Using Search Engine Query Data to Track Pharmaceutical Utilization: A Study of Statins

Nathaniel M. Schuster, BS; Mary A. M. Rogers, PhD, MS; and Laurence F. McMahon Jr, MD, MPH

Objective: To examine temporal and geographic associations between Google queries for health information and healthcare utilization benchmarks.

Study Design: Retrospective longitudinal study.

Methods: Using Google Trends and Google Insights for Search data, the search terms *Lipitor* (atorvastatin calcium; Pfizer, Ann Arbor, MI) and *simvastatin* were evaluated for change over time and for association with Lipitor revenues. The relationship between query data and community-based resource use per Medicare beneficiary was assessed for 35 US metropolitan areas.

Results: Google queries for *Lipitor* significantly decreased from January 2004 through June 2009 and queries for *simvastatin* significantly increased ($P < .001$ for both), particularly after Zocor came off patent ($P < .001$ for change in slope). The mean number of Google queries for *Lipitor* correlated ($r = 0.98$) with the percentage change in Lipitor global revenues from 2004 to 2008 ($P < .001$). Query preference for *Lipitor* over *simvastatin* was positively associated ($r = 0.40$) with a community's use of Medicare services. For every 1% increase in utilization of Medicare services in a community, there was a 0.2-unit increase in the ratio of *Lipitor* queries to *simvastatin* queries in that community ($P = .02$).

Conclusions: Specific search engine queries for medical information correlate with pharmaceutical revenue and with overall healthcare utilization in a community. This suggests that search query data can track community-wide characteristics in healthcare utilization and have the potential for informing payers and policy makers regarding trends in utilization.

(Am J Manag Care. 2010;16(8):e215-e219)

For author information and disclosures, see end of text.

In February 2009, Google researchers demonstrated that Google search engine query data could be used to map the incidence of influenza in space and time.[1] They demonstrated that a model based on the frequency of search terms that included *influenza complication* and *cold/flu remedy* accurately predicted Centers for Disease Control and Prevention (CDC) data on physician visits for influenza-like symptoms. Moreover, the Google search provided more timely information because the query data were available with a 1-day reporting lag, while the CDC data had a 1-week to 2-week lag.[1]

In April 2009, Google researchers released a report demonstrating that search engine query data could be used to improve econometric models of motor vehicle sales, home sales, and tourist arrivals.[2] Given these prior examples, we postulated that Google search behavior might also correlate with healthcare utilization, as patients often consult the Internet for health-related information. In particular, we expected that there would be a change in Google search behavior for medications during a time of change in patent protection with the emergence of lower-cost generic alternatives. Accordingly, we examined whether Google searches for cholesterol-lowering 3-hydroxy-3-methylglutaryl–coenzyme A reductase inhibitors (statins) were associated with specific revenues for one of these pharmaceuticals and whether such searches provided an insight into a community's overall healthcare spending.

METHODS

Two statins were studied, Lipitor (atorvastatin calcium; Pfizer, Ann Arbor, MI) and simvastatin. Lipitor is under patent until 2011. Simvastatin was marketed by Merck (Darmstadt, Germany) as Zocor until it came off patent on June 23, 2006. We hypothesized that just as patients who are confronted with a new illness (eg, influenza) turn to the Internet for more flu-related information, patients confronted with a potential change in their medication regimen might also turn to the Internet for more information about these drugs. We chose to study the search terms *Lipitor* and the generic alternative *simvastatin*. We knew the specific date (June 23, 2006) on which Zocor came off patent and could precisely evaluate this period of transition.

In this article
Take-Away Points / e216
Published as a Web Exclusive
www.ajmc.com

METHODS

Take-Away Points

Search engine query data can be used to track real-time trends in pharmaceutical utilization.

- Google query data for the search terms *Lipitor* (atorvastatin calcium; Pfizer, Ann Arbor, MI) and *simvastatin* correlated with changes in Lipitor revenues over time and with geographic variations in healthcare costs.
- Search engine query data provide a source for readily accessible real-time data on utilization trends, with the potential to inform payers and policy makers.
- Future uses could include mapping trends in utilization of pharmaceuticals, devices, and procedures; tracking real-time responses to policy changes and copayment adjustments; and predicting demand for healthcare services.

To assess temporal trends, we obtained weekly search query indexes from Google Trends[3] for *Lipitor* and *simvastatin* from January 4, 2004, through June 28, 2009, within the United States. Google Trends data are initially calculated by dividing the total number of searches (for a given term) in a particular geographic region by the total number of searches in that region at a point in time. This calculation was then normalized so that trends over time could be evaluated to yield a query index using 0 as the start date on January 1, 2004. Therefore, a query index of 1.5 represents a 50% increase above the reference start date in the proportionate volume of searches for a given search term.

Annual Lipitor global revenues were obtained, and the percentage change from the previous year was calculated using the 2004 to 2008 Pfizer Annual Shareholder Reports.[4] To obtain the mean annual query data, the weekly query indexes for *Lipitor* and *simvastatin* searches within the United States from January 4, 2004, through December 28, 2008, were averaged.

Resource utilization was also measured by service use per Medicare beneficiary (as a percentage of the national mean) for 35 US metropolitan areas and was obtained from a December 2009 Medicare Payment Advisory Commission report.[5] Service use per Medicare beneficiary is a measure that reflects Medicare spending but adjusts for enrollment, differences in the mean health status of beneficiaries, and variations in Medicare payment rates across geographic regions.[5] The purpose of this adjustment is to provide a Medicare spending measure in which volume and intensity of services can be compared in different areas of the country. Geographic variation in service use per Medicare beneficiary is of interest because it may reflect differences in physician practice patterns and in patient preferences and behaviors.[5] In our analyses, service use per Medicare beneficiary was divided by the national mean; therefore, service use per Medicare beneficiary of 1.2 indicates resource utilization that is 20% above the national mean.

To evaluate the association between search behavior and resource utilization, the search query indexes for *Lipitor* and *simvastatin* for June 2006 through June 2008 in 35 US metropolitan areas were obtained using Google Insights for Search.[6] The community-specific ratio of *Lipitor* queries to *simvastatin* queries was analyzed as a measure of drug preference, with values greater than 1.0 indicating a preference of searching for *Lipitor*.

Generalized linear models were used for evaluation of relationships among the following: (1) Google drug queries over time, (2) Google queries and change in annual revenues, and (3) Google queries and resource utilization as measured by community-based service use per Medicare beneficiary. For the latter, standard errors were calculated incorporating clustering by state. Pearson product moment correlation coefficients were also calculated. The 2-tailed alpha level was set at .05. Analyses were performed in STATA/SE 10.0 (StataCorp LP, College Station, TX). Bayesian information criterion and Akaike information criterion were used for model selection, with lower values indicating a better fit.[7]

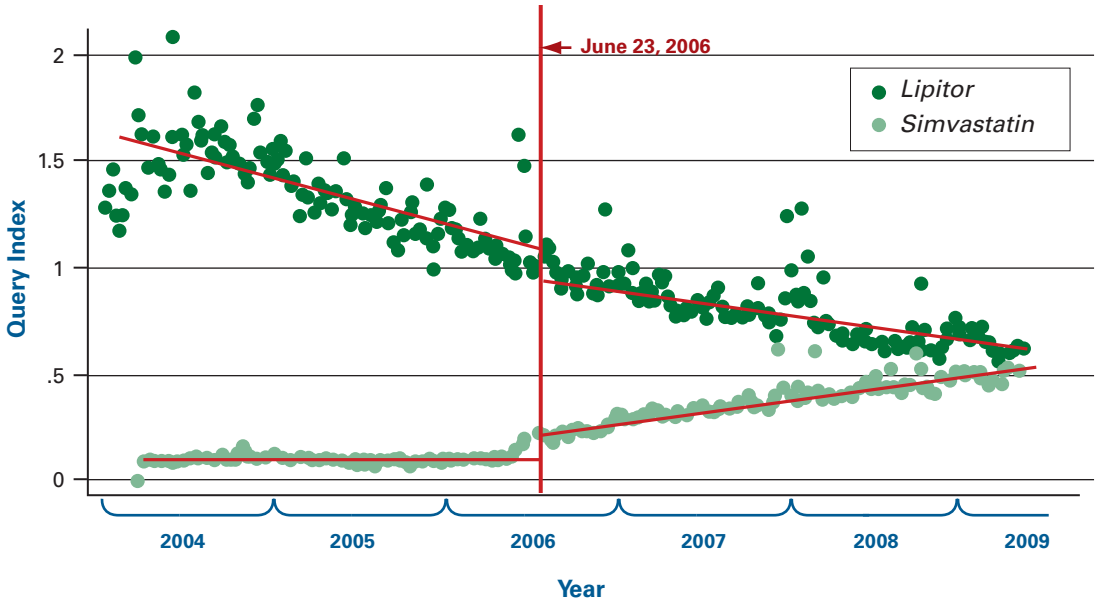
RESULTS

The number of Google search queries for *Lipitor* decreased significantly from January 2004 through June 2009 (–0.00323 slope), while the number of Google search queries for *simvastatin* increased (0.00176 slope) (P <.001 for both). The difference between these slopes was significant (P <.001). Moreover, when Zocor came off patent on June 23, 2006, there was a significant change in trends for *Lipitor* and *simvastatin* queries. The number of queries for *simvastatin* was low before the off-patent date (0.00007 slope) but increased afterward (0.00202 slope) (P <.001 for change in slope). The number of queries for *Lipitor* was declining even before the off-patent date (–0.00366 slope) and then declined to a slower degree afterward (–0.00212 slope) (P <.001 for change in slope). Trends over time are shown in **Figure 1**.

The mean Google query index for *Lipitor* decreased over time (2004-2008) and correlated (r = 0.98) with the percentage change in Lipitor global revenues during this period (P <.001). As the queries declined, so did the annual percentage change in revenues (which were 18% in 2004 but were declining by 2% in 2008). These trends are shown in **Figure 2**.

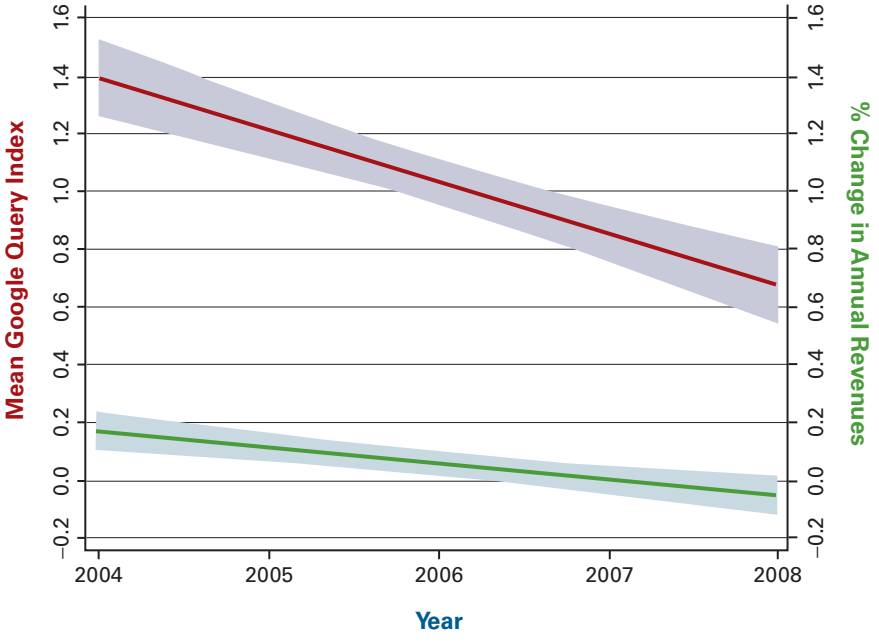
The relative preference for *Lipitor* queries (over *simvastatin* queries) varied from 82% to 139% across the 35 US metropolitan areas (ie, 0.82-1.39 in **Figure 3**, with 1.0 indicating no preference and 1.39 indicating a 39% increased preference

Figure 1. Trends in Google Queries for *Lipitor* and *Simvastatin* Over Time



for Lipitor). Metropolitan areas that had greater resource utilization, as measured by service use per Medicare beneficiary, tended to have a greater preference of searching for *Lipitor* than for *simvastatin*; the correlation coefficient for these measures was 0.40. The results from the regression model indicated that, for every 1% increase in utilization of Medicare services in a community, there was a 0.2-unit increase in the ratio of *Lipitor* Google queries to *simvastatin* Google queries (P = .02). The model using the ratio of *Lipitor* queries to *simvastatin* queries fit slightly better (lower Bayesian information criterion and Akaike information criterion) than separate models for *Lipitor* queries and *simvastatin* queries.

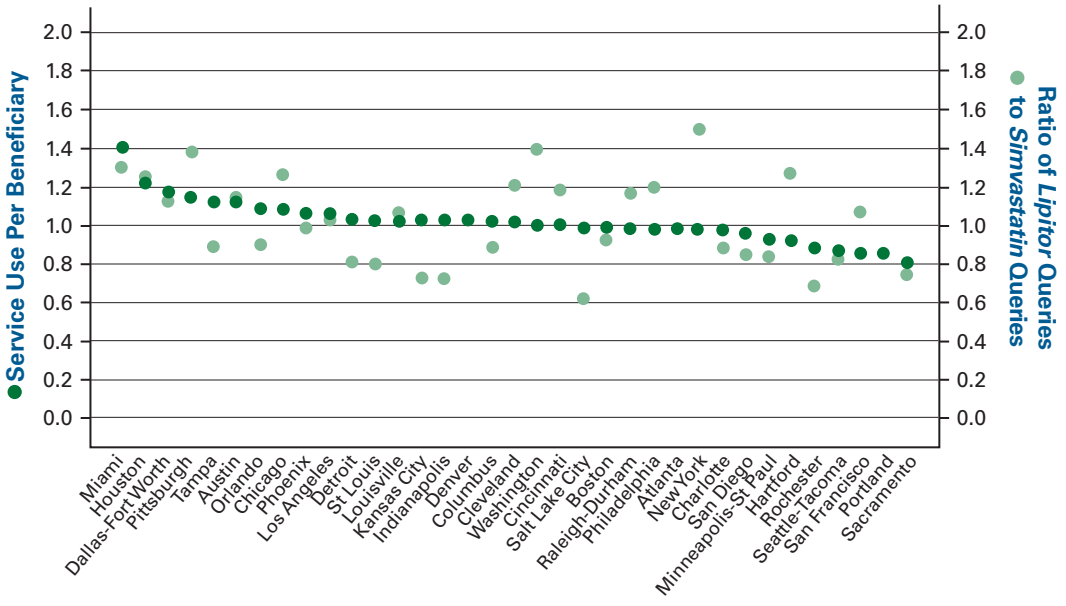
Figure 2. Mean Google Query Index for Lipitor and Percentage Change in Annual Lipitor Global Revenues From 2004 to 2008



DISCUSSION

Google search behavior for information regarding cholesterol-lowering medications was associated with pharmaceutical revenues and with metropolitan-area healthcare utilization. There was a significant change in search behavior before and after Zocor came off patent on June 23, 2006. This is not surprising, as search queries would be expected to reflect new users of a pharmaceutical. Moreover, in metropolitan areas with higher Medicare costs, Google users were significantly more likely to have searched for the trade-name *Lipitor*, while in communities with lower Medicare costs, Google users were less likely to search for *Lipitor*. This suggests that resident searching habits may reflect similar practices as the overall utilization patterns in their

Figure 3. Service Use per Beneficiary and Statin Preference in 35 US Metropolitan Areas



community, resulting in more searching for resource-intensive services (ie, higher-cost Lipitor) in resource-intensive areas. This could be due to patient preferences or physician predilection for such services in certain regions of the country. Considerable variations in practice and utilization of medical services throughout the United States have been demonstrated, with some areas using significantly more services than others.[8] Moreover, Medicare beneficiaries who move to high-utilization areas tend to assimilate the higher use patterns of their new residential area and receive more tests, diagnoses, and services.[9] This underscores the effect of community-level practices and behaviors.

One issue of concern relates to the "noise" created by searches that may not be due to the behavior of interest. For example, patients might search for *Lipitor* not only after their physicians prescribe it to them but also when a new research study involving Lipitor receives media coverage. Additional studies are needed to investigate methods that could possibly reflect the intent of the user.

Although Google users may not be representative of all users of medical services, this should not be critical if the query data sufficiently track with community-specific indicators. Although resource utilization herein was measured in an older Medicare population (with proportionately fewer computer users), it significantly correlated with Google queries for medication use (most likely representing a younger population). The value of query data for health services research is not to replace these indicators but rather to provide information in real time, possibly months or years before the community-specific indicators become available (just as they have been shown to do with community influenza tracking[1] and motor vehicle sales[2]). Although a limitation of query data is that they cannot be used to establish causation (eg, patients may search for the statin before or after the physician prescribes it), this is of less importance when the value of query data is in the ability to track utilization in real time.

When search engine queries reflect temporal and geographic variation, such data may be used to build real-time maps of utilization similar to the Google Flu Tracker.[1] Studying trends in utilization may help to inform payers and to guide efforts at changing patient and provider behavior through mechanisms such as physician education or copayment adjustment. In addition, query data may be used to educate policy makers regarding various trends in healthcare such as mapping uninsured Americans or tracking demand for primary care physicians in real time.

In conclusion, the results of our study indicate that Internet search engine queries for drug information exhibit temporal and geographic patterns of healthcare utilization. Our findings further show that search engine query data may prove helpful in providing payers and policy makers with a new window into healthcare utilization in our communities.

Acknowledgments

This research was presented at the American Medical Association Medical Student Section and Resident and Fellow Section Joint Research Symposium, 2009 American Medical Association Interim Meeting, in Houston, Texas, on November 6, 2009. We thank Benjamin Schuster, BBmE, Ariel Hecht, MS, and Eric Livak-Dahl, MS, for their comments on the manuscript.

Author Affiliations: From the School of Medicine (NMS), Department of Internal Medicine (MAMR, LFM), University of Michigan, Ann Arbor, MI; and Patient Safety Enhancement Program (MAMR, LFM), Ann Arbor Veterans Affairs Medical Center, Ann Arbor, MI.

Funding Source: None reported.

Author Disclosures: The authors (NMS, MAMR, LFM) report no relationship or financial interest with any entity that would pose a conflict of interest with the subject matter of this article.

Authorship Information: Concept and design (NMS, LFM); acquisition of data (NMS); analysis and interpretation of data (NMS, MAMR, LFM); drafting of the manuscript (NMS, LFM); critical revision of the manuscript for important intellectual content (NMS, MAMR, LFM); statistical analysis (NMS, MAMR, LFM); administrative, technical, or logistic support (MAMR, LFM); and supervision (LFM).

Address correspondence to: Laurence F. McMahon Jr, MD, MPH, Department of Internal Medicine, University of Michigan, 300 N Ingalls, Rm NI7C27, Box 0429, Ann Arbor, MI 48109-0429. E-mail: lmcmahon@umich.edu.

REFERENCES

1. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature.* 2009;457(7232):1012-1014.

2. Choi H, Varian H. Predicting the present with Google Trends. April 10, 2009. http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/en/us/googleblogs/pdfs/google_predicting_the_present.pdf. Accessed May 23, 2010.

3. Google Trends. http://www.google.com/trends. Accessed November 26, 2009.

4. Pfizer financial reports. http://www.pfizer.com/investors/financial_reports/financial_reports.jsp. Accessed November 21, 2009.

5. Medicare Payment Advisory Commission. Report to the Congress: measuring regional variation in service use. December 2009. http://www.medpac.gov/documents/Dec09_RegionalVariation_report.pdf. Accessed July 3, 2010.

6. Google Insights for Search. http://www.google.com/insights/search/#. Accessed November 26, 2009.

7. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference.* 2nd ed. New York, NY: Springer; 2002.

8. Dartmouth Institute for Health Policy and Clinical Practice. The Dartmouth Atlas of Health: understanding of the efficiency and effectiveness of the health care system. http://www.dartmouthatlas.org. Accessed May 20, 2010.

9. Song Y, Skinner J, Bynum J, Sutherland J, Wennberg JE, Fisher ES. Regional variations in diagnostic practices. *N Engl J Med.* 2010;363(1):45-53. ■