INFERRING HEALTH INFORMATION FROM NON-HEALTH SOURCES

Ingemar J. Cox[1,2]

[1] University College London, UK
[2] University of Copenhagen, Denmark

Collectively, people now create enormous quantities of digital data. Some is explicitly created, on, for example, social networks such as Twitter. Other data is unconsciously created as people interact with digital systems. For example, each user query to a web search engine is stored in a query search log, which records, amongst other things, the location of the query, the time and date of the query, and the words constituting the query.

While this data is not directly generated for health purposes, research has shown that it can be used for such. Examples include estimating the prevalence of influenza in a population, measuring the effectiveness of a vaccination campaign, and portmarket drug surveillance.

The advantages of using non-health sources depends on the circumstances, but can include (i) ease of data collection, (ii) timeliness, i.e. the lag between data creation, collection and analysis can be very short, (iii) the behavioural information inferred from the data is often unique or at least very difficult to acquire from alternative sources, and (iv) the number of participants is usually much greater than in traditional epidemiological studies. Digital data from non-health sources can complement traditional health data when it is harder to collect data in the physical world, or people have a difficulty reporting associations.

In this talk we will describe how digital data from non-health sources can be used for a variety of purposes related to health and medicine. The methods are based on statistical natural language processing and machine learning. A number of examples from our's and other's work will be given.

Keywords: computational epidemiology, statistical natural language processing, machine learning

References:

Eysenbach, G., (2006). Tracking flu-related searches on the Web for syndromic surveillance, *AMIA Annu Symp Proc.*, 244-248.

Lampos, V., Yom-Tov, E., Pebody, R., Cox, I.J., (2015). Assessing the impact of a health intervention via user-generated Internet content, *Data Mining and Knowledge Discovery*, *29*, 5, 1434-1457.

Yom-Tov, T., Gabrilovich, E (2013). Postmarket Drug Surveillance Without Trial Costs: Discovery of Adverse Drug Reactions Through Large-Scale Analysis of Web Search Queries, *Journal Medical Internet Research*, *15*, (6):e124.