NEWSLETTER
HEALTH AND AGEING

THE GENEVA ASSOCIATION

No. 32, April 2015

Social Epidemiology

By Séverine Rion Logean[+]

Introduction

Web-based data mining of chronic or infectious disease provides real-time data on a global level, a fact that revolutionises health surveillance and opens new opportunities for the health and insurance industries, and for public health. Typically, online collected epidemiological data can fill gaps when company internal data is scarce, as for example when business is expanded to new markets or to complement available data sets from other sources.

If social media platforms provide a different picture of global health to that using traditional health surveys, we have identified invaluable data sources for a new generation of health surveillance systems.

Within Swiss Re we have developed a proprietary tool, the Chronic Disease Map (CDM) in order to observe regional real-time trends on different facets of chronic disease experience based on Twitter messages. As well as finding new data sources for smart analytics, Swiss Re has developed new in-house capabilities and expertise as we overcome challenges in producing the CDM Tool.

Background

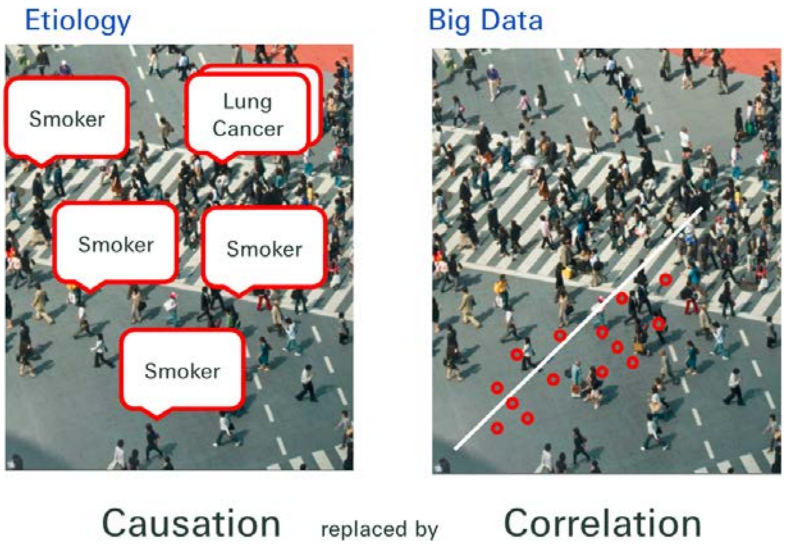
In Life and Health reinsurance the availability of experience data, epidemiological data and mortality tables is crucial when it comes to product development or pricing. Working in Research and Development, we are often confronted with requests for health or disease data for regions where internal data is scarce or public data is not easily accessible. Epidemiological data may only be published several years after its collection, which limits its usefulness in judging recent trends. Data collection and processing by government agencies vary from country to country and may be difficult for external parties to judge, whereas the systematic collection of social media real-time data allows for new insights into global health and disease dynamics.

Digital epidemiology

Epidemiology is concerned with the dynamics of health and disease in humans. It takes into account risk factors and behaviours that influence the health outcomes, and aims to discover causal relationships between risk factors and disease. Traditionally, epidemiological data has been collected by public health agencies through hospitals, doctors or other healthcare institutions; a very time- and resource-consuming task. Thanks to digital communication technologies, data can nowadays be collected directly from individuals leaving online traces on public domains [1]. Data collection is thus much more efficient and covers many more data points. This offers new possibilities for data analysis: while traditional epidemiology seeks to reduce disease through reducing exposure to risk factors, big data analysis promotes correlation over causation. Correlation does not imply causation, but correlations in a large enough database can indicate a predictive relationship, which can be exploited in practice.

[+] Head Life & Health R&D Europe, Swiss Re, Zurich.

Picture 1: Big data analysis promotes correlation over causation [2].

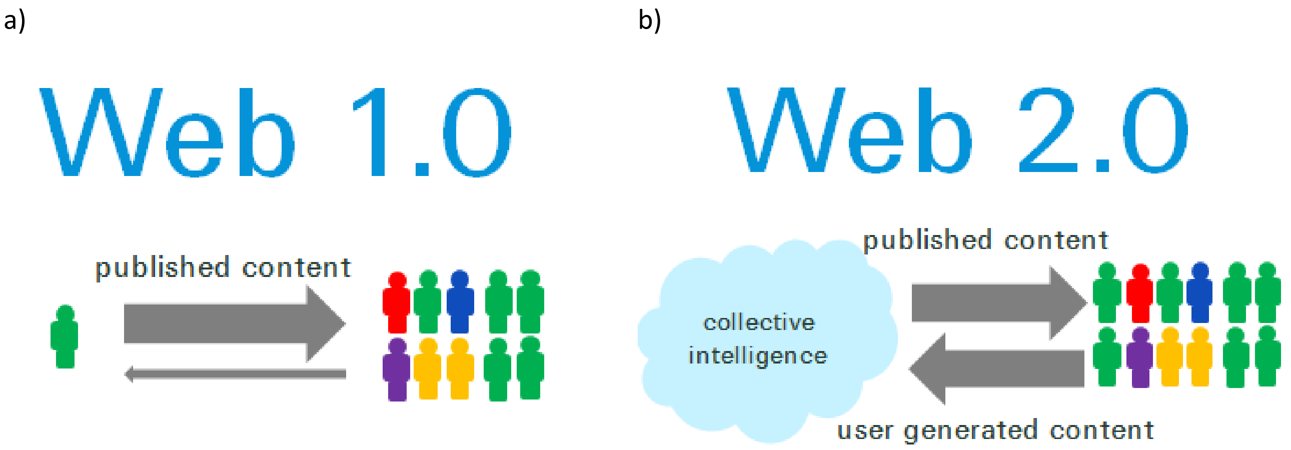


The continuous revolution in the way people communicate and the increased use of electronic devices have increased dramatically the volumes of data being created and shared continuously. We are able to focus on different countries or regions and digital communication channels facilitate the accessibility, storage and analysis of the data [3]. Relevant information on disease epidemiology can be extracted from what we do and say on public platforms and analysed. Extracting meaningful information is challenging but provides huge opportunities [3].

Short history of the new data opportunity

Some of you will remember when Tim Berners-Lee, back in 1986, announced the arrival of the internet, the so-called Web 1.0. It was a revolutionary technology, allowing individual users to search and read information on the World Wide Web. Web 2.0 provides individuals with the opportunity to produce content and share that content through many channels. Examples are sharing text on blogs, videos on YouTube, or pictures on Instagram. Web 2.0 allows bidirectional communication, where every individual can communicate with many others. A collective intelligence is being created, allowing you to find the news and the news to find you (e.g. by following your favourite editors on Twitter).

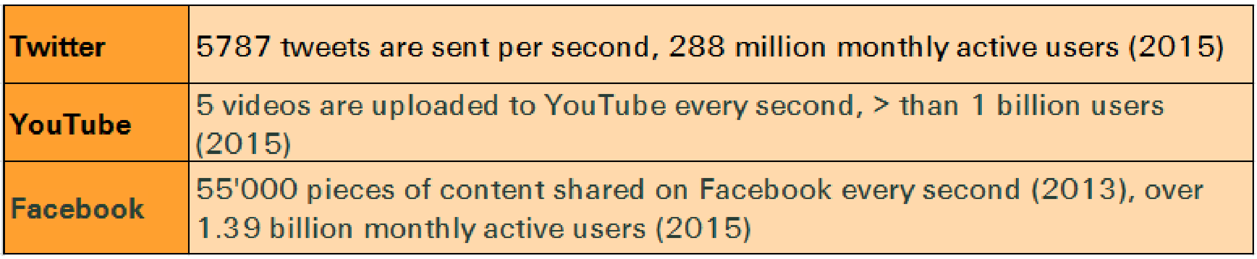
Picture 2: a) The Web 1.0 refers to the World Wide Web, an online unidirectional way of communication. b) The Web 2.0, allowing a bidirectional way of communication, building a cloud of collective intelligence.



Today we are at the brink of the next innovation, Web 3.0 or the internet of things. Not only individuals will be sharing content on the net, but sensors integrated in all kind of devices, houses, cars, etc., will be collecting data and share that data with other devices via the web. Fridges that keep themselves fully stocked without troubling you.

The amount of data that is being uploaded to the cloud is almost beyond comprehension and continuously growing. The numbers below shall give you an insight into how many messages or videos are shared on some of the most popular social media platforms.

Picture 3: Amount of data being shared on selected social media platforms: Twitter, YouTube and Facebook [4].



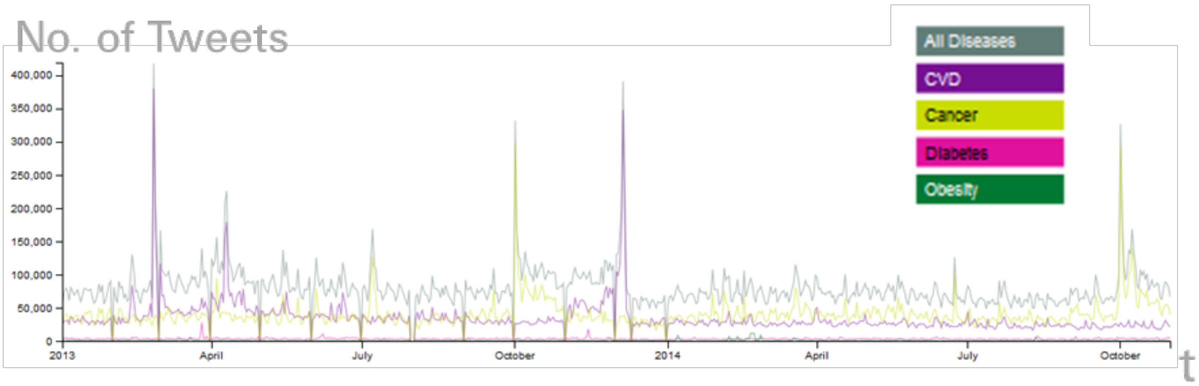
Twitter	5787 tweets are sent per second, 288 million monthly active users (2015)
YouTube	5 videos are uploaded to YouTube every second, > than 1 billion users (2015)
Facebook	55'000 pieces of content shared on Facebook every second (2013), over 1.39 billion monthly active users (2015)

Obviously not all these messages are related to chronic diseases. However, a PricewaterhouseCoopers study on social media and healthcare published in 2012 found that 15–30 per cent of consumers use social media to share health symptoms and comment on their own and others' experiences and behaviours [5]. Further 30–45 per cent of health decisions are affected by information from social media, such as coping with chronic conditions or pain, approach to diet, exercise, stress management, or taking medications [5]. As a recent controversial study from Facebook illustrated, social networks have the ability to influence emotions without direct interaction between people [6]. Further, 'social media medical doctors' that listen to your tweets and spot changes in your health might message you with recommendations, and may motivate some to share health data.

Chronic Disease Map Tool

The Chronic Disease Map Tool accumulates tweets over time, related to cancer, heart disease, diabetes and obesity on a geographical map. Specific search terms are chosen that are either descriptive of the condition, e.g. overweight for obesity or the treatments that might be used.

Picture 4: Chronic Disease Map Tool timeline representing time on the x-axis, and number of tweets for cardiovascular disease, cancer, diabetes and obesity, including all diseases, on the y-axis.

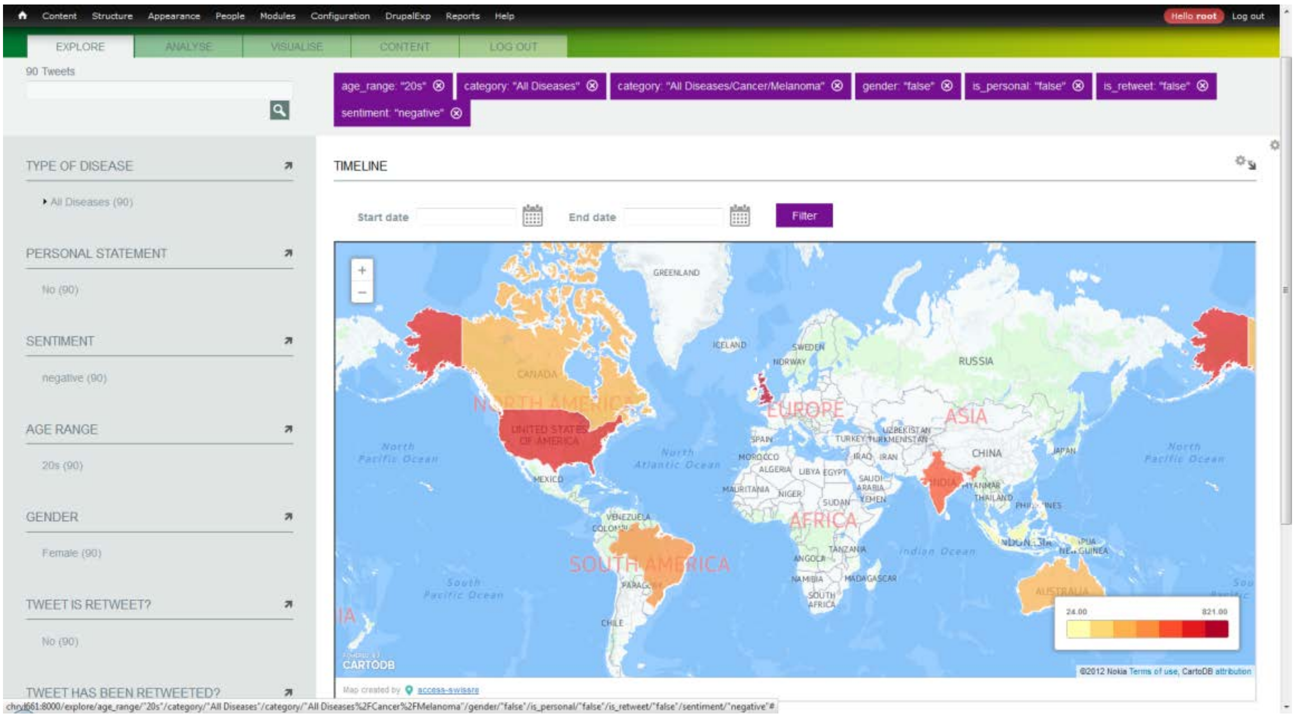


Source: Swiss Re

In order to relate the messages to personal disease, tweets are selected that include a personal attribute like "I have…" or "my"; World Wide Web links are not considered, as they are often related to publicity or news.

Challenges were encountered in selecting tweeting populations that were representative of re-/insurance portfolios. Algorithms have been developed for Twitter population characterisation. Based on these, the Twitter population is automatically classified according to age and gender. The geolocation of the tweets is provided by GNIP, one of the official Twitter data providers.

Picture 5: Chronic Disease Map Tool, overview.



90 Tweets

age_range: "20s" ⊗ category: "All Diseases" ⊗ category: "All Diseases/Cancer/Melanoma" ⊗ gender: "false" ⊗ is_personal: "false" ⊗ is_retweet: "false" ⊗ sentiment: "negative" ⊗

TYPE OF DISEASE
- All Diseases (90)

PERSONAL STATEMENT
No (90)

SENTIMENT
negative (90)

AGE RANGE
20s (90)

GENDER
Female (90)

TWEET IS RETWEET?
No (90)

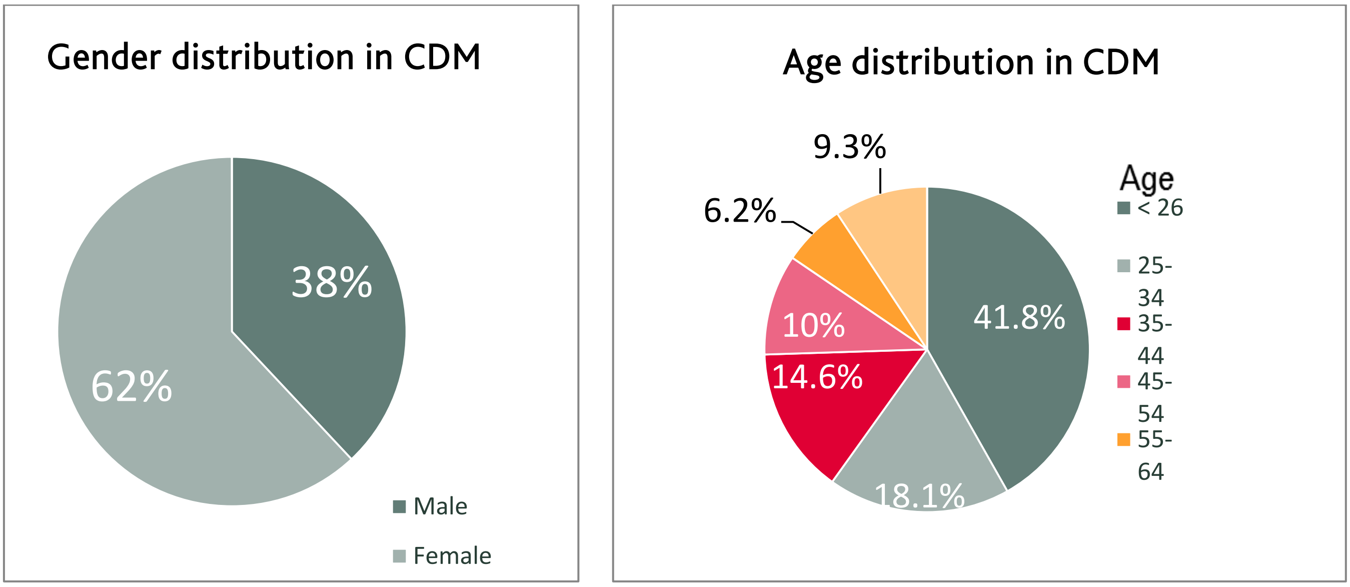
TWEET HAS BEEN RETWEETED?

chryl661:8000/explore/age_range/"20s"/category/"All Diseases"/category/"All Diseases%2FCancer%2FMelanoma"/gender/"false"/is_personal/"false"/is_retweet/"false"/sentiment/"negative"#

Source: Swiss Re

So far, the CDM data shows a gender split of 62 per cent female (f) and 38 per cent male (m) tweeters. A publication by Beevolve from 2012 suggested a split of 53 per cent (f) versus 47 per cent (m) across Twitter. This is consistent with the expectation that more women are tweeting about their or their families' health. Gender classification of tweets is based on the content of the messages. Words or phrases related to emotional and social processes are mainly used by females (e.g. "excited", "love you", "best friend"), whereas males mention more swear words and object references, and fewer emoticons.

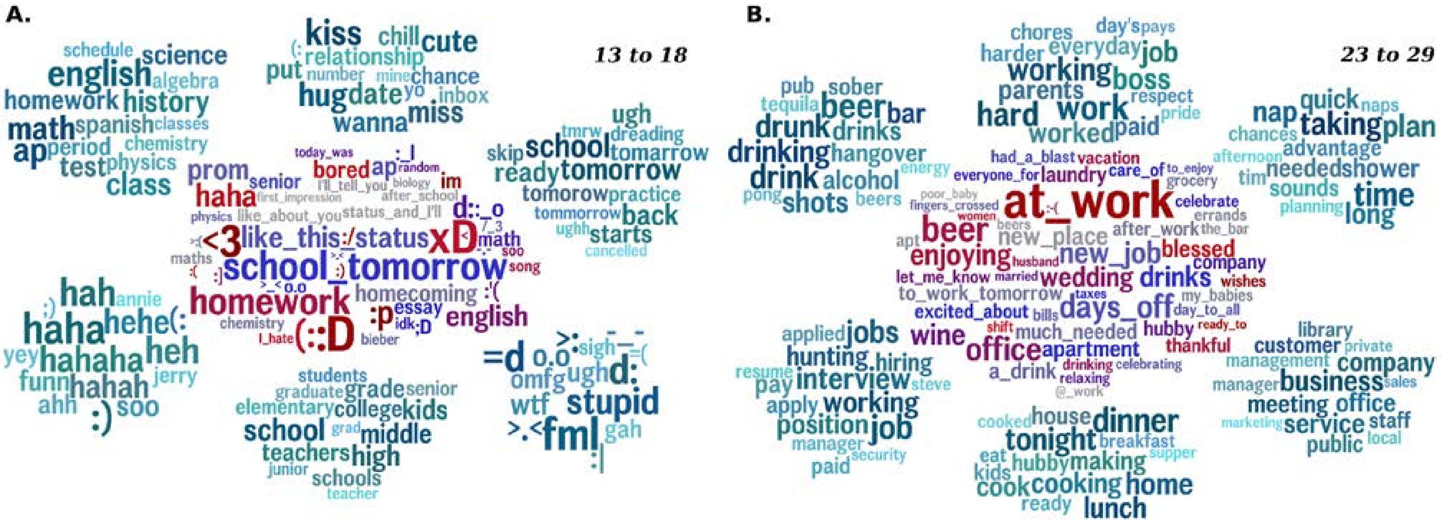
Picture 6: Chronic Disease Map Tool, gender and age distribution of Twitter messages.



Source: Swiss Re

With regard to the age of tweeters, the elderly are overly represented in the CDM Tool (see Picture 6), which is consistent with greater healthcare needs, but at the same time it is clear that Twitter spans the generations. Content of the messages can also be used to identify the age of tweeters. Whereas teenagers talk about education and school, those in their twenties focus on work, drinking, household chores, and time management.

Picture 7: Chronic Disease Map Tool, word clouds for teenagers A., and for those in their twenties B.



Source: Swiss Re

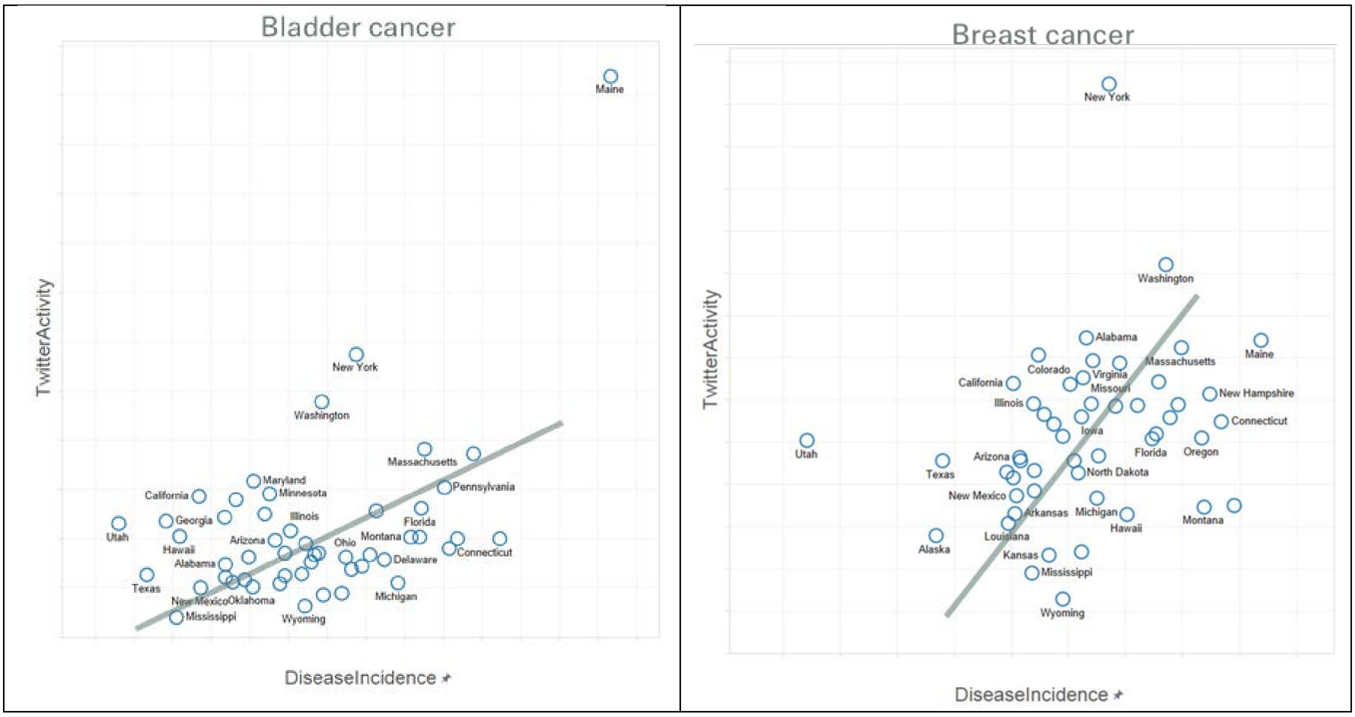
We found for both age and gender that in-house classification approaches worked better than commercial solutions. We are able to customise the classifiers to the domain of chronic disease, whereas the commercial approaches only offer domain independent classifiers (hence less specialised).

Application and relevance for the re/insurance industry

People's behaviour or concerns expressed online give indications on disease trends and encourage the insurance industry to develop new products, which meet new societal challenges. Marketing campaigns can be timely and focused on local topics of concern. Developing trends in claims can be identified and addressed. With regard to health behaviour, the uptake of new tobacco products at local and national level, for example, can be captured early and integrated into forward-looking scenario modelling. Furthermore, sentiment analysis on different treatments can identify individual concerns and likely drug adherence.

In one case study, we compared the Twitter activity on different cancers in several U.S. States with cancer diagnoses by the American Cancer Society (ACS). Comparisons can be made on the correlation between Twitter activity and cancer incidence, such as illustrated here for bladder cancer and breast cancer.

Picture 8: Twitter activity versus cancer incidence for bladder and breast cancer in different U.S. States.



Source: Swiss Re

In another, we looked at sentiment analysis on different classes of diabetic drugs to provide a picture on how individuals feel about their diabetes treatment. When looking for side effects, more complaints were linked to GPL (Glucagon-like peptide)-1- agonists than to insulins or thiazolidindiones. This is an interesting insight into diabetes management as it links to drug adherence which may limit secondary disease.

Conclusion and insights from the project

Social media is a valuable data source for Big Data and Smart Analytics applications for the re/insurance industry with potential predictive value. Respect of data privacy is and must be paramount, and data is strictly analysed on an aggregated basis. Distrust of social media analytics must be addressed by transparency over data sources and analytical techniques.

Potential use cases could include:

- defining targeted marketing campaigns that address local topics of concern
- supporting targeted product development through understanding consumer needs and behaviours
- supporting claims management and preparedness

Multidisciplinary groups of topic experts and data scientists must be involved to dig into the data and find genuine meaning. This is time-consuming work; the interpretation of Big Data comes not just by storing it. The same can be said in getting a system like the CDM running from scratch. Significant human effort was needed before the automated processing of data was possible. A clear challenge was determining the location of tweets, and classifying tweets according to age and gender.

The potential online sources of information are manifold. They include, but are not limited to, blogs, chat rooms, online news media, or other social networks like Facebook, YouTube, or web search records comparable to the Google Trends suite. The inclusion of different language searches and analytics broadens the access to regions where Twitter activity is not limited to the English language.

References

[1] Eysenbach G. (2009) 'Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet', *J Med Internet* Res 11: e11. doi: 10.2196/jmir.1157.

[2] Mayer-Schönberger, V. and Cukier, K. (2014) *Big Data, A Revolution That Will Transform How We Live, Work and Think*, John Murray Publishers, ISBN 978-1-8-84854-792-6, p. 50 ff.

[3] Salathé, M., Bengtsson, L., Bodnar, T.J., Brewer, D. D. Brownstein, J.S., Buckee, C., Campbell, E.M., Cattuto, C., Khandelwal, S., Mabry, P.L., Vespignani, A. (2012) 'Digital Epidemiology', *PLoS Comput Biol* 8(7), published 26 July 2012, DOI: 10.1371/journal.pcbi.1002616, Featured in PLOS Collections

[4] https://about.twitter.com/company, https://www.youtube.com/yt/press/statistics.html, http://newsroom.fb.com/company-info/#statistics

[5] PricewaterhouseCoopers (2012) *Social media 'likes' healthcare: From marketing to social business*.

[6] Kramer, A.D.I., Guillory, J.E. and Hancock, J.T. (2014) Experimental evidence of massive-scale emotional contagion through social networks, 8788–8790, doi: 10.1073/pnas.1320040111, http://www.pnas.org/content/111/24/8788.long