Environmental Research xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Google search patterns monitoring the daily health impact of heatwaves in England: How do the findings compare to established syndromic surveillance systems from 2013 to 2017?

Helen K. Green[a], Obaghe Edeghere[a], Alex J. Elliot[a], Ingemar J. Cox[b,c], Roger Morbey[a], Richard Pebody[d], Angie Bone[e], Rachel A. McKendry[f], Gillian E. Smith[a,*]

[a] Real-time Syndromic Surveillance Team, Public Health England, Birmingham, United Kingdom
[b] Department of Computer Science, University College London, London, United Kingdom
[c] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
[d] Respiratory Diseases Department, Public Health England, London, United Kingdom
[e] Extreme Events, Public Health England, London, United Kingdom
[f] London Centre for Nanotechnology and Division of Medicine, University College London, London, United Kingdom

ARTICLE INFO

Keywords:
Temperature
Heat-related illness
Heat-waves
Syndromic surveillance
Public health

ABSTRACT

One of the implications of climate change is a predicted increase in frequent and severe heatwaves. The impact of heatwaves on the health of the population is captured through real-time syndromic healthcare surveillance systems monitored daily in England during the summer months. Internet search data could potentially provide improved timeliness and help to assess the wider population health impact of heat by capturing a population sub-group who are symptomatic but do not seek healthcare. A retrospective observational study was carried out from June 2013 to September 2017 in England to compare daily trends in validated syndromic surveillance heat-related morbidity indicators against symptom-based heatwave related Google search terms. The degree of correlation was determined with Spearman correlation coefficients and lag assessment was carried out to determine timeliness. Daily increases in frequency in Google search terms during heatwave events correlated well with validated syndromic indicators. Correlation coefficients between search term frequency and syndromic indicators from 2013 to 2017 were highest with the telehealth service NHS 111 (range of 0.684–0.900 by search term). Lag analysis revealed a similar timeliness between the data sources, suggesting Google data did not provide a delayed or earlier signal in the context of England's syndromic surveillance systems. This work highlights the potential benefits for countries which lack established public health surveillance systems to monitor heat-related morbidity and the use of internet search data to assess the wider population health impact of exposure to heat.

1. Background

Heatwaves can significantly impact on the health of the population, with outcomes ranging from dehydration and sunburn through to heatstroke and death. The highest vulnerability is typically seen in the elderly (Benmarhnia et al., 2015), children (Li et al., 2015), and in those with pre-existing medical conditions (Bouchama et al., 2007), although all age groups may be affected, particularly in hot countries, where people are working outdoors without adequate protection (Nelson et al., 2011) or cannot seek shelter (e.g. homeless, Ramin, 2009).

The Heatwave Plan for England was launched in 2004 (PHE, 2015) in response to the 2003 European heatwave which resulted in an estimated 70,000 deaths across Europe (Robine et al., 2008). As part of this Plan, from 1st June to 15th September each year, Public Health England (PHE) routinely monitors on a daily basis heat-related morbidity across a suite of syndromic surveillance systems (Triple S Project, 2011; PHE, 2017). Syndromic surveillance is the near real-time collection, analysis, interpretation and dissemination of health-related data to provide an early warning of threats requiring public health action, or conversely provide reassurance about the absence of impact of such threats. Indicators monitored within these systems in England have been demonstrated to provide a sensitive measurement of the health impact of heat (Smith et al., 2016a, 2016b; Elliot et al., 2014) and

* Correspondence to: Real-time Syndromic Surveillance Team, National Infection Service, Public Health England, 5 St Philip's Place, Birmingham B3 2PW, United Kingdom.
E-mail address: gillian.smith@phe.gov.uk (G.E. Smith).

https://doi.org/10.1016/j.envres.2018.04.002
Received 30 October 2017; Received in revised form 31 March 2018; Accepted 3 April 2018

information can be used to facilitate optimal public health response during the heatwave through ensuring appropriate messaging (Josseran et al., 2010; Pascal et al., 2012; Lall et al., 2017).

The use of novel data sources in health surveillance, such as social media and internet search queries is gaining increasing interest (Nuti et al., 2014). While syndromic surveillance provides a broader measure of illness through a clinical diagnosis (as opposed to laboratory confirmation, which is a more specific end-point for infectious diseases), current syndromic surveillance systems in use in England still rely on an individual consulting healthcare. The potential advantages of online user-generated content include the provision of more timely information and identification of illness in symptomatic individuals, including those who do not seek healthcare. For data sources such as Google Trends (Google, 2017), this comes at a minimal cost and may therefore be useful in countries with sufficient internet coverage but without an established public health surveillance infrastructure. However, there are notable limitations with these data sources, including uncertainty of the underlying reason for an individual searching and of the characteristics of the person searching.

The potential utility of internet search data to monitor heat-related morbidity was demonstrated in Shanghai (Li et al., 2016), with strong correlation between internet searches for *heat stroke* and heatstroke deaths and hospitalised cases. However, availability of health outcome data was not as timely as in England and it is not known how internet search data for heat-related illness compares to outputs from real-time syndromic surveillance systems. During 2017 in England, there was a Level 2 heatwave alert issued (defined as a risk of reaching high temperatures) from 15 to 21st of June, with areas reaching high temperatures (a Level 3 heatwave alert) across most of the country from 17 to 20th June (Met Office, 2017). This article outlines the findings from a retrospective observational study comparing trends in heat-related syndromic surveillance indicators against selected heatwave related internet search terms during this period and for previous heatwaves. This information assists in determining the utility and timeliness of internet search data as a means of assessing population health impacts from heatwaves.

2. Methods

2.1. Data

National data for England for each of the Google search terms and heat-related syndromic indicators was extracted on a daily timescale from 1 June 2013 to 15 September 2017.

Anonymised daily syndromic data for heat impact indicators (Smith et al., 2016a) was accessed from the four national PHE syndromic surveillance systems; NHS 111 (a telehealth service available to the population of England), an in-hours general practitioner (GP) system (GPIH; denominator population of 35 million registered patients across England), an out of hours GP system (GPOOH; approximately 70% of out of hours activity across England), and a sentinel emergency department (ED) surveillance system (35 sentinel EDs across England and Northern Ireland) as previously described (Smith et al., 2016a, 2016b; Harcourt et al., 2017, PHE 2017).

Google search volume data was available from the Google Health Trends Application Programming Interface (API), a tool to explore Google search data similar to the publicly available Google Trends website (Google, 2017). Data is taken from a uniformly distributed random sample of 10–15% of Google web searches updated daily. Daily probabilities were defined as the probability of searching for a specified term on a given day in England, multiplied by ten million to be human readable.

Most disease surveillance studies looking at Google Trends data typically focus on one or two relevant search terms. We sought to expand on this by utilising the clinical codes underlying the heat-related syndromic morbidity indicators to develop a list of search terms that people may use to search Google. Relevant search terms were identified through the following process:

Table 1
Selection of Google search terms.

Terms identified from syndromic surveillance	Additional terms/comments from Google Trends
Heat exhaustion	Heat stroke
Heat rash	Prickly heat
Heatstroke	Heat stroke
	Sunstroke
Heat stroke	No additional relevant terms identified
Heatwave	No additional relevant terms identified
Prickly heat	Heat rash
Sunburn	No additional relevant terms identified
Sunstroke	No additional relevant terms identified
Sun stroke	Heat stroke
	Sunstroke
Heat syncope	Not enough searches carried out in Google to provide data
Heat fatigue	Not enough searches carried out in Google to provide data
Heat oedema	Not enough searches carried out in Google to provide data
Heat prostration	Not enough searches carried out in Google to provide data

1. Based on experience with clinical codes used for existing syndromic surveillance systems, descriptions of such codes corresponding to heat-related syndromic morbidity indicators (e.g. *heatstroke*) were collated and synthesised into a list of code terms (Table 1).
2. Each code term was entered into Google Trends and the related Google internet search terms retained if also searched for at least half as often (Google, 2017) to identify further terms used (e.g. Google users searching for *heatstroke* also searching for *sunstroke*, Table 1) and associated search terms (e.g. *symptoms*). This cut off was used based on the observed distribution of frequency of related search terms and the relevance of these terms.
3. Where multiple variations of the search terms were available (e.g. *heatstroke*, *heat stroke*), there was generally one that had a markedly larger search probability, and this was retained.

The final list of search terms was as follows:

a) Heat exhaustion
b) Heat rash
c) Heat stroke
d) Heatwave
e) Prickly heat
f) Sunburn
g) Sunstroke

2.2. Analysis

Daily time series plots of national syndromic surveillance and Google search data were produced to visually compare patterns from 2013 to 2017 using daily data and seven day moving averages to minimise the impact of day of the week effects (Buckingham-Jeffery et al., 2017). Correlation was quantified through Spearman correlation coefficients as the data was not normally distributed. This comparison was also separately done for 2017 from 8 June 2017 to 27 June 2017 to focus on the known heatwave event, encompassing one week either side of the official heatwave alerts. Lagged time series plots were produced for the 2017 heatwave to assess the time lag/lead at which the two datasets best correlated. The Google time series was shifted forwards or backwards in time up to ten days to determine if search activity was