International Journal of Business and Economics, 2017, Vol. 16, No. 1, 101-108

Awareness of Sexually Transmitted Disease and Economic Misfortune Using Search Engine Query Data

Daniel Farhat*

Department of Economics, Radford University, U.S.A.

Abstract

This brief study extracts a measure of general interest in sexually transmitted disease for the United States (2004 – 2012) using Google keyword search data. Trends in this measure are then compared to periods of decline in the US economy.

Key words: sexually transmitted disease awareness; health and the economy; search engine query data

JEL classification: I1; Z1

It is now easy to learn about ailments, treatment options, and specialist care without ever stepping foot in a clinic. The ability to quickly and anonymously self-diagnose symptoms using any number of search engines and medical websites can reduce treatment costs for both patients and providers while simultaneously improving the overall quality of healthcare.[1] For a particular class of afflictions, '*anonymous'* plays a key role. Sexually transmitted diseases [STDs] are stigmatized illnesses; people often feel embarrassed to discuss them as they are indicative of 'careless' or 'shameful' behavior. Thus, some remain ignorant of symptoms and prevention leading to greater transmission. Others delay seeking medical attention, making treatment and eradication efforts more extensive and expensive. The internet allows free and private access to STD-related information on demand, potentially mitigating these costs.

Conveniently, the keywords used in internet searches are often recorded. *Google Trends*,[2] for example, is an online tool that shows how often a particular keyword has been searched for on the Google search engine (relative to all Google searches) over time. In the area of health, this data has already been shown to be useful for predicting epidemiological events such as flu (Ginsberg et al., 2008; Polgreen et al., 2008; Carneiro and Mylonakis, 2009) and HIV infection (Jena et al., 2013). Although this data cannot distinguish between those searching for specific information out of necessity (e.g. to self-diagnose a condition) and those searching

*Correspondence to: PO Box 6952, Radford, VA 24142, United States, tel: +1 (540) 831-5191, email: dfarhat@radford.edu.

out of curiosity (e.g. to write an essay for a high school health class), it does provide a measure of general interest (Ripberger, 2011). Given the stigma associated with STDs and that a fraction of searchers are in fact interested in self-diagnosis, using this sort of data to craft a metric for STD awareness seems particularly valuable for health analytics and for helping policymakers use their ever-tightening medical budgets in a more efficient manner.

To create this STD-awareness index for the US, we can use Google Trends data for 9 STD keywords: chlamydia, genital warts, gonorrhea, hepatitis, herpes, HIV, HPV, syphilis, and trichomoniasis. The data is restricted to the theme of "Health". Weekly search volume data (January 2004 to January 2013) is collected and averaged into monthly frequency. An aggregate measure is constructed computationally by identifying a sequence, $S = \{s_{2004m1}, s_{2004m2}, ..., s_{2013m2}\}$, that maximizes the summed correlation (in absolute value) between itself and each of the 9 individual series.[3]

The contemporaneous correlation between S and its components is shown to be strong[4] for the entire sample period (see Table 1). When the sample is divided into three sub-periods (Early: 2004m1 – 2007m12; Middle: 2008m1 – 2009m5; Late: 2009m6 – 2013m2), the relationship between S and each of the STD series remains strong (particularly for chlamydia, gonorrhea, herpes, HIV, syphilis and trichomoniasis). Although the correlation between S and three of the STD series does weaken[5] slightly when the sample is divided, it can be argued that S is a reasonable proxy for overall STD awareness amongst internet-using Americans.

Alone, this STD interest index has several salient properties. Table 3 (column 1) shows that S exhibits persistence ($corr[s_t,s_{t-1}]$ is strong). Table 3 also shows two monthly cycles per year in STD interest (columns 2 – 3): (1) a small peak in November followed by a trough in December and (2) a large peak in April followed by a deep trough in August. When the monthly pattern is removed, we see that S trended downward from 2004 to 2007, followed by an upswing from 2008 to the present. The turning point in S coincides with the onset of the most recent economic downturn (December 2007).

Table 1. Correlation: Common Trend (S) and Google Trends Search Indices for 9 STDs*

Correlation with Common STD Trend (S)	Full 2004m1 - 2013m2	Early 2004m1 - 2007m12	Mid 2008m1 - 2009m5	Late 2009m6 - 2013m2
Chlamydia	0.684	0.721	0.864	0.920
Genital warts	0.453	0.901	-0.01	0.779
Gonorrhea	0.610	0.717	0.860	0.670
Hepatitis	0.302	0.866	0.386	0.282
Herpes	0.656	0.511	0.531	0.882
HIV	0.419	0.869	0.641	0.380
HPV	-0.515	-0.686	-0.061	0.302
Syphilis	0.855	0.867	0.922	0.798
Trichomoniasis	0.544	0.611	0.818	0.559
95% Confidence Bands	±0.191	±0.289	±0.485	±0.298

* Google Trends (2013).

Table 2. Correlation: Common Economic Trend (E) and 12 Economic Variables

Correlation with Common Economic Trend (E)	Full 2004m1 - 2012m12	Early 2004m1 - 2007m12	Mid 2008m1 - 2009m5	Late 2009m6 - 2012m12
Unemployment Rate*	0.809	0.365	0.815	0.529
Quits**	-0.834	-0.704	-0.851	-0.626
Job Openings**	-0.907	-0.826	-0.907	-0.842
Hires**	-0.769	-0.641	-0.794	-0.547
Average Weekly Hours***	-0.673	-0.420	-0.663	-0.686
Average Hourly Earnings***	0.642	-0.397	0.919	-0.690
Growth Rate of Average Hourly Earnings***	-0.289	-0.515	-0.146	-0.419
Dow Jones Index†	-0.471	-0.314	-0.955	-0.692
S&P 500 Index†	-0.626	-0.359	-0.966	-0.686
St. Louis FRB Financial Stress Indicator†	0.694	0.209	0.888	0.476
CPI Inflation Rate††	-0.370	-0.349	-0.566	-0.251
Total Retail Sales†††	0.143	-0.016	-0.529	-0.249
95% Confidence Bands	±0.192	±0.289	±0.485	±0.305

* Bureau of Labor Statistics (2013c).
** Bureau of Labor Statistics (2013d).
*** Bureau of Labor Statistics (2013b). Production and non-supervisory employees.
† Federal Reserve Bank of St. Louis (2013).
†† Bureau of Labor Statistics (2013a).
††† United States Census Bureau (2013).

Table 3. OLS Regressions, Full Sample (2004m1 – 2013m2)

Dependent variable:		s_t	
	(1)	(2)	(3)
constant	12.548***	—	—
s_{t-1}	0.846***	—	0.922***
January	—	81.543***	8.673***
February	—	83.097***	7.940**
March	—	83.339***	6.988**
April	—	86.405***	9.592***
May	—	85.599***	5.961*
June	—	78.689***	-0.206
July	—	78.292***	5.766*
August	—	76.880***	4.719
September	—	79.266***	8.408***
October	—	81.209***	8.151**
November	—	83.284***	8.434**
December	—	78.522***	1.761
R^2	0.716	0.177	0.879

* Significant at the 10% level.
**Significant at the 5% level.
*** Significant at the 1% level.

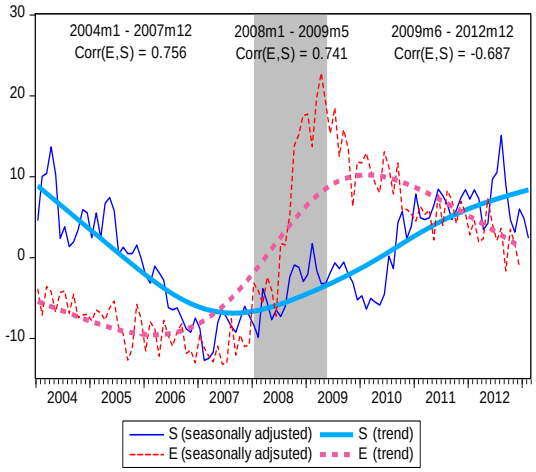
The timing appears consistent with other studies. Recent research shows that keyword search for health-related topics surged during the most recent economic downturn (Askitas and Zimmerman, 2011; Tefft, 2011; Farhat and Viiten, 2015). To further evaluate this for STD interest, a second composite measure is crafted to represent broad economic performance. A sequence, $E = \{e_{2004m1}, e_{2004m2}, ..., e_{2012m12}\}$, is created in the same manner as S using data on labor market outcomes, financial status, and spending.[6] Table 2 describes the relationship between E and its 12-component series. E is positively correlated with the 'bads' (unemployment and financial stress) and negatively correlated with the 'goods' (quit rates, job opening rates, hiring rates, average weekly working hours, the growth rate of average earnings and stock market indices) which suggests that E measures 'economic misfortune'.[7] As expected, E rises sharply from December 2007 to June 2009: the recent recession as dated by the NBER (2013).

To compare movements in E and S, we look at both the de-seasoned data (omitting monthly patterns) and a long-term trend extracted using the Hodrick-Prescott filter (see Figure 1). Economic misfortune increases in early 2007 and STD rises in mid- to late-2007: the movements of the two series appear positively correlated with STD interest lagging behind the state of the economy during the Great Recession.

We can explore other correlates. Figure 2 compares the de-seasoned series S with seasonally adjusted keyword search trends for 'cheap insurance' and 'free clinic' (i.e. the relationship between STD interest and people's attempts to find ways around expensive doctors visits). No evident pattern exists between 'free clinic' and S, but 'cheap insurance' exhibits the same trend as S with an apparent lead. This pattern supports the notion that self-diagnosis occurs in greater frequency *after* insurance alternatives are explored, and that insurance interest is a leading factor for policymakers to monitor. Figure 3 shows the relationship between the measure S and two STD avoidance methods, 'abstinence' and 'condoms' (i.e. the relationship between STD interest and cost-saving prevention). Interest in 'abstinence' exhibits no clear relationship to S while searches for 'condoms' exhibit the same trend as S but with a lag. This might provide policymakers with a target: to make prevention a forethought, not an afterthought.

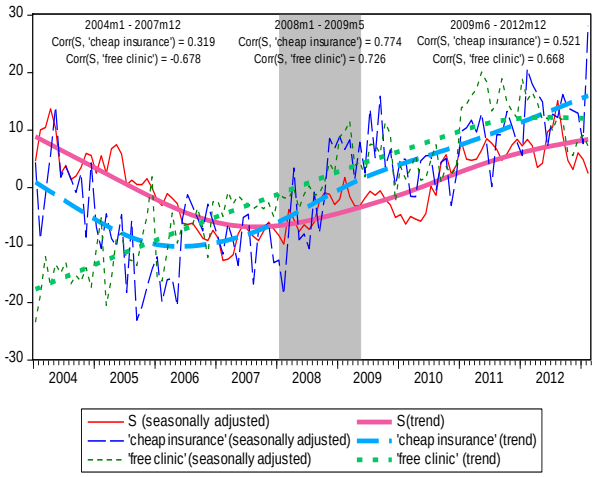
Using search engine query data to monitor public interest in health, particularly for stigmatized ailments like STDs, may prove key to forming more effective health policies. The analysis here is meant to be exploratory; correlations between STD awareness, the state of the economy, and select cost-related factors are identified by not explained. Testing potential causal links between the two is a fruitful area of future research. More elaborate studies are called for. Linking this interest to additional economic, social, and cultural trends will likely provide further insights into how internet users can become healthier people.

Figure 1. STD Interest (S) and Economic (E) Trends*, Seasonally Adjusted: Monthly, January 2004 – December 2012**



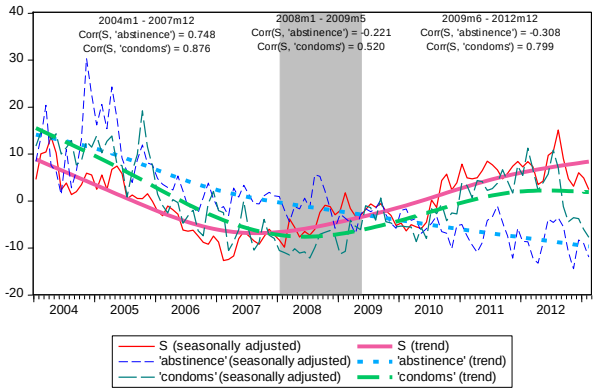
* Trend extracted using the Hodrick-Prescott Filter.
** Seasonally adjusted using monthly dummy variables.

Figure 2. STD Interest (S), 'Cheap Insurance'* and 'Free Clinic'* Trend, Seasonally Adjusted***: Monthly, January 2004 – December 2012**



* Google Trends (2013).
** Trend extracted using the Hodrick-Prescott Filter.
*** Seasonally adjusted using monthly dummy variables.

Figure 3. STD Interest (S), 'Abstinence'* and 'Condoms'* Trends, Seasonally Adjusted***: Monthly, January 2004 – December 2012**



* Google Trends (2013).
** Trend extracted using the Hodrick-Prescott Filter.
*** Seasonally adjusted using monthly dummy variables.

Notes

1. See Lanseng and Andreassen (2007) and Ryan and Wilson (2008) for descriptions and insights on the costs and benefits of online health information.
2. Google Trends is a web-based data retrieval tool (available at www.google.com/trends/). For any selected term (or terms), Google Trends will provide an estimate of the number of searches for that term relative to the total number of searches done on the Google search engine for each week (in some cases, month) from January 2004 to the present. The data is rescaled, so the period with the largest relative search volume is indexed to 100. Google Trends categorizes the search data by location, allowing us to focus our attention on interest from a particular geographic area. The data is also categorized by theme so that the meaning behind the searches can be refined (for example, people who search for "drugs" may either be looking to buy them or for medical information about them; we can look at searches related to "Shopping" and "Health" separately). Additional information about Google Trends is available at support.google.com/trends/.
3. MATLAB, a program designed for numerical analyses, is used. Denoting the search series in Figure 1 as X_i, i = 1...9, MATLAB computes S to minimize a function $F = -(\Sigma_i |\rho(S,X_i)|)$ where $\rho(S,X_i)$ is the sample correlation between S and X(i). Note that each of the 9 series is assumed to be equally important in the derivation of S and that either positive or negative correlations between S and X_i are deemed relevant. There are many more sophisticated methods for identifying common trends in time series data. More elaborate techniques are left for future research.
4. I.e., are non-zero as measured by the 2-standard error (95% confidence) bounds computed as $2/\sqrt{T}$.
5. The correlation between S and hepatitis becomes insignificant in the later periods, while the correlation between S and genital warts is high only in the early and late periods. The correlation between S and HPV switches from significantly negative in the early period to significantly positive

in the later period.

6. Data for average weekly earnings and average weekly labor hours are for production and non-supervisory employees. Monthly data on CPI inflation, unemployment, average weekly earnings, percent change in average weekly earnings, average weekly labor hours, hires, quits, and job openings are collected from the United States Bureau of Labor Statistics (2013a-d). Estimates of retail sales are provided by the United States Census Bureau (2013). Indicators of stock market activity (S&P 500 and the Dow Jones Industrial Index) and financial stress are obtained from the St. Louis Federal Reserve Bank FRED Database (2013). None of the data is seasonally adjusted to be consistent with the data from Google Trends.
7. Three anomalies appear. E is positively correlated with total average hourly earnings, perhaps due to overall persistent growth in earnings coupled with the sharp rise in E from 2007 to 2009. E is negatively correlated with CPI inflation. While a large amount of inflation is undesirable, low and stable inflation (which characterizes the US economy from 2004 – 2012) is not overtly harmful. The relationship between E and total retail sales is positive but not significant at the 95% level for the entire sample period. However, if we divide the sample, a negative relationship appears.

References

Askitas, N. and K. F. Zimmermann, (2011), "Health and Well-being in the Crisis," *IZA DP 5601*.

Bureau of Labor Statistics [BLS], (2013a), "Consumer Price Index," www.bls.gov.

Bureau of Labor Statistics [BLS], (2013b), "Current Employment Statistics Survey," www.bls.gov.

Bureau of Labor Statistics [BLS], (2013c), "Current Population Survey," www.bls.gov.

Bureau of Labor Statistics [BLS], (2013d), "Job Openings and Labor Turnover Survey," www.bls.gov.

Carneiro, H. A. and E. Mylonakis, (2009), "Google Trends: A Web-based Tool for Real-time Surveillance of Disease Outbreaks," *Clinical Infectious Diseases*, 49(10), 1557-1564.

Farhat, D. and T. Viiten, (2015), "Business 'Psych'cles: A Close Look at Mental Health across US States during the Great Recession," *Working Paper*.

Federal Reserve Bank of St. Louis, (2013), Federal Reserve Economic Data [FRED], research.stlouisfed.org/fred2.

Ginsberg J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, (2008), "Detecting Influenza Epidemics using Search Engine Query Data," *Nature*, 457(7232), 1012-1014.

Google Trends, (2013), www.google.com/trends/.

Jena, A. B., P. Karaca-Mandic, L. Weaver, and S. A. Seabury, (2013), "Predicting New Diagnoses of HIV Infection Using Internet Search Engine Data," *Clinical Infectious Diseases*, cid.oxfordjournals.org/content/early/2013/02/07/cid.cit022.short.

Lanseng, E. J. and T. W. Andreassen, (2007), "Electronic Healthcare: A Study of People's Readiness and Attitude toward Performing Self-diagnosis," *International Journal of Service Industry Management*, 18(4), 394-417.

National Bureau of Economic Research [NBER], (2013), "US Business Cycle Expansions and Contractions," www.nber.org/cycles.html.

Polgreen, P. M., Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein, (2008), "Using Internet Searches for Influenza Surveillance," *Clinical Infectious Diseases*, 47(11), 1443-1448.

Ripberger, J. T., (2011), "Capturing Curiosity: Using Internet Search Trends to Measure Public Attentiveness," *Policy Studies Journal*, 39(2), 239-259.

Ryan, A. and S. Wilson, (2008), "Internet Healthcare: Do Self-diagnosis Sites Do More Harm than Good?" *Expert Opinion on Drug Safety*, 7(3), 227-229.

Tefft, N., (2011), "Insights on Unemployment, Unemployment Insurance, and Mental Health," *Journal of Health Economics*, 30(1), 258-264.

United States Census Bureau, (2013), Monthly and Annual Retail Trade, www.census.gov/retail.