



Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

## Social Media and Health Crisis Communication During Epidemics

K. Denecke<sup>1</sup> and S. Atique<sup>2</sup>

<sup>1</sup>Bern University of Applied Sciences, Bern, Switzerland <sup>2</sup>Taipei Medical University, Taipei, Taiwan

By 2020, there will be 6.1 billion smartphone users, so it is time to get serious about digital epidemiology. Researchers have already started to develop methods and strategies for using digital epidemiology to support infectious disease monitoring and surveillance or understand attitudes and concerns about infectious diseases. But much more needs to be done to integrate digital epidemiology with existing practices and to address ethical concerns about privacy.

### 4.1 DIGITAL EPIDEMIOLOGY AND DISEASE SURVEILLANCE

A variety of factors such as population movements, behavioral changes, or food production are responsible for the continuous emergence of infectious hazards. Diseases such as *severe acute respiratory syndrome* (SARS), avian influenza, or bioterrorism caused by the deliberate release of biological agents, all represent new challenges for outbreak alert and response worldwide. Only early detection of disease activity, followed by a rapid response, can reduce the impact of epidemics and prevent harm caused by disease outbreaks [1]. The World Health Organisation defines a disease outbreak as the “occurrence of cases of disease in excess of what would normally be expected in a defined community, geographical area or season.” An outbreak may occur in a restricted geographical area or may extend over several countries. It may last for a few days or weeks, or for several years [2]. A single case of a communicable disease long absent from a population, or caused by an agent (e.g., bacterium or virus) not previously recognized in that community or area, as well as the emergence of a previously unknown disease, may also constitute an outbreak and should be reported and investigated immediately after its occurrence.

Surveillance systems support the management and early detection of disease activity [1]. Traditional surveillance systems that rely on reported diagnoses from laboratories, doctors, or hospitals are well established in all EU countries. While traditional systems can recognize trends over a long time period, and ensure a public health response to identified risks, new emerging threats such as SARS, human cases of avian influenza, might remained unrecognized. Furthermore, despite the development of new approaches for the detection of previously unknown threats (e.g., monitoring of syndromes, death rates, drug prescriptions), these are still insufficient, because signals leading to a public health alert can originate from other sources.

Today, electronic media and discussion groups are increasingly recognized as valuable sources of public health alerts. Awareness of diseases achieved through first-hand observations and “word of mouth” can influence people's behavior and reduce the risk of an outbreak and the number of infected people [3]. Therefore, gathering information from the Web now represents one important part of *Epidemic Intelligence* [1]. Epidemic intelligence combines all efforts for systematic health event detection by providing a conceptual framework into which countries may complete their public health surveillance system. The objective of Epidemic Intelligence is to complement traditional surveillance systems by going beyond traditional public health surveillance and incorporating new official and unofficial sources of structured and unstructured information [1].

The more general concept of *digital epidemiology* [4] comprises the idea that the health of a population can be assessed through digital traces, in real time. Consider the following example: many people suffer from flu every year and many of them search for relevant information in the internet, and share their health problems with others online. In this way, a description of their symptoms, time-stamped and even geo-tagged, is available through search logs, social networks, or other social media tools. Therefore, the internet provides a rather detailed picture of the health of the population, coming from digital sources, through all of our connected devices, including smartphones.

Once an outbreak has occurred, it is crucial for health experts and volunteers to have efficient means for health risk and crisis communication and assessment. Crisis communication is an ongoing process associated with the exchanging information of opinions on a crisis and

the coordination of resources including equipment, personnel, and information to avoid or reduce harm and for coordinating resources during a crisis [5,6]. It also includes the strategy to make people's behavior more rational that they could make informed decisions. The Organisation for Economic Cooperation and Development (OECD) is claiming in a report that “Social media are revolutionizing communication” [5]. They report three ways to use social media in crisis management: (1) as situation awareness tool, (2) as state communication tool, or (3) as a platform for dynamic interaction. Natural disasters such as the 2010 Haiti earthquake or flood in Thailand revealed already the utility of internet-based social media for risk and crisis communication [7]. In these contexts, social media represented an opportunity to broaden warnings to large population groups. OECD acknowledges a “great potential to support two-way crisis communication at low cost and with high efficacy” [5].

In the following sections, we will describe social media data sources and their content with respect to a use in digital epidemiology and health crisis management. We further provide an overview on approaches to digital epidemiology and analyze their strengths and weaknesses in a SWOT (strength, weakness, opportunities, and threats) analysis. Concrete experiences from one project will be summarized. Given the progress in technology, it is often easy to implement new tools for digital epidemiology. However, ethical and legal aspects must be considered carefully. Some of these aspects will be presented at the end of this chapter.

## **4.2 SOCIAL MEDIA DATA AND EPIDEMICS**

What can we find in social media and in the internet with respect to disease activity? Who is reporting on what and in which manner? How can social media support crisis communication and risk assessment? This section tries to answer these questions. As described before, digital epidemiology relies upon sensors, which are humans, that are recognizing and reporting disease activity on social media or leave other digital traces in the web.

Social media are internet-based applications that enable people to share their own information via the internet. This form of communication is more common than ever before and has gained unprecedented popularity around the world through social networking websites like Facebook or microblogging websites like Twitter. The trend is also

recognizable in the healthcare field, where people are accessing websites for medical advice, joining patient communities, and are posting information about their own health status [8].

Social media data includes various kinds of publicly available content that is produced by end-users, rather than by the operator of a website. Medical social media data is a subset of the social media data space, in which the interests of the participants are specifically devoted to medicine and health issues [8]. More specifically, with medical social media data we refer to web-based narrative text and data that contain medical content which was written by individuals (potential patients), physicians, or other healthcare professionals.

#### 4.2.1 Data Sources

In general, the content in medical social media is characterized by a mixture of expert knowledge, layman knowledge or experiences and empirical findings. We can distinguish different social media tools where this content is distributed.

*Social networking sites* with health-related content enable people with similar interests to connect. More specifically, patients who suffer from diseases can share health data in order to empathize with each other or learn about treatments, physical exercises, or medications other patients are consuming in order to improve their health status. For example, PatientsLikeMe (<https://www.patientslikeme.com/>, last accessed 17.11.2015) is a social network for patients that allows them to share health-related experiences and compare treatments. The community currently comprises more than 350,000 patient members (November, 2015). Over 2500 conditions are reported in the platform. Data is collected in a rather structured manner: for the various features such as quality of life or single symptoms, categories are predefined (e.g., quality of physical life on a scale of 4 between best and worst, see Fig. 4.1). Access to health social networks is often restricted to members, i.e., only registered members can connect to others and read through their content. For applications in digital epidemiology, posted messages would need to be collected and analyzed automatically. Given the restricted access, this is difficult if not impossible, also due to legal issues (see Section 4.6). However, social networks offer the opportunity to be used in crisis communication, either for coordinating emergency services and volunteers, or to share information inside a community.

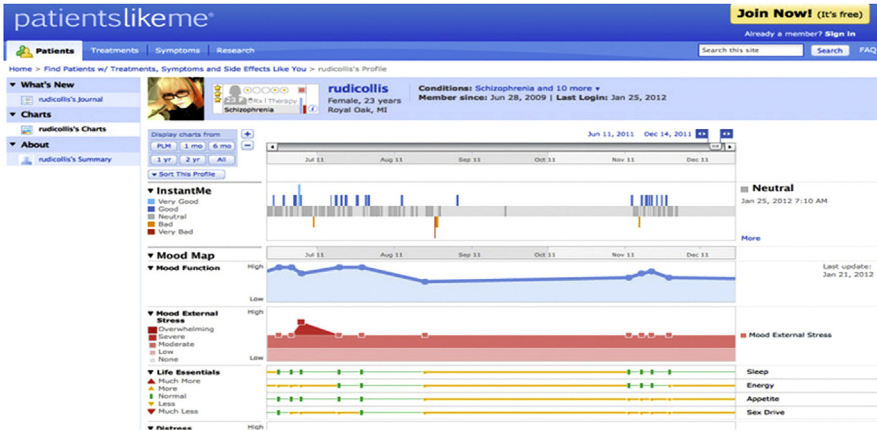


Figure 4.1 Screenshot of PatientsLikeMe. Data is directly plotted to graphs showing changes in a person's health status over time. Source: <http://trendreport.betterplace-lab.org/case/patientslikeme>.

*Content sharing media* allow anyone to upload content such as videos or pictures to be shared with everyone or with a restrictive community of users. *Collaborating knowledge sharing social media* such as forums enable users to ask questions and wait for answers coming from different users. In crisis situations such tools can be exploited for information exchange including images and videos.

*Weblogs or blogs* are similar to paper-based diaries that are normally kept by individuals and shared with others. Similar to a paper-based diary, the authors describe their personal opinions, impressions, and feelings. Online reviews of medical products are an additional source of information regarding the efficacy and adverse effects of drugs and medical devices. A microblog is a blogging platform where the amount of information that can be shared per author is very short. The most common examples of microblogs are Twitter and Tumblr. Twitter's limit is set such that a standard text message, which is limited at 160 characters, can include one entire tweet plus address information. Besides individuals, organizations are tweeting. For example, the U.S. Center of Disease Prevention and Control (<https://twitter.com/cdcflu>) and the organization Médecins sans Frontiers (<https://twitter.com/MSF>) are tweeting updates on disease activity (Table 4.1). Systems that collect information on disease activity have channels where detected activities are posted [e.g., HealthMap (<https://twitter.com/healthmap>), ProMED-mail ([https://twitter.com/ProMED\\_mail](https://twitter.com/ProMED_mail))]. Furthermore, vaccination or disease prevention campaigns are supported by information distribution through

**Table 4.1 Example of Tweets on Swine Flu, Posted on March 15, 2010**

[WEBMD] H1N1 Swine Flu Still Smoldering in U.S.—It’s no wildfire, but H1N1 swine flu continues to smolder in the... <a href="http://ow.ly/16PbAi">http://ow.ly/16PbAi</a>
[FluGov] March 12—WHO Updates International H1N1 Flu Situation. <a href="http://ow.ly/16OwoZ">http://ow.ly/16OwoZ</a>
[CDC] UPDATE: CDC Estimates of 2009 H1N1 Influenza Cases, Hospitalizations and Deaths in the United States, April 2... <a href="http://ow.ly/16Owp0">http://ow.ly/16Owp0</a>

**Table 4.2 Types of Social Media and Their Use for Health Risk and Crisis Communication and Digital Epidemiology**

Type of Social Media	Example	Use for Risk and Crisis Communication and for Digital Epidemiology
Social network	PatientsLikeMe, Facebook	Coordination among emergency services and volunteers, share information inside a community, swift update on emergency situation
Content sharing	YouTube	Situational awareness in real time through exchange of pictures and videos, launch vaccination or disease prevention campaigns
	FlickrR	
Collaborating knowledge sharing social media	Wikis	Situational awareness, dialog between victims and emergency services
	Forums	
	Message boards	
Blogging/microblogging	Blogger	Convey recommendations, warnings, share facts and rumors, source for mapping emergency information
	Twitter	
	WordPress	

twitter or other blogging platforms. Twitter has been proven a frequently updated data source and many technologies analyze twitter messages for the purpose of detecting public health threats [5].

In summary, these different social media are a source of patient-collected clinical values (e.g., blood pressure, pulse, weight), individual judgements on symptoms or efficacy of drugs and treatments, and feelings and sentiments reflecting the health status. Table 4.2 summarizes potential use cases for the four social media types in digital epidemiology and crisis communication.

**4.2.2 An Example of Twitter as Data Source**

Twitter messages have a common format: [username] [text] [date time client]. The linguistic variety goes from complete sentences to listing of keywords. Hashtags, i.e., terms that are combined with a hash (e.g., #flu) denote specific topics and are primarily utilized by experienced users.

<b>Content</b>	<b>Description</b>	<b>Example Tweets</b>
1. Resources	Resource tweets contain news, updates, or information about diseases or outbreaks. The title of the linked article might be mentioned	#schweinegrippe Neue Schweinegrippefälle in Europa: Kulmbacher Gesundheitsamt warnt vor P... <a href="http://tinyurl.com/36o4nqh">http://tinyurl.com/36o4nqh</a> #influenza #h1n1 (Translation: #swineflu New cases of swine flu in Europe: Health ministry of Kulmbach warns...)
2. Personal opinion with linking	Twitter users post their opinion on a disease, virus, symptom	Schweinegrippe : D <a href="http://yfrog.com/h3zuchtj">http://yfrog.com/h3zuchtj</a> (Translation: swineflu :D)
3. Personal opinion and information	Twitter users post only their personal feelings or health status	Hallo Freitag—Hallo Erkältung (Translation: Hello Friday—hello cold)
4. Marketing	Tweets contain an advertisement for a disease-related product or service	It's National Influenza Vaccination Week. Get vaccinated to fight flu!
5. Spam	Tweets are unrelated to diseases or symptoms, but contain mentions of diseases or symptoms	“Das soll gegen Erkältung helfen!” <a href="http://twitpic.com/3f19q0">http://twitpic.com/3f19q0</a> (via @haraldmeyer) (Translation: This is expected to help against a cold)

Referring to the study from Chew and Eysenbach [9] we categorize tweets according to their contents (Table 4.3). In more detail, Twitter messages can:

- provide information,
- express opinions,
- report personal issues.

Information or resources can be provided by authorities, individuals, news agencies, or health professions (Table 4.4). If information is provided, the authority of that information can normally not be determined, so it might be unverified information. Opinions are often expressed with humor or sarcasm and may be highly contradictory in the emotions that are expressed. Consider for example the tweet: “I feel so sick. I have Bieber fever:-).” On the one hand, it reports about the sickness which is rather negative. On the other hand, there is the smiley, which denotes that there is no serious illness, but only “Bieber fever,” which is not really a disease, but rather it is related to the pop star Justin Bieber.

Tweets that contain mentions of symptoms or diseases can be further distinguished based on their content in informing about the



**Table 4.4 Author Groups and Their Content Distributions in Twitter**

Author Group	Provided Content
Individual, nonhealth professional	Personal symptoms
	Observation of symptoms in others
	Reposting information from others including news agencies or health institutions
Individual, health professional	Observation of symptoms in others
	Reposting information from others including news agencies or health institutions
News agencies	Mainly official information
Health institutions	Official information, warnings

health status of the (1) author of the tweet, (2) a friend of the author, or (3) a prominent person. Rarely, they are reporting about health status of animals. Further, personal tweets and resources are reporting about general health information or health education, official information or advices from travel medicine. Characteristically, tweeters are using short sentences (e.g., *I have fever*) or just keywords (e.g., *fever, cough, headache*). Abbreviations are widespread and sometimes difficult to understand due to a lack of context.

When looking into the content of tweets, we can recognize that the various user groups provide different types of content through the twitter channel (Table 4.4).

Challenges for automatic processing of tweets are related to the unstructured nature of the data (free text) and layman language which hamper the connection to clinical terminology. Another issue is the volume of data that is available as well as its reliability. Associated with the reliability of data is the difficulty in interpretation and evaluation of the data. The quality of the data provided through social media tools is unknown. It can be comprehensive and helpful, and also misleading or wrong. Different terminologies and semantics complicate an automatic analysis and interpretation. Subjective information needs to be interpreted, weighted, and linked to objective clinical parameters.

### 4.3 TECHNOLOGIES

Practice showed that a substantial amount of initial outbreak reports come from unofficial informal sources distributed through social

**Table 4.5 Summary of Internet-Based Outbreak Detection Systems**

System Name	Data Sources	Technology	Description
GPHIN	Online news, websites	Taxonomy, translation, categorization	Disease outbreak warning tool
BioCaster	RSS feeds, Google news, WHO, ProMED-mail, European Media Monitor	BioCaster Ontology	Online media data monitoring system
HealthMap	Baidu News, SOSO Info, Google News, Moreover, GeoSentinel, WHO, ProMED-mail	Google Maps, Google Translate, Fisher–Robinson Bayesian filtering	Global disease alert map
EpiSpider	ProMED-Mail, medical websites	OpenCalais, UMLS	Global disease alert map
MedISys	RSS feeds from news, blogs, and official sources	Lexicon, keyword list	Monitoring tool for infectious diseases and chemical, biological, radiological, and nuclear hazards, statistics
Google Trends	Google search term statistics	Word statistics	Trend monitoring system can be used for flu and also other medical conditions

media platforms. These messages need verification. Thus, there is a need to support health officials and epidemiologists to identify relevant information related to disease outbreaks.

### 4.3.1 Social Media for Surveillance and Detection of Outbreaks

Surveillance and outbreak detection tools use different sources of web data that is checked for disease names, mentions of symptoms or other features enabling an identification of relevant postings or web content [10]. Some systems rely upon keyword lists, others on ontologies. Most of them are processing content in different languages, focusing on global disease surveillance. Interestingly, the systems are based on different knowledge resources ranging from keyword lists to taxonomies and ontologies. Even the Unified Medical Language System [11] is exploited as knowledge source by one system. Some examples of such systems are described in Table 4.5.

#### 4.3.1.1 Global Public Health Intelligence Network

The Global Public Health Intelligence Network (GPHIN [12]) is an electronic public health early warning system developed by Canada's Public Health Agency which is part of the World Health Organization's (WHO) Global Outbreak and Alert Response Network (GOARN).

This system monitors internet media, such as news and websites, in nine languages in order to help detect and report potential disease outbreaks or other health threats globally. From 2002 to 2003, this surveillance system was able to detect the Severe Acute Respiratory Syndrome (SARS) outbreak.

More specifically, GPHIN is a secure, web-based restricted access system for outbreak alert that deals with news information about public health events. In contrast to traditional surveillance systems that rely on subscriber input, GPHIN gathers information on disease outbreaks and other public health events by monitoring global media sources on a 24/7 basis. GPHIN's two main sources of outbreak information are the global news services Factiva and Al Bawaba in Arabic language. These services operate as news aggregators that provide multiple sources of information through a single access point. Factiva, for example, aggregates news information from nearly 9000 sources in 22 languages.

GPHIN works in five main steps: From collected data duplicates are eliminated; texts are translated and metadata is inserted using a taxonomy (e.g., mentions of "SARS" or "H1N1" are recognized as "human diseases"). Then, the data are categorized and a relevance score is determined. All data considered relevant is published and available for a manual analysis triage.

#### **4.3.1.2 BioCaster**

BioCaster [13] is a project aimed at providing advanced search and analysis of internet news and research literature for public health workers, clinicians, and researchers interested in communicable diseases. The system monitors many hundreds of internet newsfeeds simultaneously to detect and track infectious disease outbreaks (Fig. 4.2).

More specifically, the system continuously analyzes documents reported from over 1700 RSS feeds including Google News, WHO, ProMED-mail, and the European Media Monitor. The extracted portions of text are classified for topical relevance and plot onto a Google map using geoinformation. The system works in four main steps: topic classification, named entity recognition (NER), disease/location detection, and event recognition.

In more detail, the BioCaster system is equipped with text mining technology which continuously scans hundreds of RSS newsfeeds. The



Figure 4.2 BioCaster user interface.

text mining system has a detailed knowledge about the important concepts such as diseases, pathogens, symptoms, people, places, and drugs. This allows to semantically index relevant parts of news articles, enabling precise access to information. The knowledge underlying the text mining algorithm comes from annotated text collections, gazetteer lists of nomenclature, and the BioCaster ontology. The BioCaster system is no longer accessible online.

#### 4.3.1.3 HealthMap

HealthMap [14], available at <http://www.healthmap.org/en/>, is a platform developed by a team of researchers, epidemiologists, and software developers at Boston Children's Hospital founded in 2006. The system exploits online informal sources for disease outbreak monitoring and real-time surveillance of emerging public health dangers.

Similar to the systems described before, HealthMap collects data from different data sources, including online news, eyewitness reports, expert-curated discussions, and validated official reports. Via an automatic system, which is being updated 24 hours per day, this system monitors, organizes, integrates, filters, visualizes, and disseminates online information about emerging diseases in nine distinct languages, facilitating early detection of public health threats. Collected data is processed by means of automated filtering, and visualization of reports

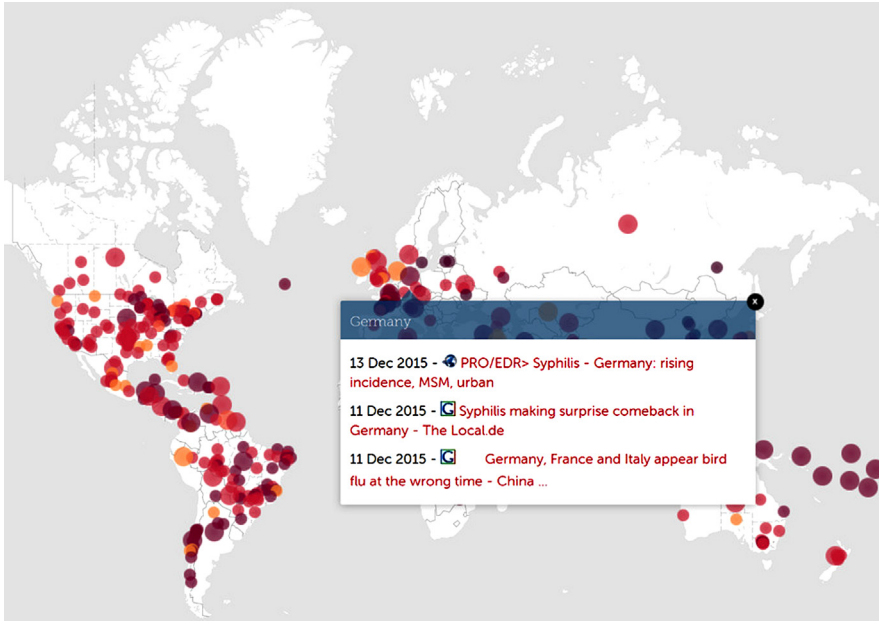


Figure 4.3 Health Map alert map. Source: <http://www.healthmap.org/en/>.

through the utilization of automated text processing algorithms that classify alerts by location and disease. A mean against information overload, the articles are further categorized for improved filtering. The additional categories include breaking news (e.g., of a newly discovered outbreak), warning, follow-up, background/context, and not disease related (Fig. 4.3).

#### 4.3.1.4 EpiSpider

EpiSpider [15] was initially developed to serve as a visualization supplement to the ProMED-mail reports (<http://www.promedmail.org/>), i.e., ProMED-mail reports were analyzed with respect to topic intensity and displayed on a map. In addition to ProMED data, EpiSpider collects information from Google, Humanitarian News, Twitter, WHO, and Daylife (cloud-based media service) and processes the data with natural language processing used to transform free text into structured information. EpiSPIDER began outsourcing some of its preprocessing and natural language processing tasks to external service providers such as OpenCalais ([www.opencalais.com](http://www.opencalais.com)) and the Unified Medical Language System (UMLS) web service for concept annotation. This action has enabled the screening of noncurated news sources



Figure 4.4 MedISys user interface. Source: <http://medusa.jrc.it/>.

as well. However, it scans articles only in English. EpiSPIDER has a timeline visualization to help users to order events in time, and a word cloud that helps users to get a sense of what topics are making headlines. Location names in reports are recognized and georeference using the georeferencing services of Yahoo Maps, Google Maps, and Geonames.

### 4.3.1.5 MedISys

MedISys [16] is an internet monitoring and analysis system that identifies potential threats to the public health (Fig. 4.4). Collected articles are grouped by disease or disease type; location names are identified. The system is based on a list of sources, including official channels, blogs, and online news. It analyzes the texts using keyword lists and identifies topics that are focused by many reports at the same time span. MedISys (<http://medusa.jrc.it/>) covers global health issues including multiple diseases and multiple locations.

### 4.3.1.6 Google Trends

Google trends (<https://www.google.com/trends/>) are another emerging tool for the detection of outbreaks. It uses search query data, i.e., frequency of search terms, and plots them over time, allowing for the

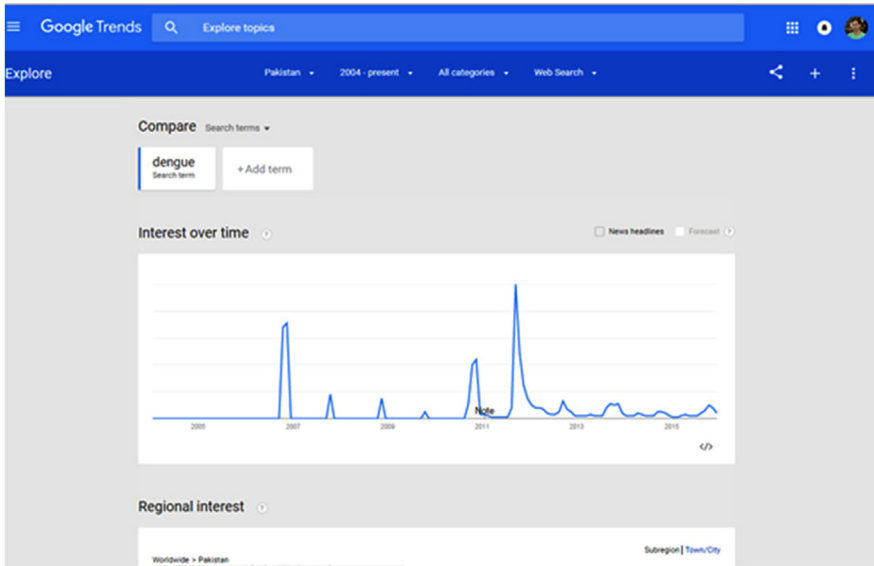


Figure 4.5 Google trends showing dengue searches from 2004 to present in Pakistan. Source: <https://www.google.com/trends/>.

recognition of potential disease activity when considering a disease-related term. As a result, a time series of the searches in a period of time is provided giving an idea how the search trend changed during defined time period (Fig. 4.5). This in turn gives an idea of possible disease activity which might have occurred. Indeed, studies have shown a correlation between search behavior and flu activity [17].

### 4.3.2 Social Media for Health Crisis Communication

Emergency management and crisis communications have become more participatory. Through Social Media channels information on disease activity or natural disasters can be distributed quickly to large groups of individuals. The OECD claims: “Social Media can enhance risk and crisis communication in several ways: (1) they are collaborative and participatory, and thus can improve situation awareness, (2) they are decentralized, thus, information can circulate quickly, and (3) they are geographically traceable and thus allow for the monitoring of a crisis” [5].

No specific system is required for social media-based health crisis communication—the use of the existing social media tools such as



social networks (e.g., Facebook) or microblogging systems is sufficient for crisis communication.

In order to exploit social media for situational awareness, the health officials need to locate social media content that contain crisis-relevant information during mass emergency situations. The systems described in the previous sections are able to support in this task. Another possibility is to search Twitter or other potential social media sources directly which can be realized by conventional manually edited keywords, location-based searches, or relying upon lexicons [18].

Regarding circulation of information, there are four main possible applications: distribution of information to the public, information exchange among staff and volunteers, and acquisition of volunteers. Analysis of twitter data sets collected related to crisis situations revealed that there are several categories of content of messages. Informative postings can contribute contextual information to better understand the situation. They include status messages of users, explanations of particular problems, and precise data of the crisis, e.g., number of victims. The content can be predicting or forecasting, instrumental or conformational [19].

CDC officials use various social media channels to inform the public and provide health and safety information. The CDC Emergency feed (<https://twitter.com/CDCemergency>) is the official feed of CDC's Office of Public Health Preparedness and Response. It provides latest information on emergencies, preparedness tips, and real-time updates and health alerts for the public during an emergency. Additionally, email updates are provided with information on recent outbreaks and incidents, radiation emergencies, or public health matters. More specifically, the E-Mail alerts are generated when new information on the corresponding CDC website is available.

During the Ebola outbreak in 2014, social media supported the communication between healthcare providers, local and national health authorities, and international health agencies. Furthermore, social media has been proven to be able to replace traditional communication systems during crisis situations [5]. During the 2010 earthquake in Haiti, traditional communication systems were down. People started to use social media quickly as communication channel [20].



Additionally, social media can be used to indicate willingness to help in the event of an emergency and thus may help in mobilizing volunteers [21]. For example, indicating in the “status” of the personal Facebook profile, the availability and skills for both professionals and volunteers could be a way for public authorities to know in real time whom to mobilize in a given area of disaster.

## 4.4 THE M-Eco System and Lessons Learnt

The EU-funded project M-Eco: Medical Ecosystem was conducted between 2010 and 2012 with seven project partners from Austria, Italy, Germany, Czech Republic, and Denmark, including the German health organization Robert Koch Institute and with support of representatives of various health organizations including the World Health Organisation, European Centre of Disease Prevention and control, and Institute de Veille Sanitaire. In this section, we briefly summarize the architecture of the M-Eco system and its functionalities, as well as report on experiences in its evaluation and testing. The M-Eco system could so far not been established into regular use by health organizations due to missing personnel resources from implementation. More details about the technology and studies can be found in papers by Denecke et al. [22] and Velasco et al. [10].

### 4.4.1 Overview

The M-Eco system exploited data from multiple sources for public health monitoring purposes. The system:

- monitored social media, TV, radio, and online news,
- aggregated the contents into signals,
- visualized the signals using geographic maps, time series, and tag clouds,
- allowed searching and filtering signals along various criteria (location, time, medical condition).

The system was intended to support in health monitoring during mass gathering events in a cross-country setting and in health monitoring on a national level. Signals pointed the user to relevant information and their sources which allowed to analyze its relevance and need for interaction through health officials. Automatically generated time series supported the monitoring of disease activity over a longer time period. Tag clouds summarized the related information in a visual manner and supported

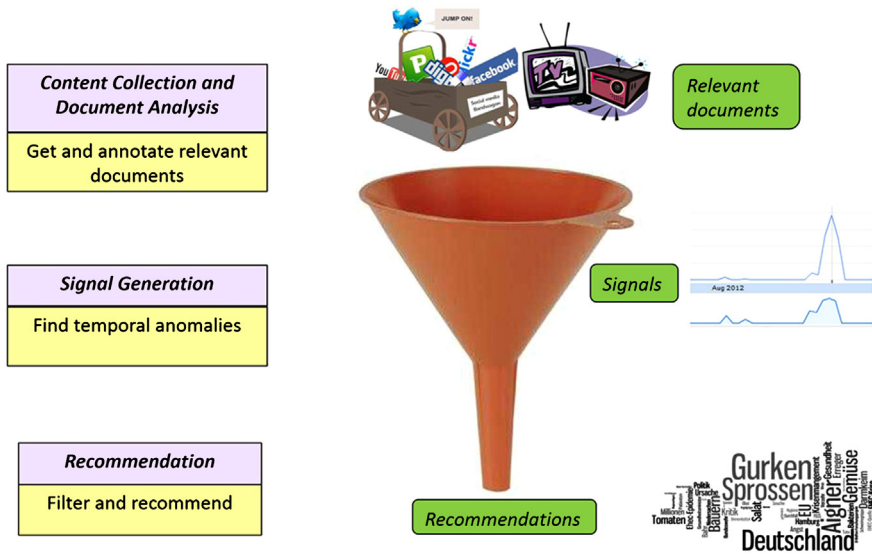


Figure 4.6 The M-Eco approach in a nutshell.

navigation through signals. The plotting of signals to geographic maps allowed to localize disease outbreaks. M-Eco offered (1) additional information through social media monitoring, (2) perception of recommendation and user behavior, and (3) visualization and support for risk assessment.

#### 4.4.2 Architecture

To realize these steps, the M-Eco system consisted of a set of web services that cover four areas depicted in Fig. 4.6. These were (1) content collection, (2) signal generation, (3) user modeling and recommendation, as well as (4) visualization in a user interface. The services worked in a pipeline fashion and were triggered automatically four times a day.

##### Content collection and document analysis component

The information database of the system was filled continuously by the content collector and document analysis component. It collected data from various sources by means of web crawling and streaming APIs (e.g., the Twitter API), and made them accessible to other components. The collection focused on broadcast news from TV and radio, news data from MedISys [16], and social media content from blogs, forums, and Twitter. The TV and radio data was collected via satellite and transcribed to written text by SAILS Media

Mining Indexing System [23]. About 1300 names of symptoms and diseases were used as keywords for collecting data extended by existing language resources such as WordNet, GermaNet, or the OpenOffice thesaurus. The data was tokenized and part-of-speech tagged by the Tree Tagger and parsed by the Stanford Parser. All texts were also semantically annotated with geo tags, disease or symptom tag and temporal expressions as well as with information on the affected organism. In the following, the term “text” is used to refer to some piece of text which can be, e.g., a tweet, a blog posting, or even a transcript of a TV or radio transmission.

### **Signal generation component**

The event detection and signal generation component exploited the annotated texts provided by the content collection and document analysis component to generate signals. A signal was a hint to some anomalous event. The component produced signals with associated information on the disease or symptom the signal is referring to and a location that has been extracted for that signal. In a first step, sentences were classified as relevant or irrelevant by the method presented before. For all relevant sentences, entity pairs (location, disease) were exploited to produce time series for each entity pair occurring in sentences of texts published within 1 week.

The time series provided the input for statistical methods for signal generation, CUSUM and Farrington. These two statistical methods had originally been developed for indicator-based surveillance [24]. Cumulative sum or CUSUM methods focused on several consecutive periods and sum up the aberrations in one particular direction. The Farrington approach fitted a regression model to the data over several years, allowing for a secular trend. Outbreaks in the past were automatically identified and removed, and the statistical distribution fitted either to rare counts or to frequent counts.

The user interface allowed to select the algorithm used to calculate signals. Between 0 and 50 signals were generated by this procedure every night. The exact number depended on several variables or factors that influence the generation of signals such as the type of considered data (e.g., Twitter’s update frequency was much higher than of a blog).

### **Recommendation component**

The recommendation component got as input the generated signals and either selected those that were of interest for a user according to his profile or ranked the signals appropriately. The component

also supported users with personalized presentation options (e.g., tag clouds, list of recommendations) that were visualized in the user interface. In this way, information or alerts were filtered before being presented to a user, which in turn reduced information overload. The recommendation component required a user profile that consisted of information on user behavior from interactions with the system (e.g., ratings, tags, search terms).

The personalization and recommendation of signals mainly relied upon the tagging behavior of a user. Tags were potential indicators of user preference. For recommending items to the user, tags assigned by him to his texts of interest were compared against the tags assigned to candidate and unknown texts. In order to help users navigating through a vast collection of texts and finding new items, a tag cloud component provided a visual representation of texts. Besides indexing texts in the corpus, each tag helped users to find new related information of interest.

#### **User interface and visualization**

The user interface allowed a user to search for disease names or symptoms and to assess the related signal information by means of a geographic map, a tag cloud, or a timeline. The geographic map plotted the signals to a map. It enabled users to select specifically signals related to locations that are of interest to them. The timeline showed the text volume referring to a specific disease or symptom (or the corresponding signal, respectively) over time. This allowed users to learn about the progress of a disease outbreak as reflected in social media and also about seasonal differences. The tag clouds provided a quick overview on the content of the texts associated with a signal. They enabled the user to quickly decide about the relevance of a signal. Access to the original sources that contributed to the signal generation was provided, as well as filtering capabilities (e.g., selecting a time span). Furthermore, user feedback options were included into the user interface. With “Thumbs up—thumbs down” and a rating scale for signals, users could judge the relevancy of the presented signal. This information was fed back to the recommendation process and considered for ranking and filtering.

#### **4.4.3 Lessons Learnt**

The M-Eco system results were analyzed in several studies [10,22]. They revealed characteristics of social media that are relevant for disease surveillance. First, the texts that contributed to signals rated as relevant

by the epidemiologist were often linked to media reports or so-called secondary reports. This suggests that there might be a trend in social media whereby users tend to write less often about their personal specific symptoms, but most often forward information from reliable sources such as news sites or prevention efforts from authorities.

Second, most signals were generated from Twitter data. The volume of relevant Twitter data that was processed by the system was much higher than that from any other source considered as input. Beyond, it is unclear who is providing relevant health information via social media, which age groups, personal background of persons might play a role, geographic coverage, etc. This means that relevant information from segments of the populations are not coming through these channels. Another challenge is the quality of content collected from social media and the difficulty to automatically decide whether it is a real outbreak or not. Many of the social media texts present vague reports of illnesses. It is difficult to judge the seriousness of the reported information.

In contrast to initial expectations, the signals were not generated from clustered reports on personally reported symptoms, but on news reports that were fed into social media, and replicated or forwarded by interested users. Therefore, M-Eco was not the first instance to detect the public health event, because there were local actors who had already detected and reported the event. But, M-Eco brought such reports quickly to a broader attention. It was not possible to present an example where M-Eco was the first to detect an outbreak by a clustering of social media contributions with similar symptoms in space and time, and where the outbreak was afterwards confirmed by the traditional notification system. Another lesson learnt is not to underestimate the legal and ethical issues related to IT solutions for digital epidemiology. We discuss relevant issues in [Section 4.6](#).

## **4.5 SWOT ANALYSIS**

By means of an analysis of the strengths, weaknesses, opportunities, and challenges, we are studying the potential of social media for disease surveillance and crisis communication and prevention.

The objective is to identify future perspectives and open issues to make social media a useful tool in this context. Strengths and weaknesses are mainly internal factors while opportunities and threats

<b>Table 4.6 SWOT Analysis Results</b>		
	<b>Positive</b>	<b>Negative</b>
Internal	Strengths <ul style="list-style-type: none"> <li>• Easy to use</li> <li>• Availability without any installation</li> <li>• Low costs</li> <li>• Fast and timely</li> <li>• Enables direct contact to public and volunteers</li> <li>• Enables a large coverage through a broad distribution</li> <li>• Can be used by professionals and laymen</li> <li>• Reception of outbreaks that would be otherwise recognized too late</li> </ul>	Weaknesses <ul style="list-style-type: none"> <li>• Information overload</li> <li>• Entirely subjective information, unverifiable</li> <li>• Difficult to assess and check validity</li> <li>• Unauthorized information</li> <li>• Unknown source of information; even location can be unknown</li> <li>• Wrong or misleading information can be posted</li> </ul>
	External	Opportunities <ul style="list-style-type: none"> <li>• Enables interactions among health organizations and volunteers</li> <li>• Engagement of the population in prevention</li> <li>• Increased availability of internet in rural areas</li> <li>• Large distribution of smart and mobile phones also in low income countries</li> <li>• Health officials use social media for global disease surveillance and prevention campaigns</li> </ul>

generally relate to external factors. For the analysis, we collected answers to the following questions (Table 4.6):

### **Strength**

- What advantages does social media have for digital epidemiology?

### **Weaknesses**

- What can be improved in current tools and their usage? What should be avoided?

### **Opportunities**

- What interesting trends and opportunities can be identified?

### **Threats**

- What obstacles can be faced when implementing using social media for disease surveillance and crisis communication?
- Do quality standards or ethical and legal aspects threaten the use in practice?

It can be seen that there are many positive aspects supporting the use of social media for disease surveillance and crisis communication. More timely information and networking among the “helping hands” could become the most driving factor. Challenges are related to the increased information overload, including the amount of unauthorized information which is information not officially confirmed by health officials, which might be addressed by technology, e.g., by including sophisticated filtering algorithms to prefilter the information before showing to the user. A challenge is the high risk of manipulation, in particular the risk of analyzing postings containing wrong or misleading information. The usefulness of social media for disease surveillance depends clearly on the willingness of people to share (correct) information online and to use online tools. In particular, we will need in future ethical and legal standards to ensure that people will continuously use these tools for reporting on disease activity. In the next section, unintended consequences, in particular ethical and legal issues will be discussed in more detail.

#### **4.6 UNINTENDED CONSEQUENCES OF SOCIAL MEDIA USAGE IN PUBLIC HEALTH**

Even though useful, social media usage for prevention and detection of epidemics provides some unwanted or unintended consequences with regards to technical, functional, and formal issues.

Formal problems include quality and reliability of content, payment models, as well as ethical and legal issues. The latter are related to the usage of data posted through social media tools for research or epidemiological purposes. In this context, it is important to clarify responsibilities. Imagine a health status monitoring tool that identifies a group of sick persons by analyzing social media conversation. In which manner should a health organization react that becomes aware of this conversation? The current interaction or reaction processes are often not foreseeing a reaction of the national health organizations, but on a local level. This means, processes need to be adapted when considering social media as source of information or for crisis communication. These and similar questions need to be answered before such applications go online. When using social media or online traces for surveillance purposes, the right for individual self-determination—what happens with my data—is weighted in crisis situations against the

wellbeing of the society. The objective of data collection and analysis needs to be specified, i.e., whether the data is analyzed and collected for treatment, care, prevention, or crisis management.

In both cases, disease surveillance and crisis communication, it needs to be ensured that the technology is robust against errors and abuse in order not to overload the health officials with information, but also to prevent population groups from social stigmatizations and prejudices due to false alarms. Corresponding measures (e.g., integrating spam detection methods) should be implemented or personnel needs to be aware of misleading information.

Technical and functional challenges are related to the data volume and an increased risk of generating false alarms. Even though social media data provide a new source of information to hint to public health threats, their analysis and interpretation is challenging. Language is ambiguous and automatic interpretation becomes difficult when symptoms are used in different contexts than expected (e.g., “football fever” could produce an alert since the keyword « fever » is used). Intended as support for epidemiologists and healthcare workers, there is a risk of an additional workload due to large numbers of such false alarms. Comprehensive filtering algorithms need to be established keeping a good balance between sensitivity and specificity of generated alarms. Another option is to use social media tools for active reporting on disease activity by the population. However, even this method is prone to errors when misleading information is posted.

Another issue is the quality and reliability of data as well as localizing the outbreak. As reported by Goff et al. [25], sometimes misinformation regarding infectious diseases is disseminated through Twitter. An additional limitation is that majority of users of social media are younger people from developed countries. This makes social media-based information biased spreading misleading information as reported by Paul et al. [26]. Dyar et al. [27] found out that Twitter leads to activation of certain searches and sharing of information about outbreaks globally rather than locally. This can become misleading as the outbreak can occur in a certain part of the world while information is being shared to other parts.

Maintaining traditional media and reporting in the crisis communication strategies and for disease surveillance are relevant to



ensure inclusion of all segments of the population. Beyond, there are other measures to be taken to make the best use out of social media in digital epidemiology and crisis communication. When developing a concrete social media application in epidemiology and for the detection of epidemics, it is crucial to determine the scope of the system under development, i.e., it needs to be clarified which users are involved, which application area is concerned and on which dimension it is operated. Questions include:

- Who is affected by the analysis and application of medical social media data and how should they be affected by it?
- Who is compelled to act on the new knowledge?
- What action is appropriate based on the information learned as a result of the analysis?
- Who is responsible when a predictive analysis is incorrect?

Answering those questions before implementing a system in practice and even addressing these questions in the development phase will help in producing useful applications, limiting the risks of social media usage for prevention of epidemics. However, there is still a need for guidelines, standard operating procedures, and best practices in digital epidemiology to ensure that harm is prevented.

## REFERENCES

- [1] Paquet C, Coulombier D, Kaier R, Ciotti M. Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. *Euro Surveill* 2006;11(12). Available from: <<http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=665>>.
- [2] WHO: Health topics. Available from: <[http://www.who.int/topics/disease\\_outbreaks/en/](http://www.who.int/topics/disease_outbreaks/en/)> [last accessed 12.11.15].
- [3] Funk S, Gilad E, Watkins C, Jansen VAA. The spread of awareness and its impact on epidemic outbreaks. *Proc Natl Acad Sci* 2009.
- [4] Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital epidemiology. *PLoS Comput Biol* 2012;8(7):e1002616. Available from: <<http://dx.doi.org/10.1371/journal.pcbi.1002616>>.
- [5] OECD. The changing face of strategic crisis management. Paris: OECD Publishing; 2015. Available from: <http://dx.doi.org/10.1787/9789264249127-en>.
- [6] Whitney Holmes. Crisis communications and social media: advantages, disadvantages and best practices. CCI Symposium. Available from: <<http://trace.tennessee.edu/cgi/viewcontent.cgi?article=1003&context=ccisymposium>> [last accessed 15.02.16].
- [7] Smith BG. Socially distributing public relations: Twitter, Haiti, and interactivity in social media. *Public Relations Rev* 2010;36(4):329–35.
- [8] Denecke K. Health web science. Social media data for healthcare. Springer International Publishing, Switzerland; 2015.

- [9] Chew C, Eysenbach G. Pandemics in the age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS One* 2010;5(11):e14118. Available from: <http://dx.doi.org/10.1371/journal.pone.0014118>.
- [10] Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: a systematic review. *Milbank Q* 2014;92(1):7–33.
- [11] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. National Library of Medicine, Bethesda, MD. *Methods Inf Med* 1993;32(4):281–91.
- [12] Mykhalovskiy E, Weir L. The global public health intelligence network and early warning outbreak detection. *Can J Public Health* 2006;97(1).
- [13] Collier N, Doan S, Kawazoe A, Goodwin RM, Conway Mike, Tateno Y, et al. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics* 2008;24(24):2940–1.
- [14] Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Surveillance sans frontieres: internet-based emerging infectious disease intelligence and the healthmap project. *PLoS Med* 2008;5(7):e151.
- [15] Keller M, Blench M, Tolentino H, Freifeld CC, Mandl KD, Mawudeku A, et al. Use of unstructured event-based reports for global infectious disease surveillance. *Emerg Infect Dis* 2009;15(5):689.
- [16] Linge JP, Steinberger R, Fuart F, Bucci S, Belyaeva J, Gemo M, et al. MedISys: medical information system. *Advanced ICTs for disaster management and threat detection: collaborative and distributed frameworks*. IGI Global Press; 2010.
- [17] Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google Flu trends performance in the United States during the 2009 Influenza Virus A (H1N1) pandemic. *PLoS One* 2011;6(8):e23610. Available from: <http://dx.doi.org/10.1371/journal.pone.0023610>.
- [18] Olteanu A, Castillo C, Diaz F, Vieweg S. CrisisLex: a Lexicon for collecting and filtering microblogged communications in crises. *Proceedings of the AAAI conference on weblogs and social media (ICWSM'14)*. Ann Arbor, MI: AAAI Press; 2014.
- [19] Sreenivasan ND, Lee CS, Goh DH-L. Tweet me home: exploring information use on Twitter in crisis situations. *Online Communities and Social Computing: 4th international conference, OCSC'11, Held as Part of HCI International 2011, Orlando, FL, USA, July 9–14, 2011*. Proceedings. Springer Berlin/Heidelberg. p. 120–9.
- [20] Cambié S. Medecins sans frontieres: social media lessons from the Haiti crisis. Available from: <https://www.simply-communicate.com/case-studies/company-profile/medecins-sans-frontieres-social-media-lessons-haiti-crisis> [last accessed 20.12.15].
- [21] Wendling C, Radisch J, Jacobzone S. The use of social media in risk and crisis communication. *OECD Working Papers on Public Governance No. 24*, OECD Publishing; 2013.
- [22] Denecke K, Kriek M, Otrusina L, Smrz P, Dolog P, Nejd W, et al. How to exploit twitter for public health monitoring? *Methods Inf Med* 2013;52(4):326–39.
- [23] Backfried G, Schmidt C, Pfeiffer M, Quirschmayr G, Glanzer M, Rainer K. Open source intelligence for disaster management. *intelligence and security informatics conference (eisc)*; 2012. p. 254–58.
- [24] Hoehle M. Surveillance: an r package for the surveillance of infectious diseases. *Comput Stat* 2007;22(4):571–82.
- [25] Goff DA, Kullar R, Newland JG. Review of twitter for infectious diseases clinicians: useful or a waste of time? *Clin Inf Dis* 2015;60(10):1533–40.
- [26] Paul MJ, Dredze M. You are what you Tweet: analyzing Twitter for public health. In: *ICWSM*; 2011. p. 265–272.
- [27] Dyar OJ, Castro-Sánchez E, Holmes AH. What makes people talk about antibiotics on social media? A retrospective analysis of Twitter use. *J Antimicrob Chemother* 2014; dku165.