Increasing Utility of Google Trends in Monitoring Cardiovascular Disease

Conor Senecal

Mayo Clinic

Madeline Mahowald

Mayo Clinic

Lilach Lerman

Mayo Clinic

Francisco Lopez-Jimenez

Mayo Clinic

Amir Lerman (✉ Lerman.Amir@mayo.edu)

Mayo Clinic

Research Article

Keywords: Cardiovascular disease, Google Trends

Posted Date: March 1st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-251637/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Introduction: Cardiovascular disease is the most common cause of morbidity and mortality in the United States and in the world. Patients are increasingly using internet search to find health-related information, including searches for cardiovascular diseases and risk factors. We sought to evaluate the change in the state by state correlation of cardiovascular disease and risk factors with Google Trends search volumes.

Methods: Data on cardiovascular disease hospitalizations and risk factor prevalence were obtained from the publically available CDC website from 2006-2018. Google Trends data were obtained for matching conditions and time periods. Simple linear regression was performed to evaluate for an increase in correlation over time.

Results: Hospitalizations for six separate cardiovascular disease conditions showed moderate to strong correlation with online search data in the last period studied (heart failure (0.58, $P<0.001$), atrial fibrillation (0.57, $P<0.001$), coronary heart disease (0.58, $P<0.001$), myocardial infarction (0.70, $P<0.001$), stroke (0.62, $P<0.001$), cardiac dysrhythmia (0.46, $P<0.001$)). All diseases studied showed a positive increase in correlation throughout the time period studied ($P<0.05$). All five of the cardiovascular risk factors studied showed strong correlation with online search data; diabetes ($R=0.78$, $P<0.001$), cigarette use ($R=0.79$, $P<0.001$), hypertension ($R=0.81$, $P<0.001$), high cholesterol ($R=0.59$, $P<0.001$), obesity ($R=0.80$, $P<0.001$). Three of the five showed an increasing correlation over time.

Conclusion: The prevalence of and hospitalizations for cardiovascular conditions in the United States strongly correlate with online search volumes nationwide and when analyzed by state. This relationship has progressively strengthened or been strong and stable over recent years for these conditions. Google Trends represents an increasingly valuable tool for evaluating the burden of cardiovascular disease and risk factors in the United States.

Introduction:

Cardiovascular disease is the leading cause of morbidity and mortality in the United States as a whole as well as in the majority of states. This includes the majority of gender and racial subgroups and geographically the majority of the individual states. It is estimated that one-half of adults in the country experience some form of cardiovascular disease, and it spares no gender or racial subgroup (1). The burden of these diseases disproportionately affects the elderly with greater than 70% of individuals over 60 years old experiencing some form of cardiovascular disease. It also disproportionately affects black Americans and those of lower socioeconomic status (1–3).

Internet search has rapidly become the first place people turn when searching for healthcare information due to its convenience and availability(4). 80% of Americans report searching the internet for health information, with 80% of those people starting at a search engine for their queries(5). More than half of people presenting to the emergency room performed a Google search directly related to the ER complaint in the week prior to their visit(6). In fact, an estimated 7% of total Google searches are healthcare-related,

approximately 70,000 searches per minute, and more than 1 billion per day(7). Access to healthcare information via the internet has been shown to increase patient engagement and understanding about their health(8). However, disparities in internet access exist that in many respects are similar to the disparities seen in cardiovascular disease including the elderly, racial minorities, and those of low socioeconomic status(9). Fortunately, these disparities have narrowed over the last 15 years. Since 2006 the percentage of adults 65 and above using the internet has more than doubled to 73%. Similarly, black Americans have increased from 59–85% and low-income families have increased from 52–82% in the same time frame(10).

Google Trends is a publically available information source that provides information about the relative search volume (RSV) of queried terms. The available data has been used to study a wide variety of health conditions including infectious and psychiatric diseases, and a more recently chronic health condition such as cancer and cardiovascular disease (11–13). However, concerns remain about the validity of it as a useful epidemiological tool(14).

Cardiovascular disease, with its high prevalence, would seem uniquely suited to be followed by search engine volume. Accounting for a relatively recent rise in online health search and the recent gains in online access to many of the population subgroups that are disproportionately affected by cardiovascular disease the accuracy of Google trends for monitoring cardiovascular disease has probably increased. We hypothesized that Google trends has become increasingly accurate for monitoring the relative prevalence of cardiovascular disease and its risk factors on a state by state basis.

Methods:

Cardiovascular Disease Prevalence

The Centers for Disease Control and Prevention (CDC) provides public online access to hospitalizations and mortality through the Interactive Heart Disease and Stroke Atlas(15). Hospitalizations for Medicare beneficiaries (age ≥ 65) are reported on a 3-year moving interval from 2005 to 2017. They are reported separately by primary ICD-10 diagnosis code for coronary heart disease (ICD-10: I20-I25), acute myocardial infarction (ICD-10: I21-I22), cardiac dysrhythmia (ICD-10: I47-I49), atrial fibrillation (ICD-10: I48.0-I48.2), heart failure (ICD: I09.81, I11.0, I13.0, I13.2, I50.1-I50.4, I50.8-I50.9), and stroke (ICD-10: I60-I69). Data were obtained for each of the above diagnoses in every interval available for all U.S. states and Washington D.C.; data for other U.S. territories are not consistently available and thus were not included.

Cardiovascular Disease Risk Factor Prevalence

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services(16). They estimate the state-level prevalence of a variety of cardiovascular risk factors including obesity, hypertension, diabetes, hypertension, elevated cholesterol, and cigarette use. The data are provided as a cross-sectional survey for each year. However,

questions vary on an annual basis and not all chronic health conditions are included each year. For example, survey results for hypertension are available for 2007, 2009, 2011, 2013, 2015, 2017; while results are available for 2006–2017 for diabetes. Survey results are publically available via the CDC website.

Google Trends

Google Trends is an unbiased sample of Google search data that is publically available for query(17). Specific topics (i.e. Myocardial Infarction) can be searched and a RSV is provided. All searches are indexed to 100, reflecting the maximum search interest (geographically or chronologically) for the time and location specified. It also provides a RSV for smaller geographic areas enclosed in a search area, for example, individual states in a search with a location of the United States. Data can be analyzed for specific words entered by the searcher or for a search topic. Search topics are a group of search terms that share the same concept in any language, for example the topic “myocardial infarction” would include searches for “heart attack” and “infarto de miocardio” and the topic hypertension would include “high blood pressure” and “elevated blood pressure”. For this study, search topics were used whenever possible in order to include multiple languages and the variety of terminology used to describe medical conditions. Search topics do not perfectly reflect terminology used in the CDC data, thus appropriate substitutions were required made for example, coronary artery disease as opposed to coronary heart disease. The only exception to this methodology was obesity, which has previously been shown to be more closely reflected by searches for weight loss(18). Table 1 contains the details of the search queries used for this analysis.

Analysis

For each of the CDC cardiovascular diseases and the BRFSS risk factors, state-level data at all time periods available were obtained from 2006–2018. Results of corresponding Google Trends queries were obtained, as displayed in Table 1, for each state during each time period. All analysis was performed using JMP 14.1.0. Pearson correlation coefficients comparing RSV value with reported prevalence on a state by state basis were calculated for each pairing, as has previously been done in analyses using Google trends(12). An example is displayed in Fig. 1, which shows representative choropleth maps for BRFSS hypertension prevalence and Google Trends hypertension RSVs for 2017. For each condition, a simple linear regression was then performed evaluating the relationship between year and correlation coefficient. An alpha level of 0.05 was used to evaluate for significance. In order to assess states that showed the least correlation with Google Trends data, the last observation period for each cardiovascular disease and risk factor was analyzed using a linear regression. Studentized residuals were calculated for each state in each model. The ten samples for each state were then averaged to obtain an average residual for each state in an effort to reflect over and underestimation of disease using search data. No IRB approval was obtained for this study given that all data used in this analysis are publicly available and anonymous.

Results:

Six separate categories of cardiovascular disease were available for evaluation, each with eleven separate three-year time intervals. At the earliest interval studied (2005–2007) coronary artery disease ($R = 0.30$, $P < 0.05$), acute myocardial infarction ($R = 0.49$, $P < 0.05$), heart failure ($R = 0.34$, $P < 0.05$) and stroke ($R = 0.38$, $P < 0.05$) all had a modest but significant correlation with the corresponding Google Trends RSV values. Cardiac dysrhythmia and atrial fibrillation showed a non-significant positive correlation ($R = 0.21$, $P = 0.05$). In the final interval studied (2015–2017) all six categories of disease had a strong positive correlation with the corresponding Google Trends states by state RSV; heart failure (0.58 , $P < 0.001$), atrial fibrillation (0.57 , $P < 0.001$), coronary heart disease (0.58 , $P < 0.001$), myocardial infarction (0.70 , $P < 0.001$), stroke (0.62 , $P < 0.001$), cardiac dysrhythmia (0.46 , $P < 0.001$). Complete results are available in Fig. 2.

Five separate cardiovascular disease risk factors were available for evaluation in one-year intervals with a minimum of six and a maximum of thirteen intervals. Four of the risk factors had at least modest positive significant correlation with the corresponding Google Trends RSV on a state-by-state comparison at the earliest studied interval. Diabetes ($R = 0.52$, $P < 0.001$), hypertension ($R = 0.63$, $P < 0.001$) and obesity ($R = 0.74$, $P < 0.001$) all had strong correlation, while cigarette smoking showed a modest correlation ($R = 0.34$, $P = 0.01$). High cholesterol showed no correlation. By the final interval studied, all five cardiovascular risk factors evaluated showed a strong association with the corresponding Google Trends RSV; diabetes ($R = 0.78$, $P < 0.001$), cigarette use ($R = 0.79$, $P < 0.001$), hypertension ($R = 0.81$, $P < 0.001$), high cholesterol ($R = 0.59$, $P < 0.001$), obesity ($R = 0.80$, $P < 0.001$). Complete results are available in Fig. 2.

Five of the six cardiovascular diseases showed increasing strength of the correlation over time, with only cardiac dysrhythmia showing a non-significant relationship. Atrial fibrillation showed the strongest relationship ($R^2 = 0.94$, $P < 0.001$), but every model evaluated had an $R^2 > 0.47$ except for cardiac dysrhythmia. Complete summaries of the linear models are available in Table 2.

All five risk factors had positive correlations with time, three of which were statistically significant; diabetes ($R^2 = 0.60$, $P = 0.005$), cigarette use ($R^2 = 0.69$, $P < 0.001$), and high cholesterol ($R^2 = 0.83$, $P = 0.01$). Hypertension and obesity did not show a significant trend of increasing correlation with time. Complete summaries of the linear models are available in Table 2.

The average studentized residual for each state is displayed in Fig. 4. In general states located in the Southern and upper Midwestern United States have positive (red) average residuals meaning their burden of disease is higher than would be predicted using Google Trends alone to predict. Alternatively, states located in the mountain west and northeastern United States consistently showed negative (green) average residuals representing a lower burden of disease than would be predicted by Google Trends alone.

Discussion:

In this study, we show hospitalizations for cardiovascular disease and prevalence of cardiovascular disease risk factors in the United States strongly correlate with Google Trends RSVs on a state-by-state basis. This relationship has progressively strengthened or been strong and stable over recent years for these conditions. Google trends represents an increasingly valuable tool for evaluating the burden of cardiovascular disease and risk factors in the United States that can be expected to improve as more users use internet search to gather health-related information.

The consistent moderate to strong correlation between search data and cardiovascular disease hospitalizations is in line with the growing use of the internet and specifically search engines for patients and their families to find information about health conditions(9). The strength of these relationships has uniformly increased throughout the period studied, a period in which an increasing number of patients at elevated risk for cardiovascular disease have gained access to the internet (9). The hospitalization data only includes Medicare beneficiaries limiting it to patients 65 and above. Since an estimated 27% of this age group does not actively use the internet, it's foreseeable that this relationship will continue to strengthen over the next several years as people using the internet enter this age group or those 65 and above become active on the internet(10). Of the six disease states evaluated, cardiac dysrhythmia, while still significant, had the poorest association. This may be because cardiac dysrhythmia is an umbrella term comprised of several diagnoses with completely separate terminology (ex. supraventricular tachycardia or atrial flutter), as opposed to the other conditions, which have more overlap in terminology (heart failure with preserved ejection fraction and heart failure with reduced ejection fraction).

All of the cardiovascular risk factors assessed had a strong correlation with the survey data from BRFSS. Hypertension and diabetes did not show a significant increase in correlation over time. However, this is predominately because they exhibited a strong correlation from the earliest interval studied and thus had little room to improve. The strength of the correlation was consistently higher in cardiovascular risk factor comparisons as opposed to hospitalizations. There are several possible etiologies for this. The first is that survey data includes American adults over eighteen years of age as opposed to solely Medicare beneficiaries in the hospitalization data. The second is that hospitalization data is coded based on the primary diagnosis for the hospitalization without accounting for commonly concurrent conditions such as atrial fibrillation and heart failure while BRFSS data asks about cardiovascular risk factors independently. Thirdly, by virtue of being a survey, BRFSS depends on participants having knowledge of their conditions to report them, this is also required to search for health information on the topic. Hospitalization data can be reported without the patient ever fully grasping their specific diagnosis.

In addition to closely correlating with both hospitalizations and risk factor prevalence, online search data has the distinct advantage of being free of cost and nearly real-time(19). Both CDC hospitalizations data and the BRFSS provided increased granularity and demographic subgrouping that are extremely valuable in the study of cardiovascular disease. However, data from these sources typically requires years before it becomes available for study. If online search data continues to improve in its accuracy, as would be expected based on current trends, it may eventually provide an accurate enough tool to allow for early investigation in to changes in cardiovascular disease prevalence. This would allow for more rapid

mobilization of resources to help combat the leading cause of morbidity and mortality in the country. Another possible benefit is the use of online search data to investigate cardiovascular conditions not currently included national statistics, providing preliminary data on prevalence.

This study highlights areas of the country in which Google Trends may over or underestimate the burden of cardiovascular disease. The reasons for these discrepancies are unclear however there are several possibilities. These include internet access, socioeconomic disparities, differing demographics with regards to age and race, and educational disparities. Another possibility would be access to care, with people having more difficulty accessing care potentially utilizing internet search more frequently. Further work evaluating these possibilities is required.

Patient awareness of cardiovascular conditions is associated with improved outcomes(20). It is interesting to consider search data as a possible surrogate for patient awareness, as awareness is required to perform an online search. By comparing this surrogate for awareness with measures of cardiovascular disease and risk factor control it may be possible to identify areas of high disease burden but relatively low search volume that should be targeted for more intensive cardiovascular health education interventions on a statewide scale. Similarly, areas with high relative search volume as compared to disease burden may be identified as models and studied to help look for transferable practices that can lead to increased patient awareness. This study also highlights the important need for accurate disease-specific materials available online, given the volume of people using this as information source and the diagnostic specificity with which they are searching. Recent work has shown that among prominent cardiovascular societies, patient specific online materials are consistently written above recommended grade levels(21). Every effort should be made to make these resources applicable to a broad audience.

There are clear limitations to this study given the nature of online search data. Online search data is anonymous and thus demographic data is not available. Similarly, specific health information is about the searcher is not available to confirm the presence of a condition and motivation for the online search. This study relies on the correlation between national data sets that have an inherent margin of error which could skew the results of correlation with online trends in either direction.

Overall, our study highlights the strong correlation of search engine data with cardiovascular conditions at a state level as well as the increasing strength of these relationships throughout the period studied. Based, on this improvement online search data can be expected to become an increasingly valuable tool in cardiovascular research in providing nearly real-time information without public cost.

Abbreviations:

RSV: Relative search volume

CDC: Centers for disease control

BRFSS: Behavioral Risk Factor Surveillance System

ICD: International Disease Classification

Declarations:

Conflict of Interest: The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References:

1. Virani SS, Alonso A, Benjamin EJ et al. Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. *Circulation* 2020;141:e139-e596.
2. Carnethon MR, Pu J, Howard G et al. Cardiovascular Health in African Americans: A Scientific Statement From the American Heart Association. *Circulation* 2017;136:e393-e423.
3. Schultz WM, Kelli HM, Lisko JC et al. Socioeconomic Status and Cardiovascular Outcomes: Challenges and Interventions. *Circulation* 2018;137:2166-2178.
4. Chu JT, Wang MP, Shen C, Viswanath K, Lam TH, Chan SSC. How, When and Why People Seek Health Information Online: Qualitative Study in Hong Kong. *Interact J Med Res* 2017;6:e24.
5. Weaver J. More People search for Health Online. *Telemedicine*. NBCNEWS.com: NBC NEWS, 2020.
6. Asch JM, Asch DA, Klinger EV, Marks J, Sadek N, Merchant RM. Google search histories of patients presenting to an emergency department: an observational study. *BMJ Open* 2019;9:e024791.
7. Dr Google will see you now: Search giant wants to cash in on your medical queries. *The Telegraph*.
8. Tan SS-L, Goonawardene N. Internet Health Information Seeking and the Patient-Physician Relationship: A Systematic Review. *J Med Internet Res* 2017;19:e9.
9. Greenberg-Worisek AJ, Kurani S, Finney Rutten LJ, Blake KD, Moser RP, Hesse BW. Tracking Healthy People 2020 Internet, Broadband, and Mobile Device Access Goals: An Update Using Data From the Health Information National Trends Survey. *J Med Internet Res* 2019;21:e13300.
10. Internet/Broadband Fact Sheet 2019. Pew Research Center.
11. Schootman M, Toor A, Cavazos-Rehg P et al. The utility of Google Trends data to examine interest in cancer screening. *BMJ Open* 2015;5:e006678.
12. Senecal C, Widmer RJ, Lerman LO, Lerman A. Association of Search Engine Queries for Chest Pain With Coronary Heart Disease Epidemiology. *JAMA Cardiology* 2018;3:1218-1221.
13. Effenberger M, Kronbichler A, Shin JI, Mayer G, Tilg H, Perco P. Association of the COVID-19 pandemic with Internet Search Volumes: A Google Trends™ Analysis. *International Journal of Infectious Diseases* 2020;95:192-197.
14. Cervellin G, Comelli I, Lippi G. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. *Journal of Epidemiology and Global Health* 2017;7:185-189.

15. Interactive Atlas of Heart Disease and Stroke. Centers for Disease Control and Prevention, Division for Heart Disease and Stroke Prevention, 2020.
16. (CDC) CfDCaP. Behavioral Risk Factor Surveillance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.
17. Google Trends.
18. Pawar A, Nagpal S, Pawar N, Lerman L, Eirin A. Information Seeking Pattern of Public on Topics Related to Obesity: Google Trends Analysis (Preprint), 2020.
19. Arora VS, McKee M, Stuckler D. Google Trends: Opportunities and limitations in health and health policy research. *Health Policy* 2019;123:338-341.
20. Riegel B, Moser DK, Buck HG et al. Self-Care for the Prevention and Management of Cardiovascular Disease and Stroke: A Scientific Statement for Healthcare Professionals From the American Heart Association. *J Am Heart Assoc* 2017;6.
21. Kapoor K, George P, Evans MC, Miller WJ, Liu SS. Health Literacy: Readability of ACC/AHA Online Patient Education Material. *Cardiology* 2017;138:36-40.

Tables:

Table 1
Google Trends Search Queries and Corresponding National Data Set Terminology, Time Periods and Geography

Trends Search Topic	CDC Term	Time Period(s)	Geography
Heart Failure	Heart Failure	2006–2008, 2007–2009, 2008–2010, 2009–2011, 2010–2012, 2011–2013, 2012–2014, 2013–2015, 2014–2016, 2015–2017	United States
Heart Arrhythmia	Cardiac Dysrhythmia	2006–2008, 2007–2009, 2008–2010, 2009–2011, 2010–2012, 2011–2013, 2012–2014, 2013–2015, 2014–2016, 2015–2017	United States
Atrial Fibrillation	Atrial Fibrillation	2006–2008, 2007–2009, 2008–2010, 2009–2011, 2010–2012, 2011–2013, 2012–2014, 2013–2015, 2014–2016, 2015–2017	United States
Coronary Artery Disease	Coronary Heart Disease	2006–2008, 2007–2009, 2008–2010, 2009–2011, 2010–2012, 2011–2013, 2012–2014, 2013–2015, 2014–2016, 2015–2017	United States
Myocardial Infarction	Acute Myocardial Infarction	2006–2008, 2007–2009, 2008–2010, 2009–2011, 2010–2012, 2011–2013, 2012–2014, 2013–2015, 2014–2016, 2015–2017	United States
Stroke	Stroke	2006–2008, 2007–2009, 2008–2010, 2009–2011, 2010–2012, 2011–2013, 2012–2014, 2013–2015, 2014–2016, 2015–2017	United States
Trends Search Topic	BRFSS Term	Time Period(s)	Geography
Hypertension	Hypertension	2007, 2009, 2011, 2013, 2015, 2017	United States
Diabetes	Diabetes	2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016	United States
Cigarettes	Cigarette Use	2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018	United States
Hypercholestermia	High Cholesterol	2007, 2009, 2011, 2013, 2015, 2017	United States
Weight Loss	Obesity	2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018	United States

Table 2
Time and Correlation Linear Model Summaries

	R Square	RMSE	F	Intercept	Coefficient	P
Heart Failure	0.88	0.03	64.20	-48.94	0.02	< .0001*
Atrial Fibrillation	0.94	0.05	141.42	-111.61	0.06	< .0001*
Coronary Heart Disease	0.48	0.08	8.40	-42.70	0.02	0.0177*
Myocardial infarction	0.81	0.05	39.63	-54.82	0.03	0.0001*
Stroke	0.83	0.04	43.32	-54.42	0.03	0.0001*
Cardiac Dysrhythmia	0.24	0.12	2.81	-37.40	0.02	0.1278
	R Square	RMSE	F	Intercept	Coefficient	P
Diabetes	0.60	0.08	13.77	-56.84	0.03	0.0048*
Obesity	0.02	0.07	0.17	-3.46	0.00	0.6852
Cigarette Use	0.69	0.12	24.21	-86.62	0.04	0.0005*
Hypertension	0.46	0.07	3.39	-30.87	0.02	0.1393
High Cholesterol	0.83	0.10	20.02	-105.11	0.05	0.011*
* indicates P < 0.05						

Figures

Figure 1

(A) Choropleth map of the prevalence of hypertension estimated by BRFSS for each states in 2017 (B) Choropleth map of the RSV of hypertension in Google Trends for each state in 2017 (C) Scatter chart of the relationship between RSV for hypertension and BRFSS estimated prevalence for hypertension in 2017. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Figure 2

Online Search Correlations with Cardiovascular Diseases over Time

Figure 3

Online Search Correlations with Cardiovascular Risk Factors over Time