Research Paper

What can Google and Wikipedia can tell us about a disease? Big Data trends analysis in Systemic Lupus Erythematosus

Savino Sciascia*, Massimo Radin

Center of Research of Immunopathology and Rare Diseases- Coordinating Center of the Network for Rare Diseases of Piedmont and Aosta Valley, Department of Clinical and Biological Sciences, University of Turin, Turin, Italy

ARTICLE INFO

ABSTRACT

Objective: To investigate trends of Internet search volumes linked to Systemic Lupus Erythematosus (SLE), on-going clinical trials and research developments associated to the disease, using Big Data monitoring and data mining.

Methods: We performed a longitudinal analysis based on the large amount of data generated by Google Trends, scientific search tools (SCOPUS, Medline/Pubmed/ClinicalTrails.gov) considering 'SLE', and 'lupus' in a 5-year web-based research. Wikipedia page views were also analysed using WikiTrends and the results were compared with the search volumes generated by Google Trends.

Results: We observed an overall higher distribution of search volumes from Google Trends in United States, South America, Canada, South Africa, Australia and Europe (mainly Italy, United Kingdom, Spain, France, Germany), showing a geographically heterogeneity in insight into health-related behaviour of the different populations towards SLE. By comparing the search volumes analysing the Wikipedia page views of both SLE and belimumab, we found a close peak trend, reflecting the knowledge translation after the approval of belimumab for the treatment of SLE. When focusing on search volumes of Google Trends, we noticed that the highest peaks were related to news headlines that involved celebrities affected by SLE, also when comparing to the peak generated by the approval of belimumab.

Conclusion: This new approach, able to investigate health information seeking, might give an estimate of the health-related demand and even of the health-related behaviour of SLE, bringing new light to unanswered questions.

1. Introduction

Imagine being a researcher, or a scientist, or a public health official and having a tool reporting you in real-time what people with a specific condition or diseases are doing, experiencing or feeling, much as economists can look at the Dow Jones stock index as a real-time measure of what people are buying or selling in the global market.

In this study we explore new real-time metrics (presented as maps, graphics, and indices) of public behaviour, opinion, and knowledge, related to a specific health condition, based on textual data harvested from the Internet. The volume of user-generated data on Internet research tools, social media and other Internet-based venues has made measurable what was previously immeasurable [1]. This approach has opened up a fascinating area of research and new development with the potential to systematically mine, aggregate, and analyse unstructured data, aiming to address questions difficult or impossible to answer with traditional research tools.

The global epidemiology of many low-prevalence conditions is still to be determined, and available data showed wide ranges across different studies. This is the case of Systemic Lupus Erythematosus (SLE). The overall estimated prevalence for this condition in the population is 20–150 cases per 100,000; in women, prevalence rates vary from 164 (white) to 406 (African American) per 100,000. The estimated incidence rates are 1–25 per 100,000 in North America, South America, Europe and Asia [2,3].

"*Infodemiology*" and "*infoveillance*" are two recent terms created to describe a new developing approach for public health, based on Big Data monitoring and data mining, applicable to provide new insights into unmet needs, such as the epidemiology of uncommon diseases.

Infodemiology data can be collected and analysed in near real time. Examples for infodemiology applications include: the disease outbreaks prediction analysis using Internet search volumes (e.g. Influenza) [4];

monitoring peoples' status updates on social networks for syndromic surveillance; analysing disparities in health information availability and accessibility; identifying and monitoring of public health relevant publications on the Internet; seasonability of autoimmune diseases [5]; automated tools to measure information diffusion, health literacy; tracking the effectiveness of health marketing campaigns. Moreover, analysing how people (including patients, researchers, physicians) search and navigate the Internet for health-related information, as well as how they communicate and share this information, can provide valuable insights into health-related behaviour of populations [1].

In this study we aimed to investigate with this new approach possible trends of Internet search volumes linked to SLE, on-going clinical trials and research developments associated to the disease, by analysing relative search volumes generated by Google trends (GT) and scientific search tools.

2. Materials and methods

We longitudinally analysed the large amount of data generated by GT, based on search terms 'SLE', and 'Lupus' in a 5-year web-based research (from January 2011 to January 2016). We then compared the relative search volumes generated by GT with data generated with the same queries from scientific search tools (SCOPUS, Medline/Pubmed) and with a registry and result database of publicly and privately supported clinical studies of human participants conducted globally (ClinicalTrials.gov), considering 'SLE' and 'clinical trials'. For those countries where English is not the first language, additional analyses were conducted by including search terms in the local language. In order to analyse peak-related media information, we longitudinally analysed the Wikipedia page views using WikiTrends and compared the results with the search volumes generated by Google Trends.

3. Results

The geo-infodemiology for clinical trials distribution and search volumes for SLE-related terms is shown in Fig.1.

We observed an overall higher distribution of search volumes from GT in United States, South America, Canada, South Africa, Australia and Europe (mainly Italy, United Kingdom, Spain, France, Germany), showing a geographically heterogeneity in insight into health-related behaviour of the different populations towards SLE.

Data from Medline/Pubmed, SCOPUS and ClinicalTrials.gov were also analysed in order to monitor public health relevant publications and on-going trials. When comparing these results to the distribution of search volumes of GT, we observed an overall similar distribution of Big Data for United States and Europe, while South America, Canada, Australia and South Africa were less represented. A misbalanced between search volumes for GT compared to Medline/Pubmed, SCOPUS and ClinicalTrials.gov might highlight the growing interest on SLE in some areas whereas the number of clinical trials and research in the field is lacking. On the other hand, we observed a higher distribution of search volumes from Medline/Pubmed, SCOPUS and ClinicalTrials.gov in China, while the search volumes from GT were less representative. This might be explained by the limited access to Google search tool in some areas of the world. In Fig. 2 a rappresentation of the global internet use in 2015 is shown. However, one should consider that this approach, when applied to search engines like GT, presents intrinsic limitations, such as age, disability, income or preferred search engine:these limitations might entail a non-representative sampling bias.

United states is the country with the highest number of research publications and on-going trials in the field of SLE, and the highest search volumes in GT.

3.1. Other applications of infodemiology to SLE

3.1.1. SLE and belimumab

This new approach can also be applied to gain further geo-infodemiological information, such as "knowledge translation" or "innovation diffusion metrics", that can be applied to investigate population behaviour towards, for instance, a new drug. By analysing with GT co-occurrences of the key world "SLE" and "lupus" with the approval of belimumab by the FDA and EMA (http://www.fda.gov; http://www.ema.europa.eu) in March 2011 for the treatment of SLE, we observed a rising trend of several peak-related search volumes (Fig. 3).

This approval was a cornerstone for the treatment of SLE: Belimumab was the first drug licenced for lupus in over 50 years and peak-related search volumes confirm the global interest of such innovation. By comparing the search volumes analysing the Wikipedia page views of both SLE and belimumab, we noticed a closer peak trend, reflecting the knowledge translation after the approval of belimumab for the treatment of SLE, indicating a strong acknowledgement by the general population and a strong health-related population information seeking towards the biological drug (Fig. 3).

3.1.2. SLE and the media

Information prevalence is deeply influenced by media. By conducting a longitudinal worldwide search of the key worlds "SLE" and "lupus" with GT, we observed several peak-related news headlines (Fig. 3). The highest peaks of Google Trends search volumes corresponded to news highlights that involved celebrities affected by SLE, also when comparing to the peak generated by the approval of belimumab. These trends were confirmed when searching with GT for the rising related topics to the search term "Systemic Lupus Erythematosus" in a 5 year worldwide search (Fig. 4). These results are in line with other studies [6] and feature the influence of media on the health-related information seeking.

4. Discussion

The interest (assessed both in terms of generated search volumes from biomedical search engines and from public reliable tools) in rare conditions such as SLE is not equally distributed globally.

However, when comparing the observations of Google Trends search volumes with public health relevant publications and on-going trials support the idea that there is a vivid and growing research interest in the field of SLE. In fact, when stratifying for article type data from Medline/PubMed and SCOPUS we observed that original articles represented about 63.8% of publications in the field of SLE.

One could not ignore that, geopolitically differences represent a limitation when attempting to estimate the prevalence of SLE, particularly in some areas like Africa where the epidemiology of SLE is largely undetermined. Similarly, the difficulties in designing worldwide investigations in SLE might be influenced by the imbalanced interest and/or information availability in this condition.

New approaches able to investigate health information seeking might however give an estimate of the health-related demand and even of the health-related behaviour of populations, bringing new light to unanswered questions.

Scientific networking collaborative efforts will help to further investigate SLE prevalence and incidence along with expanding clinical trials research. In a near future infodemiology might be used to assess, not only epidemiological information, but also to obtain health-related behaviour of populations towards newly investigated drugs, giving a unique perspective of the geo-epidemiological innovation diffusion.