BMJ 2017;358:j3159 doi: 10.1136/bmj.j3159 (Published 2017 July 03)

THE ART OF RISK COMMUNICATION

Can search engine data predict pancreatic cancer?

Gerd Gigerenzer discusses how search engines use big data analytics to “diagnose” your state of health

Gerd Gigerenzer *director*

Max Planck Institute for Human Development and Harding Center for Risk Literacy, Berlin, Germany

Imagine this warning popping up on your search engine page: “Attention! There are signs that you might have pancreatic cancer. Please visit your doctor immediately.” Just as search engines use big data analytics to detect your book and music preferences, they may also “diagnose” your state of health.

Microsoft researchers have claimed that web search queries could predict pancreatic adenocarcinoma.[1] A retrospective study of 6.4 million users of Microsoft’s search engine Bing identified first person searches suggestive of a recent diagnosis, such as “I was told I have pancreatic cancer, what to expect.” Then the researchers went back months before these queries were made and looked for earlier ones indicating symptoms or risk factors, such as blood clots and unexplained weight loss. They concluded that their statistical classifiers “can identify 5% to 15% of cases, while preserving extremely low false positive rates (0.00001 to 0.0001),” and that “this screening capability could increase five year survival.” The *New York Times* reported that the study “suggests that early screening can increase the five year survival rate of patients with pancreatic cancer to 5 to 7%, from just 3%.”[2]

Thus it appears that Microsoft researchers have found a low cost, high coverage surveillance system that produces almost no false alarms and saves lives—an improvement over previous diagnostic attempts using biomarkers or imaging. In this column, I do not deal with the typical problems of big data, such as intransparent algorithms and the danger of overfitting noise, or with the ethics of not soliciting users’ consent to having their personal data (albeit anonymously) analysed. Rather, I take the results as given and focus on how the presentation of these results invites several potential misunderstandings.

Firstly, consider the prospect of increased survival rates, which suggests that surveillance saves lives. In fact, in the context of screening, the correlation between increases in survival rates and decreases in mortality rates is approximately zero for the 20 most common solid tumours over the past 50 years.[3] One reason for this is called “lead time bias.” Early detection implies that diagnosis occurs at an earlier stage of a disease, which leads to higher five year survival rates from the time of diagnosis—even if the patients ultimately do not live any longer in terms of absolute age. Reporting survival instead of mortality rates misleads the reader about the benefits of cancer screening.[4]

Secondly, consider the extremely low false positive rates. Does that mean that Bing users who get the news that they are positive are seldom falsely alarmed? To answer that, let us go through a simple example. Assume 100 000 users, 10 of whom have undetected pancreatic cancer.[5] Given a sensitivity of 10% (the average of 5% and 15%), we expect that one user correctly tests positive and the other nine are missed. Given a false positive rate of 1 in 10 000 (or 0.0001), we expect that about 10 users test positive even though they do not have cancer. Thus, we expect a total of 11 users to test positive, of whom 10 do not have pancreatic cancer. The general point is that even with low false positive rates, the proportion of false alarms among all users who test positive can be high if the disease is rare.

The authors note that surveillance by Bing does not replace a physician. Yet their presentation of the statistical results easily invites systematic misunderstandings by both patients and doctors. In the *New York Times,* one of the authors of the research, Eric Horvitz, mentions his hope that the study will stimulate quite a bit of interesting conversation. My response is that in order to demonstrate the clinical usefulness of big data analytics, the first step should be towards more transparency and fewer misleading statistics. Not doing so recalls the rise and fall of Google Flu Trends, which in 2009 was trumpeted as being able to predict influenza but disappeared from sight after years of failing to meet its own projected rates of predictive accuracy. Big data are known for their fanfare and hype. In this case, all has been quiet since the buzz last summer.

Competing interests: None declared.

1 Paparrizos J, White RW, Horvitz E. Screening for pancreatic adenocarcinoma using signals from web search logs: feasibility study and results. *J Oncol Pract* 2016;358:737-44.pmid:27271506.

gigerenzer@mpib-berlin.mpg.de

2 Markoff J. Microsoft finds cancer clues in search queries. *New York Times* 2016. www.nytimes.com/2016/06/08/technology/online-searches-can-identify-cancer-victims-study-finds.html.

3 Welch HG, Schwartz LM, Woloshin S. Are increasing 5-year survival rates evidence of success against cancer?*JAMA* 2000;358:2975-8. doi:10.1001/jama.283.22.2975 pmid:10865276.

4 Gigerenzer G. Breast cancer screening pamphlets mislead women. *BMJ* 2014 25;348:g2636.

5 Gravano L. An interview with John Paparrizos: hunting early signs of pancreatic cancer in web searches. 2016. www.cs.columbia.edu/2016/web-searches-as-an-early-warning-system-for-pancreatic-cancer.