N1 flu pandemic from 2009 to 2010, the affected time zone experienced an enormous spike in searches. This may have resulted in deviations when calculating the correlation between symptoms and flu.

Fig 5. GT interest in influenza, nasal congestion, cough, sore throat, and fever since 2004.

Table 1. Correlation coefficients between symptoms and diseases.

<table>
<thead>
<tr>
<th></th>
<th>Nasal congestion</th>
<th>Cough</th>
<th>Sore throat</th>
<th>Fever</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common cold</td>
<td>2004-2018</td>
<td>.938**</td>
<td>.929**</td>
<td>.872**</td>
</tr>
<tr>
<td></td>
<td>2008-2013</td>
<td>.892**</td>
<td>.887**</td>
<td>.778**</td>
</tr>
<tr>
<td></td>
<td>2013-2018</td>
<td>.925**</td>
<td>.925**</td>
<td>.853**</td>
</tr>
<tr>
<td>Influenza</td>
<td>2004-2018</td>
<td>0.019</td>
<td>0.065</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td>2008-2013</td>
<td>0.001</td>
<td>-0.088</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>2013-2018</td>
<td>.765**</td>
<td>.775**</td>
<td>.650**</td>
</tr>
</tbody>
</table>

(**With correlation is significant at the 0.01 level in the two-tailed test. *With correlation is significant at the 0.05 level in the two-tailed test.)

Discussion
Principle results
Symptoms are an important element in the field of medicine. They are used as firsthand information gathered from patients without any need for procedures like physical examinations, blood tests, and radiology that might cause them physical or mental discomfort.
The correlation between symptoms and diseases is a key factor in diagnosis; numerous highly correlated symptoms can lead to a specific diagnosis.
To quantify the relationship between symptoms and diseases, most earlier studies relied on hospitals, which are restricted by region and amount of data. Using GT's big data, a huge amount of information is available for us to calculate the correlation between symptoms and diseases.

Limitations
This method, however, may have some restrictions. In the event of a pandemic outbreak, a disease occurs in greater numbers than expected in a community or region or during a season. Online queries regarding the disease may increase tremendously, causing an imbalance in the correlation, leading to an inaccurate correlation coefficient. This might have been the case with the H1N1 outbreak.
Another restriction might be the type of disease. In this study, we have used the common cold and flu, which are both seasonal infectious diseases, making it easier to observe their correlation with symptoms. However, it might be a challenge to observe diseases with stable prevalence rates.

Conclusions
We used GT's big data to quantify the relationship between the common cold and its symptoms. Based on the past five years of data, cough and nose congestion both have a high correlation coefficient of 0.925 to the common cold. Fever, however, has a slightly lower correlation coefficient of 0.626.
Information about symptoms is crucial to diagnosing diseases, and by being able to quantify the relationship between symptoms and diseases, the education of physicians as well as the future of artificial intelligence can be improved.
Acknowledgments and Conflicts of Interest
References


