

Original Paper

Integrating Google Trends Search Engine Query Data Into Adult Emergency Department Volume Forecasting: Infodemiology Study

Jesus Trevino^{1*}, MD, MBA; Sanjeev Malik^{2*}, MD; Michael Schmidt^{2*}, MD

¹Department of Emergency Medicine, The George Washington University School of Medicine & Health Sciences, Washington, DC, United States

²Department of Emergency Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

* all authors contributed equally

Corresponding Author:

Jesus Trevino, MD, MBA

Department of Emergency Medicine

The George Washington University School of Medicine & Health Sciences

2120 L Street NW

Suite 450

Washington, DC, 20037

United States

Phone: 1 202 741 2904

Email: jtrevino@mfa.gwu.edu

Abstract

Background: The search for health information from web-based resources raises opportunities to inform the service operations of health care systems. Google Trends search query data have been used to study public health topics, such as seasonal influenza, suicide, and prescription drug abuse; however, there is a paucity of literature using Google Trends data to improve emergency department patient-volume forecasting.

Objective: We assessed the ability of Google Trends search query data to improve the performance of adult emergency department daily volume prediction models.

Methods: Google Trends search query data related to chief complaints and health care facilities were collected from Chicago, Illinois (July 2015 to June 2017). We calculated correlations between Google Trends search query data and emergency department daily patient volumes from a tertiary care adult hospital in Chicago. A baseline multiple linear regression model of emergency department daily volume with traditional predictors was augmented with Google Trends search query data; model performance was measured using mean absolute error and mean absolute percentage error.

Results: There were substantial correlations between emergency department daily volume and Google Trends “hospital” ($r=0.54$), combined terms ($r=0.50$), and “Northwestern Memorial Hospital” ($r=0.34$) search query data. The final Google Trends data-augmented model included the predictors Combined 3-day moving average and Hospital 3-day moving average and performed better (mean absolute percentage error 6.42%) than the final baseline model (mean absolute percentage error 6.67%)—an improvement of 3.1%.

Conclusions: The incorporation of Google Trends search query data into an adult tertiary care hospital emergency department daily volume prediction model modestly improved model performance. Further development of advanced models with comprehensive search query terms and complementary data sources may improve prediction performance and could be an avenue for further research.

(*JMIR Infodemiology* 2022;2(1):e32386) doi: [10.2196/32386](https://doi.org/10.2196/32386)

KEYWORDS

infodemiology; patient volume forecasting; emergency medicine; digital health; Google Trends; infoveillance; social media; prediction models; emergency department

Introduction

Background

Internet-based technologies and web-based services have facilitated new ways of seeking and communicating health-related information. A valuable aspect of web-based information transactions is the record of communication itself, which, in aggregate, may reflect population-level behaviors. For example, researchers have used search engine queries and volumes, such as Google Trends, to attempt to recognize population behavior-based patterns. Examples of this research are found in many industries, such as finance [1] and criminology [2].

The emerging field of infodemiology is defined by Eysenbach [3] as “the science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy.” The major debut application of infodemiology within the health care industry involved monitoring the seasonal emergence and peak of influenza with Google Flu Trends [4], which initially outperformed the extant gold standard FluNet from the Centers for Disease Control and Prevention; however, Google Flu Trends later suffered from poor predictions attributed to model overfitting, among other reasons [5].

The field of infodemiology has grown substantially in the past decade, in terms of disease applications and data sources. In early infodemiology research, the majority of papers involved the study of influenza; more recent reviews [6,7] detail an expanded scope of subject matter, such as influenza, multiple sclerosis, suicide, prescription drug abuse, and e-cigarettes, and the most common data sources included Twitter (45%), Google (24.6%), other websites (13.9%), blogs (10.1%), and Facebook (8.9%). In addition to research applications, one review [8] described the following practical applications of infodemiology by health care organizations: infoveillance, dissemination of health information, misinformation management, and health interventions. Most recently, during the COVID-19 pandemic, researchers have used infodemiology to study public opinion toward COVID-19 vaccines [9] and public health containment measures [10], capture the most frequently asked questions regarding COVID-19 vaccines [11], augment the performance of conventional prediction models for COVID-19 infections [12], and characterize the partisan differences of US legislators in the initial phase of this pandemic [13].

Prior Work

In infodemiology, data reflecting the use of the internet in seeking health information have been used to improve emergency department patient volume predictions and optimize emergency department resource allocation [14-16]. A Swedish study [14] of emergency department patient volume found that the use of a popular public health website's traffic volume as a predictor yielded an impressive mean absolute percentage error (MAPE) of 4.8%, which demonstrated that web-based information seeking behaviors can be a useful leading indicator of acute care encounters [14]. A study in the United States found that 86% of participants, who had been recruited from an

emergency department waiting room, utilized Google search in the week prior to their emergency department visit; 15% of their searches had been health-related and two-thirds of these searches had been either related to their current chief complaint or for information related to the emergency department and hospital [15]. In addition, internet health information-seeking behavior has been described as a method for patients to prepare questions for upcoming medical appointments with health care providers [16].

Prior studies have used Google Trends search query data to forecast influenza-like illness cases [17] and pediatric daily volumes [18]; however, no studies have evaluated the ability of Google Trends search query data related to chief complaints and health care facilities to predict the overall daily volume in an adult emergency department.

Study Goal

The ability to predict deviations in typical weekly patterns of emergency department patient volumes could provide emergency department administrators with a valuable tool to optimize resource allocation. We explored the use of Google Trends search query data of chief complaints and health care facilities to improve the prediction performance of adult emergency department daily patient volume.

Methods

Emergency Department Encounter Data

Emergency department daily patient volume data were collected from Northwestern Memorial Hospital, a tertiary care adult center located in Chicago, Illinois with an annual volume of 88,000 patient encounters. Data were collected retrospectively from the institution's databases and included 159,769 emergency department patient encounters that occurred in the period from July 1, 2015 to June 30, 2017. These data included patient arrival date and time, and Emergency Severity Index (levels 1 through 5 in decreasing order of case urgency) [19]. For analysis, data were aggregated by date and Emergency Severity Index.

Environmental Data

To develop prediction models to be used as a point of reference, we used calendar day (ie, day of week, month) and weather-related variables to derive a traditional emergency department forecasting model. Daily weather data were obtained from the National Centers for Environmental Information and included average temperature, maximum temperature, minimum temperature, precipitation (categorical), and snow (categorical) [20].

Google Trends Data

Data Collection

Google Trends search query data were accessed from the Google Trends API service on June 19, 2018 [21].

Keyword Selection

Based on clinical experience and expert opinion, we generated a list of Google Trends terms that would be relevant to an individual seeking health information (ie, terms that would be

part of their search engine query) prior to a health care encounter. The terms, which included “emergency department,” “Northwestern Memorial Hospital,” “hospital,” “WebMD,” “chest pain,” “back pain,” “abdominal pain,” “stomach pain,” “side pain,” “fever,” “cough,” “shortness of breath,” “headache,” “numbness,” “weakness,” “blood urine,” and “blood stool,” corresponded to 3 broad categories: health care facility, reputable website, and general chief complaints encountered in the emergency department.

Region and Period Selection

Google Trends search query data were limited to the Chicago metropolitan area by constraining the API request to the Chicago Nielsen Designated Market Area (code 602) and to daily relative search frequencies from July 1, 2015 to June 30, 2017.

Feature Engineering

To engineer a feature that reflected a more precise region around the study hospital, we derived an independent variable: the search query ratio of “Northwestern Memorial Hospital” over “hospital.” We also created a combined variable, which aggregated all Google Trends search query data into a single measure. We performed the following transformations on Google Trends search query variables to explore temporal associations and to engineer features that smooth out short-term fluctuations: 1-day lag, 1-day percentage change, 3-day moving average, and 7-day moving average. After these transformations, a total of 85 Google Trends search query terms were included in the candidate set of predictor variables. Given the difference in scales, Google Trends search query data were standardized before their inclusion as predictor variables in the regression analysis.

Exploratory and Correlational Analysis

We performed visual analysis of Google Trends search query data and calculated Pearson correlation coefficients between emergency department daily volume and Google Trends variables.

Model Development and Evaluation

We utilized multiple linear regression, one of the most common methods for emergency department patient volume forecasting and for predictive modeling with Google Trends search query data [22], to create separate predictive models for overall emergency department patient volume and for patient volume by Emergency Severity Index (ie, 1 through 5).

We also created a baseline model with traditional variables, such as calendar day and weather, similar to prior literature [23]. Predictor variable selection was performed using recursive feature elimination, which is a type of backward selection algorithm that offers a systematic approach to variable selection by constructing multiple models with permutations of predictor

variables and selecting a parsimonious model that optimizes a prediction performance metric [24]. To evaluate the ability of Google Trends search query data to improve forecasting performance, we augmented the baseline model with Google Trends variables and used recursive feature elimination to identify the highest impact predictor variables.

Models were trained using 10-fold cross-validation, and model performance was assessed using mean absolute error (MAE) and MAPE of prediction values in relation to actual values. Analysis was conducted using R software (version 4.1.0; The R Project) and utilized the caret package (version 6.0-88; Max Kuhn) [25].

Ethics

This study was considered exempt from review by the Northwestern Memorial Hospital Institutional Review Board because emergency department data were deidentified and contained no protected health information.

Results

Exploratory Analysis

The median total emergency department daily volume over this period was 242 patients per day (range 152-305 patients per day; Emergency Severity Index 1: 4043/159,769, 2.5%; Emergency Severity Index 2: 63,611/159,769, 39.8%; Emergency Severity Index 3: 64,091/159,769, 40.1%; Emergency Severity Index 4: 23,773/159,769, 15.0%; Emergency Severity Index 5: 2300/159,769, 1.4%; Emergency Severity Index not available: 1951/159,769, 1.2%).

The daily Google Trends relative frequency for most terms demonstrated properties of a normal distribution, with the exception of those for “shortness of breath,” “hospital,” or for all terms combined (Figure 1). The relative search frequencies for “hospital” and all terms combined exhibited a bimodal distribution; the bimodal distribution for “hospital” data was largely explained by weekday and weekend differences (Figure 2). A similar pattern was evident in emergency department daily volume (Figure 3). Two terms, “blood stool” and “blood urine,” did not yield any relative frequency data and, therefore, were excluded from subsequent analyses. When search terms occur infrequently, Google does not share these data in order to safeguard user privacy.

Visual analysis of Google Trends search query data time series demonstrated 3 patterns (Figure 4): seasonal, for example, “hospital” and “fever” data exhibited weekly and annual periodicity, respectively; a declining trend, such as that for “WebMD,” and random (ie, white noise), such as that exhibited by “Northwestern Memorial Hospital” and “emergency department.”

Figure 1. Histograms of candidate Google Trends search query data. N: count; NMH: Northwestern Memorial Hospital.

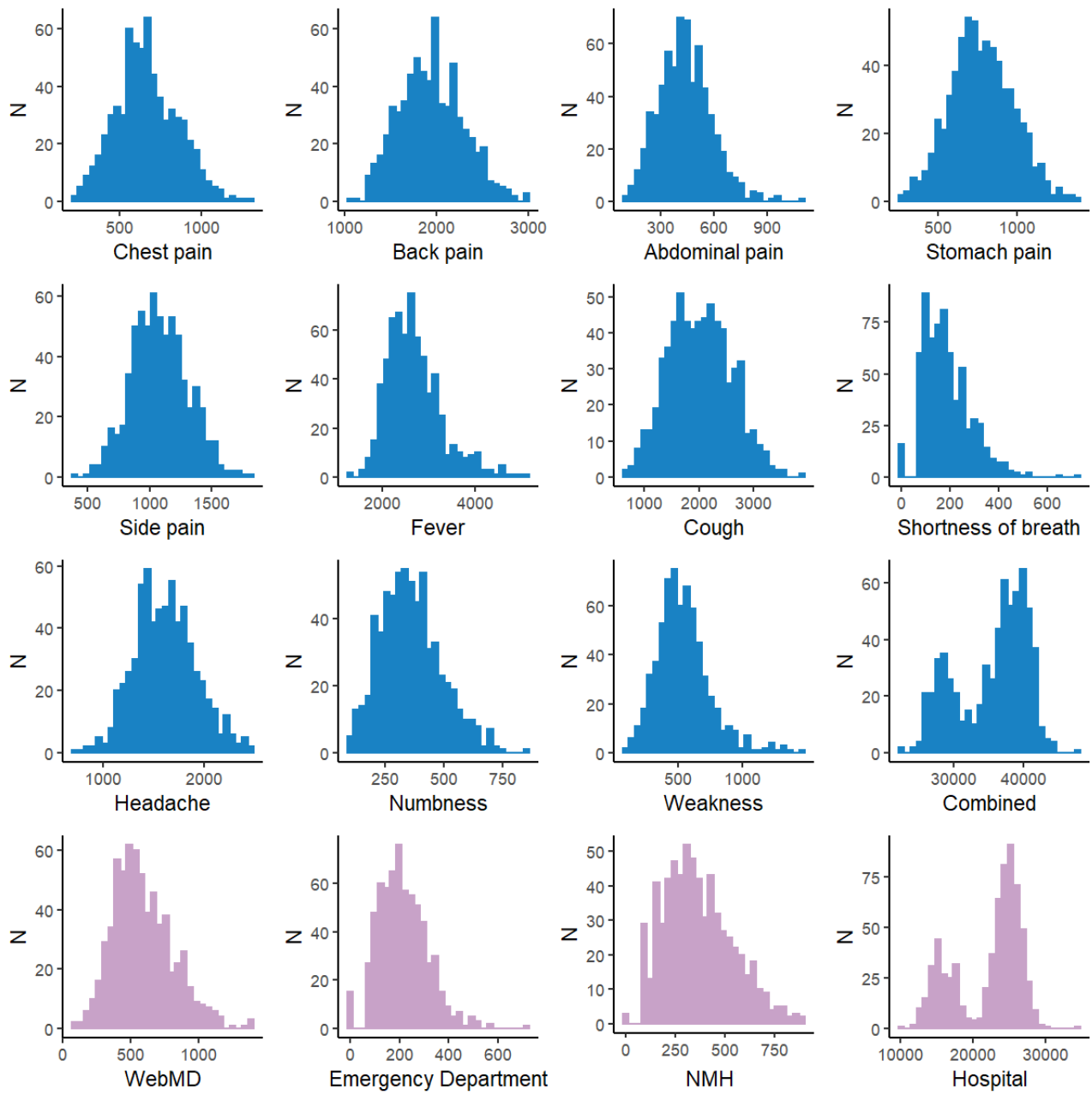


Figure 2. Density plot of Google Trends “hospital” data.

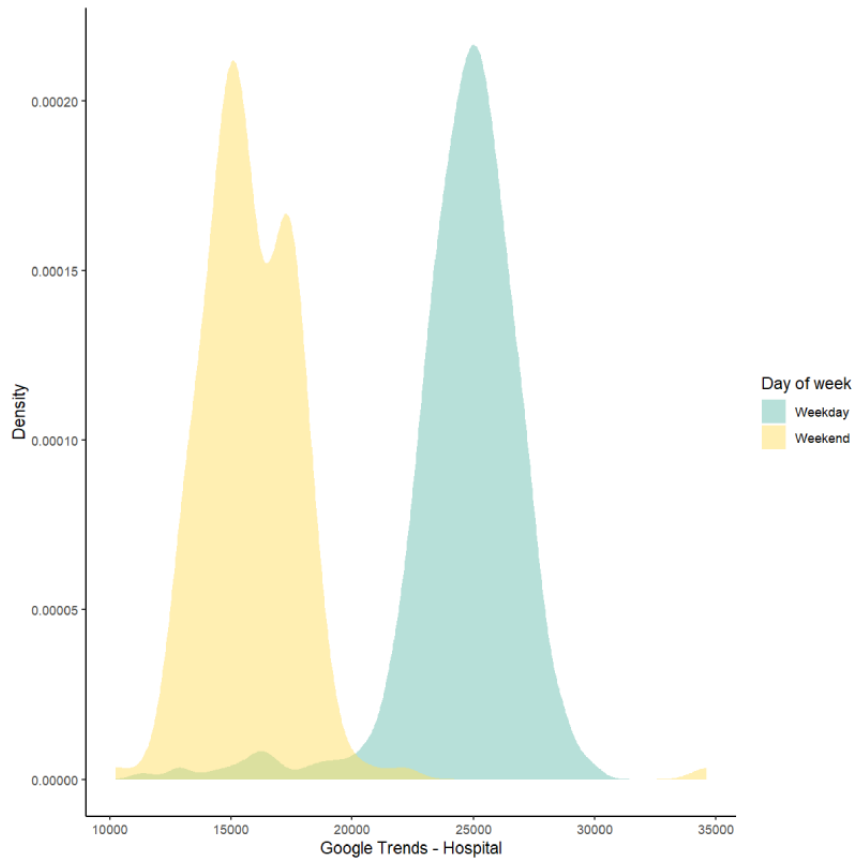


Figure 3. Density plot of emergency department (ED) daily volume.

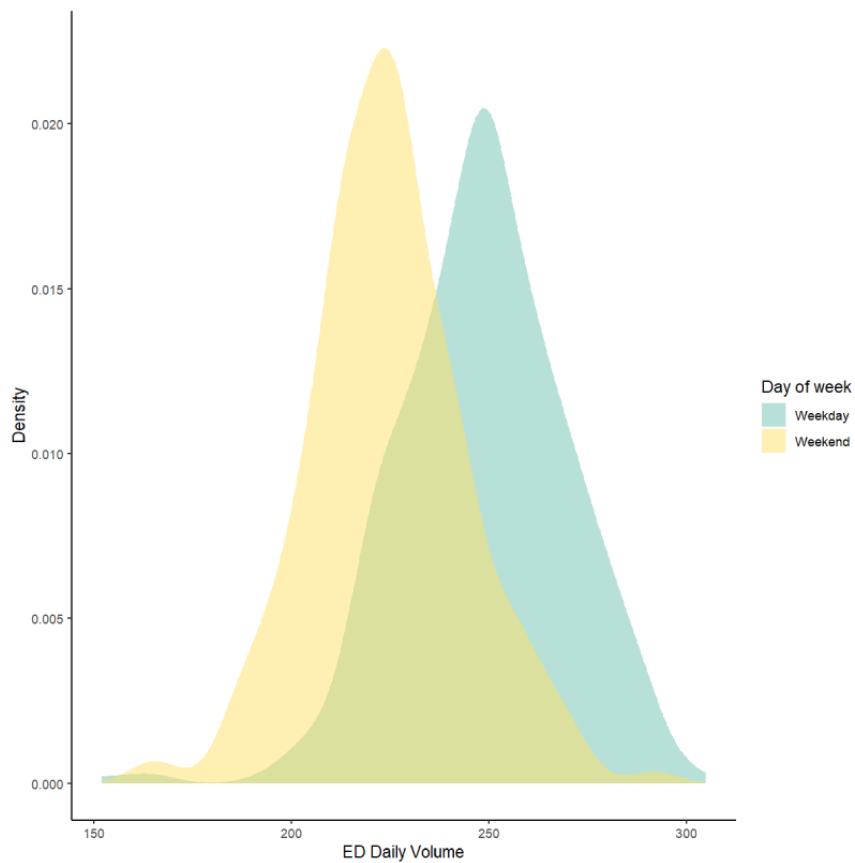
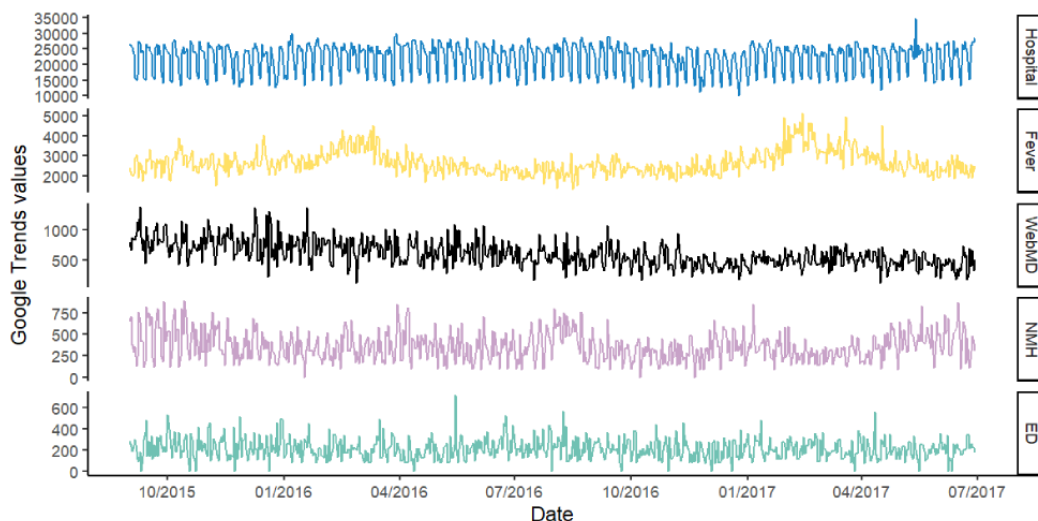


Figure 4. Google Trends search query data time series for the terms "hospital" (blue), "fever" (yellow), "WebMD" (black), "Northwestern Memorial Hospital" (NMH, pink), "emergency department" (ED, green).



Correlation Analysis

Emergency department daily volume data were moderately correlated with "hospital" ($r=0.54, P<.001$) and combined

($r=0.50, P<.001$) Google Trends search query data and were weakly correlated with "Northwestern Memorial Hospital" ($r=0.34, P<.001$) Google Trends search query data (Table 1).

Table 1. Pearson correlations between Google Trends data and emergency department daily volume.

Google Trends	None		1-day lag		1-day percentage change		3-day moving average		7-day moving average	
	<i>r</i>	<i>P</i> value	<i>r</i>	<i>P</i> value	<i>r</i>	<i>P</i> value	<i>r</i>	<i>P</i> value	<i>r</i>	<i>P</i> value
Chest pain	0.00	.98	0.01	.89	0.00	.91	0.05	.18	0.05	.25
Back pain	0.11	.005	0.16	<.001	-0.04	.31	0.05	.20	0.09	.02
Abdominal pain	-0.04	.26	0.00	.92	-0.01	.77	-0.02	.63	-0.02	.53
Stomach pain	0.00	.95	0.04	.32	-0.03	.38	0.03	.40	0.04	.30
Side pain	0.05	.17	0.01	.86	0.01	.81	0.05	.22	0.06	.16
Fever	-0.06	.10	-0.03	.48	-0.02	.56	0.03	.49	0.05	.20
Cough	-0.21	<.001	-0.18	<.001	-0.01	.89	-0.03	.44	0.01	.89
Shortness of breath	0.02	.54	0.02	.54	0.03	.44	0.03	.42	0.02	.70
Headache	-0.04	.27	0.05	.17	-0.08	.04	-0.05	.21	-0.01	.76
Numbness	0.15	<.001	0.05	.19	0.05	.17	0.11	.003	0.09	.02
Weakness	0.11	.004	0.10	.007	0.00	.95	0.09	.03	0.07	.07
Combined	0.50	<.001	-0.04	.35	0.49	<.001	0.52	<.001	0.52	<.001
WebMD	0.10	.007	0.09	.03	0.02	.53	0.11	.006	0.09	.02
Emergency department	0.02	.61	0.01	.81	-0.05	.19	-0.04	.30	-0.03	.43
Hospital	0.54	<.001	-0.04	.27	0.51	<.001	0.53	<.001	0.53	<.001
NMH ^a	0.34	<.001	0.02	.53	0.30	<.001	0.34	<.001	0.30	<.001
NMH share	0.12	.002	0.04	.25	0.07	.06	0.09	.02	0.08	.048

^aNMH: Northwestern Memorial Hospital.

The transformations of Google Trends search query data to explore lagging or leading indicators did not uncover hidden correlations with emergency department daily volume.

Predictive Model Development

The application of recursive feature elimination to the candidate set of traditional variables resulted in an optimal model that

utilized the Day-of-week predictor; with Sunday as the reference level, this traditional model is characterized by decreasing magnitudes of regression coefficients as the week progresses from Monday to Sunday (Table 2).

Table 2. Regression coefficients for traditional and Google Trend data-augmented linear regression models for total emergency department daily volume.

Variable	Traditional, beta (95% CI)	Google Trends, beta (95% CI)
Intercept	223 (219, 227)	242 (240, 243)
Day of week		
Sunday	Reference	Reference
Monday	40 (35, 46)	— ^a
Tuesday	27 (22, 33)	—
Wednesday	18 (12, 23)	—
Thursday	17 (11, 23)	—
Friday	25 (19, 30)	—
Saturday	3.1 (–2.5, 8.7)	—
Northwestern Memorial Hospital	—	3.5 (1.8, 5.1)
Hospital 1-day percentage change	—	5.5 (–4.0, 15)
Hospital 3-day moving average	—	17 (4.4, 29)
Combined 1-day percentage change	—	1.6 (–8.3, 11)
Combined 3-day moving average	—	–11 (–24, 1.7)

^aThe predictor was not included in the model.

For the model augmented with Google Trends predictor variables, the application of recursive feature elimination yielded a model that excluded *Day of week* and contained the *Northwestern Memorial Hospital*, *Hospital 1-day percentage change*, *Hospital 3-day moving average*, and *Combined 1-day percentage change* Google Trends predictors (Table 2). When comparing the traditional and Google Trends data-augmented models, the y-axis intercepts were largely similar, although the

y-axis intercept of the Google Trends data-augmented model was identical to the median emergency department daily volume of this data set.

For emergency department daily volume predictions by Emergency Severity Index level, recursive feature elimination produced models that utilized *Combined 3-day moving average* for every level and *Hospital 3-day moving average* for level 2 (Table 3).

Table 3. Regression coefficients for traditional and Google Trends data–augmented linear regression models for daily volume by Emergency Severity Index.

Model and variable	ESI ^a 1, beta (95% CI)	ESI 2, beta (95% CI)	ESI 3, beta (95% CI)	ESI 4, beta (95% CI)	ESI 5, beta (95% CI)
Traditional model					
Intercept	5.9 (5.3, 6.5)	83 (80, 85)	93 (91, 95)	36 (34, 37)	3.3 (2.9, 3.7)
Day of week					
Sunday	Reference	Reference	Reference	Reference	Reference
Monday	1.0 (0.20, 1.9)	24 (21, 28)	12 (8.4, 15)	2.2 (–0.11, 4.5)	0.44 (–0.16, 1.0)
Tuesday	–0.08 (–0.92, 0.75)	19 (16, 23)	6.9 (3.6, 10)	–0.40 (–2.7, 1.9)	0.30 (–0.30, 0.90)
Wednesday	0.46 (–0.37, 1.3)	16 (12, 20)	2.4 (–0.90, 5.7)	–1.1 (–3.4, 1.3)	0.21 (–0.39, 0.81)
Thursday	0.06 (–0.78, 0.90)	15 (11, 19)	2.8 (–0.53, 6.1)	–1.2 (–3.5, 1.1)	0.38 (–0.22, 1.0)
Friday	0.29 (–0.55, 1.1)	19 (15, 23)	4.7 (1.4, 8.0)	0.30 (–2.0, 2.6)	0.27 (–0.33, 0.88)
Saturday	–0.30 (–1.1, 0.54)	1.8 (–1.8, 5.4)	–0.62 (–3.9, 2.7)	1.7 (–0.59, 4.1)	–0.24 (–0.85, 0.36)
Augmented model					
Intercept	6.1 (5.9, 6.3)	96 (95, 97)	97 (96, 98)	36 (35, 37)	3.5 (3.3, 3.6)
Combined 3-day moving average	0.30 (0.08, 0.53)	–7.0 (–12, –1.6)	3.8 (2.9, 4.7)	0.33 (–0.30, 1.0)	0.18 (0.01, 0.34)
Hospital 3-day moving average	— ^b	15 (9.2, 20)	—	—	—

^aESI: Emergency Severity Index.

^bThe predictor was not included in the model.

Model Performance

We observed that Google Trends data–augmented models generally had superior prediction performance compared to the

traditional model, when based on MAE; however, these improvements were minimal (Table 4).

Table 4. Predictive performance of total and Emergency Severity Index daily volume for traditional and Google Trends data–augmented models.

Model	Traditional model, mean absolute error ^a	Augmented model, mean absolute error	Change (%)
All visits	15.69	15.21	–3.1
Emergency Severity Index			
1	2.52	2.41	–4.7
2	10.37	10.66	2.8
3	9.55	9.12	–4.5
4	6.93	6.85	–1.2
5	1.80	1.74	–3.6

^aIn units of patients/day.

The MAPE of the traditional model was 6.67%, and the MAPE of the Google Trends data–augmented model was 6.42%; MAPE was not calculated for models by Emergency Severity Index since they contained records with 0 daily volume, which would produce an undefined result (ie, the denominator would have been 0 in these instances).

Discussion

Principal Results

The goal of this study was to evaluate the potential of Google Trends search query data of healthcare facilities and chief

complaints to improve the prediction performance of ED daily volume of a large-volume, tertiary-care, adult hospital. The use of Google Trends search query data to forecast emergency department daily volume resulted in a marginal improvement (MAE 3.1%) in prediction performance compared to that of a traditional prediction model. This is a small but notable improvement; when one considers that the original Google Flu Trends model included data from a set of 45 unique search queries, the ability of this study's narrow list of Google Trends terms to produce forecast results similar to traditional models highlights the potential for this alternative real-time data source to be honed further with more advanced models and a more

expansive set of Google Trends term candidates [4]. Alternatively, one may conclude that the prediction capabilities of traditional and Google Trends data-augmented models were roughly similar. The finding that Google Trends search query data alone reproduced similar predictions to those made with conventional calendar day variables demonstrates the utility of Google Trends search query data in signaling health information-seeking behavior from prospective emergency department patients.

A notable strength of this study was the use of daily emergency department encounter data. A common obstacle that infodemiology researchers face is the lack of accessible, high-frequency, and recent hospital data, which constrains their ability to leverage the real-time and high-volume attributes of Google Trends and other social media data sources (ie, big data). As more and more collaborations leverage health care organization databases for service operations data, researchers will accelerate the development of nowcasting services that have the potential to inform and optimize service operations decisions. For example, a robust nowcasting service for emergency department daily volumes could provide hospital administrators with advanced notice of impending emergency department overcrowding and trigger the coordination of earlier mitigating responses throughout the hospital.

Unexpectedly, model coefficients for the *Combined 3-day moving average* variable were negative in the Google Trends data-augmented models of total volume ($\beta=-11.0$, 95% CI -24 to 1.7) and Emergency Severity Index 2 ($\beta=-7.0$, 95% CI -12 to -1.6). Negative coefficients may reflect that sicker patients present rapidly to emergency departments and do not have time to contemplate their illness and search the internet for information. Although, this negative coefficient result was not found in the Google Trends data-augmented Emergency Severity Index 1 model, we suspect this could be due to the small proportion of Emergency Severity Index 1 encounters that were available in this data set (4043/159,769, 2.5%). Analysis of a data set with more Emergency Severity Index 1 encounters could show results consistent with other Emergency Severity Index levels. Given that low-acuity encounters (Emergency Severity Index 3, 4, and 5) were the majority, with approximately 60% (95,861/159,769), the implication that individuals with high-acuity cases may not consult the internet prior to arriving at the emergency department would not have applied to a majority of emergency department encounters at this study site. Alternatively, these counterintuitive results of a negative coefficient value for the *Combined 3-day moving average* variable may be explained by the proximity of coefficients' 95% confidence intervals to 0; nonetheless, it is important to present these model outputs to highlight the unbiased results from a systematic approach to model generation. Altogether, these findings of illogical regression coefficients remind us of the need exercise caution with data mining exercises and predictive models that emphasize error metrics while overlooking meaningful causal relationships.

Comparison With Prior Work

The traditional model using a day-of-week predictor in a prior study [18] that explored forecasting daily volume at an academic

children's hospital emergency department in Boston had a larger error (MAPE 10.99%) and their Google Trends search query data-based model had a smaller improvement (MAPE 1.67%) than those found in our study (traditional day-of-week model: MAPE 6.67%; improvement: MAPE 3.1%). Although the reason for the differences in MAPE for models that employed day-of-week predictors is not obvious, we hypothesize that the differences in the impact of Google Trends search query data could be due to a greater utilization of the internet to understand symptoms of an acute illness among adults compared to pediatric patients and their adult guardians. There may be a population subset whose health activity is better measured by internet and social media activity data such as in the case of suicide surveillance among 25- to 34-year-old adults in the United Kingdom [26].

Limitations

There are several limitations that are important to consider. We only utilized the emergency department daily volume from a single hospital in Chicago, whereas the Google Trends search query data pertained to the entire metropolitan area; we may have failed to identify more meaningful predictive relationships between Google Trends search query data and emergency department daily volume since we did not include the metropolitan-wide emergency department daily volume data, nor could we identify Google Trends search queries that occurred within our study site's geographic service area.

Moreover, we only analyzed Google Trends search query data in English, which limits our ability to extrapolate these results to regions of the country where there may be greater segments of the population that use search engines in non-English languages.

Similar limitations exist in regions of the country that face barriers to internet access, such as rural areas, although a recent survey [27] found that the gap in home broadband internet between rural and nonrural homes has decreased from 16% to 7% and overall smartphone ownership has increased from 81% to 85% between 2019 to 2021; in addition, 72% of nonbroadband users reported the ability of smartphones to accomplish all desired internet tasks [27]. Therefore, as market penetration of home broadband and smartphone ownership increases, limitations due to barriers to the internet may become less prominent.

In addition, we did not attempt to predict emergency department daily volume by type of chief complaint (eg, cardiac, respiratory, neurologic). Given the difference in scale of the Google Trends search query data across types of chief complaints, future work should focus on predicting daily volumes of categories of chief complaints using an expanded set of symptom-specific Google Trends search query data.

Lastly, we only leveraged a single source of internet data, which may have only provided a glimpse into health information-seeking behaviors from prospective emergency department patients. Other data sources, such as news media and social media platforms could be incorporated [28,29]. While more resources would be required to leverage additional data sources for more complex and potentially more accurate

prediction models of emergency department daily volumes, the ability for health care systems to anticipate increased demand for emergency department services would be valuable in terms of reduced health care expenses and improved patient experiences. For instance, the potential for health care systems to identify when and where low-acuity emergency department encounters may occur could guide the strategic expansion of clinic appointment availability and required advertisements to divert potential emergency department patients into less costly and more convenient venues of care.

It is worth discussing the ongoing debate regarding the ability of infodemiology data such as Google Trends search queries to reliably supplement or entirely replace traditional epidemiological data. While Google Trends search query data offers an enticing value proposition in providing insights into a population's internet health information-seeking behaviors in a cost-efficient manner compared with traditional epidemiology data-gathering processes, it is important to remain critical of this emerging source of population health data. In some instances, Google Trends search query data reasonably mirror traditional epidemiology data. For example, tobacco use search query data were well correlated with findings from Youth Risk Behaviors Surveillance System and National Survey on Drug Use and Health data in the United States [30], and Google Trends search query data for chest pain were found to be strongly correlated with hospital admission data for coronary heart disease from the US Centers for Disease Control and Prevention Atlas of Heart and Stroke Statistics [31]. More recently, it was also demonstrated that Google Trends COVID-19 symptom search query data were significantly correlated with new cases and deaths from this disease [32,33]. However, potential confounders such as media influence have

been found to effect Google Trends data. For instance, correlations between Google Trends search query data for anosmia and ageusia and COVID-19 cases were inconsistent early in the COVID-19 pandemic, and Google Trends search query volumes showed a marked increase following the beginning of the media's coverage of these two prominent symptoms of COVID-19 [34]. In addition, COVID-19 Google Trends search query data from Europe were poorly correlated with COVID-19 epidemiological measures and were well correlated with the occurrence of pandemic-related press releases from the World Health Organization [35]. Overall, the ability to use Google Trends search query data for epidemiologic purposes remains an active area of inquiry, and these types of data must be used cautiously for such purposes.

Conclusion

Emergency department daily volume prediction models augmented with Google Trends search query data performed similarly to baseline models utilizing traditional variables; error metrics demonstrated modest improvements in model accuracy for overall volume and nearly all Emergency Severity Index volumes. Our results suggest that even greater improvements in emergency department daily volume predictions can be attained with a more comprehensive set of Google Trends search query terms or the addition of complementary internet data sources such as social media.

The potential for these types of prediction models to leverage near real-time information to capture health information-seeking behavior preceding emergency department encounters and to be used as a tool for health care system administrators to better anticipate patient demands and optimize resource allocation warrants further investigation.

Acknowledgments

We thank Stephanie Gravenor for her assistance in collecting the emergency department daily volume data used in this analysis.

Conflicts of Interest

None declared.

References

1. Preis T, Moat HS, Stanley HE. Quantifying trading behavior in financial markets using Google Trends. *Sci Rep* 2013;3:1684 [FREE Full text] [doi: [10.1038/srep01684](https://doi.org/10.1038/srep01684)] [Medline: [23619126](https://pubmed.ncbi.nlm.nih.gov/23619126/)]
2. Gamma A, Schleifer R, Weinmann W, Buadze A, Liebreuz M. Could Google Trends be used to predict methamphetamine-related crime? an analysis of search volume data in Switzerland, Germany, and Austria. *PLoS One* 2016;11(11):e0166566 [FREE Full text] [doi: [10.1371/journal.pone.0166566](https://doi.org/10.1371/journal.pone.0166566)] [Medline: [27902717](https://pubmed.ncbi.nlm.nih.gov/27902717/)]
3. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J Med Internet Res* 2009;11(1):e11 [FREE Full text] [doi: [10.2196/jmir.1157](https://doi.org/10.2196/jmir.1157)] [Medline: [19329408](https://pubmed.ncbi.nlm.nih.gov/19329408/)]
4. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature* 2009 Feb 19;457(7232):1012-1014. [doi: [10.1038/nature07634](https://doi.org/10.1038/nature07634)] [Medline: [19020500](https://pubmed.ncbi.nlm.nih.gov/19020500/)]
5. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. *Science* 2014 Mar 14;343(6176):1203-1205. [doi: [10.1126/science.1248506](https://doi.org/10.1126/science.1248506)] [Medline: [24626916](https://pubmed.ncbi.nlm.nih.gov/24626916/)]
6. Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *J Med Internet Res* 2013;15(7):e147 [FREE Full text] [doi: [10.2196/jmir.2740](https://doi.org/10.2196/jmir.2740)] [Medline: [23896182](https://pubmed.ncbi.nlm.nih.gov/23896182/)]
7. Mavragani A. Infodemiology and infoveillance: scoping review. *J Med Internet Res* 2020 Apr 28;22(4):e16206 [FREE Full text] [doi: [10.2196/16206](https://doi.org/10.2196/16206)] [Medline: [32310818](https://pubmed.ncbi.nlm.nih.gov/32310818/)]

8. Chen J, Wang Y. Social media use for health purposes: systematic review. *J Med Internet Res* 2021 May 12;23(5):e17917 [FREE Full text] [doi: [10.2196/17917](https://doi.org/10.2196/17917)] [Medline: [33978589](https://pubmed.ncbi.nlm.nih.gov/33978589/)]
9. Hu T, Wang S, Luo W, Zhang M, Huang X, Yan Y, et al. Revealing public opinion towards COVID-19 vaccines with Twitter data in the United States: spatiotemporal perspective. *J Med Internet Res* 2021 Sep 10;23(9):e30854 [FREE Full text] [doi: [10.2196/30854](https://doi.org/10.2196/30854)] [Medline: [34346888](https://pubmed.ncbi.nlm.nih.gov/34346888/)]
10. Chum A, Nielsen A, Bellows Z, Farrell E, Durette P, Banda JM, et al. Changes in public response associated with various COVID-19 restrictions in Ontario, Canada: observational infoveillance study using social media time series data. *J Med Internet Res* 2021 Aug 25;23(8):e28716 [FREE Full text] [doi: [10.2196/28716](https://doi.org/10.2196/28716)] [Medline: [34227996](https://pubmed.ncbi.nlm.nih.gov/34227996/)]
11. Sajjadi NB, Shepard S, Ottwell R, Murray K, Chronister J, Hartwell M, et al. Examining the public's most frequently asked questions regarding COVID-19 vaccines using search engine analytics in the United States: observational study. *JMIR Infodemiology* 2021;1(1):e28740 [FREE Full text] [doi: [10.2196/28740](https://doi.org/10.2196/28740)] [Medline: [34458683](https://pubmed.ncbi.nlm.nih.gov/34458683/)]
12. Rabiolo A, Alladio E, Morales E, McNaught AI, Bandello F, Afifi AA, et al. Forecasting the COVID-19 epidemic by integrating symptom search behavior into predictive models: infoveillance study. *J Med Internet Res* 2021 Aug 11;23(8):e28876 [FREE Full text] [doi: [10.2196/28876](https://doi.org/10.2196/28876)] [Medline: [34156966](https://pubmed.ncbi.nlm.nih.gov/34156966/)]
13. Guntuku SC, Purtle J, Meisel ZF, Merchant RM, Agarwal A. Partisan differences in Twitter language among United States legislators during COVID-19: a cross-sectional study. *J Med Internet Res* 2021 Apr 16;23(6):e27300 [FREE Full text] [doi: [10.2196/27300](https://doi.org/10.2196/27300)] [Medline: [33939620](https://pubmed.ncbi.nlm.nih.gov/33939620/)]
14. Ekström A, Kurland L, Farrokhnia N, Castrén M, Nordberg M. Forecasting emergency department visits using internet data. *Ann Emerg Med* 2015 Apr;65(4):436-442.e1. [doi: [10.1016/j.annemergmed.2014.10.008](https://doi.org/10.1016/j.annemergmed.2014.10.008)] [Medline: [25487026](https://pubmed.ncbi.nlm.nih.gov/25487026/)]
15. 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 Feb 20;9(2):e024791 [FREE Full text] [doi: [10.1136/bmjopen-2018-024791](https://doi.org/10.1136/bmjopen-2018-024791)] [Medline: [30787088](https://pubmed.ncbi.nlm.nih.gov/30787088/)]
16. Tan S, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. *J Med Internet Res* 2017 Jan 19;19(1):e9 [FREE Full text] [doi: [10.2196/jmir.5729](https://doi.org/10.2196/jmir.5729)] [Medline: [28104579](https://pubmed.ncbi.nlm.nih.gov/28104579/)]
17. Araz OM, Bentley D, Muelleman RL. Using Google Flu Trends data in forecasting influenza-like-illness related ED visits in Omaha, Nebraska. *Am J Emerg Med* 2014 Sep;32(9):1016-1023. [doi: [10.1016/j.ajem.2014.05.052](https://doi.org/10.1016/j.ajem.2014.05.052)] [Medline: [25037278](https://pubmed.ncbi.nlm.nih.gov/25037278/)]
18. Tideman S, Santillana M, Bickel J, Reis B. Internet search query data improve forecasts of daily emergency department volume. *J Am Med Inform Assoc* 2019 Dec 01;26(12):1574-1583 [FREE Full text] [doi: [10.1093/jamia/ocz154](https://doi.org/10.1093/jamia/ocz154)] [Medline: [31730701](https://pubmed.ncbi.nlm.nih.gov/31730701/)]
19. Fernandes C, Tanabe P, Gilboy N, Johnson L, McNair R, Rosenau A, et al. Five-level triage: a report from the ACEP/ENA five-level triage task force. *J Emerg Nurs* 2005 Feb;31(1):39-50; quiz 118. [doi: [10.1016/j.jen.2004.11.002](https://doi.org/10.1016/j.jen.2004.11.002)] [Medline: [15682128](https://pubmed.ncbi.nlm.nih.gov/15682128/)]
20. Data tools. National Centers for Environmental Data. URL: <https://www.ncdc.noaa.gov/cdo-web/datatools/selectlocation> [accessed 2017-09-22]
21. Health trends API getting started guide. Google. URL: <https://sites.google.com/a/google.com/health-trends-api-getting-started-guide/> [accessed 2018-10-01]
22. Sampri A, Mavragani A, Tsagarakis KP. Evaluating Google Trends as a tool for integrating the 'smart health' concept in the Smart Cities' governance in USA. *Procedia Engineering* 2016;162:585-592. [doi: [10.1016/j.proeng.2016.11.104](https://doi.org/10.1016/j.proeng.2016.11.104)]
23. Wargon M, Casalino E, Guidet B. From model to forecasting: a multicenter study in emergency departments. *Acad Emerg Med* 2010 Sep;17(9):970-978 [FREE Full text] [doi: [10.1111/j.1553-2712.2010.00847.x](https://doi.org/10.1111/j.1553-2712.2010.00847.x)] [Medline: [20836778](https://pubmed.ncbi.nlm.nih.gov/20836778/)]
24. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn* 2002 Jan;46:389-422 [FREE Full text] [doi: [10.1023/A:1012487302797](https://doi.org/10.1023/A:1012487302797)]
25. Kuhn M. Building predictive models in R using the caret package. *J Stat Soft* 2008;28(5):1-26. [doi: [10.18637/jss.v028.i05](https://doi.org/10.18637/jss.v028.i05)]
26. Arora VS, Stuckler D, McKee M. Tracking search engine queries for suicide in the United Kingdom, 2004-2013. *Public Health* 2016 Aug;137:147-153. [doi: [10.1016/j.puhe.2015.10.015](https://doi.org/10.1016/j.puhe.2015.10.015)] [Medline: [26976489](https://pubmed.ncbi.nlm.nih.gov/26976489/)]
27. Perrin A. Mobile technology and home broadband 2021. Pew Research Center. 2021 Jun 03. URL: <https://www.pewresearch.org/internet/2021/06/03/mobile-technology-and-home-broadband-2021/> [accessed 2021-09-30]
28. Wilson K, Brownstein JS. Early detection of disease outbreaks using the internet. *CMAJ* 2009 Apr 14;180(8):829-831 [FREE Full text] [doi: [10.1503/cmaj.090215](https://doi.org/10.1503/cmaj.090215)] [Medline: [19364791](https://pubmed.ncbi.nlm.nih.gov/19364791/)]
29. Lu FS, Hou S, Baltrusaitis K, Shah M, Leskovec J, Sosic R, et al. Accurate influenza monitoring and forecasting using novel internet data streams: a case study in the Boston metropolis. *JMIR Public Health Surveill* 2018 Jan 09;4(1):e4 [FREE Full text] [doi: [10.2196/publichealth.8950](https://doi.org/10.2196/publichealth.8950)] [Medline: [29317382](https://pubmed.ncbi.nlm.nih.gov/29317382/)]
30. Cavazos-Rehg PA, Krauss MJ, Spitznagel EL, Lowery A, Grucza RA, Chaloupka FJ, et al. Monitoring of non-cigarette tobacco use using Google Trends. *Tob Control* 2015 May 05;24(3):249-255 [FREE Full text] [doi: [10.1136/tobaccocontrol-2013-051276](https://doi.org/10.1136/tobaccocontrol-2013-051276)] [Medline: [24500269](https://pubmed.ncbi.nlm.nih.gov/24500269/)]
31. Senecal C, Widmer RJ, Lerman LO, Lerman A. Association of search engine queries for chest pain with coronary heart disease epidemiology. *JAMA Cardiol* 2018 Dec 01;3(12):1218-1221 [FREE Full text] [doi: [10.1001/jamacardio.2018.3459](https://doi.org/10.1001/jamacardio.2018.3459)] [Medline: [30422176](https://pubmed.ncbi.nlm.nih.gov/30422176/)]

32. Higgins T, Wu A, Sharma D, Illing E, Rubel K, Ting J, Snot Force Alliance. Correlations of online search engine trends with coronavirus disease (COVID-19) incidence: infodemiology study. *JMIR Public Health Surveill* 2020 May 21;6(2):e19702 [[FREE Full text](#)] [doi: [10.2196/19702](https://doi.org/10.2196/19702)] [Medline: [32401211](https://pubmed.ncbi.nlm.nih.gov/32401211/)]
33. Mavragani A. Tracking COVID-19 in Europe: infodemiology approach. *JMIR Public Health Surveill* 2020 Apr 20;6(2):e18941 [[FREE Full text](#)] [doi: [10.2196/18941](https://doi.org/10.2196/18941)] [Medline: [32250957](https://pubmed.ncbi.nlm.nih.gov/32250957/)]
34. Sousa-Pinto B, Anto A, Czarlewski W, Anto J, Fonseca J, Bousquet J. Assessment of the impact of media coverage on COVID-19-related google trends data: infodemiology study. *J Med Internet Res* 2020 Aug 10;22(8):e19611 [[FREE Full text](#)] [doi: [10.2196/19611](https://doi.org/10.2196/19611)] [Medline: [32530816](https://pubmed.ncbi.nlm.nih.gov/32530816/)]
35. Szmuda T, Ali S, Hetzger TV, Rosvall P, Sloniewski P. Are online searches for the novel coronavirus (COVID-19) related to media or epidemiology? a cross-sectional study. *Int J Infect Dis* 2020 Aug;97:386-390 [[FREE Full text](#)] [doi: [10.1016/j.ijid.2020.06.028](https://doi.org/10.1016/j.ijid.2020.06.028)] [Medline: [32535297](https://pubmed.ncbi.nlm.nih.gov/32535297/)]

Abbreviations

MAE: mean absolute error

MAPE: mean absolute percentage error

Edited by T Mackey; submitted 05.08.21; peer-reviewed by A Rodriguez, M Lotto, A Mavragani, Y Nishimura; comments to author 28.09.21; revised version received 05.10.21; accepted 07.12.21; published 22.04.22

Please cite as:

Trevino J, Malik S, Schmidt M

Integrating Google Trends Search Engine Query Data Into Adult Emergency Department Volume Forecasting: Infodemiology Study
JMIR Infodemiology 2022;2(1):e32386

URL: <https://infodemiology.jmir.org/2022/1/e32386>

doi: [10.2196/32386](https://doi.org/10.2196/32386)

PMID:

©Jesus Trevino, Sanjeev Malik, Michael Schmidt. Originally published in *JMIR Infodemiology* (<https://infodemiology.jmir.org>), 22.04.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Infodemiology*, is properly cited. The complete bibliographic information, a link to the original publication on <https://infodemiology.jmir.org/>, as well as this copyright and license information must be included.