Valerie Racine's Commentary on "Big Data & Public Health"

Commentary On
Case: Big Data & Public Health

The fictional scenario described above is loosely based on a recent initiative by Google. In 2009, research scientists at Google published a study in *Nature*, describing their methods for tracking seasonal and pandemic influenza outbreaks using data generated from monitoring health-seeking behaviour on Internet search engines (Ginsberg *et al.* 2009). They had developed tools to track outbreaks in real-time in order to improve upon the traditional methods used by the Center for Disease Control and Prevention (CDC), which take approximately two weeks to gather and analyze data. The algorithms developed by the scientists at Google led to the creation of Google Flu Trends (GFT), a web service launched in 2008 to track flu outbreaks. The service is no longer publishing its results, but its data are made available to other researchers.

The 2009 *Nature* paper is often used as a paradigm example to illustrate the emergence of a new field referred to as digital epidemiology, or digital disease detection (DDD) (Brownstein *et al.* 2009; Salathe *et al.* 2012; Vayena *et al.* 2015). This field shares the goals and objectives of traditional epidemiology (e.g. public health surveillance, disease outbreak detection, etc.), but makes use of electronic information sources, such as internet search engines, mobile devices, and other social media platforms, which can generate data related to public health but that are not explicitly designed for collecting public health-related data. The motivation behind DDD initiatives, like Global Flu Trends, is to mine large datasets in order to accelerate the process of tracking and responding to outbreaks of infectious diseases.

In 2013, Google’s program to track influenza outbreaks was heavily criticized for mis-estimating the prevalence of influenza outbreaks (Butler 2013, Lazer et al. 2014, Lazer & Kennedy 2015). Its first big mistake occurred in 2009, when it underestimated the Swine Flu (H1N1) pandemic (Butler 2013; White 2015), due to

changes in people's search behaviour with respect to the categories of "influenza complications" and "term for influenza" given the non-typical seasonal outbreak of H1N1 during the summer months (Cook *et al.* 2011). Then, in 2013, *Nature* reported that GFT significantly over-estimated outbreaks of influenza (Butler 2013; Lazer *et al.* 2014). In a comment published in *Science* in 2014, Lazer *et al.* reported that GFT had been consistently over-estimating the prevalence of flu outbreaks before then, inaccurately predicting the prevalence of flu cases in 100 of 108 weeks during the 2011-2012 flu seasons (Lazer *et al.* 2014).

GFT's track record of mis-estimations has been described as "big data hubris" – "the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis" (Lazer *et al.* 2014, 1203). In epidemiology, traditional data collection and analysis involves gathering data from structured interviews, archives, censuses, and surveys, and then to look for patterns and trends in the data. However, most scientists commenting on the case of GFT have insisted that, despite its failures, the use of big data in epidemiology can be extremely valuable for public health surveillance (Lazer et al. 2014, Lazer & Kennedy 2015, White 2015).

The GFT case has invoked many epistemological questions about how to improve Google's flu algorithms, and big data analytics more generally, and how public health policy and decision-makers ought to use these tools. But, it has also engendered ethical concerns at "the nexus of ethics and methodology" (Vayena *et al.* 2015).

For example, there can be harmful consequences when such models are woefully inaccurate or imprecise. False identification of outbreaks or inaccurate and imprecise predictions of outbreak trajectories could place undue stress on limited health resources (Vayena *et al.* 2015). Wrong results or predictions might also undermine the public's trust in scientific findings, and worse, might lead to the public's dismissal of public health warnings.

In addition to worries about maintaining the public's trust on issues of public health, researchers developing models aimed at detecting outbreaks must consider that their results risk harming individuals, businesses, communities, and even entire regions or countries (Vayena *et al.* 2015). This harm may take the form of stigmatization of groups, and financial loss due to prejudice or restrictions on travel to tourist destinations. It can also restrict the freedom of individuals in the form of

imposed travel restrictions or quarantines. Consequently, ethicists have stressed that “methodological robustness” with respect to digital epidemiology is “an ethical, not just a scientific, requirement” (Vayena *et al.* 2015, 4).

As with other instances of big data collection and use in the life sciences, the use of big data gathered online in social or commercial contexts for public health purposes raises ethical issues about an individual’s right to privacy and notions of informed consent when that data is used for research purposes. However, in this context, it has been suggested that private corporations that have access to relevant data might have a moral obligation to share that data for matters related to public health and public health research. This consideration raises questions about how to regulate private-public partnerships with regards to data ownership within a global context in order to uphold the values of transparency, global justice, and the common good in public health research (Vayena *et al*. 2015).