Original Research

Tracking search engine queries for suicide in the United Kingdom, 2004—2013

V.S. Arora [a,b,*], D. Stuckler [c], M. McKee [a,b]

[a] *Department of Public Health and Policy, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK*
[b] *European Observatory on Health Systems and Policies, Houghton Street, London WC2A 2AE, UK*
[c] *Department of Sociology, University of Oxford, Manor Road Building, Oxford OX1 3UQ, UK*

ABSTRACT

Objectives: First, to determine if a cyclical trend is observed for search activity of suicide and three common suicide risk factors in the United Kingdom: depression, unemployment, and marital strain. Second, to test the validity of suicide search data as a potential marker of suicide risk by evaluating whether web searches for suicide associate with suicide rates among those of different ages and genders in the United Kingdom.

Study design: Cross-sectional.

Methods: Search engine data was obtained from Google Trends, a publicly available repository of information of trends and patterns of user searches on Google. The following phrases were entered into Google Trends to analyse relative search volume for suicide, depression, job loss, and divorce, respectively: 'suicide'; 'depression + depressed + hopeless'; 'unemployed + lost job'; 'divorce'. Spearman's rank correlation coefficient was employed to test bivariate associations between suicide search activity and official suicide rates from the Office of National Statistics (ONS).

Results: Cyclical trends were observed in search activity for suicide and depression-related search activity, with peaks in autumn and winter months, and a trough in summer months. A positive, non-significant association was found between suicide-related search activity and suicide rates in the general working-age population (15—64 years) ($\rho = 0.164$; $P = 0.652$). This association is stronger in younger age groups, particularly for those 25—34 years of age ($\rho = 0.848$; $P = 0.002$).

Conclusions: We give credence to a link between search activity for suicide and suicide rates in the United Kingdom from 2004 to 2013 for high risk sub-populations (i.e. male youth and young professionals). There remains a need for further research on how Google Trends can be used in other areas of disease surveillance and for work to provide greater geographical precision, as well as research on ways of mitigating the risk of internet use leading to suicide ideation in youth.

© 2015 The Royal Society for Public Health. Published by Elsevier Ltd. All rights reserved.

Introduction

Suicide rates have long been viewed as an early warning sign of wider social, political, or economic distress.[1–3] Yet delays inherent in collection and analysis of suicide surveillance often take several years before such data become available. For example, suicide rates for 2013 in England and Wales were not published until February 2015[4] while, in other countries, delays are even longer. The most recent data for the entire Belgian population are from 2008.[5]

Recently efforts have been made to take advantage of new data created by people's engagement with social media and search engines. Infectious disease epidemiologists have drawn on internet search data to provide early warning of potential outbreaks,[6,7] such as by tracking search engine queries relevant to infectious diseases, including influenza,[8] HIV,[9] and gastroenteritis.[10] These applications have now gone beyond infectious disease, from monitoring interest in abortion[11] to electronic cigarettes.[12]

Prior academic evidence suggests that this approach may also be fruitful for mental health and suicide epidemiology.[13,14] Using web search engines for premeditating suicide is not unheard of in the general public. Even in the case of the tragic Germanwings plane crash in the French Alps this spring, German prosecutors revealed that the pilot, Andreas Lubitz, had conducted internet research on 'cockpit doors' and 'suicides' days before the crash.[15]

However, for search data to be useful in population health research, there are remaining questions about their validity. In particular, there are concerns about representativeness of data, especially if they fail to capture the intentions of vulnerable groups that may not have routine access to search engines. A recent American study by McCarthy[14] matched Google internet search engine activity to US suicide rates reported by the Centers of Disease Control (CDC), finding a strong negative correlation with suicide rates in the general US population but a strong positive correlation with rates among youth. McCarthy[14] also found a cyclical trend in search activity for suicide and searches for suicide risk factors (e.g. depression) that correlated with seasonal patterns in suicides, but only in individuals aged 15–24 years of age. In addition, there was no acknowledgement that search activity may vary by gender. Lastly, no tests were employed to determine the significanceof association between search activity and suicide rates.

Thus far, McCarthy's results have not been replicated in other countries and specifically, Western industrialised countries. In this paper we have two specific aims. First, we sought to determine if the cyclical trend seen in the USA was also evident for search activity related to suicide and three common suicide risk factors in the United Kingdom: depression,[16,17] unemployment,[2,18–21] and marital strain.[22–24] Second, we test the validity of suicide search data as a potential marker of suicide risk by evaluating whether web searches for suicide associate with suicide rates among those of different ages and genders in the United Kingdom.

Methods

We obtained search engine data from Google Trends, a publicly available repository of information of trends and patterns of user searches on Google. Data can be disaggregated by geographic region and time period although, clearly, they do not provide information on the demographic characteristics of those undertaking the searches. While Google Trends does not show the absolute number of searches, it calculates a *query share* for a search term. In practice, this means that Google calculates the number of searches for a given term as a proportion of the total number of searches in each location and at each time. These calculations are then normalised to a Google Trends relative search volume (RSV) index between 0 and 100, where an RSV index of 100 designates the date when there was the highest amount of search activity for a given term. Thus, a search index of 40 equates to 40% of the most intense search activity at any time.

Within these analyses, users can easily stratify findings on a national, regional, or city level if necessary. Multiple searches are possible, combining multiple search terms using a '+' symbol. An obvious issue is how to differentiate terms that may have multiple meanings in various fields. For example, inputting the search term 'depression' can correspond to mental illness, but it could also pick up search activity relating to the Great Depression. To help deal with this, Google is able to use natural language processing (NLP) methods and index web pages to allocate search engine queries into one of 25 specific categories (e.g. Finance, Health, Population & Society, etc.) and over 300 subcategories. Details of NLP methods have been provided elsewhere.[25]

We entered the following phrases into Google Trends to analyse relative search volume for suicide, depression, job loss, and divorce, respectively: 'suicide', 'depression + depressed + hopeless', 'unemployed + lost job', 'divorce'. As with McCarthy,[14] we used lay terms when inputting search terms since medical terms did not generate sufficient search volume for useful analysis. While a term like 'divorce' may not fully capture relationship breakdown, combinations of divorce with other terms (e.g. 'abuse', 'domestic violence', and 'abuse') do not alter search trends in the United Kingdom (data not shown). This suggests that the search term for divorce has a much greater relative search volume in defining relationship breakdown. The United Kingdom was set as the geographic location for this study, with a time period of January 2004 to December 2013, the most recent date for which suicide data are currently available. Since Google Trends reports data either in a weekly or monthly format while suicide data are reported annually, indices were averaged to form yearly estimates of relative search volume.

To determine the feasibility of using search queries to measure suicide rates, yearly relative search volume estimates for suicide were compared against publicly available official suicide rates (per 100,000 population) from the Office of National Statistics (ONS) for 2004 to 2013 for the working-age population (15–64 years), with a stratification based on sex and the following age cohorts: 15–24 years, 25–34 years, 35–44 years, 45–54 years, and 55–64 years. In a secondary