Dig Dis Sci (2017) 62:562–563
DOI 10.1007/s10620-017-4458-5

EDITORIAL

# The Premise and Promise of Big Data for Tracking Population Health: Big Deal or Big Disappointment?

**Emad Mansoor**[1] · **Sadeer G. Al-Kindi**[1]

Published online: 20 January 2017


*"We are drowning in information but starved for knowledge"*
*John Naisbitt (1982)*

It is estimated that 60% of adults in the USA search for health information online, with the majority searching for a specific disease or treatment [1]. The patterns of these searches can provide important temporal and spatial data that reflect health indicators at the individual and the population levels. As examples, Google searches for a multitude of keywords accurately predicted influenza outbreaks in real-time [2]; searches for specific symptoms preceded searches for terms strongly suggestive of lung cancer [3].

In this issue of *Digestive Diseases and Sciences,* Hassid and colleagues [4] offer a pertinent study of the utilization of worldwide web search engine queries to study the prevalence of common gastrointestinal symptoms. The authors used Google search engine data (Google Trends) to evaluate the volume of searches for dysphagia, vomiting, and diarrhea in the USA over three years, correlating the relative changes in the search volume for these symptoms with an inpatient (National Inpatient Sample) and an outpatient (National Hospital Ambulatory Medical Care Survey) clinical dataset. The authors reported that the changes in Google search volume for dysphagia ($r = 0.5$, $P = 0.002$), diarrhea ($r = 0.79$, $P < 0.001$), and vomiting ($r = 0.76$, $P < 0.001$) correlated significantly with the inpatient data. Both Google Trends and NIS data indicated that the prevalence of these symptoms increased over the study period, with concordant seasonal variations.

Hassid's study, as part of the recent explosion of the application of big data analytics of the output of web search engines and social media platforms for disease epidemiology and surveillance [5], is among the first to investigate the use of web search trends to identify the prevalence of non-communicable diseases. Although this fresh approach to disease epidemiology has raised cautious optimism for the use of real-time examination and analysis of freely available online data, its relationship to tangible clinical data has not yet been fully investigated, particularly in the field of gastroenterology.

Hassid's study emphasizes that terms used by the public (e.g., diarrhea, vomiting) may be more accurately tracked through public web searches than terms used mostly by healthcare professionals (e.g., dysphagia), as evidenced by the difference in correlation strengths. One opportunity is to model multiple search terms with different weights, as originally used in the Google Flu Trends to identify relationships to clinical data, which are now made publicly available through Google Correlate and Google Insights for search. The temporal use of web searches for symptoms may also prove useful in the early detection of serious diseases. Such analyses will provide not only descriptive trends of diagnoses, but also hold promise to introduce new information acquired from the public.

This study reports that the frequency of at least a few gastroenterological symptoms prevalent in the population can be tracked through web-based search engines with reasonable accuracy, providing a template for the utilization of open-access large datasets in the examination of variation in non-communicable disease epidemiology, thereby helping assign timely resources for efficient patient management of these diseases in inpatient and outpatient

✉ Sadeer G. Al-Kindi
sadeer.alkindi@uhhospitals.org

[1] Department of Medicine, Case Western Reserve University/University Hospitals Cleveland Medical Center, 11100 Euclid Avenue, Cleveland, OH 44106, USA

settings. Furthermore, studying the seasonal variations and geographic sources of these searches may help uncover pathophysiologic triggers that may relate to viral outbreaks, ambient weather, and air pollution, important relationships that await validation of Google searches as surrogates of disease epidemiology.

The promise of web search analytics, however, is far from being fulfilled since analysis and interpretation of these data remains uncertain and inefficient. Big data mining continues to be a challenge due to the enormous amounts of data combined with the lack of efficient analytic methods. Big data analysis may also be subject to false associations, which may be more pronounced in the analysis of search engine patterns. For example, when developing Google Flu Trends, some search terms which significantly correlated with influenza were clearly coincidental (e.g., "high school basketball") [2]. Given these and other challenges, big data is often considered as complementary to, rather than a replacement for traditional systematic research approaches.

In summary, the potential of big data, with its remarkable volume, variety, and velocity, will undoubtedly substantially influence twenty-first-century medicine and gastroenterology. Nevertheless, new computational algorithms, improved machine learning, and improved quality of data are needed for the promise of big data to be fully realized.

**Compliance with ethical standards**

**Conflict of interest** None.

## References

1. Fox S, Duggan M. Health online 2013. *Health (NY)*. 2013;1–55.
2. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457:1012–1014.
3. White RW, Horvitz E. Evaluation of the feasibility of screening patients for early signs of lung carcinoma in web search logs. *JAMA Oncol.* 2016. doi:10.1001/jamaoncol.2016.4911.
4. Hassid BG, Day LW, Awad MA, Sewell JL, Osterberg EC, Breyer BN. Using search engine query data to explore the epidemiology of common gastrointestinal symptoms. *Dig Dis Sci*. (Epub ahead of print). doi:10.1007/s10620-016-4384-y.
5. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection—harnessing the Web for public health surveillance. *N Engl J Med*. 2009;360:2153–2157.