Norovirus Disease Surveillance Using Google Internet Query Share Data

Rishi Desai,1 Aron J. Hall,1 Benjamin A. Lopman,1 Yair Shimshoni,2 Marcus Rennick,1 Niv Efron,2 Yossi Matias,2 Manish M. Patel,1 and Umesh D. Parashar1

1Centers for Disease Control and Prevention, Atlanta, Georgia; 2Google Labs, Tel-Aviv, Israel; and 3Boston Public Health Commission, Massachusetts

Google Internet query share (IQS) data for gastroenteritis-related search terms correlated strongly with contemporaneous national ($R^2 = 0.70$) and regional ($R^2 = 0.74$) norovirus surveillance data in the United States. IQS data may facilitate rapid identification of norovirus season onset, elevated peak activity, and potential emergence of novel strains.

Norovirus is the leading cause of gastroenteritis in the United States, contributing to approximately 21 million illnesses annually, the majority occurring during winter months [1, 2]. Unfortunately, prompt, publicly available norovirus surveillance data are scarce in the United States. Syndromic surveillance approaches are limited because approximately 90% of individuals with norovirus disease do not seek medical attention [3]. Because both the timing and magnitude of norovirus activity may vary from year to year [4, 5], a timely, publicly accessible, and economical surveillance tool may allow public health officials to establish situational awareness and offer targeted health education campaigns.

Google Insights for Search allows users to easily track Internet search queries over specified time intervals and geographic regions [6, 7]. Internet-based public health surveillance is a new approach that can be performed using syndromic and disease-specific terms [8]. To assess whether Internet search trends are appropriate for monitoring norovirus disease, we compared Internet query share (IQS) for gastroenteritis-related syndromic search terms with existing norovirus disease surveillance data in the United States. Although these data sources may not fully capture norovirus disease burden, they are appropriate for time-series analysis and timeliness.

METHODS

Data Sources

Norovirus outbreak surveillance data and estimated hospital discharges for norovirus were used to represent national norovirus trends. Outbreak surveillance data were based on monthly reports of suspected and confirmed norovirus outbreaks occurring between 1 January 2007 and 30 April 2010 from 30 US states [5]. Norovirus-associated hospital discharges among all ages were estimated from time-series regression models constructed using all-cause gastroenteritis-associated hospitalizations between 1 January 2004 and 30 June 2007 from a nationally representative data set [9].

Prospective citywide syndromic surveillance data on proportion of all emergency department visits due to gastrointestinal illness were obtained from the Boston Public Health Commission (BPHC) for the period 1 January 2006 through 2 July 2011. The Boston Syndromic Surveillance System receives daily information on all Boston-based hospital emergency department (n = 10) visits, with information on visit date, facility identification, a unique visit identification, chief complaint, age, gender, race, and zip code of residence (billing zip code). BPHC epidemiologists use the Early Aberration Reporting System to categorize chief complaints into customizable syndromes. Acute gastrointestinal illness syndrome is defined as “non-chronic diarrhea or vomiting in the chief complaint or an enteric illness (food poisoning, salmonella).”

Google stores information on all search queries performed using the Google search engine [6]. These data are used to calculate an IQS, which is the number of search queries for the term of interest in a given time period and geographic region, divided by the total number of queries from the same time period and region. IQS data are scaled, which means that weekly data are reported as a proportion of the peak week query share in a time period. The scaled query share values are plotted over time, generating a time series. The norovirus-related IQS definition was initially created using the following 10 search terms for norovirus disease: norovirus, vomiting, diarrhea, nausea, abdominal pain, stomach virus, food poisoning, gastroenteritis, Norwalk virus, and rotavirus. Each term was separately grouped along with any other strongly correlated search term(s). The combined query share of each group
was evaluated both with and without a filter that only includes searches related to the category “Gastro Esophageal Reflux Disease and Digestive Disorders (GERD).” The IQS definition with the strongest correlation with US norovirus outbreak surveillance data was selected.

**Data Analysis**

Monthly IQS data were generated for 30 selected US states (for which outbreak data were available) and for the entire US population from 1 January 2004 through 30 April 2010. Norovirus IQS data for 30 US states were compared with norovirus outbreak surveillance data from 1 January 2007 through 30 April 2010, and norovirus IQS data for the entire US population were compared with estimated norovirus-associated hospital discharges from 1 January 2004 through 30 June 2007. We defined the norovirus seasonal year as 1 May through 30 April of the following year in order to capture the winter seasonal peak. Months with the highest level of seasonal activity were identified for each norovirus season, and Pearson correlation coefficients were calculated. At the local level, we obtained weekly IQS data for the metro population of Boston from 1 January 2006 through 2 July 2011. Pearson correlation coefficients between BPHC surveillance data and IQS data were calculated for a range of ±26-week lead and lag times, and the lead or lag time that maximized the correlation coefficient was identified.

**RESULTS**

Between 1 January 2007 and 1 May 2010, the group of norovirus syndromic search terms with query share data that had the strongest correlation ($R^2 = 0.95$) with US norovirus outbreak surveillance data included: “stomach virus,” “stomach flu,” “stomach illness,” “stomach bug,” “stomach sickness,” and “stomach sick,” using a GERD filter (Supplementary Table). When compared with US norovirus outbreak surveillance data, IQS data had the same peak month in 2 seasons (January in 2007–2008 and February in 2009–2010). In the 2008–2009 season, IQS data peaked in December, whereas the US outbreak surveillance data peaked in January. Between 1 January 2004 and 1 May 2007, the correlation between IQS data and estimated norovirus-associated hospital discharges was moderately strong ($R^2 = 0.70$). When compared with estimated norovirus-associated hospital discharges data, IQS data had the same peak month in 2 seasons (January 2004–2005 and January 2006–2007). In the 2005–2006 season, IQS data peaked in December, whereas the estimated norovirus-associated hospital discharges data peaked in March (Figure 1).

In Boston between 1 January 2006 and 2 July 2011, correlation was strongest ($R^2 = 0.74$) between IQS data from 2 weeks prior to emergency department surveillance (ie, IQS data from week$-2$ compared with norovirus emergency department surveillance data from week$_n$) (Figure 2).

**DISCUSSION**

US IQS data for norovirus correlated strongly with 2 indicators of national norovirus activity and at the metropolitan level in Boston. This suggests a unique role for this timely surveillance tool at various levels of public health. Utility of some traditional norovirus surveillance systems is limited by substantial reporting delays, which diminish their ability to offer meaningful information about season onset or magnitude. As

![Figure 1](cid201255_15october_briefreport.png)

Figure 1. (A) Norovirus Internet query share (IQS) data (blue) for the entire US population were compared with data on estimated norovirus-associated hospital discharges (red) from 1 January 2004 through 30 June 2007. (B) Norovirus IQS data (blue) for 30 US states were compared with norovirus outbreak surveillance data (red) from 1 January 2007 through 30 April 2010.
a real-time, automated global data source, IQS data can offer an early indication of increased norovirus disease, which may guide local clinical and diagnostic protocols. It may also allow public health authorities to rapidly identify norovirus season onset and elevated peak activity, which may suggest potential emergence of new norovirus variants [4].

The interpretation of IQS data is hampered by unknown user search motives, possibly reflecting both active illness and general media-driven interest in norovirus disease. Interestingly, in a Swedish study, media reports did not correlate with either search queries or laboratory-confirmed cases of norovirus [10]. Another concern is the potential for an outbreak of another disease that causes a similar illness to influence search patterns. In our analysis, search terms related to gastroenteritis, but not specific to norovirus, were used. Fortunately, because norovirus disease is the most common cause of gastrointestinal disease in the United States, the IQS data correlated well with national norovirus surveillance data. Additionally, any large national or local event that increases or decreases the number of Internet searches disproportionate to the number of norovirus queries would affect the IQS data. Finally, this approach may not work well among small populations with limited Internet access.

In this analysis, IQS data were compared with 3 data sources, each with its own limitations. The norovirus outbreak data relied upon voluntary state and local public health agency reporting practices, which vary from region to region and may fluctuate over time. Norovirus-associated hospital discharge data were modeled using seasonal residuals of cause-unspecified gastroenteritis data rather than laboratory-confirmed episodes of norovirus. The 2005–2006 season had a high rotavirus disease burden, causing the norovirus-associated hospital discharge data to have an erratic pattern; likely explaining its 3-month discordance with the IQS data. Fortunately, the diminished burden of rotavirus disease in the postvaccine era combined with an effective national rotavirus surveillance system can help distinguish these etiologies from one another moving forward. Although the BPHC data were not specific to norovirus, they provide valuable syndromic trend data to municipal public health and local clinicians. Optimal association between IQS and BPHC data occurred with a 2-week IQS lead time, suggesting that Internet searches for norovirus may be occurring prior to emergency department visits for gastroenteritis. Unfortunately, such a temporal association cannot be further evaluated because IQS data do not provide individual-level information.

Internet-based search data hold promise in settings where traditional surveillance data are limited, absent, or simply lack timeliness, but Internet access is widespread. The ability to rapidly gather and disseminate information about disease trends in their community is relevant for clinical providers and for individuals. Additionally, this approach has been successfully used for other infectious diseases including influenza and rotavirus [8, 11]. As new national surveillance systems for norovirus outbreaks in the United States become better established [1], additional data will become available to further evaluate IQS data as a complement to traditional surveillance.

Supplementary Data

Supplementary materials are available at Clinical Infectious Diseases online (http://cid.oxfordjournals.org). Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

Notes

Acknowledgments. The authors acknowledge Julia Gunn from the Boston Public Health Commission in Boston, Massachusetts for her assistance in obtaining local public health data and her critical review of the manuscript.

Disclaimer. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

Potential conflicts of interest. Three coauthors, Y. S., N. E., and Y. M., work at Google Incorporated. No compensation was given to any of the aforementioned individuals or Google Incorporated. No funding body or sponsor had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript. R. D. had full access to all data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. All other authors report no potential conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.
References