BMJ Open

The utility of Google Trends data to examine interest in cancer screening

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2014-006678
Article Type:	Research
Date Submitted by the Author:	22-Sep-2014
Complete List of Authors:	Schootman, Mario; Saint Louis University, Department of Epidemiology, College for Public and Social Justice; Toor, Aroona; Saint Louis University, Department of Epidemiology Cavazos-Rehg, Patricia; Washington University, Department of Psychiatry Jeffe, Donna; Washington University, Department of Internal Medicine McQueen, Amy; Washington University, Department of Internal Medicine Eberth, Jan M.; University of South Carolina, Davidson, Nicholas; Washington University, Department of Internal Medicine
Primary Subject Heading:	Public health
Secondary Subject Heading:	Epidemiology, Gastroenterology and hepatology, Health informatics
Keywords:	EPIDEMIOLOGY, PUBLIC HEALTH, Adult gastroenterology < GASTROENTEROLOGY

The utility of Google Trends data to examine interest in cancer screening

Schootman M,[1,2] Toor A,[1] Cavazos-Rehg P,[3] Jeffe DB,[2,4] McQueen A, [2,4] Eberth J,[5] Davidson NO [2,6]

[1] Saint Louis University College for Public Health and Social Justice
Department of Epidemiology
Saint Louis, MO

[2] Alvin J. Siteman Cancer Center at Barnes-Jewish Hospital and Washington University School of Medicine
Saint Louis, MO

[3] Washington University School of Medicine
Department of Psychiatry
Saint Louis, MO

[4] Washington University School of Medicine
Department of Psychiatry
Saint Louis, MO

[5] Department of Epidemiology and Biostatistics
Arnold School of Public Health
University of South Carolina, SC

[6] Washington University School of Medicine
Department of Medicine
Division of Gastroenterology
Saint Louis, MO

Address for correspondence:
Mario Schootman, PhD
Saint Louis University
College for Public Health and Social Justice
Department of Epidemiology
3545 Lafayette Ave
Saint Louis, MO 63104
e-mail: schootm@slu.edu
phone: 314-977-8133

Abstract

Background

Traditional surveillance system that collect self-reported data about cancer screening use are ill equipped to deal with a rapidly changing digital world with a need for timely health data.

Methods

We examined the utility of Google Trends data from information searches for cancer screenings and preparations as a complement to population screening data, which are traditionally estimated through costly population-level surveys. State-level Google Trends data were correlated with state-level, self-reported Behavioral Risk Factor Surveillance System (BRFSS) breast, cervical, colorectal, and prostate cancer screening rates. Google Trends provides relative search volume (RSV) data scaled to the highest search proportion week (RSV100) for search terms over time since 2004 and across different geographical locations. We also examined RSV of new screening tests, free/low cost screening for breast and colorectal cancer, and new preparations for colonoscopy (Prepopik™).

Results

Correlations between Google Trends and BRFSS data ranged from 0.55 for ever having had a colonoscopy to 0.14 for having a Pap smear within the past three years. RSV varied for different screening tests and was highest for prostate vs. other cancer screening sites. Free-low cost mammography and colonoscopy showed higher RSV during their respective cancer awareness months. RSV for Miralax remained stable, while interest in Prepopik increased over time. RSV for lung cancer screening, virtual colonscopy and 3D mammography was low.

Conclusions

Google Trends data provides enormous scientific possibilities, but are not a suitable substitute for, but may complement, traditional data collection and analysis about cancer screening and related interests.

Article summary – strengths and limitations of this study

- Google Trends data help identify developing interests in new cancer screening tests or related aspects of specific screening tests.
- Internet searches can be an important source for generating hypotheses about public awareness and interest in cancer screening, evaluating changes in information seeking after targeted interventions or media coverage, and directing new communication campaigns to explain the evidence base for screening tests.
- An evaluation that occurs almost immediately after an intervention may inform policy makers of the associated costs and benefits when there is still interest to make modifications to, or expand, any policy changes.
- The utility of Google Trends to help evaluate interventions depends on the area where the intervention is implemented, since data is only available for states and selected metropolitan areas, limiting its use in rural areas or areas with a low search volume.
- Google Trends data are anonymous, which limits its utility in examining specific subpopulations and disparities among populations. Also, Google Trends data represent only searches done using Google.

What is already known on this subject?

Traditional surveillance systems are ill equipped to deal with a rapidly changing digital world with a need for timely cancer screening data for public health and medical professionals, policy makers, and the public who influence policy choices. Recent technological advances in data acquisition, such as Google Trends, may allow for more timely data collection. Depending on its utility, Google Trends may complement existing surveillance systems that could further improve screening use, possibly resulting in early diagnosis and prevention of cancer.

What this study adds?

Google Trends data provides enormous scientific possibilities that may complement traditional data collection and analysis about cancer screening and related interests. The strengths of Google Trends to provide data about the public's interests in cancer screening, despite its inability to provide cancer screening usage data, can foster provision of timely feedback about interventions aimed at increasing interest in cancer screening.

Introduction

Cancer screening is a cornerstone of public health aimed at promoting early diagnosis and, in some instances, prevention of cancer. There are several surveillance systems that monitor self-reported cancer screening utilization, including the Behavioral Risk Factor Surveillance System (BRFSS),[1 2] the National Health Interview Survey (NHIS),[3] and the Health Information National Trends Survey (HINTS).[4] These databases have been invaluable in identifying determinants of screening use and describing trends and disparities over time.

These traditional surveillance systems are ill equipped to deal with a rapidly changing digital world with a need for timely health data for public health and medical professionals, policy makers, and the public who influence policy choices. They are expensive to maintain due to their use of survey interview methods for data collection and the time required to aggregate the data; require participation of a large study population to estimate screening use reliably; rely on self-report resulting in potential recall bias; and, for the BRFSS and HINTS, participants include only persons with landline telephones and, more recently, mobile phones, leaving the door open for potential selection bias. They often do not capture new screening modalities (e.g., virtual colonoscopy for colorectal cancer, magnetic resonance imaging for breast cancer detection, or low-dose spiral computed tomography for lung cancer screening among persons at high risk for lung cancer) especially when use is still low. As a result, population-based prevalence of newer screening methods is unknown.

Recent technological advances in data acquisition, such as Google Trends, may allow for more timely data collection to learn about trends in interest in various health-related topics, including cancer screening. Google Trends is a keyword research tool that provides near real-time trend data regarding interest as operationalized by internet search volume. Both Google and Yahoo! search engines have been used to analyze different types of search queries, for example about cancer incidence,[5] cancer mortality,[5] kidney stones,[6] non-cigarette tobacco use,[7] sexually transmitted infections,[8] and flu trends.[9 10] However, the value of Google Trends in illuminating

trends reflecting interest in cancer screening and related topics has not yet been examined. Depending on its utility, Google Trends may complement existing surveillance systems that could further improve screening use, possibly resulting in early diagnosis and prevention of cancer.

Here, we examined the utility of Google Trends relative to the BRFSS, focusing on cancer screening. Specifically, we examined 1) the correlation between Google Trends and self-reported breast, cervical, colorectal, and prostate cancer screening in the BRFSS, and 2) interest in possible new and developing screening modalities and preparations not currently captured in existing surveillance systems.

Methods

Data sources about screening use

Breast cancer screening (mammography and breast self-exam), cervical cancer (Pap smear), colorectal cancer (fecal occult blood test [FOBT], colonoscopy), prostate cancer screening using prostate screening antigen (PSA) test, were all obtained from the BRFSS database.[11] The BRFSS is one of the largest annual telephone health-survey database systems in the world. The survey provides state-level prevalence data of the major behavioral risks among adults associated with premature morbidity and mortality among adults. Data are collected from all 50 U.S. states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, Guam, American Samoa, and Palau. Questions about cancer screening use have been validated.[12] Prevalence of screening use was obtained from the Centers for Disease Control and Prevention: http://apps.nccd.cdc.gov/brfss/. In this study, we included data from the 50 U.S. states.

State-level Google Trends data (http://www.google.com/trends/explore#cmpt=q) were obtained and then compared with state-level rates of breast cancer (mammography, breast self exam), cervical cancer (Pap smear), colorectal cancer (FOBT, colonoscopy), and prostate cancer (PSA) reported on the BRFSS. Google Trends, based on Google Search, the most widely used internet search engine, offers search volume data for search terms over time since 2004 and

across different geographical locations. Google Trends shows how often search terms are entered in Google relative to the total search volume in a region or globally. Google Trends produces relative search volume (RSV) scaled to the highest search proportion week. RSV values are by definition always less than 100 and demonstrate how other weekly search proportions compared to the highest (RSV=100) search proportion. For example, RSV=50 represents 50% of the highest observed search proportion during the study period. RSV indirectly corrects for population size and Internet access, both of which increased during the study period and would bias any absolute search volume measure. However, RSV allows for directly comparing search volume across search terms.

Google Trends can compile search volume for up to 30 words. We selected search terms based on their face validity for the term's relationship to the screening test of interest. Google Trends allows up to four strata for different trend data. If these the relative interest score, We included additional search terms in our main search if these additional strata increased at least RSV by at least 1 point.

We also added search terms based on popular "related terms" suggested by Google Trends. We included singular and plural forms of the search terms. Appendix 1 shows the specific search terms used for each screening test and associated terms relevant to specific tests (e.g., Miralax for colonoscopy). In addition to obtaining search volume data about screening interest, we examined search volume data regarding new screening tests (virtual colonoscopy, lung cancer screening using computed tomography [CT], 3D mammography), free/low cost screening for breast and colorectal cancer, and new preparations to cleanse the colon for colonoscopy (Prepopik™). Prepopik™ was approved on July 16, 2012 by the Food and Drug Administration to help cleanse the colon in adults preparing for colonoscopy.[13]

Statistical analysis

We used the Pearson correlation coefficient to examine the associations between state-level Google Trends RSV and BRFSS state-level screening prevalence for each of the five cancer screening tests. We weighted these correlations by the 2011 state population estimates from the Bureau of the Census using weighted regression. Weighted estimates provide more weight to states with larger populations. We used Stata 13.1 to calculate weighted correlations using the wls0 command.

We used the joinpoint methodology to identify significant changes in weekly RSV over time for each of the screening tests and associated interests.[14 15] Linear trends in search volume were summarized using the estimated annual percentage change (EAPC). The EAPC was calculated by fitting a linear regression to the natural logarithm of the weekly RSV, using week as a regression variable. Joinpoint regression tests were used to identify an inflection point (hereafter, called joinpoint) with a significant change in the slope of the trend.[14 15] For our analysis, a minimum of four weeks between two joinpoints was required, and a maximum of three joinpoints was allowed to describe the data.

Results

Figure 1a shows the weekly Google Trends RSV for colorectal cancer screening using colonoscopy between January, 2004 and April, 2014. The average RSV was 61.9 in 2004 and increased to 85.8 during the last 52 weeks of data. During the first 3 years, RSV per week remained stable, but then increased 0.2 percent per week (95% CI: 0.1; 0.2). Starting at week 308 (November, 2009), RSV increased 0.09 percent per week (95% CI: 0.07; 0.11). RSV was lowest during December of each year and slightly higher during March of each year (average: 74.3). The weighted correlation between ever having had a colonoscopy based on 2012 BRFSS data and 2004-2012 Google Trends colonoscopy data was 0.55 using state estimates. During 2007, the average RSV/week for virtual colonoscopy was 22.5, but RSV decreased 0.30 percent per week (-

0.33; -0.27) starting in January 2008 (Figure 1b). RSV/week for Miralax as a colon cleanser declined 0.50 percent per week (95% CI: -0.69; -0.30) during January 2009 through August 2010, after which RSV about Miralax remained stable until April 2014 (Figure 1c). RSV/week for Prepopik, a newer colon cleanser approved by the FDA in July 2012, increased rapidly over time and approached RSV for Miralax (70) in April 2014 (Prepopik: 57). Google Trends data was available for only eight states due to low search volume, and a correlation between BRFSS data about FOBT use and Google Trends RSV could not be calculated. Prior to 2009, RSV about FOBT was zero and remained stable starting in 2009. During 2009-2014, RSV/week for colonoscopy was 85, while it was only 1 for FOBT. Also during this period, RSV for the cost for colonoscopy was substantially higher (average: 77) than for cancer treatment cost (average: 49).

Figure 2a shows RSV/week for mammography over time. Peaks were present during October each year and about 10 points higher than during December, the month with the lowest RSV. In November 2009, mammography RSV was highest during this 10-year period. Weighted correlation between Google Trends RSV and BRFSS-based mammography use was 0.36. Figure 2b shows Google Trends RSV/week in free/low-cost mammography, which peaked in October every year. RSV/week was much higher for free/low-cost mammography (average: 31) compared to free/low-cost colonoscopy (average: 3) during the 10-year period. RSV for digital mammography (average: 54) was higher than for free/low-cost mammography (average: 4) and also showed peaks during October. RSV for 3D mammography has been increasing since 2011, but is still much lower than overall RSV for mammography (average: 63; 3D mammography: 0). Breast self-examination RSV (average: 3) over the study period was much lower than for mammography (average: 51). RSV for mammography (average: 32) was substantially lower than for colonoscopy (average: 78).

Figure 3 shows that during week 1-137, Pap smears RSV/week increased slightly (0.08 percent per week; 95% CI: 0.03; 0.13), remained stable during weeks 137-208, increased during weeks 208-426 (0.13 percent per week; 95% CI: 0.11; 0.16), but then decreased starting in week

426 (-0.11 percent per week; 95% CI: -0.18; -0.04). Weighted correlation between 2012 BRFSS prevalence of Pap smear use within the past three years and RSV for Pap smears during 2010-2012 was 0.14. Pap smear RSV (average: 33) was very similar to mammography RSV (average: 32) but much lower than for colonoscopy (average: 78) during the ten year period.

RSV for PSA declined very slowly (0.05 percent per week) starting in 2004 (95% CI: -0.06; -0.05) until October 2009 (week 302), after which the decline became steeper at 0.20 percent per week (95% CI: -0.30; -0.11) until December 2010 (week 364), then there were three weeks during which RSV remained stable (Figure 4). Starting in January 2011 (week 367), RSV declined 0.05 percent per week (95% CI: -0.07; -0.03). However, between January 2010 and April 2014, RSV for prostate cancer screening (average: 77.4) was highest of any of the other four types of cancer screening (colonoscopy: 47.9, Pap smear: 19.9, mammography: 18.9, lung cancer: 0.2). RSV for PSA was highest for week 272 (March, 2009). Correlation between Google Trends and BRFSS data was 0.42.

Between January 2007 and July 2010, RSV about lung cancer screening declined 1.1 percent per month (95% CI: -1.7; -0.5), but then increased 2.8 percent per month (95% CI: 2.3; 3.4) until April 2014 (Figure 5). There was a peak in RSV about lung cancer screening during November 2010 (month 47). RSV for lung cancer has been very low relative to other types of screening across the study period.

Discussion

We examined the utility of Google Trends relative to the BRFSS, one of the existing surveillance systems focusing on cancer screening. Correlations between Google Trends and BRFSS data ranged from a high of 0.55 for ever having had a colonoscopy to a low of 0.14 for having a Pap smear within the past three years. Although self-reported screening use is less than perfect,[12] these modest correlations indicate that they are measuring different constructs. Awareness and interest in cancer screening is a necessary, but not sufficient, determinant of

screening behavior.[16 17] Search volume data using Google Trends enabled us to measure the public's awareness and interest in possible new and developing screening modalities (e.g., virtual colonoscopy, digital mammography, 3D mammography, computed tomography for lung cancer screening) and screening test preparations (e.g., Prepopik versus Miralax and Suprep), which are not currently captured in existing surveillance systems. But Google Trends and existing surveillance systems are measuring different things: Google Trends provides estimates of the public's interest in learning more about cancer screening tests; the BRFSS and other surveillance systems provide estimates of self-reported use of these tests.

Based on our findings and those of other studies,[18-20] there appears to be complementary utility of Google Trends data relative to existing surveillance systems to monitor cancer screening. By harnessing real-time search-engine data around community-based interventions, programs can be evaluated as they are implemented, generating timely feedback to assess the effectiveness of interventions to increase interest in cancer screening and other public health recommendations. Such adaptive designs using accumulating data to modify the intervention's course [21 22] have been used infrequently in community-based evaluations. Adaptive interventions that can be evaluated using interest and awareness may be especially useful. It appears that in some instances an increase in public interest in cancer screening is associated with the timing of news reports and advertisements.[23] For example, the increase in search volume each October coincides with news stories and advertisements during Breast Cancer Awareness Month. Search volume for colon cancer screening was also slightly higher during Colon Cancer Awareness Month. Google Trends also identified a large interest in November 2009 when search volume about mammography increased dramatically likely in response to critics citing health care rationing in response to new mammography recommendations from the US Preventive Services Task Force.[24] The panel recommended that most women wait until age 50 to start routine mammography, then get the exam every two years instead of annually. Increases in RSV were also seen after periods of extensive media coverage. For example, in March 2009 RSV for prostate cancer screening

increased following coverage of two studies showing that prostate cancer screening did not reduce the risk of death.[25] Also, in November 2010, RSV for lung cancer screening increased after trials reported its potential to reduce the risk of death among heavy smokers.[26] The utility of Google Trends to help adapt interventions depends on the area where the intervention is implemented, since data is only available for states and selected metropolitan areas, limiting its use in rural areas or areas with a low search volume.

Internet searches using Google Trends can guide the development of traditional surveillance systems surveys, such as the BRFSS, NHIS, and HINTS, by vetting the inclusion of questions on surveys. Google Trends data help identify developing interests in new cancer screening tests (e.g., virtual colonoscopy) or related aspects of specific screening tests (e.g., about preparation for colonoscopy). For example, Google Trends showed that interest in lung cancer screening and virtual colonoscopy is still very low, while interest in prostate cancer screening is very high even though PSA tests have been shown to be not very effective in reducing risk of death.[27] Interest in virtual colonoscopy, despite showing promise as a screening tool relative to traditional colonoscopy,[28] was very low. Thus, awareness should be increased for virtual colonoscopy to become a standard part of the armamentarium of colorectal cancer screening with concomitant increase in availability. Internet searches can be an important source for generating hypotheses about public awareness and interest in cancer screening, evaluating changes in information seeking after targeted interventions or media coverage, and directing new communication campaigns to explain the evidence base for screening tests.

The inclusion of questions on surveillance system surveys is constrained by anticipated effects of participant burden including respondent fatigue and non-response. In contrast, Google Trends can systematically generate hundreds of possible outcomes rather than arbitrary selection of a few outcomes that can be included in traditional surveillance systems, recognizing that both data sources measure related but different constructs.

Search query results may also be politically relevant. Since policy changes often require public support, evaluation strategies that take years to perform may not provide relevant feedback to public interest groups and voters. Instead, an evaluation that occurs almost immediately after the policy change may inform policy makers and their supporters of the associated costs and benefits when there is still interest to make modifications to, or expand, the policy change.[29] For example, the interest in and implementation of free/low cost breast and colorectal cancer screening can be evaluated. The CDC and local organizations implemented free/low cost mammograms starting in the 1990s across the United States followed by free/low cost colonoscopies in selected locations to eligible participants. The potential need for and likely success of the expansion of these interventions could be gleaned through Google Trends data, much earlier than traditional evaluation strategies.

The utility of Google Trends data should be viewed in light of its limitations. Google Trends data are anonymous, which limits its utility in examining specific subpopulations and disparities among populations. Also, Google Trends data represent only searches done using Google. However, Google accounts for an estimated 65 percent of all internet searches.[30] Google Trends data may have sampling biases. However, such biases are increasingly eroding at the population level as more and more people search for information online. Google Trends uses a certain threshold of traffic volume so that very new search terms are assigned a value of zero, but this could change very quickly. The motivation of Google users is not known. As a corollary, the data obtained from Google trends cannot be independently verified. Also, the researcher has no control over the data, making quality control difficult. Finally, search terms entered in other languages were not captured by us, but could be used to examine interest among non-English speaking populations.

Although Google Trends' "big data" approach provides enormous scientific possibilities, they are not a substitute for, but rather complement, traditional data collection and analysis. The strengths of Google Trends to provide data about the public's interests in cancer screening,

despite its inability to provide cancer screening usage data, can foster provision of timely feedback about interventions aimed at increasing interest in cancer screening and other public health recommendations.

Funding

This work was supported by grants from the National Cancer Institute at the National Institutes of Health (grant number CA112159); and the Health Behavior, Communication and Outreach Core; the Core is supported in part by the National Cancer Institute Cancer Center Support Grant (grant number P30 CA91842) to the Alvin J. Siteman Cancer Center at Washington University School of Medicine and Barnes-Jewish Hospital in St. Louis, Missouri. Dr. Davidson was supported in part through grants HL-38180, DK-56260, and Digestive Disease Research Core Center DK-52574. We thank the Alvin J. Siteman Cancer Center at Barnes-Jewish Hospital and Washington University School of Medicine in St. Louis, Missouri, for the use of the Health Behavior, Communication, and Outreach Core.

Data sharing statement: No additional data are available.

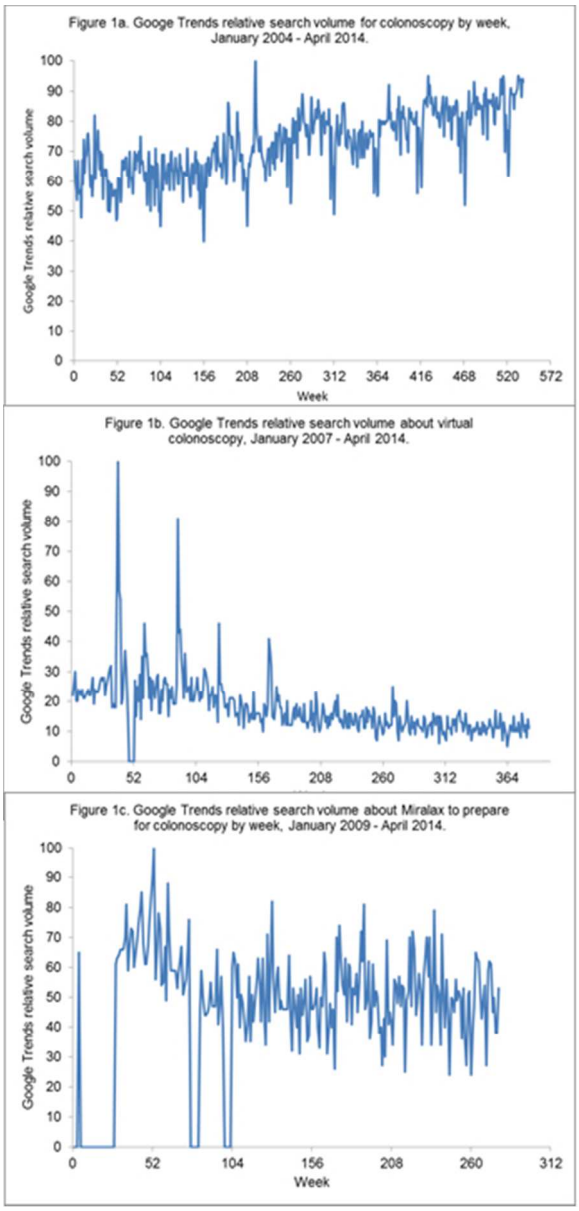
Authorship contributions

Author MS was the principal investigator of the study. MS and AT designed and conceptualized the study, with MS overseeing data collection. MS performed the statistical analysis pertaining to the trends over time. MS and AT and wrote sections of the manuscript. AT helped to conceptualize the study, performed some of the data analysis, and edited the manuscript. PC, DJ, AM, JE, and ND helped to conceptualize the study, interpreted the results, and edited the manuscript. AM and JE provided insight into the use of behavioral theories that could help explain the findings. ND provided clinical insight into screening issues pertaining to colorectal cancer.

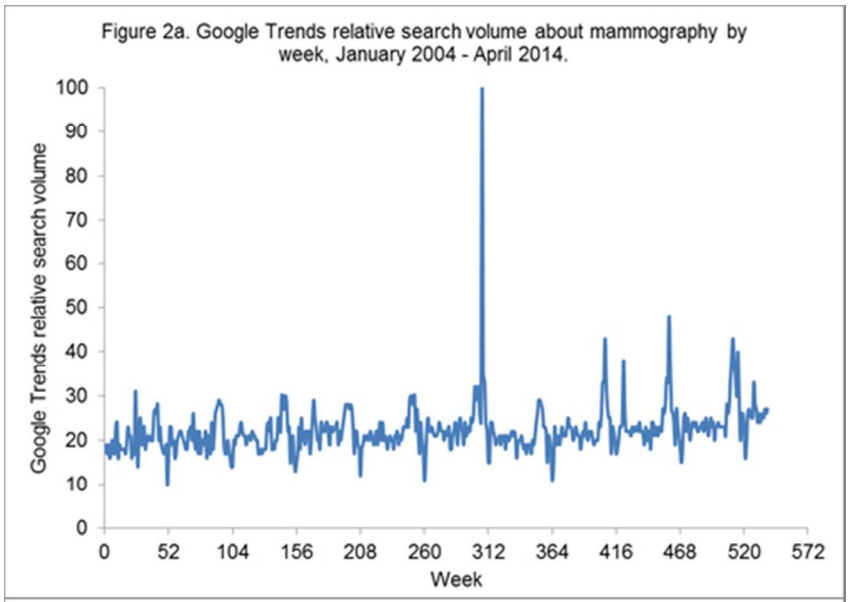
References

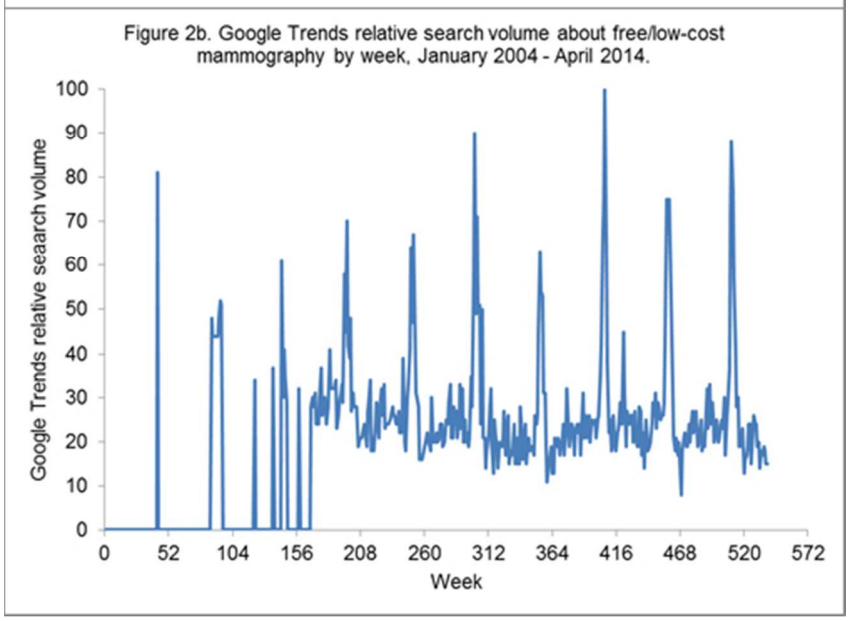
1. Joseph DA, King JB, Miller JW, et al. Prevalence of colorectal cancer screening among adults--Behavioral Risk Factor Surveillance System, United States, 2010. MMWR Morbidity and mortality weekly report 2012;**61 Suppl**:51-6.
2. Miller JW, King JB, Joseph DA, et al. Breast cancer screening among adult women--Behavioral Risk Factor Surveillance System, United States, 2010. MMWR Morbidity and mortality weekly report 2012;**61 Suppl**:46-50.
3. Hiatt RA, Klabunde C, Breen N, et al. Cancer screening practices from National Health Interview Surveys: past, present, and future. Journal of the National Cancer Institute 2002;**94**(24):1837-46.
4. Ashok M, Berkowitz Z, Hawkins NA, et al. Recency of Pap testing and future testing plans among women aged 18-64: analysis of the 2007 Health Information National Trends Survey. Journal of women's health (2002) 2012;**21**(7):705-12.
5. Cooper CP, Mallon KP, Leadbetter S, et al. Cancer Internet Search Activity on a Major Search Engine, United States 2001-2003. J Med Int Res 2005;**7**(3):e36.
6. Breyer BN, Sen S, Aaronson DS, et al. Use of Google Insights for Search to Track Seasonal and Geographic Kidney Stone Incidence in the United States. Urology 2011;**78**(2):267-71.
7. Cavazos-Rehg PA, Krauss MJ, Spitznagel EL, et al. Monitoring of non-cigarette tobacco use using Google Trends. Tobacco control 2014.
8. Johnson AK, Mehta SD. A Comparison of Internet Search Trends and Sexually Transmitted Infection Rates Using Google Trends. Sexually Transmitted Diseases 2014;**41**(1):61-63 10.1097/OLQ.0000000000000065.
9. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 2009;**49**(10):1557-64.
10. Pervaiz F, Pervaiz M, Abdur Rehman N, et al. FluBreaks: early epidemic detection from Google flu trends. Journal of medical Internet research 2012;**14**(5):e125.
11. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System, 2014: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2014.
12. Pierannunzi C, Hu SS, Balluz L. A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004-2011. BMC Medical Research Methodology 2013;**13**(1):49.
13. Administration FaD. FDA approves new colon-cleansing drug for colonoscopy prep, July 17, 2012.
14. Kim HJ, Fay MP, Feuer EJ, et al. Permutation tests for joinpoint regression with applications to cancer rates. Statistics in Medicine 2000;**19**:335-51.
15. Joinpoint Regression Program, Version 4.1.0.
[program], 2014.
16. Fishbein M. The role of theory in HIV prevention. AIDS Care 2000;**12**(3):273-8.
17. Weinstein ND. The precaution adoption process. Health Psychology 1988;**7**(4):355-86.
18. Butler D. When Google got flu wrong. Nature 2013;**494**(7436):155-56.
19. Olson DR, Konty KJ, Paladini M, et al. Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. PLoS Comput Biol 2013;**9**(10):e1003256.
20. Resnick B. Why Google Flu Trends will not replace the CDC anytime soon. *National Journal* January 25, 2013.

21. Chang M, Chow SC, Pong A. Adaptive design in clinical research: issues, opportunities, and recommendations. Journal of biopharmaceutical statistics 2006;**16**(3):299-309; discussion 11-2.
22. Coffey CS, Levin B, Clark C, et al. Overview, hurdles, and future work in adaptive designs: perspectives from a National Institutes of Health-funded workshop. Clinical trials (London, England) 2012;**9**(6):671-80.
23. Glynn RW, Kelly JC, Coffey N, et al. The effect of breast cancer awareness month on internet search activity--a comparison with awareness campaigns for lung and prostate cancer. BMC cancer 2011;**11**:442.
24. Stein R. In wake of mammography guidelines, U.S. health task force faces new scrutiny.: Washington Post, ecember 20, 2009.
25. Parker-Pope T. For men, to screen or not to screen: The New York Times, March 23, 2009.
26. McCook A. More signs lung cancer screening could save lives: Reuters, December 28, 2010.
27. U.S. Preventive Services Task Force. Screening for Prostate Cancer, Topic Page. . Secondary Screening for Prostate Cancer, Topic Page. 2012. http://www.uspreventiveservicestaskforce.org/prostatecancerscreening.htm.
28. Kim DH, Pickhardt PJ, Taylor AJ, et al. CT colonography versus colonoscopy for the detection of advanced neoplasia. The New England journal of medicine 2007;**357**(14):1403-12.
29. Ayers JW, Ribisl K, Brownstein JS. Using Search Query Surveillance to Monitor Tax Avoidance and Smoking Cessation following the United States' 2009 “SCHIP” Cigarette Tax Increase. PLoS ONE 2011;**6**(3):e16777.
30. Sullivan D. Google still world’s most popular search engine by far, but share of unique searchers dips slightly, February 11, 2013.

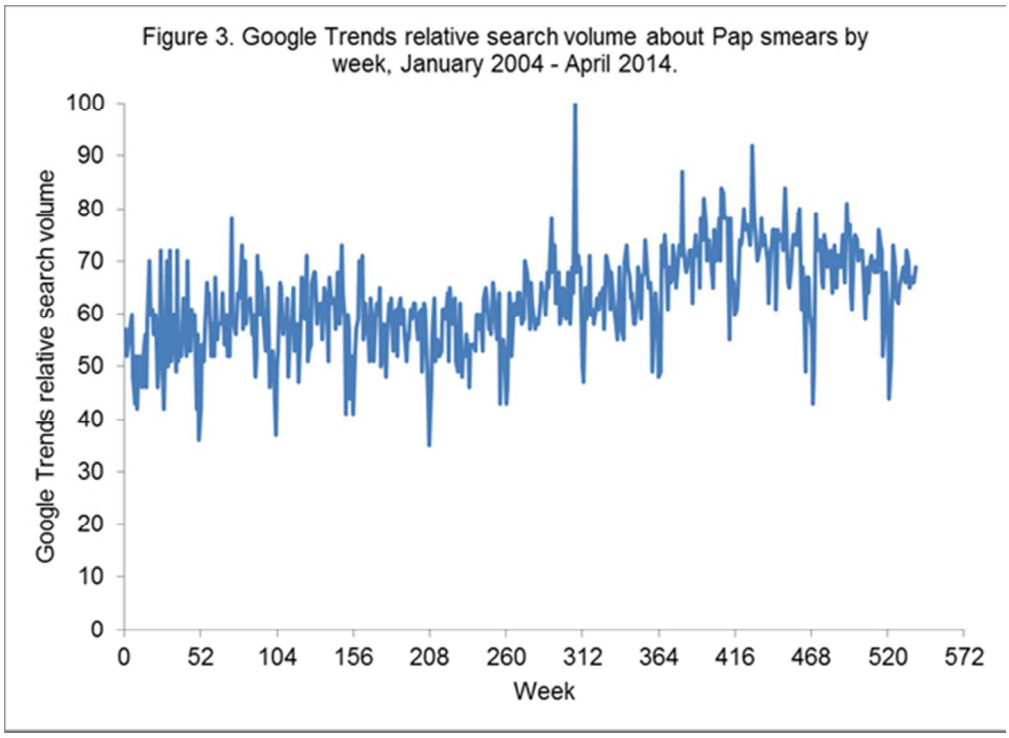


92x190mm (96 x 96 DPI)

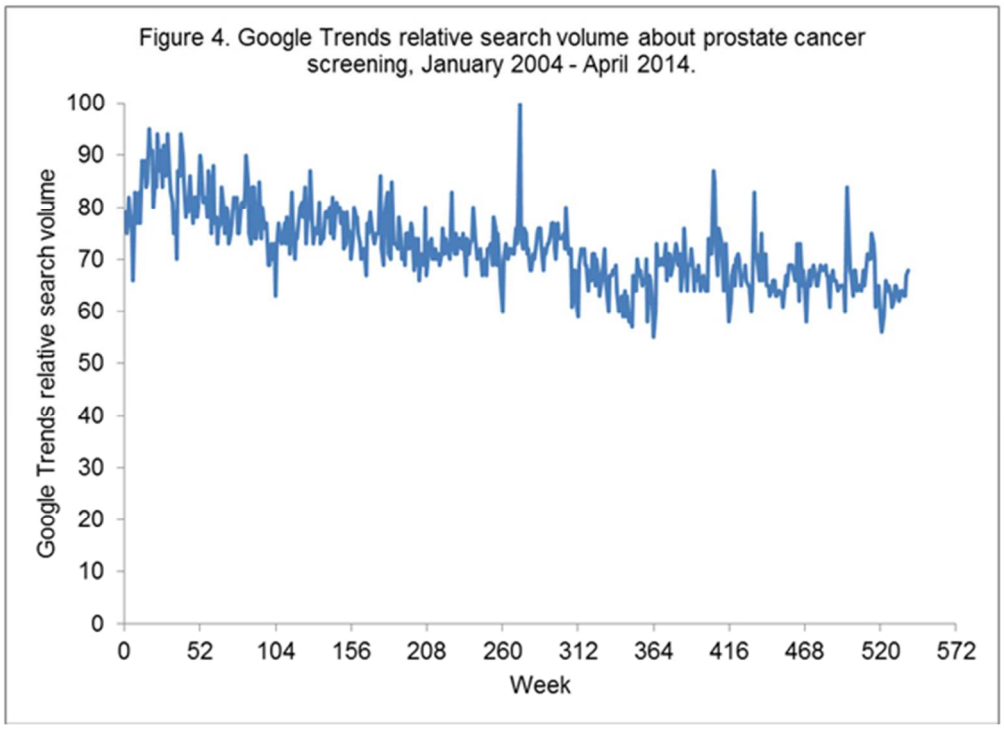




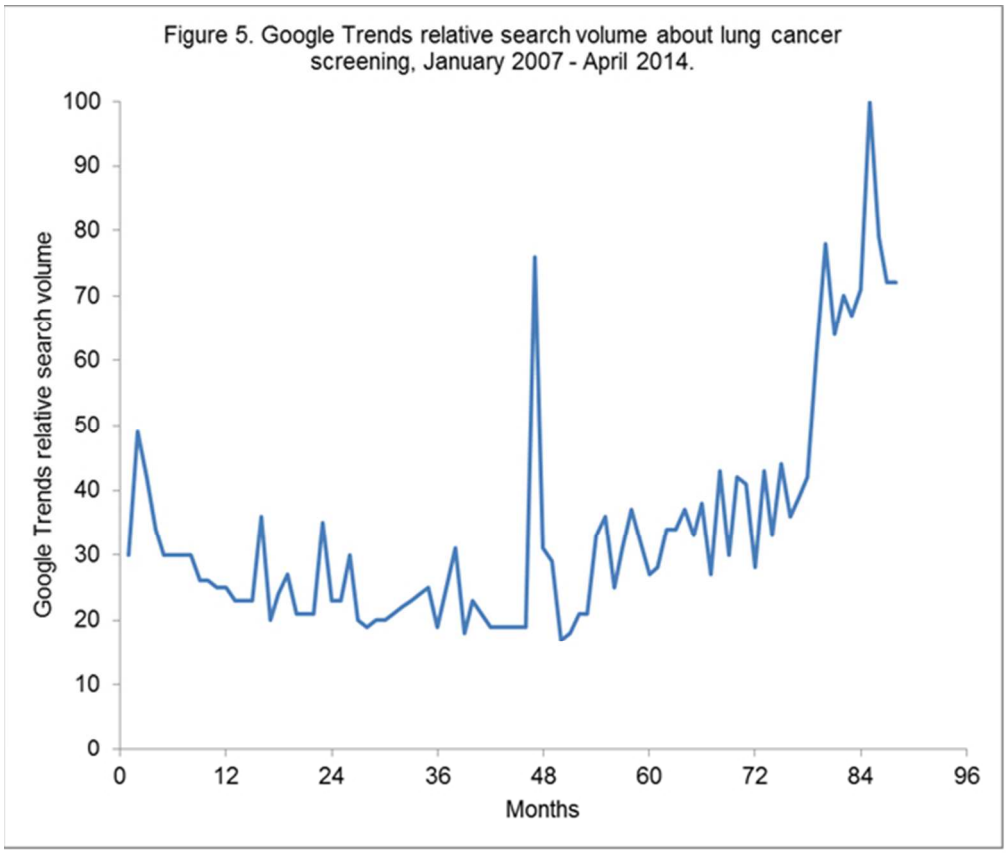
133x190mm (96 x 96 DPI)



164x119mm (96 x 96 DPI)



166x119mm (96 x 96 DPI)



168x142mm (96 x 96 DPI)

Appendix 1: Google Trends search terms used for each screening test and associated interests and Behavioral Risk Factor Surveillance Survey question.

Concept	Google Trends Search terms used	Behavioral Risk Factor Surveillance Survey question (BRFSS)
Screening for colorectal cancer		
Colonoscopy	Colonoscopy+ colonoscopy procedure +virtual colonoscopy+endoscopy +miralax prep+colonoscopy procedure+endoscopy procedure+what is colonoscopy+prep for colonoscopy+miralax+bowel prep+colonoscopy prep+colon cancer screening+colon cancer test	1. Sigmoidoscopy and colonoscopy are exams in which a tube is inserted in the rectum to view the colon for signs of cancer or other health problems. Have you ever had either of these exams? 2. For a SIGMOIDOSCOPY, a flexible tube is inserted into the rectum to look for problems. A COLONOSCOPY is similar, but uses a longer tube, and you are usually given medication through a needle in your arm to make you sleepy and told to have someone else drive you home after the test. Was your MOST RECENT exam a sigmoidoscopy or a colonoscopy? 3. How long has it been since you had your last sigmoidoscopy or colonoscopy?
Virtual colonoscopy	virtual colonoscopy+ct colonography+virtual colonoscopy cost+ct colonoscopy	Not asked in BRFSS
Miralax to cleanse colon for colonoscopy	miralax colonoscopy+colonoscopy prep miralax+miralax dosage colonoscopy+ colonoscopy miralax prep+miralax for colonoscopy+miralax and colonoscopy+miralax gatorade colonoscopy+colonoscopy preparation miralax+miralax before colonoscopy+miralax bowel prep	Not asked in BRFSS
Prepopik to cleanse colon for colonoscopy	prepopik+prepopik dosage+prepopik prep+side effects prepopik+prepopik colonoscopy+colonoscopy prep prepopik	Not asked in BRFSS

Suprep to cleanse colon for colonoscopy	suprep+colonoscopy+colonoscopy prep suprep+suprep dosage colonoscopy+ colonoscopy suprep+suprep for colonoscopy+suprep and colonoscopy+suprep gatorade colonoscopy+colonoscopy preparation suprep +suprep before colonoscopy+suprep bowel prep	Not asked in BRFSS
Free/low-cost colonoscopy	free colonoscopy+low cost colonoscopy+free colonoscopy screening	Not asked in BRFSS
Cost for colonoscopy	cost of colonoscopy+colonoscopy cost+average colonoscopy cost+endoscopy cost	Not asked in BRFSS
Fecal Occult Blood Test (FOBT)	fobt+blood test for colon cancer+colon cancer blood test+screening for colon cancer with blood test+fobt test+fecal occult blood test+blood stool test for cancer	1. A blood stool test is a test that may use a special kit at home to determine whether the stool contains blood. Have you ever had this test using a home kit? 2. How long has it been since you had your last blood stool test using a home kit?
Screening for breast cancer		
Mammography	mammography+breast mammography+breast cancer screening+mammography+mammograms+screening mammography+breast mammogram+breast cancer mammogram+mammo+mammogram screening+mammogram+mammogram results+free mammogram+digital mammography	1. A mammogram is an x-ray of each breast to look for breast cancer. Have you ever had a mammogram? 2. How long has it been since you had your last mammogram?
Digital mammography	digital mammography+mammogram+mammography +digital mammograms+digital mammography screening	Not asked in BRFSS

3D mammography	3D mammography+3D mammogram	Not asked in BRFSS
Free/low-cost mammography	free mammogram+free mammography+low cost mammogram+low cost mammography+free mammogram screening+free mammograms	Not asked in BRFSS
Screening for cervical cancer		
Pap smear	pap test+cervical cancer screening+pap smear test+the pap test+pap smears+free pap smear+free pap+pap tests+pap smear+cervical cancer test+cervical smear+pap testing+papanicolaou	1. A Pap test is a test for cancer of the cervix. Have you ever had a Pap test? 2. How long has it been since you had your last Pap test?
Breast self exam	Breast self exam	Not asked in BRFSS
Screening for prostate cancer		
PSA test	psa test+prostate cancer test+psa testing+prostate test+psa test cancer+prostate+cancer tests+prostate specific antigen test+ prostate psa+ prostate cancer screening tests	1. A Prostate-Specific Antigen test, also called a PSA test, is a blood test used to check men for prostate cancer. Has a doctor EVER recommended that you have a PSA test? 2. Have you EVER HAD a PSA test? 3. How long has it been since you had your last PSA test?
Screening for lung cancer		
Lung cancer screening	lung cancer screening+screening for lung cancer+lung cancer screening CT+CT lung cancer screening	Not asked in BRFSS

BMJ Open

The utility of Google Trends data to examine interest in cancer screening

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2014-006678.R1
Article Type:	Research
Date Submitted by the Author:	16-Apr-2015
Complete List of Authors:	Schootman, Mario; Saint Louis University, Department of Epidemiology, College for Public and Social Justice; Toor, Aroona; Saint Louis University, Department of Epidemiology Cavazos-Rehg, Patricia; Washington University, Department of Psychiatry Jeffe, Donna; Washington University, Department of Internal Medicine McQueen, Amy; Washington University, Department of Internal Medicine Eberth, Jan M.; University of South Carolina, Davidson, Nicholas; Washington University, Department of Internal Medicine
Primary Subject Heading:	Public health
Secondary Subject Heading:	Epidemiology, Gastroenterology and hepatology, Health informatics
Keywords:	EPIDEMIOLOGY, PUBLIC HEALTH, Adult gastroenterology < GASTROENTEROLOGY

The utility of Google Trends data to examine interest in cancer screening

Schootman M,[1,2] Toor A,[1] Cavazos-Rehg P,[3] Jeffe DB,[2,4] McQueen A, [2,4] Eberth J,[5] Davidson NO [2,6]

[1] Saint Louis University College for Public Health and Social Justice
Department of Epidemiology
Saint Louis, MO

[2] Alvin J. Siteman Cancer Center at Barnes-Jewish Hospital and Washington University School of Medicine
Saint Louis, MO

[3] Washington University School of Medicine
Department of Psychiatry
Saint Louis, MO

[4] Washington University School of Medicine
Department of Medicine
Division of General Medical Sciences
Saint Louis, MO

[5] Department of Epidemiology and Biostatistics
Arnold School of Public Health
University of South Carolina, SC

[6] Washington University School of Medicine
Department of Medicine
Division of Gastroenterology
Saint Louis, MO

Address for correspondence:
Mario Schootman, PhD
Saint Louis University
College for Public Health and Social Justice
Department of Epidemiology
3545 Lafayette Ave
Saint Louis, MO 63104
e-mail: schootm@slu.edu
phone: 314-977-8133

Authorship contributions

Author MS was the principal investigator of the study. MS and AT designed and conceptualized the study, with MS overseeing data collection. MS performed the statistical analysis pertaining to the trends over time. MS and AT and wrote sections of the manuscript. AT helped to conceptualize the study, performed some of the data analysis, and edited the manuscript. PC, DJ, AM, JE, and ND helped to conceptualize the study, interpreted the results, and edited the manuscript. AM and JE provided insight into the use of behavioral theories that could help explain the findings. ND provided clinical insight into screening issues pertaining to colorectal cancer.

Abstract

Objectives

We examined the utility of January 2004 – April 2014 Google Trends data from information searches for cancer screenings and preparations as a complement to population screening data, which are traditionally estimated through costly population-level surveys.

Setting

State-level data across the United States

Participants

Persons who searched for terms related to cancer screening using Google and persons who participated in the Behavioral Risk Factor Surveillance System (BRFSS).

Primary and secondary outcome measures

1) State-level Google Trends data, providing relative search volume (RSV) data scaled to the highest search proportion per week (RSV100) for search terms over time since 2004 and across different geographical locations. 2) RSV of new screening tests, free/low cost screening for breast and colorectal cancer, and new preparations for colonoscopy (Prepopik™). 3) State-level breast, cervical, colorectal, and prostate cancer screening rates.

Results

Correlations between Google Trends and BRFSS data ranged from 0.55 for ever having had a colonoscopy to 0.14 for having a Pap smear within the past three years. RSV varied for different screening tests and was highest for prostate vs. other cancer screening sites. Free-low cost mammography and colonoscopy showed higher RSV during their respective cancer awareness

months. RSV for Miralax remained stable, while interest in Prepopik increased over time. RSV for lung cancer screening, virtual colonoscopy and 3D mammography was low.

Conclusions

Google Trends data provides enormous scientific possibilities, but are not a suitable substitute for, but may complement, traditional data collection and analysis about cancer screening and related interests.

Article summary – strengths and limitations of this study

- Google Trends data help identify developing interests in new cancer screening tests or related aspects of specific screening tests.
- Internet searches can be an important source for generating hypotheses about public awareness and interest in cancer screening, evaluating changes in information seeking after targeted interventions or media coverage, and directing new communication campaigns to explain the evidence base for screening tests.
- An evaluation that occurs almost immediately after an intervention may inform policy makers of the associated costs and benefits when there is still interest to make modifications to, or expand, any policy changes.
- The utility of Google Trends to help evaluate interventions depends on the area where the intervention is implemented, since data is only available for states and selected metropolitan areas, limiting its use in rural areas or areas with a low search volume.
- Google Trends data are anonymous, which limits its utility in examining specific subpopulations and disparities among populations. Also, Google Trends data represent only searches done using Google.

Introduction

Cancer screening is a cornerstone of public health aimed at promoting early diagnosis and, in some instances, prevention of cancer. There are several surveillance systems that monitor self-reported cancer screening utilization, including the Behavioral Risk Factor Surveillance System (BRFSS),[1 2] the National Health Interview Survey (NHIS),[3] and the Health Information National Trends Survey (HINTS).[4] These databases have been invaluable in identifying determinants of screening use and describing trends and disparities over time.

These traditional surveillance systems are ill equipped to deal with a rapidly changing digital world with a need for timely health data for public health and medical professionals, policy makers, and the public who influence policy choices. Traditional surveillance approaches are expensive to maintain due to their use of survey interview methods for data collection and the time required to aggregate the data. In addition, these older methods require participation of a large study population to estimate screening use accurately, they rely on self-report resulting in potential recall bias; and, for the BRFSS and HINTS, participants include only persons with landline telephones and, more recently, mobile phones and a mailing address to complete a self-administered questionnaire, leaving the door open for potential selection bias. Other limitations of traditional surveillance approaches include the failure to capture new and emerging screening modalities (e.g., virtual colonoscopy for colorectal cancer, magnetic resonance imaging for breast cancer detection, or low-dose spiral computed tomography for lung cancer screening among persons at high risk for lung cancer) especially when use is still low. As a result, population-based prevalence of newer screening methods is unknown.

Recent technological advances in data acquisition, such as Google Trends, may allow for more timely data collection to learn about trends in interest in various health-related topics, including cancer screening. Google Trends is a keyword research tool that provides near real-time trend data regarding interest as operationalized by internet search volume. Both Google and Yahoo! search engines have been used to analyze different types of search queries, for example

about cancer incidence,[5] cancer mortality,[5] kidney stones,[6] non-cigarette tobacco use,[7] sexually transmitted infections,[8] and flu trends.[9 10] However, the value of Google Trends in illuminating search trends reflecting interest in cancer screening and related topics has not yet been examined. Depending on its utility, Google Trends may complement existing surveillance systems that monitor screening use.

Here, we examined the utility of Google Trends relative to the BRFSS, focusing on cancer screening. Specifically, we examined 1) the correlation between 2012 Google Trends and self-reported breast, cervical, colorectal, and prostate cancer screening in the 2012 BRFSS, and 2) interest in possible new and developing screening modalities and preparations not currently captured in existing surveillance systems since 2004.

Methods

Data sources about screening use

Prevalence data about breast cancer screening (mammography and breast self-exam), cervical cancer screening (Pap smear), colorectal cancer screening (fecal occult blood test [FOBT], colonoscopy), and prostate cancer screening using prostate screening antigen (PSA) test were all obtained from the 2012 BRFSS database http://apps.nccd.cdc.gov/brfss/.[11] The BRFSS is one of the largest annual telephone health-survey database systems in the world. The survey provides state-level prevalence data of the major behavioral risks among adults associated with premature morbidity and mortality among adults. Data are collected from all 50 U.S. states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, Guam, American Samoa, and Palau. Questions about cancer screening use have been validated.[12] In this study, we included BRFSS data from all 50 U.S. states to calculate correlations between reported screening use and Google Trends search volume. Mammography use in the past two years was calculated among women aged 40 or older. Pap smear use among women aged 18 or older was estimated within the past three years. FOBT use in the past two years was calculated among men and women aged 50 or older. Colonoscopy

use was defined as having ever had a colonoscopy among men and women aged 50 or older. PSA testing prevalence was defined as a PSA test within the past two years among men aged 40 or older.

Google Trends (http://www.google.com/trends/explore#cmpt=q), based on Google Search, the most widely used internet search engine, offers search volume data for search terms over time since 2004 and across different geographical locations. Google Trends shows how often search terms are entered in Google relative to the total search volume in a region or globally. Google Trends produces relative search volume (RSV) scaled to the highest search proportion week. RSV values are by definition always less than 100 and demonstrate how other weekly search proportions compared to the highest (RSV=100) search proportion. For example, RSV=50 represents 50% of the highest observed search proportion during the study period. RSV indirectly corrects for population size and Internet access, both of which increased during the study period and would bias any absolute search volume measure. However, RSV allows for directly comparing search volume across search terms.

Google Trends can compile search volume for up to 30 words. We selected search terms a priori based on their face validity for the term's relationship to the screening test of interest. Google Trends allows up to four strata for different trend data. We included additional search terms in our main search if these additional strata increased RSV by at least 1 point. We also added search terms based on popular "related terms" suggested by Google Trends. We included singular and plural forms of the search terms. Appendix 1 shows the specific search terms used for each screening test and associated terms relevant to specific tests (e.g., Miralax for colonoscopy). In addition to obtaining search volume data about interest in existing screening tests, we examined search volume data regarding new screening tests (virtual colonoscopy, lung cancer screening using computed tomography [CT], 3D mammography), free/low cost screening for breast and colorectal cancer, and new preparations to cleanse the colon for colonoscopy (Prepopik™).

Prepopik™ was approved on July 16, 2012 by the Food and Drug Administration to help cleanse the colon in adults preparing for colonoscopy.[13]

Statistical analysis

We used the Pearson correlation coefficient to examine the associations between state-level Google Trends RSV and BRFSS state-level screening prevalence for each of the five cancer screening tests. We weighted these correlations by the 2011 state population estimates from the Bureau of the Census using weighted regression because such estimates provide more weight to states with larger populations. We used Stata 13.1 to calculate weighted correlations using the wls0 command.

We used the joinpoint methodology to identify significant changes in weekly RSV over time for each of the screening tests and associated interests.[14 15] The joinpoint methodology is ideally suited to examine trends over time and to test whether an apparent change in trend is statistically significant, which other methods (e.g., autoregressive integrated moving average [ARIMA] analysis) may miss. Linear trends in search volume were summarized using the estimated annual percentage change (EAPC). The EAPC was calculated by fitting a linear regression to the natural logarithm of the weekly RSV, using week as a regression variable. Joinpoint regression tests were used to identify an inflection point (hereafter, called joinpoint) with a significant change in the slope of the trend.[14 15] For our analysis, a minimum of four weeks between two joinpoints was required, and a maximum of three joinpoints was allowed to describe the data.

Results

Colorectal cancer screening

The weighted correlation between ever having had a colonoscopy based on 2012 BRFSS data and 2004-2012 Google Trends colonoscopy data was 0.55. Figure 1a shows the weekly Google Trends RSV for colorectal cancer screening using colonoscopy between January, 2004

and April, 2014. The average RSV was 61.9 in 2004 and increased to 85.8 during the last 52 weeks of data. During the first 3 years, RSV per week remained stable, but then increased 0.2 percent per week (95% CI: 0.1; 0.2). Starting at week 308 (November, 2009), RSV increased 0.09 percent per week (95% CI: 0.07; 0.11). RSV was lowest during December of each year and slightly higher during March of each year (average: 74.3).

During 2007, the average RSV/week for virtual colonoscopy was 22.5, but RSV decreased 0.30 percent per week (-0.33; -0.27) starting in January 2008 (Figure 1b). RSV/week for Miralax as a colon cleanser declined 0.50 percent per week (95% CI: -0.69; -0.30) during January 2009 through August 2010, after which RSV about Miralax remained stable until April 2014 (Figure 1c). The RSV/week for Prepopik, a newer colon cleanser approved by the FDA in July 2012, increased rapidly over time.

For FOBT use, Google Trends data was available for only eight states due to low search volume, and a correlation between BRFSS data about FOBT use and Google Trends RSV could not be calculated.

Breast cancer screening

The weighted correlation between Google Trends RSV and BRFSS-based mammography use was 0.36. Figure 2a shows RSV/week for mammography over time. Peaks were present during October each year and about 10 points higher than during December, the month with the lowest RSV. In November 2009, mammography RSV was highest during this 10-year period. Figure 2b shows Google Trends RSV/week for free/low-cost mammography, which peaked in October every year.

Cervical cancer screening

The weighted correlation between 2012 BRFSS-based Pap smear use and RSV for Pap smears during 2010-2012 was 0.14. Figure 3 shows that during week 1-137, RSV/week for pap

smear increased slightly (0.08 percent per week; 95% CI: 0.03; 0.13), remained stable during weeks 137-208, increased during weeks 208-426 (0.13 percent per week; 95% CI: 0.11; 0.16), but then decreased starting in week 426 (-0.11 percent per week; 95% CI: -0.18; -0.04).

Prostate cancer screening

The weighted correlation between Google Trends and BRFSS-based PSA use was 0.42. RSV for PSA declined very slowly (0.05 percent per week) starting in 2004 (95% CI: -0.06; -0.05) until October 2009 (week 302), after which the decline became steeper at 0.20 percent per week (95% CI: -0.30; -0.11) until December 2010 (week 364), then there were three weeks during which RSV remained stable (Figure 4). Starting in January 2011 (week 367), RSV declined 0.05 percent per week (95% CI: -0.07; -0.03). RSV for PSA was highest for week 272 (March, 2009).

Lung cancer screening

Between January 2007 and July 2010, RSV about lung cancer screening declined 1.1 percent per month (95% CI: -1.7; -0.5), but then increased 2.8 percent per month (95% CI: 2.3; 3.4) until April 2014 (Figure 5). There was a peak in RSV about lung cancer screening during November 2010 (month 47).

Discussion

We examined the utility of Google Trends relative to the BRFSS, one of the existing surveillance systems focusing on cancer screening. Correlations between Google Trends and BRFSS data ranged from a high of 0.55 for ever having had a colonoscopy to a low of 0.14 for having a Pap smear within the past three years. Although self-reported screening use is a less than perfect measure of behavior,[12] these modest correlations between data sources indicate that they are measuring different constructs: Google Trends provides estimates of the public's interest in learning more about cancer screening tests; the BRFSS and other surveillance systems provide

estimates of self-reported use of these tests. However, correlations between the two data sources varied across screening types. One reason for the lower correlation related to cervical cancer screening may be that Pap smear use is very common and often part of routine primary care visits, resulting in lower information seeking.[16]

Based on our findings, there appears to be some utility of Google Trends data relative to existing surveillance systems to monitor cancer screening. Awareness and interest in cancer screening is a necessary, but not sufficient, determinant of screening behavior.[17 18] Search volume data using Google Trends enabled us to measure the public's awareness and interest in possible new and developing screening modalities (e.g., virtual colonoscopy, digital mammography, 3D mammography, computed tomography for lung cancer screening) and screening test preparations (e.g., Prepopik versus Miralax), which are not currently captured in existing surveillance systems. By harnessing real-time search-engine data around national media-based interventions (e.g., CDC's Tips from Former Smokers), programs can be evaluated as they are implemented, generating timely feedback to assess the effectiveness of interventions to increase interest in cancer screening, prevention, and other public health recommendations. Such adaptive designs using accumulating data to modify the intervention's course [19 20] have been used infrequently in community-based evaluations. Adaptive interventions that can be evaluated using interest and awareness may be especially useful. It appears that in some instances an increase in public interest in cancer screening is associated with the timing of news reports, celebrity cancer diagnosis, and advertisements.[21] For example, the increase in search volume each October coincides with news stories and advertisements during Breast Cancer Awareness Month. Search volume for colon cancer screening was also slightly higher during March, Colon Cancer Awareness Month. Google Trends also identified a large interest in November 2009 when search volume about mammography increased dramatically likely in response to critics citing health care rationing in response to new mammography guidelines from the US Preventive Services Task Force.[22] The panel recommended that most women wait until age 50 to start routine mammography, then get

the exam every two years instead of annually. For example, in March 2009 RSV for prostate cancer screening increased following coverage of two studies showing that prostate cancer screening did not reduce the risk of death.[23] Also, in November 2010, RSV for lung cancer screening increased after trials reported its potential to reduce the risk of death among heavy smokers.[24] The utility of Google Trends to help adapt interventions is limited by the area where the intervention is implemented, since data is only available at the state-level and for selected large metropolitan areas, limiting its use in rural areas or areas with a low search volume. Consequently, disparities in cancer screening are difficult to examine using these data. Additionally, Google Trends data is unable to evaluate interventions using outcomes such as behaviors or disease development.

Internet searches using Google Trends can guide the development of traditional surveillance systems surveys, such as the BRFSS, NHIS, and HINTS, by vetting the inclusion of questions on surveys. Google Trends data can also gauge developing awareness and interests in new cancer screening tests (e.g., virtual colonoscopy) or related aspects of specific screening tests (e.g., about preparation for colonoscopy). For example, Google Trends showed that interest in lung cancer screening and virtual colonoscopy is still very low, while interest in prostate cancer screening is very high even though PSA tests have been shown to be not very effective in reducing risk of death.[25] Interest in virtual colonoscopy, despite showing promise as a screening tool relative to traditional colonoscopy,[26] was very low. For the most part, screening colonoscopy remains the first-line strategy for the detection of adenomas, with a lower miss rate than virtual colonoscopy, no radiation exposure, and offers therapeutic removal of polyps as well.[27] Internet searches can be an important source of information for generating hypotheses about public awareness and interest in cancer screening, evaluating changes in information seeking after targeted interventions or media coverage, and directing new communication campaigns to explain the evidence base for screening tests.

Search query results may also be politically relevant. Since policy changes often require public support, evaluation strategies that take years to perform may not provide relevant feedback to public interest groups and voters. Instead, an evaluation that occurs almost immediately after the policy change may inform policy makers and their supporters of the associated costs and benefits when there is still interest to make modifications to, or expand, the policy change.[28] For example, the interest in and implementation of free/low cost breast and colorectal cancer screening can be evaluated. The CDC and local organizations implemented free/low cost mammograms starting in the 1990s across the United States followed by free/low cost colonoscopies in selected locations to eligible participants. The potential need for and likely early success (e.g., awareness) of the expansion of these interventions could be gleaned through Google Trends data, much earlier than traditional evaluation strategies.

The utility of Google Trends data should be viewed in light of its limitations. Google Trends data are anonymous, which limits its utility in examining specific subpopulations and disparities among populations. Also, Google Trends data represent only searches done using Google. However, Google accounts for an estimated 65 percent of all internet searches.[29] Google Trends data may have sampling biases. However, such biases are increasingly eroding at the population level as more and more people search for information online. Google Trends does eliminate repeated queries from the same user over a short period of time to reduce counts of continued searching. Google Trends uses a certain threshold of traffic volume so that very new search terms are assigned a value of zero, but this could change very quickly. The motivation of Google users is not known. As a corollary, the data obtained from Google trends cannot be independently verified. Also, the researcher has no control over the data, making quality control difficult. Understanding local health information-seeking behaviors also may be important, but Google Trends data may not be available for geographic areas smaller than at the state level depending on search volume. Additionally, a user option to download Google Trends data for different time periods (e.g., by month or seasonal) is not currently available. One remedy that would circumvent many of these

limitations is the release of actual search volume data rather than relative search volume data. Finally, search terms entered in other languages were not captured by this study, but could be used to examine interest among non-English speaking populations.

Although Google Trends' "big data" approach provides enormous scientific possibilities, they are not a substitute for, but may complement, traditional data collection and analysis of cancer preventive behavior. The strengths of Google Trends to provide data about the public's interests in cancer screening, despite its inability to provide cancer screening usage data, can foster provision of timely feedback about interventions aimed at increasing interest in cancer screening and other public health recommendations.

Funding

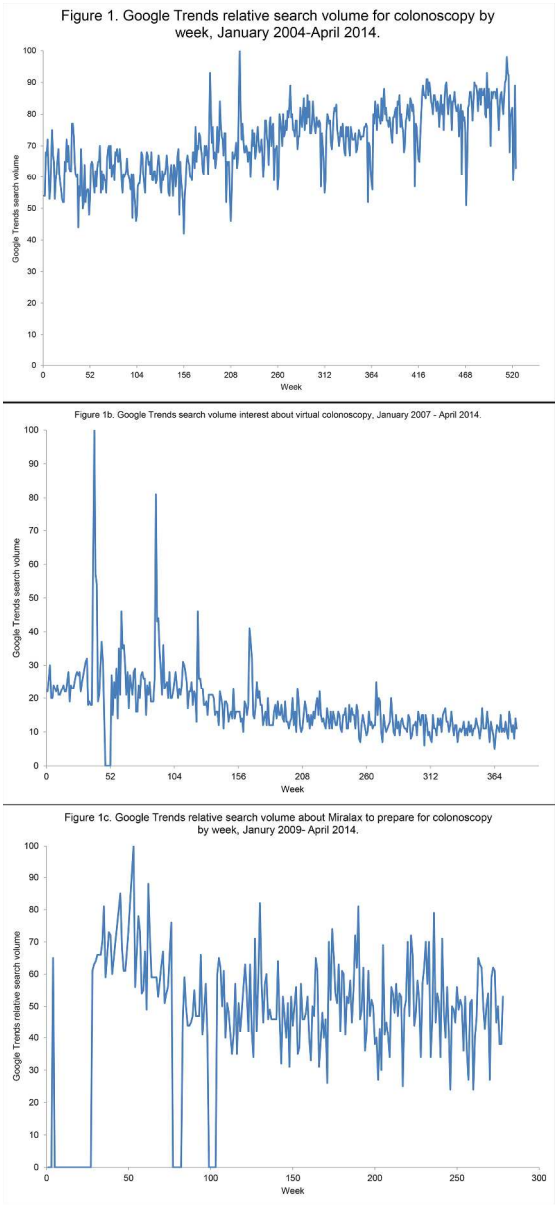
This work was supported by grants from the National Cancer Institute at the National Institutes of Health (grant number R01 CA112159). We thank the Alvin J. Siteman Cancer Center at Barnes-Jewish Hospital and Washington University School of Medicine in St. Louis, Missouri, for the use of the Health Behavior, Communication, and Outreach Core, which is supported in part by the National Cancer Institute Cancer Center Support Grant (grant number P30 CA91842) to the Alvin J. Siteman Cancer Center. Dr. Davidson was supported in part through grants HL-38180, DK-56260, and Digestive Disease Research Core Center DK-52574.

Data sharing statement: No additional data are available.

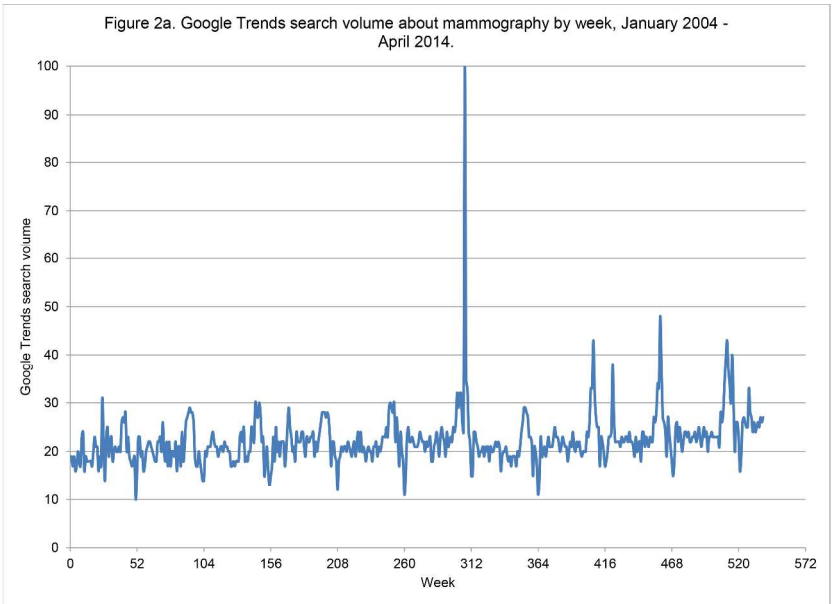
References

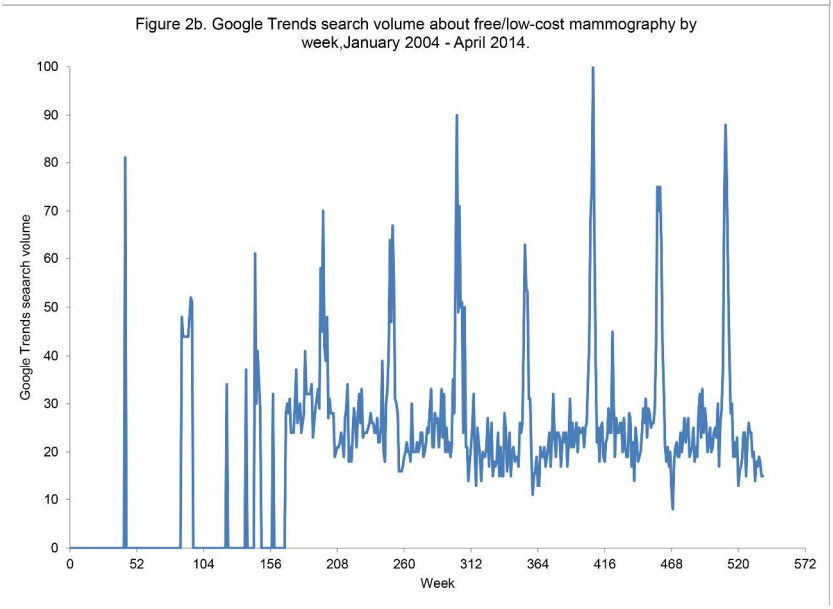
1. Joseph DA, King JB, Miller JW, et al. Prevalence of colorectal cancer screening among adults--Behavioral Risk Factor Surveillance System, United States, 2010. MMWR Morbidity and mortality weekly report 2012;**61 Suppl**:51-6.
2. Miller JW, King JB, Joseph DA, et al. Breast cancer screening among adult women--Behavioral Risk Factor Surveillance System, United States, 2010. MMWR Morbidity and mortality weekly report 2012;**61 Suppl**:46-50.
3. Hiatt RA, Klabunde C, Breen N, et al. Cancer screening practices from National Health Interview Surveys: past, present, and future. Journal of the National Cancer Institute 2002;**94**(24):1837-46.
4. Ashok M, Berkowitz Z, Hawkins NA, et al. Recency of Pap testing and future testing plans among women aged 18-64: analysis of the 2007 Health Information National Trends Survey. Journal of women's health (2002) 2012;**21**(7):705-12.
5. Cooper CP, Mallon KP, Leadbetter S, et al. Cancer Internet Search Activity on a Major Search Engine, United States 2001-2003. J Med Int Res 2005;**7**(3):e36.
6. Breyer BN, Sen S, Aaronson DS, et al. Use of Google Insights for Search to Track Seasonal and Geographic Kidney Stone Incidence in the United States. Urology 2011;**78**(2):267-71.
7. Cavazos-Rehg PA, Krauss MJ, Spitznagel EL, et al. Monitoring of non-cigarette tobacco use using Google Trends. Tobacco control 2014.
8. Johnson AK, Mehta SD. A Comparison of Internet Search Trends and Sexually Transmitted Infection Rates Using Google Trends. Sexually Transmitted Diseases 2014;**41**(1):61-63 10.1097/OLQ.0000000000000065.
9. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 2009;**49**(10):1557-64.
10. Pervaiz F, Pervaiz M, Abdur Rehman N, et al. FluBreaks: early epidemic detection from Google flu trends. Journal of medical Internet research 2012;**14**(5):e125.
11. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System, 2014: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2014.
12. Pierannunzi C, Hu SS, Balluz L. A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004-2011. BMC Medical Research Methodology 2013;**13**(1):49.
13. Administration FaD. FDA approves new colon-cleansing drug for colonoscopy prep, July 17, 2012.
14. Kim HJ, Fay MP, Feuer EJ, et al. Permutation tests for joinpoint regression with applications to cancer rates. Statistics in Medicine 2000;**19**:335-51.
15. Joinpoint Regression Program, Version 4.1.0. [software program], 2014.
16. Redmond N, Baer HJ, Clark CR, et al. Sources of health information related to preventive health behaviors in a national study. American Journal of Preventive Medicine 2010;**38**(6):620-27.e2.
17. Fishbein M. The role of theory in HIV prevention. AIDS Care 2000;**12**(3):273-8.
18. Weinstein ND. The precaution adoption process. Health Psychology 1988;**7**(4):355-86.
19. Chang M, Chow SC, Pong A. Adaptive design in clinical research: issues, opportunities, and recommendations. Journal of biopharmaceutical statistics 2006;**16**(3):299-309; discussion 11-2.
20. Coffey CS, Levin B, Clark C, et al. Overview, hurdles, and future work in adaptive designs: perspectives from a National Institutes of Health-funded workshop. Clinical trials (London, England) 2012;**9**(6):671-80.

21. Glynn RW, Kelly JC, Coffey N, et al. The effect of breast cancer awareness month on internet search activity--a comparison with awareness campaigns for lung and prostate cancer. BMC cancer 2011;**11**:442.
22. Stein R. In wake of mammography guidelines, U.S. health task force faces new scrutiny.: Washington Post, ecember 20, 2009.
23. Parker-Pope T. For men, to screen or not to screen: The New York Times, March 23, 2009.
24. McCook A. More signs lung cancer screening could save lives: Reuters, December 28, 2010.
25. U.S. Preventive Services Task Force. Screening for Prostate Cancer, Topic Page. . Secondary Screening for Prostate Cancer, Topic Page. 2012. http://www.uspreventiveservicestaskforce.org/prostatecancerscreening.htm.
26. Kim DH, Pickhardt PJ, Taylor AJ, et al. CT colonography versus colonoscopy for the detection of advanced neoplasia. The New England journal of medicine 2007;**357**(14):1403-12.
27. Than M, Witherspoon J, Shami J, et al. Diagnostic miss rate for colorectal cancer: an audit. Annals of gastroenterology : quarterly publication of the Hellenic Society of Gastroenterology 2015;**28**(1):94-98.
28. Ayers JW, Ribisl K, Brownstein JS. Using Search Query Surveillance to Monitor Tax Avoidance and Smoking Cessation following the United States' 2009 “SCHIP” Cigarette Tax Increase. PLoS ONE 2011;**6**(3):e16777.
29. Sullivan D. Google still world’s most popular search engine by far, but share of unique searchers dips slightly, February 11, 2013.

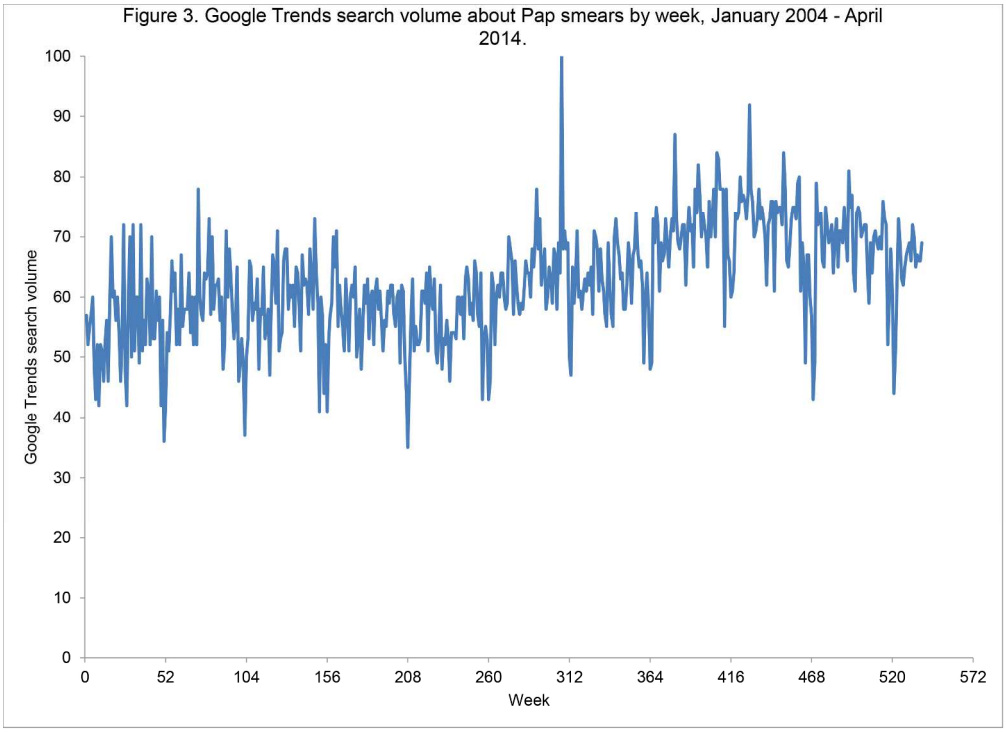


152x332mm (300 x 300 DPI)

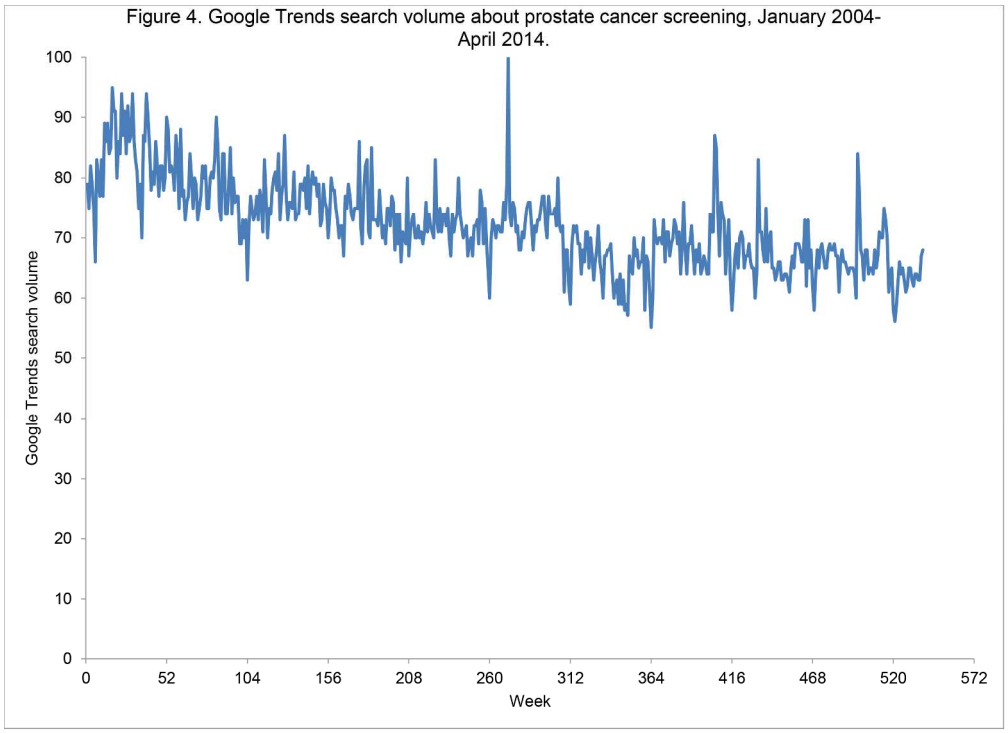




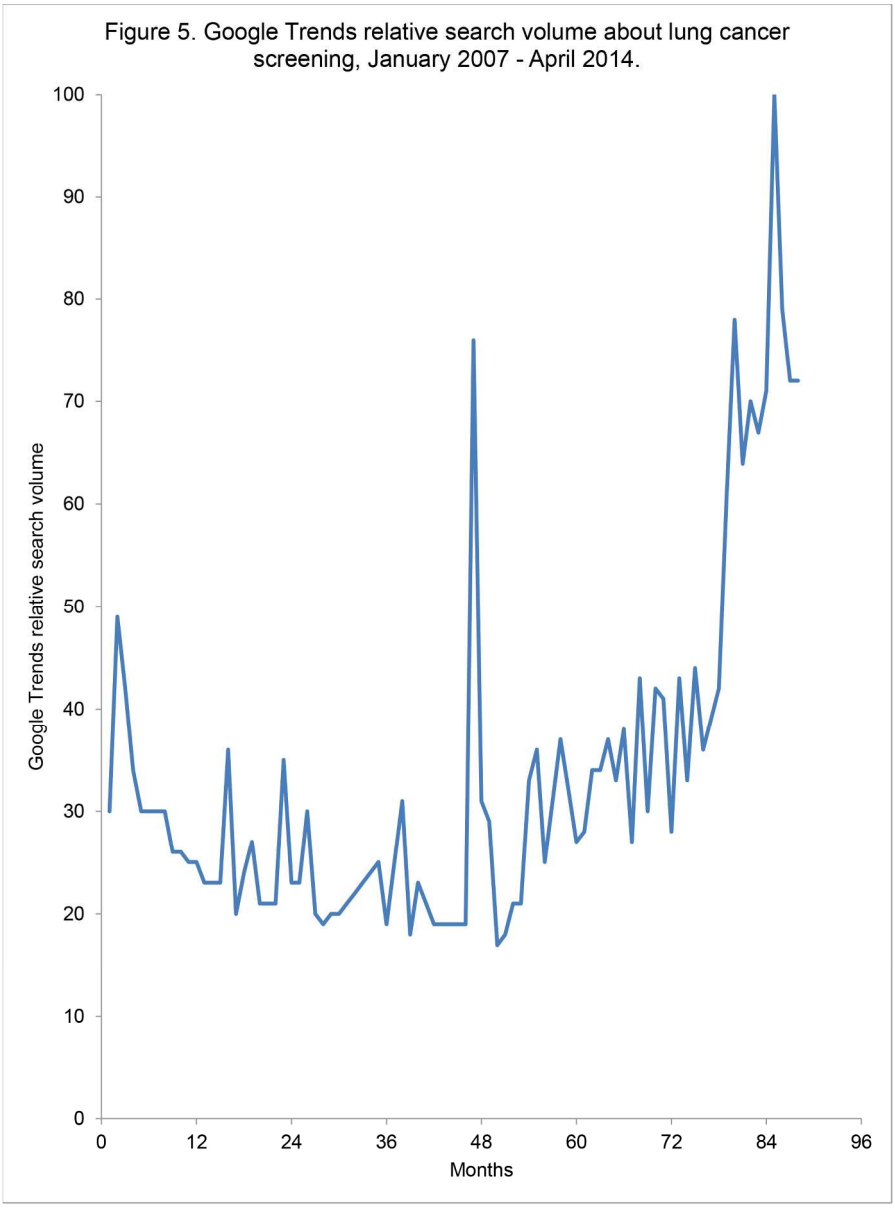
241x349mm (300 x 300 DPI)



241x175mm (300 x 300 DPI)



241x175mm (300 x 300 DPI)



177x239mm (300 x 300 DPI)

Appendix 1: Google Trends search terms used for each screening test and associated interests and Behavioral Risk Factor Surveillance Survey question.

Concept	Google Trends Search terms used	Behavioral Risk Factor Surveillance Survey question (BRFSS)
Screening for colorectal cancer		
Colonoscopy	Colonoscopy+ colonoscopy procedure +virtual colonoscopy+endoscopy +miralax prep+colonoscopy procedure+endoscopy procedure+what is colonoscopy+prep for colonoscopy+miralax+bowel prep+colonoscopy prep+colon cancer screening+colon cancer test	1. Sigmoidoscopy and colonoscopy are exams in which a tube is inserted in the rectum to view the colon for signs of cancer or other health problems. Have you ever had either of these exams? 2. For a SIGMOIDOSCOPY, a flexible tube is inserted into the rectum to look for problems. A COLONOSCOPY is similar, but uses a longer tube, and you are usually given medication through a needle in your arm to make you sleepy and told to have someone else drive you home after the test. Was your MOST RECENT exam a sigmoidoscopy or a colonoscopy? 3. How long has it been since you had your last sigmoidoscopy or colonoscopy?
Virtual colonoscopy	virtual colonoscopy+ct colonography+virtual colonoscopy cost+ct colonoscopy	Not asked in BRFSS
Miralax to cleanse colon for colonoscopy	miralax colonoscopy+colonoscopy prep miralax+miralax dosage colonoscopy+ colonoscopy miralax prep+miralax for colonoscopy+miralax and colonoscopy+miralax gatorade colonoscopy+colonoscopy preparation miralax+miralax before colonoscopy+miralax bowel prep	Not asked in BRFSS
Prepopik to cleanse colon for colonoscopy	prepopik+prepopik dosage+prepopik prep+side effects prepopik+prepopik colonoscopy+colonoscopy prep prepopik	Not asked in BRFSS

Suprep to cleanse colon for colonoscopy	suprep+colonoscopy+colonoscopy prep suprep+suprep dosage colonoscopy+ colonoscopy suprep+suprep for colonoscopy+suprep and colonoscopy+suprep gatorade colonoscopy+colonoscopy preparation suprep +suprep before colonoscopy+suprep bowel prep	Not asked in BRFSS
Free/low-cost colonoscopy	free colonoscopy+low cost colonoscopy+free colonoscopy screening	Not asked in BRFSS
Cost for colonoscopy	cost of colonoscopy+colonoscopy cost+average colonoscopy cost+endoscopy cost	Not asked in BRFSS
Fecal Occult Blood Test (FOBT)	fobt+blood test for colon cancer+colon cancer blood test+screening for colon cancer with blood test+fobt test+fecal occult blood test+blood stool test for cancer	1. A blood stool test is a test that may use a special kit at home to determine whether the stool contains blood. Have you ever had this test using a home kit? 2. How long has it been since you had your last blood stool test using a home kit?
Screening for breast cancer		
Mammography	mammography+breast mammography+breast cancer screening+mammography+mammograms+screening mammography+breast mammogram+breast cancer mammogram+mammo+mammogram screening+mammogram+mammogram results+free mammogram+digital mammography	1. A mammogram is an x-ray of each breast to look for breast cancer. Have you ever had a mammogram? 2. How long has it been since you had your last mammogram?
Digital mammography	digital mammography+mammogram+mammography +digital mammograms+digital mammography screening	Not asked in BRFSS

3D mammography	3D mammography+3D mammogram	Not asked in BRFSS
Free/low-cost mammography	free mammogram+free mammography+low cost mammogram+low cost mammography+free mammogram screening+free mammograms	Not asked in BRFSS
Screening for cervical cancer		
Pap smear	pap test+cervical cancer screening+pap smear test+the pap test+pap smears+free pap smear+free pap+pap tests+pap smear+cervical cancer test+cervical smear+pap testing+papanicolaou	1. A Pap test is a test for cancer of the cervix. Have you ever had a Pap test? 2. How long has it been since you had your last Pap test?
Breast self exam	Breast self exam	Not asked in BRFSS
Screening for prostate cancer		
PSA test	psa test+prostate cancer test+psa testing+prostate test+psa test cancer+prostate+cancer tests+prostate specific antigen test+ prostate psa+ prostate cancer screening tests	1. A Prostate-Specific Antigen test, also called a PSA test, is a blood test used to check men for prostate cancer. Has a doctor EVER recommended that you have a PSA test? 2. Have you EVER HAD a PSA test? 3. How long has it been since you had your last PSA test?
Screening for lung cancer		
Lung cancer screening	lung cancer screening+screening for lung cancer+lung cancer screening CT+CT lung cancer screening	Not asked in BRFSS

BMJ Open

The utility of Google Trends data to examine interest in cancer screening

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2014-006678.R2
Article Type:	Research
Date Submitted by the Author:	12-May-2015
Complete List of Authors:	Schootman, Mario; Saint Louis University, Department of Epidemiology, College for Public and Social Justice; Toor, Aroona; Saint Louis University, Department of Epidemiology Cavazos-Rehg, Patricia; Washington University, Department of Psychiatry Jeffe, Donna; Washington University, Department of Internal Medicine McQueen, Amy; Washington University, Department of Internal Medicine Eberth, Jan M.; University of South Carolina, Davidson, Nicholas; Washington University, Department of Internal Medicine
Primary Subject Heading:	Public health
Secondary Subject Heading:	Epidemiology, Gastroenterology and hepatology, Health informatics
Keywords:	EPIDEMIOLOGY, PUBLIC HEALTH, Adult gastroenterology < GASTROENTEROLOGY

The utility of Google Trends data to examine interest in cancer screening

Schootman M,[1,2] Toor A,[1] Cavazos-Rehg P,[3] Jeffe DB,[2,4] McQueen A, [2,4] Eberth J,[5] Davidson NO [2,6]

[1] Saint Louis University College for Public Health and Social Justice
Department of Epidemiology
Saint Louis, MO

[2] Alvin J. Siteman Cancer Center at Barnes-Jewish Hospital and Washington University School of Medicine
Saint Louis, MO

[3] Washington University School of Medicine
Department of Psychiatry
Saint Louis, MO

[4] Washington University School of Medicine
Department of Medicine
Division of General Medical Sciences
Saint Louis, MO

[5] Department of Epidemiology and Biostatistics
Arnold School of Public Health
University of South Carolina, SC

[6] Washington University School of Medicine
Department of Medicine
Division of Gastroenterology
Saint Louis, MO

Address for correspondence:
Mario Schootman, PhD
Saint Louis University
College for Public Health and Social Justice
Department of Epidemiology
3545 Lafayette Ave
Saint Louis, MO 63104
e-mail: schootm@slu.edu
phone: 314-977-8133

Authorship contributions

Author MS was the principal investigator of the study. MS and AT designed and conceptualized the study, with MS overseeing data collection. MS performed the statistical analysis pertaining to the trends over time. MS and AT and wrote sections of the manuscript. AT helped to conceptualize the study, performed some of the data analysis, and edited the manuscript. PC, DJ, AM, JE, and ND helped to conceptualize the study, interpreted the results, and edited the manuscript. AM and JE provided insight into the use of behavioral theories that could help explain the findings. ND provided clinical insight into screening issues pertaining to colorectal cancer.

Abstract

Objectives

We examined the utility of January 2004 – April 2014 Google Trends data from information searches for cancer screenings and preparations as a complement to population screening data, which are traditionally estimated through costly population-level surveys.

Setting

State-level data across the United States

Participants

Persons who searched for terms related to cancer screening using Google and persons who participated in the Behavioral Risk Factor Surveillance System (BRFSS).

Primary and secondary outcome measures

1) State-level Google Trends data, providing relative search volume (RSV) data scaled to the highest search proportion per week (RSV100) for search terms over time since 2004 and across different geographical locations. 2) RSV of new screening tests, free/low cost screening for breast and colorectal cancer, and new preparations for colonoscopy (Prepopik™). 3) State-level breast, cervical, colorectal, and prostate cancer screening rates.

Results

Correlations between Google Trends and BRFSS data ranged from 0.55 for ever having had a colonoscopy to 0.14 for having a Pap smear within the past three years. Free-low cost mammography and colonoscopy showed higher RSV during their respective cancer awareness months. RSV for Miralax remained stable, while interest in Prepopik increased over time. RSV for lung cancer screening, virtual colonoscopy and 3D mammography was low.

Conclusions

Google Trends data provides enormous scientific possibilities, but are not a suitable substitute for, but may complement, traditional data collection and analysis about cancer screening and related interests.

Article summary – strengths and limitations of this study

- Google Trends data help identify developing interests in new cancer screening tests or related aspects of specific screening tests.
- Internet searches can be an important source for generating hypotheses about public awareness and interest in cancer screening, evaluating changes in information seeking after targeted interventions or media coverage, and directing new communication campaigns to explain the evidence base for screening tests.
- An evaluation that occurs almost immediately after an intervention may inform policy makers of the associated costs and benefits when there is still interest to make modifications to, or expand, any policy changes.
- The utility of Google Trends to help evaluate interventions depends on the area where the intervention is implemented, since data is only available for states and selected metropolitan areas, limiting its use in rural areas or areas with a low search volume.
- Google Trends data are anonymous, which limits its utility in examining specific subpopulations and disparities among populations. Also, Google Trends data represent only searches done using Google.

Introduction

Cancer screening is a cornerstone of public health aimed at promoting early diagnosis and, in some instances, prevention of cancer. There are several surveillance systems that monitor self-reported cancer screening utilization, including the Behavioral Risk Factor Surveillance System (BRFSS),[1 2] the National Health Interview Survey (NHIS),[3] and the Health Information National Trends Survey (HINTS).[4] These databases have been invaluable in identifying determinants of screening use and describing trends and disparities over time.

These traditional surveillance systems are ill equipped to deal with a rapidly changing digital world with a need for timely health data for public health and medical professionals, policy makers, and the public who influence policy choices. Traditional surveillance approaches are expensive to maintain due to their use of survey interview methods for data collection and the time required to aggregate the data. In addition, these older methods require participation of a large study population to estimate screening use accurately, they rely on self-report resulting in potential recall bias; and, for the BRFSS and HINTS, participants include only persons with landline telephones and, more recently, mobile phones and a mailing address to complete a self-administered questionnaire, leaving the door open for potential selection bias. Other limitations of traditional surveillance approaches include the failure to capture new and emerging screening modalities (e.g., virtual colonoscopy for colorectal cancer, magnetic resonance imaging for breast cancer detection, or low-dose spiral computed tomography for lung cancer screening among persons at high risk for lung cancer) especially when use is still low. As a result, population-based prevalence of newer screening methods is unknown.

Recent technological advances in data acquisition, such as Google Trends, may allow for more timely data collection to learn about trends in interest in various health-related topics, including cancer screening. Google Trends is a keyword research tool that provides near real-time trend data regarding interest as operationalized by internet search volume. Both Google and Yahoo! search engines have been used to analyze different types of search queries, for example

about cancer incidence,[5] cancer mortality,[5] kidney stones,[6] non-cigarette tobacco use,[7] sexually transmitted infections,[8] and flu trends.[9 10] However, the value of Google Trends in illuminating search trends reflecting interest in cancer screening and related topics has not yet been examined. Depending on its utility, Google Trends may complement existing surveillance systems that monitor screening use.

Here, we examined the utility of Google Trends relative to the BRFSS, focusing on cancer screening. Specifically, we examined 1) the correlation between 2012 Google Trends and self-reported breast, cervical, colorectal, and prostate cancer screening in the 2012 BRFSS, and 2) interest in possible new and developing screening modalities and preparations not currently captured in existing surveillance systems since 2004.

Methods

Data sources about screening use

Prevalence data about breast cancer screening (mammography and breast self-exam), cervical cancer screening (Pap smear), colorectal cancer screening (fecal occult blood test [FOBT], colonoscopy), and prostate cancer screening using prostate screening antigen (PSA) test were all obtained from the 2012 BRFSS database http://apps.nccd.cdc.gov/brfss/.[11] The BRFSS is one of the largest annual telephone health-survey database systems in the world. The survey provides state-level prevalence data of the major behavioral risks among adults associated with premature morbidity and mortality among adults. Data are collected from all 50 U.S. states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, Guam, American Samoa, and Palau. Questions about cancer screening use have been validated.[12] In this study, we included BRFSS data from all 50 U.S. states to calculate correlations between reported screening use and Google Trends search volume. Mammography use in the past two years was calculated among women aged 40 or older. Pap smear use among women aged 18 or older was estimated within the past three years. FOBT use in the past two years was calculated among men and women aged 50 or older. Colonoscopy

use was defined as having ever had a colonoscopy among men and women aged 50 or older. PSA testing prevalence was defined as a PSA test within the past two years among men aged 40 or older.

Google Trends (http://www.google.com/trends/explore#cmpt=q), based on Google Search, the most widely used internet search engine, offers search volume data for search terms over time since 2004 and across different geographical locations. Google Trends shows how often search terms are entered in Google relative to the total search volume in a region or globally. Google Trends produces relative search volume (RSV) scaled to the highest search proportion week. RSV values are by definition always less than 100 and demonstrate how other weekly search proportions compared to the highest (RSV=100) search proportion. For example, RSV=50 represents 50% of the highest observed search proportion during the study period. RSV indirectly corrects for population size and Internet access, both of which increased during the study period and would bias any absolute search volume measure. However, RSV allows for directly comparing search volume across search terms.

Google Trends can compile search volume for up to 30 words. We selected search terms a priori based on their face validity for the term's relationship to the screening test of interest. Google Trends allows up to four strata for different trend data. We included additional search terms in our main search if these additional strata increased RSV by at least 1 point. We also added search terms based on popular "related terms" suggested by Google Trends. We included singular and plural forms of the search terms. Appendix 1 shows the specific search terms used for each screening test and associated terms relevant to specific tests (e.g., Miralax for colonoscopy). In addition to obtaining search volume data about interest in existing screening tests, we examined search volume data regarding new screening tests (virtual colonoscopy, lung cancer screening using computed tomography [CT], 3D mammography), free/low cost screening for breast and colorectal cancer, and new preparations to cleanse the colon for colonoscopy (Prepopik™).

Prepopik™ was approved on July 16, 2012 by the Food and Drug Administration to help cleanse the colon in adults preparing for colonoscopy.[13]

Statistical analysis

We used the Pearson correlation coefficient to examine the associations between state-level Google Trends RSV and BRFSS state-level screening prevalence for each of the five cancer screening tests. We weighted these correlations by the 2011 state population estimates from the Bureau of the Census using weighted regression because such estimates provide more weight to states with larger populations. We used Stata 13.1 to calculate weighted correlations using the wls0 command.

We used the joinpoint methodology to identify significant changes in weekly RSV over time for each of the screening tests and associated interests.[14 15] The joinpoint methodology is ideally suited to examine trends over time and to test whether an apparent change in trend is statistically significant, which other methods (e.g., autoregressive integrated moving average [ARIMA] analysis) may miss. Linear trends in search volume were summarized using the estimated annual percentage change (EAPC). The EAPC was calculated by fitting a linear regression to the natural logarithm of the weekly RSV, using week as a regression variable. Joinpoint regression tests were used to identify an inflection point (hereafter, called joinpoint) with a significant change in the slope of the trend.[14 15] For our analysis, a minimum of four weeks between two joinpoints was required, and a maximum of three joinpoints was allowed to describe the data.

Results

Colorectal cancer screening

The weighted correlation between ever having had a colonoscopy based on 2012 BRFSS data and 2004-2012 Google Trends colonoscopy data was 0.55. Figure 1a shows the weekly Google Trends RSV for colorectal cancer screening using colonoscopy between January, 2004

and April, 2014. The average RSV was 61.9 in 2004 and increased to 85.8 during the last 52 weeks of data. During the first 3 years, RSV per week remained stable, but then increased 0.2 percent per week (95% CI: 0.1; 0.2). Starting at week 308 (November, 2009), RSV increased 0.09 percent per week (95% CI: 0.07; 0.11). RSV was lowest during December of each year and slightly higher during March of each year (average: 74.3).

During 2007, the average RSV/week for virtual colonoscopy was 22.5, but RSV decreased 0.30 percent per week (-0.33; -0.27) starting in January 2008 (Figure 1b). RSV/week for Miralax as a colon cleanser declined 0.50 percent per week (95% CI: -0.69; -0.30) during January 2009 through August 2010, after which RSV about Miralax remained stable until April 2014 (Figure 1c). The RSV/week for Prepopik, a newer colon cleanser approved by the FDA in July 2012, increased rapidly over time.

For FOBT use, Google Trends data was available for only eight states due to low search volume, and a correlation between BRFSS data about FOBT use and Google Trends RSV could not be calculated.

Breast cancer screening

The weighted correlation between Google Trends RSV and BRFSS-based mammography use was 0.36. Figure 2a shows RSV/week for mammography over time. Peaks were present during October each year and about 10 points higher than during December, the month with the lowest RSV. In November 2009, mammography RSV was highest during this 10-year period. Figure 2b shows Google Trends RSV/week for free/low-cost mammography, which peaked in October every year.

Cervical cancer screening

The weighted correlation between 2012 BRFSS-based Pap smear use and RSV for Pap smears during 2010-2012 was 0.14. Figure 3 shows that during week 1-137, RSV/week for pap

smear increased slightly (0.08 percent per week; 95% CI: 0.03; 0.13), remained stable during weeks 137-208, increased during weeks 208-426 (0.13 percent per week; 95% CI: 0.11; 0.16), but then decreased starting in week 426 (-0.11 percent per week; 95% CI: -0.18; -0.04).

Prostate cancer screening

The weighted correlation between Google Trends and BRFSS-based PSA use was 0.42. RSV for PSA declined very slowly (0.05 percent per week) starting in 2004 (95% CI: -0.06; -0.05) until October 2009 (week 302), after which the decline became steeper at 0.20 percent per week (95% CI: -0.30; -0.11) until December 2010 (week 364), then there were three weeks during which RSV remained stable (Figure 4). Starting in January 2011 (week 367), RSV declined 0.05 percent per week (95% CI: -0.07; -0.03). RSV for PSA was highest for week 272 (March, 2009).

Lung cancer screening

Between January 2007 and July 2010, RSV about lung cancer screening declined 1.1 percent per month (95% CI: -1.7; -0.5), but then increased 2.8 percent per month (95% CI: 2.3; 3.4) until April 2014 (Figure 5). There was a peak in RSV about lung cancer screening during November 2010 (month 47).

Discussion

We examined the utility of Google Trends relative to the BRFSS, one of the existing surveillance systems focusing on cancer screening. Correlations between Google Trends and BRFSS data ranged from a high of 0.55 for ever having had a colonoscopy to a low of 0.14 for having a Pap smear within the past three years. Although self-reported screening use is a less than perfect measure of behavior,[12] these modest correlations between data sources indicate that they are measuring different constructs: Google Trends provides estimates of the public's interest in learning more about cancer screening tests; the BRFSS and other surveillance systems provide

estimates of self-reported use of these tests. However, correlations between the two data sources varied across screening types. One reason for the lower correlation related to cervical cancer screening may be that Pap smear use is very common and often part of routine primary care visits, resulting in lower information seeking.[16]

Based on our findings, there appears to be some utility of Google Trends data relative to existing surveillance systems to monitor cancer screening. Awareness and interest in cancer screening is a necessary, but not sufficient, determinant of screening behavior.[17 18] Search volume data using Google Trends enabled us to measure the public's awareness and interest in possible new and developing screening modalities (e.g., virtual colonoscopy, digital mammography, 3D mammography, computed tomography for lung cancer screening) and screening test preparations (e.g., Prepopik versus Miralax), which are not currently captured in existing surveillance systems. By harnessing real-time search-engine data around national media-based interventions (e.g., CDC's Tips from Former Smokers), programs can be evaluated as they are implemented, generating timely feedback to assess the effectiveness of interventions to increase interest in cancer screening, prevention, and other public health recommendations. Such adaptive designs using accumulating data to modify the intervention's course [19 20] have been used infrequently in community-based evaluations. Adaptive interventions that can be evaluated using interest and awareness may be especially useful. It appears that in some instances an increase in public interest in cancer screening is associated with the timing of news reports, celebrity cancer diagnosis, and advertisements.[21] For example, the increase in search volume each October coincides with news stories and advertisements during Breast Cancer Awareness Month. Search volume for colon cancer screening was also slightly higher during March, Colon Cancer Awareness Month. Google Trends also identified a large interest in November 2009 when search volume about mammography increased dramatically likely in response to critics citing health care rationing in response to new mammography guidelines from the US Preventive Services Task Force.[22] The panel recommended that most women wait until age 50 to start routine mammography, then get

the exam every two years instead of annually. For example, in March 2009 RSV for prostate cancer screening increased following coverage of two studies showing that prostate cancer screening did not reduce the risk of death.[23] Also, in November 2010, RSV for lung cancer screening increased after trials reported its potential to reduce the risk of death among heavy smokers.[24] The utility of Google Trends to help adapt interventions is limited by the area where the intervention is implemented, since data is only available at the state-level and for selected large metropolitan areas, limiting its use in rural areas or areas with a low search volume. Consequently, disparities in cancer screening are difficult to examine using these data. Additionally, Google Trends data is unable to evaluate interventions using outcomes such as behaviors or disease development.

Internet searches using Google Trends can guide the development of traditional surveillance systems surveys, such as the BRFSS, NHIS, and HINTS, by vetting the inclusion of questions on surveys. Google Trends data can also gauge developing awareness and interests in new cancer screening tests (e.g., virtual colonoscopy) or related aspects of specific screening tests (e.g., about preparation for colonoscopy). For example, Google Trends showed that interest in lung cancer screening and virtual colonoscopy is still very low, while interest in prostate cancer screening is very high even though PSA tests have been shown to be not very effective in reducing risk of death.[25] Interest in virtual colonoscopy, despite showing promise as a screening tool relative to traditional colonoscopy,[26] was very low. For the most part, screening colonoscopy remains the first-line strategy for the detection of adenomas, with a lower miss rate than virtual colonoscopy, no radiation exposure, and offers therapeutic removal of polyps as well.[27] Internet searches can be an important source of information for generating hypotheses about public awareness and interest in cancer screening, evaluating changes in information seeking after targeted interventions or media coverage, and directing new communication campaigns to explain the evidence base for screening tests.

Search query results may also be politically relevant. Since policy changes often require public support, evaluation strategies that take years to perform may not provide relevant feedback to public interest groups and voters. Instead, an evaluation that occurs almost immediately after the policy change may inform policy makers and their supporters of the associated costs and benefits when there is still interest to make modifications to, or expand, the policy change.[28] For example, the interest in and implementation of free/low cost breast and colorectal cancer screening can be evaluated. The CDC and local organizations implemented free/low cost mammograms starting in the 1990s across the United States followed by free/low cost colonoscopies in selected locations to eligible participants. The potential need for and likely early success (e.g., awareness) of the expansion of these interventions could be gleaned through Google Trends data, much earlier than traditional evaluation strategies.

The utility of Google Trends data should be viewed in light of its limitations. Google Trends data are anonymous, which limits its utility in examining specific subpopulations and disparities among populations. Also, Google Trends data represent only searches done using Google. However, Google accounts for an estimated 65 percent of all internet searches.[29] Google Trends data may have sampling biases. However, such biases are increasingly eroding at the population level as more and more people search for information online. Google Trends does eliminate repeated queries from the same user over a short period of time to reduce counts of continued searching. Google Trends uses a certain threshold of traffic volume so that very new search terms are assigned a value of zero, but this could change very quickly. The motivation of Google users is not known. As a corollary, the data obtained from Google trends cannot be independently verified. Also, the researcher has no control over the data, making quality control difficult. Understanding local health information-seeking behaviors also may be important, but Google Trends data may not be available for geographic areas smaller than at the state level depending on search volume. Additionally, a user option to download Google Trends data for different time periods (e.g., by month or season) is not currently available. Finally, it may be misleading to compare levels of

interest in different screening methods based on the way RSV values are constructed. One remedy that would circumvent many of these limitations is the release of actual search volume data rather than relative search volume data. Finally, search terms entered in other languages were not captured by this study, but could be used to examine interest among non-English speaking populations.

Although Google Trends' "big data" approach provides enormous scientific possibilities, they are not a substitute for, but may complement, traditional data collection and analysis of cancer preventive behavior. The strengths of Google Trends to provide data about the public's interests in cancer screening, despite its inability to provide cancer screening usage data, can foster provision of timely feedback about interventions aimed at increasing interest in cancer screening and other public health recommendations.

Funding

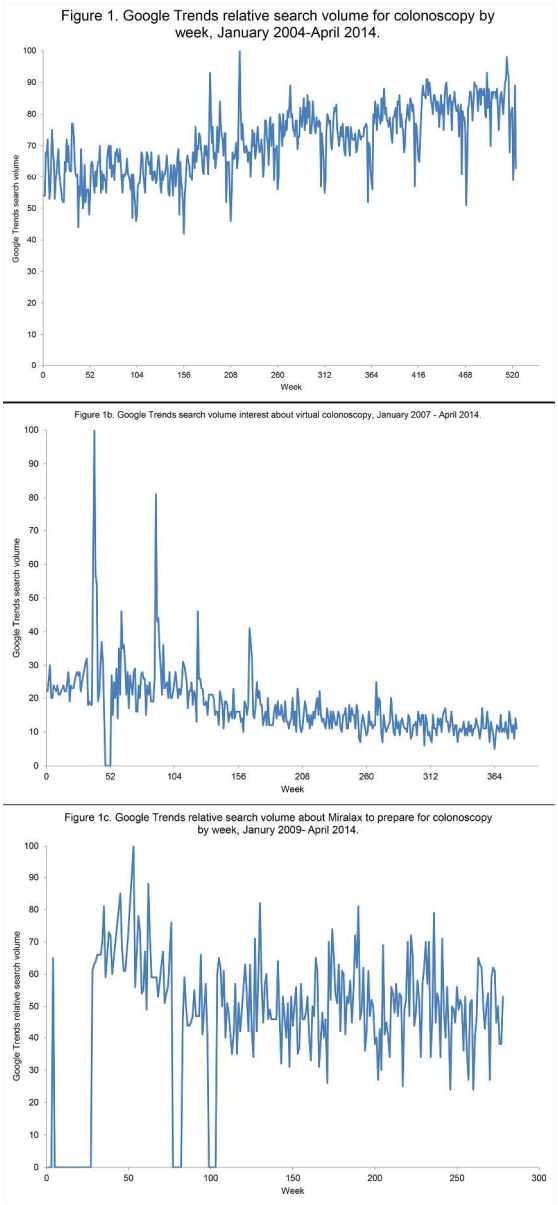
This work was supported by grants from the National Cancer Institute at the National Institutes of Health (grant number R01 CA112159). We thank the Alvin J. Siteman Cancer Center at Barnes-Jewish Hospital and Washington University School of Medicine in St. Louis, Missouri, for the use of the Health Behavior, Communication, and Outreach Core, which is supported in part by the National Cancer Institute Cancer Center Support Grant (grant number P30 CA91842) to the Alvin J. Siteman Cancer Center. Dr. Davidson was supported in part through grants HL38180, DK56260, and Digestive Disease Research Core Center DK52574.

Data sharing statement: No additional data are available.

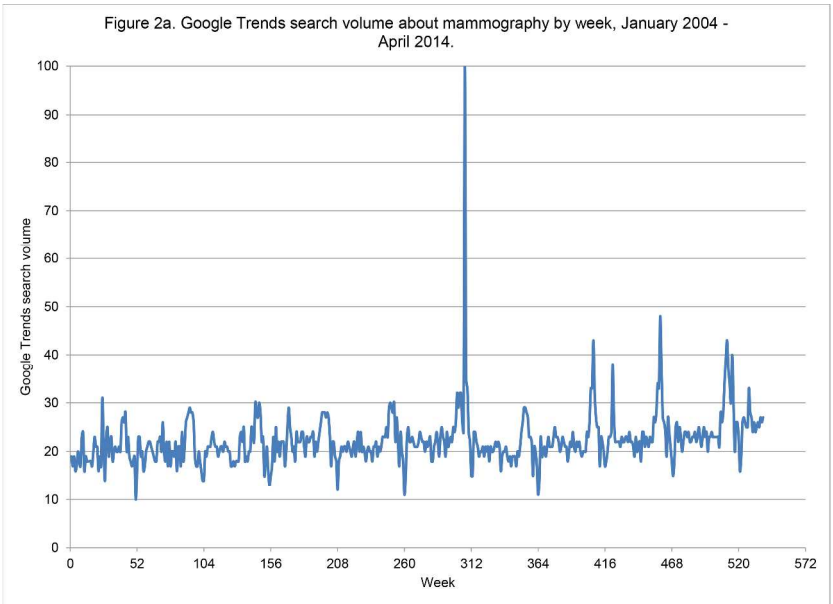
References

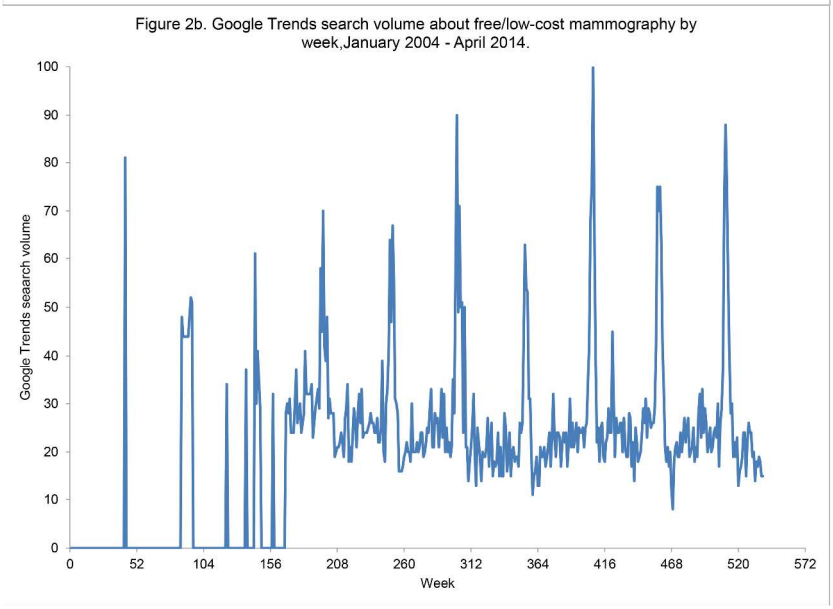
1. Joseph DA, King JB, Miller JW, et al. Prevalence of colorectal cancer screening among adults--Behavioral Risk Factor Surveillance System, United States, 2010. MMWR Morbidity and mortality weekly report 2012;**61 Suppl**:51-6.
2. Miller JW, King JB, Joseph DA, et al. Breast cancer screening among adult women--Behavioral Risk Factor Surveillance System, United States, 2010. MMWR Morbidity and mortality weekly report 2012;**61 Suppl**:46-50.
3. Hiatt RA, Klabunde C, Breen N, et al. Cancer screening practices from National Health Interview Surveys: past, present, and future. Journal of the National Cancer Institute 2002;**94**(24):1837-46.
4. Ashok M, Berkowitz Z, Hawkins NA, et al. Recency of Pap testing and future testing plans among women aged 18-64: analysis of the 2007 Health Information National Trends Survey. Journal of women's health (2002) 2012;**21**(7):705-12.
5. Cooper CP, Mallon KP, Leadbetter S, et al. Cancer Internet Search Activity on a Major Search Engine, United States 2001-2003. J Med Int Res 2005;**7**(3):e36.
6. Breyer BN, Sen S, Aaronson DS, et al. Use of Google Insights for Search to Track Seasonal and Geographic Kidney Stone Incidence in the United States. Urology 2011;**78**(2):267-71.
7. Cavazos-Rehg PA, Krauss MJ, Spitznagel EL, et al. Monitoring of non-cigarette tobacco use using Google Trends. Tobacco control 2014.
8. Johnson AK, Mehta SD. A Comparison of Internet Search Trends and Sexually Transmitted Infection Rates Using Google Trends. Sexually Transmitted Diseases 2014;**41**(1):61-63 10.1097/OLQ.0000000000000065.
9. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 2009;**49**(10):1557-64.
10. Pervaiz F, Pervaiz M, Abdur Rehman N, et al. FluBreaks: early epidemic detection from Google flu trends. Journal of medical Internet research 2012;**14**(5):e125.
11. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System, 2014: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2014.
12. Pierannunzi C, Hu SS, Balluz L. A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004-2011. BMC Medical Research Methodology 2013;**13**(1):49.
13. Administration FaD. FDA approves new colon-cleansing drug for colonoscopy prep, July 17, 2012.
14. Kim HJ, Fay MP, Feuer EJ, et al. Permutation tests for joinpoint regression with applications to cancer rates. Statistics in Medicine 2000;**19**:335-51.
15. Joinpoint Regression Program, Version 4.1.0. [software program], 2014.
16. Redmond N, Baer HJ, Clark CR, et al. Sources of health information related to preventive health behaviors in a national study. American Journal of Preventive Medicine 2010;**38**(6):620-27.e2.
17. Fishbein M. The role of theory in HIV prevention. AIDS Care 2000;**12**(3):273-8.
18. Weinstein ND. The precaution adoption process. Health Psychology 1988;**7**(4):355-86.
19. Chang M, Chow SC, Pong A. Adaptive design in clinical research: issues, opportunities, and recommendations. Journal of biopharmaceutical statistics 2006;**16**(3):299-309; discussion 11-2.
20. Coffey CS, Levin B, Clark C, et al. Overview, hurdles, and future work in adaptive designs: perspectives from a National Institutes of Health-funded workshop. Clinical trials (London, England) 2012;**9**(6):671-80.

21. Glynn RW, Kelly JC, Coffey N, et al. The effect of breast cancer awareness month on internet search activity--a comparison with awareness campaigns for lung and prostate cancer. BMC cancer 2011;**11**:442.
22. Stein R. In wake of mammography guidelines, U.S. health task force faces new scrutiny.: Washington Post, ecember 20, 2009.
23. Parker-Pope T. For men, to screen or not to screen: The New York Times, March 23, 2009.
24. McCook A. More signs lung cancer screening could save lives: Reuters, December 28, 2010.
25. U.S. Preventive Services Task Force. Screening for Prostate Cancer, Topic Page. . Secondary Screening for Prostate Cancer, Topic Page. 2012. http://www.uspreventiveservicestaskforce.org/prostatecancerscreening.htm.
26. Kim DH, Pickhardt PJ, Taylor AJ, et al. CT colonography versus colonoscopy for the detection of advanced neoplasia. The New England journal of medicine 2007;**357**(14):1403-12.
27. Than M, Witherspoon J, Shami J, et al. Diagnostic miss rate for colorectal cancer: an audit. Annals of gastroenterology : quarterly publication of the Hellenic Society of Gastroenterology 2015;**28**(1):94-98.
28. Ayers JW, Ribisl K, Brownstein JS. Using Search Query Surveillance to Monitor Tax Avoidance and Smoking Cessation following the United States' 2009 “SCHIP” Cigarette Tax Increase. PLoS ONE 2011;**6**(3):e16777.
29. Sullivan D. Google still world’s most popular search engine by far, but share of unique searchers dips slightly, February 11, 2013.

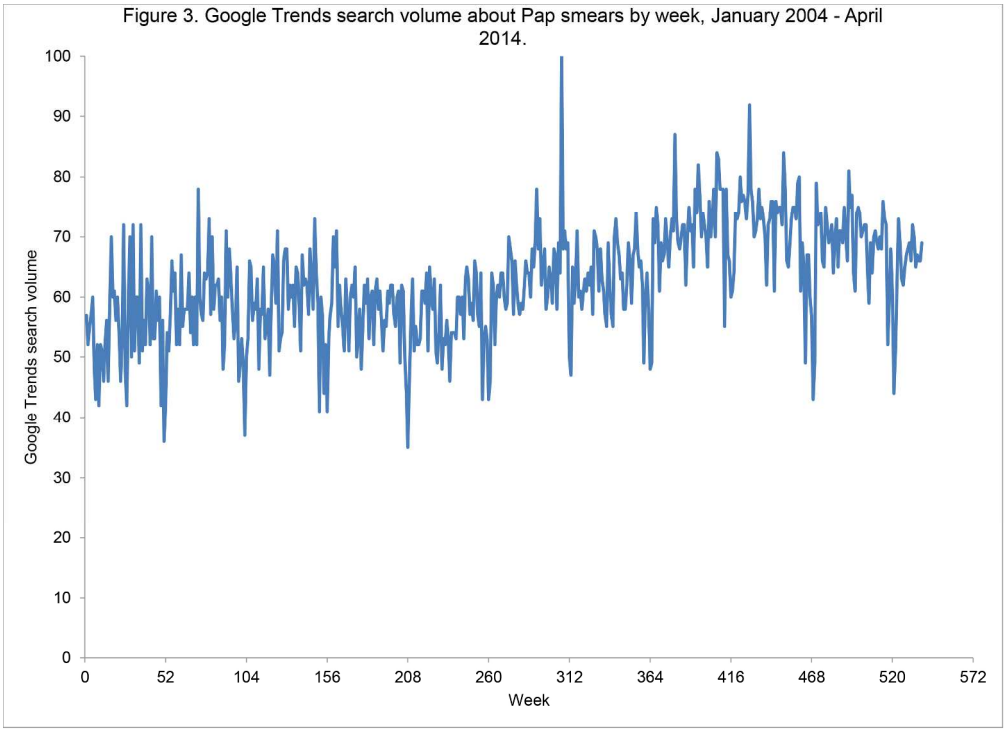


152x332mm (300 x 300 DPI)

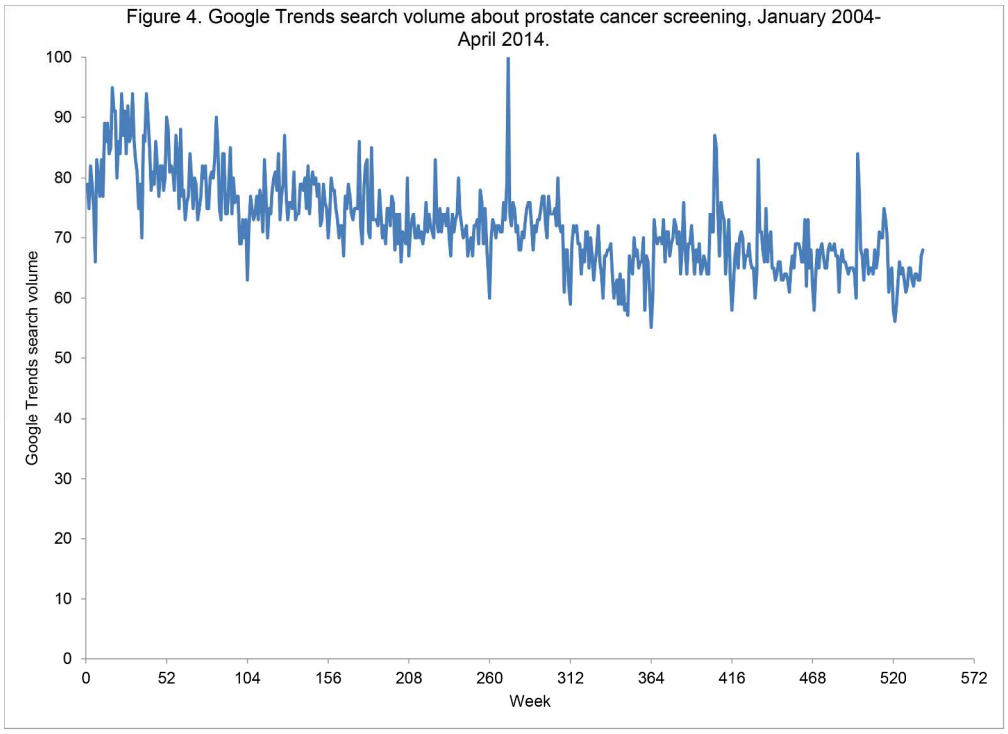




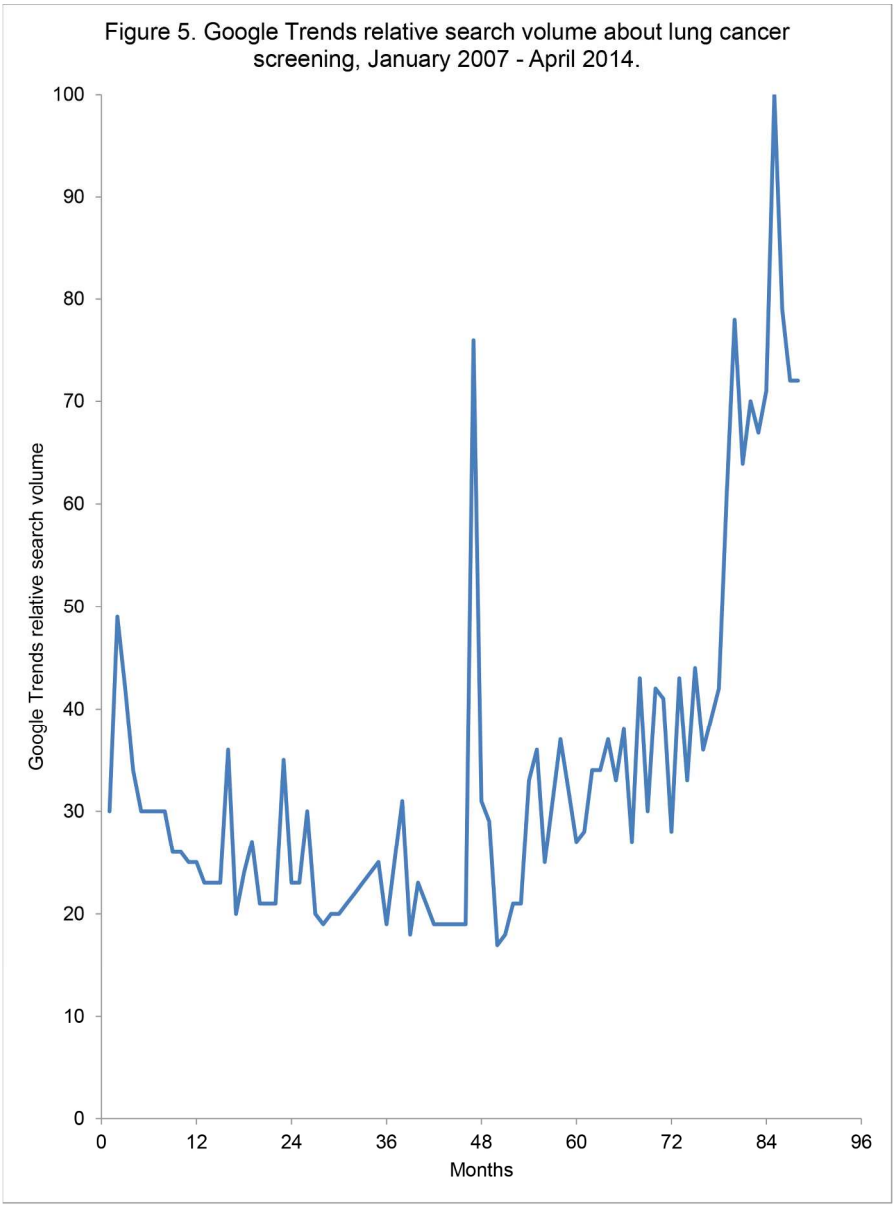
241x349mm (300 x 300 DPI)



241x175mm (300 x 300 DPI)



241x175mm (300 x 300 DPI)



177x239mm (300 x 300 DPI)

Appendix 1: Google Trends search terms used for each screening test and associated interests and Behavioral Risk Factor Surveillance Survey question.

Concept	Google Trends Search terms used	Behavioral Risk Factor Surveillance Survey question (BRFSS)
Screening for colorectal cancer		
Colonoscopy	Colonoscopy+ colonoscopy procedure +virtual colonoscopy+endoscopy +miralax prep+colonoscopy procedure+endoscopy procedure+what is colonoscopy+prep for colonoscopy+miralax+bowel prep+colonoscopy prep+colon cancer screening+colon cancer test	1. Sigmoidoscopy and colonoscopy are exams in which a tube is inserted in the rectum to view the colon for signs of cancer or other health problems. Have you ever had either of these exams? 2. For a SIGMOIDOSCOPY, a flexible tube is inserted into the rectum to look for problems. A COLONOSCOPY is similar, but uses a longer tube, and you are usually given medication through a needle in your arm to make you sleepy and told to have someone else drive you home after the test. Was your MOST RECENT exam a sigmoidoscopy or a colonoscopy? 3. How long has it been since you had your last sigmoidoscopy or colonoscopy?
Virtual colonoscopy	virtual colonoscopy+ct colonography+virtual colonoscopy cost+ct colonoscopy	Not asked in BRFSS
Miralax to cleanse colon for colonoscopy	miralax colonoscopy+colonoscopy prep miralax+miralax dosage colonoscopy+ colonoscopy miralax prep+miralax for colonoscopy+miralax and colonoscopy+miralax gatorade colonoscopy+colonoscopy preparation miralax+miralax before colonoscopy+miralax bowel prep	Not asked in BRFSS
Prepopik to cleanse colon for colonoscopy	prepopik+prepopik dosage+prepopik prep+side effects prepopik+prepopik colonoscopy+colonoscopy prep prepopik	Not asked in BRFSS

Suprep to cleanse colon for colonoscopy	suprep+colonoscopy+colonoscopy prep suprep+suprep dosage colonoscopy+ colonoscopy suprep+suprep for colonoscopy+suprep and colonoscopy+suprep gatorade colonoscopy+colonoscopy preparation suprep +suprep before colonoscopy+suprep bowel prep	Not asked in BRFSS
Free/low-cost colonoscopy	free colonoscopy+low cost colonoscopy+free colonoscopy screening	Not asked in BRFSS
Cost for colonoscopy	cost of colonoscopy+colonoscopy cost+average colonoscopy cost+endoscopy cost	Not asked in BRFSS
Fecal Occult Blood Test (FOBT)	fobt+blood test for colon cancer+colon cancer blood test+screening for colon cancer with blood test+fobt test+fecal occult blood test+blood stool test for cancer	1. A blood stool test is a test that may use a special kit at home to determine whether the stool contains blood. Have you ever had this test using a home kit? 2. How long has it been since you had your last blood stool test using a home kit?
Screening for breast cancer		
Mammography	mammography+breast mammography+breast cancer screening+mammography+mammograms+screening mammography+breast mammogram+breast cancer mammogram+mammo+mammogram screening+mammogram+mammogram results+free mammogram+digital mammography	1. A mammogram is an x-ray of each breast to look for breast cancer. Have you ever had a mammogram? 2. How long has it been since you had your last mammogram?
Digital mammography	digital mammography+mammogram+mammography +digital mammograms+digital mammography screening	Not asked in BRFSS

3D mammography	3D mammography+3D mammogram	Not asked in BRFSS
Free/low-cost mammography	free mammogram+free mammography+low cost mammogram+low cost mammography+free mammogram screening+free mammograms	Not asked in BRFSS
Screening for cervical cancer		
Pap smear	pap test+cervical cancer screening+pap smear test+the pap test+pap smears+free pap smear+free pap+pap tests+pap smear+cervical cancer test+cervical smear+pap testing+papanicolaou	1. A Pap test is a test for cancer of the cervix. Have you ever had a Pap test? 2. How long has it been since you had your last Pap test?
Breast self exam	Breast self exam	Not asked in BRFSS
Screening for prostate cancer		
PSA test	psa test+prostate cancer test+psa testing+prostate test+psa test cancer+prostate+cancer tests+prostate specific antigen test+ prostate psa+ prostate cancer screening tests	1. A Prostate-Specific Antigen test, also called a PSA test, is a blood test used to check men for prostate cancer. Has a doctor EVER recommended that you have a PSA test? 2. Have you EVER HAD a PSA test? 3. How long has it been since you had your last PSA test?
Screening for lung cancer		
Lung cancer screening	lung cancer screening+screening for lung cancer+lung cancer screening CT+CT lung cancer screening	Not asked in BRFSS