Digital Pharmacovigilance and Disease Surveillance: Combining Traditional and Big-Data Systems for Better Public Health

Marcel Salathé

The digital revolution has contributed to very large data sets (ie, big data) relevant for public health. The two major data sources are electronic health records from traditional health systems and patient-generated data. As the two data sources have complementary strengths—high veracity in the data from traditional sources and high velocity and variety in patient-generated data—they can be combined to build more-robust public health systems. However, they also have unique challenges. Patient-generated data in particular are often completely unstructured and highly context dependent, posing essentially a machine-learning challenge. Some recent examples from infectious disease surveillance and adverse drug event monitoring demonstrate that the technical challenges can be solved. Despite these advances, the problem of verification remains, and unless traditional and digital epidemiologic approaches are combined, these data sources will be constrained by their intrinsic limits.

Keywords. digital epidemiology; disease surveillance; pharmacovigilance; Twitter.

Traditional disease surveillance has been a key ingredient in any public health portfolio for many decades. Disease surveillance is widely recognized as one of the most important tools to assess, predict, and mitigate infectious disease outbreaks. Traditional disease surveillance is based on data collected by health institutions, and the data typically consist of information such as morbidity and mortality data, laboratory reports, individual case reports, field investigations, surveys, and demographic data. They are generally collected by physicians, public health laboratories, hospitals, and other health providers and institutions. The computer revolution that began in the 1970s has affected traditional disease surveillance systems by improving the accessibility of data and by increasing the speed at which data are transmitted between institutions. However, the ongoing Internet and mobile phone revolution has a qualitatively distinct effect: in addition to making epidemiologic data available faster and more broadly, new data are generated directly by the public, often on platforms not primarily designed for health purposes. These data streams of user-generated data are almost always by-passing traditional public health channels. They are the data streams on which digital epidemiology is generally based [1, 2].

One of the first and certainly the most prominent examples of digital disease surveillance was Google Flu Trends [3]. Google Flu Trends was essentially an analytical estimate of the level of weekly influenza activity based on the search queries that Google received. The analytical estimate was derived by a model selected by generating the best fit to the Centers for Disease Control and Prevention’s (CDC’s) influenza-like illness (ILI) data from a number of different US regions. The original model results obtained a mean correlation of 0.9 with the CDC data. A few years later, in summer 2015, Google decided to shut down the public website of Google Flu Trends and instead opted to give select academic and public health institutions access to the data. This announcement followed numerous reports that systematically assessed Google Flu Trends’ overestimation of influenza activity, attributing it to a combination of a phenomenon termed “big-data hubris” and algorithm dynamics. The first refers to the assumption that the novel big-data streams are a substitute, rather than a supplement, to traditional data collection efforts. The second refers to the observation that, while the Google search algorithm receives updates on a weekly or even daily basis, the Google Flu Trends model received updates only rarely. This led to a situation where the model did not keep in sync with the changing nature of the data from which it was supposed to generate predictions.

Despite the problems of Google Flu Trends, the system was an important example of the promises of digital epidemiology: to use novel data streams, often generated for purposes quite distinct from public health, to extract additional public health signals, such as those relevant for disease surveillance. But while Google makes some search pattern data available through an interface called Google Trends, the raw search-query data...
that Google Flu Trends was based on is not publicly available. In recent years, two other digital data sources have attracted the attention of digital epidemiologists: Twitter, the popular microblogging services, and Wikipedia, the world’s largest open-access encyclopedia. Twitter data are openly accessible through an application programming interface, which allows any third party to stream Twitter data in real time to their own application. Twitter has been extensively used to assess influenza activity [7, 8], but it may principally suffer from the same problems of overfitting [9] and static algorithms. Wikipedia access logs, a public data source, have recently attracted the attention of the research community as a proxy of search engine query logs because Wikipedia pages are often ranked highly in the search engine results following disease-related queries. Early analyses of Wikipedia access logs have shown great promise in providing real-time estimates (so called now-casting) of the prevalence of a number of infectious diseases [10, 11].

While search engine logs, social media posts, and Wikipedia access logs are a few examples of big-data sets that have emerged following the ongoing Internet penetration worldwide, there is also another source of data that is increasingly relevant for disease surveillance—the public itself. In contrast to classical surveillance that reports on a patient’s health status once they have accessed the health system, participatory surveillance asks patients to report symptoms and other data directly online. Web-based participatory surveillance systems have shown great promise in the case of influenza, for example. In Europe, the Influenzanet project has been successfully collecting data on ILI activity in a number of European countries [12], and in the United States, Flu Near You has emerged as a leading crowdsourced influenza surveillance system, and there are more in other parts of the world [13]. Given the widespread use of smartphones with broadband Internet access worldwide, we can expect many more participatory public health applications in the near future, complementing traditional surveillance systems.

Importantly, because many of the data streams of digital epidemiology have not been generated for the disease surveillance niche, much broader insights can be gained from these data sources. While much of the earlier work on digital epidemiology has focused on user-generated description of symptoms, later work has increasingly focused on the analysis of health behaviors and sentiments/opinions, particularly as they relate to infectious diseases. For example, Twitter data have been mined for signals of vaccine sentiments to estimate vaccine uptake rates. During the 2009 influenza A(H1N1) pandemic, vaccination sentiments measured on Twitter correlated positively with prospectively reported vaccination uptake rates across US states [14]. This indicates that these new data streams can help in the public health decision-making process, because sentiments expressed on Twitter can be measured in real time, giving those in public health practice early warning signals of possibly undervaccinated populations. Later work on the same data set investigated how negative and positive sentiments about vaccination spread across the social network, suggesting that negative sentiments are more susceptible to social contagion than positive sentiments [15]. Last but not least, data from most of these services are increasingly generated on mobile phones and other devices, increasing the probability that high-resolution geographic information is associated with the data, a phenomenon that will become increasingly important, given the spatial dynamics of disease spread.

**DIGITAL PHARMACOVIGILANCE**

The widespread use of the Internet and of social media in particular has had a dramatic effect not only on infectious disease surveillance, but also on the surveillance of drug use and related events. Perhaps even more so than traditional infectious disease surveillance, traditional surveillance of adverse drug reactions (ADRs) after drug use is slow and patchy. When reported by patients or healthcare professionals, ADRs are typically assessed by drug experts and pharmaceutical companies, and the results are then passed on to government agencies. This leads to substantial data loss and delays. A recent study in the United States showed that hospital staff did not report 86% of ADRs among patients [16]. The rate of underreporting in nonclinical settings is arguably even higher. Once government agencies receive the reports, they often release them with a delay of months or even years. The lack of speed and broad coverage has multiple causes, including the fact that a proper assessment of ADR data is both imperative and time-consuming; it is nevertheless in direct contrast with the public health importance of ADRs. In the European Union alone, ADRs are the cause of 5% of all hospital admissions and are responsible for an estimated 197 000 yearly deaths [17].

Public ADR reporting systems are largely unknown to the public, despite long-term governmental support. A recent study in Australia reported that only 10.4% of the general population was even aware of the national ADR reporting system [18]. This low awareness was comparable to the results reported in an earlier study in the United Kingdom, where only 8.5% of the adult population was aware of the United Kingdom ADR reporting system [19]. Among physicians, ADR reporting has been declining over time in both countries. Such declining reporting by physicians has been linked to ignorance, difficulty, lethargy, and insecurity (sorted here by decreasing frequency, with emphatic connection between their symptoms and the drugs they are
a combination of crowdsourced human assessment and machine-learning algorithms to identify the tweets of individual reports about ADRs with HIV drugs such as Atripla (Gilead and Bristol-Myers Squibb) and Truvada (Gilead, Foster City, California). The remaining 1642 tweets represented ADRs from single drugs or drug combinations and captured well-recognized toxicities known from clinical practice. For example, efavirenz-containing treatments (eg, Sustiva [Bristol-Myers Squibb, New York City, New York] and Atripla) were often reported in conjunction with sleep-related problems, such as nightmares or lack of sleep, a phenomenon well-documented in the clinical literature [27]. The study also analyzed the sentiment expressed in these tweets, which was mainly but not always negative, highlighting a benefit of social media studies over search query analysis. Tweets, despite their limitation of 140 characters, can still convey much more information than a search query, containing valuable information that can put potential ADRs in a specific and possibly relevant context.

Another source of user-generated content on the Internet are health forums. Leaman et al [28] mined data from the website DailyStrength and manually annotated 3600 posts relating to 4 drugs, carbamazepine, olanzapine, trazodone, and ziprasidone. The ADR incidence rates for these drugs is well established by the FDA, and the study showed that there was a strong correlation between those well-established incidence rates and the rates derived from the annotations generated from the user-generated posts. The authors also developed an automated system to identify adverse reactions by means of a primary lexical method, using 450 of the 3600 posts. When they evaluated the system against the 3150 posts not used for system development, they found that it performed well, with a precision of 78.3% and a recall (sensitivity) of 69.9%. Chee et al [29] conducted a study with a less constrained data set consisting of 27 290 public health and wellness groups on Yahoo. They used a natural language processing (NLP) approach to identify drugs that were withdrawn from the market. The identification was based on an NLP classifier trained on forum posts, allowing for further prediction of drugs that may be candidates for market withdrawal.

These studies are only a few examples of a growing literature aiming to detect ADRs through nontraditional data streams of patient-generated data. Such digital pharmacovigilance has the potential to strongly supplement pharmacovigilance based on traditional ADR reporting systems. At the same time, this new approach comes with its own set of challenges. First, access to data is often difficult or at times impossible. Given the widespread global use of Facebook, for example, ADR reports on Facebook would likely be a tremendous resource for pharmacovigilance, but the data are not accessible to the public or to researchers. Even with easier data access as provided by Twitter or scrapable websites, the terms of service of these platforms often prohibit access to the full data set and the sharing these data with others to verify and replicate results.

Digital Pharmacovigilance and Disease Surveillance • JID 2016:214 (Suppl 4) • S401
The challenge posed by the question “Are privately held data accessible for public health research?” is one of many ethical challenges surrounding the use of big data for public health. Vayena et al [30] have identified a number of challenges surrounding the use of big data for public health. Algorithmic claims of undocumented adverse effects may quickly sway public opinion in one way or another and, if the claims turn out to be wrong, might potentially taint otherwise safe drugs with a bad reputation for a long time. Of course, this problem is not limited to algorithmic suggestions alone: the now widely debunked claim that the measles, mumps, and rubella vaccine may cause autism, for example, was based on a fraudulent study whose failings were entirely noncomputational. Nevertheless, today’s availability of cheap computational power substantially reduces the ease with which algorithmic claims can be made, and we therefore need to think about a system that can weigh these claims in a way that is both scientifically sound and remains open to anyone.

CONCLUSIONS

The emergence and subsequent public withdrawal of Google Flu Trends has illuminated two potential key problems of digital epidemiology: big-data hubris and algorithm dynamics [6]. As the examples mentioned above have shown, there is tremendous potential for epidemiology in these novel data streams that have emerged during the growth of the Internet and the widespread use of smartphones. Nevertheless, these data streams are conduits to big-data hubris, a situation where these new data streams seek to supplant traditional data streams, rather than supplement them. In this context, it is worthwhile to note that the original authors of Google Flu Trends warned that “this system is not designed to be a replacement for traditional surveillance networks or supplant the need for laboratory-based diagnoses and surveillance” [3, pp 1013]. Despite traditional epidemiology’s shortcomings, it is ultimately the generator of ground-truth data against which novel, digital systems need to be validated. It will be prudent of the public health community to build on the strengths of both systems—veracity in traditional epidemiology and velocity and variety in digital epidemiology—in conjunction [31]. At the same time, traditional public health systems need to integrate novel data streams into their workflow and provide the corresponding infrastructural investment. Algorithmic intelligence in the public health domain needs to adjust to changing conditions all the time. The incentive structures in most of academia (frequent publication of novel findings) are at odds with the requirements of building long-term systems with dynamic algorithms that need to be maintained and updated regularly. The call for leveraging digital epidemiology intelligence for public health is therefore strongest when it comes from within the existing public health institutions.

Notes

Acknowledgments. I thank Antoine Flahault and two anonymous reviewers for comments on the manuscript.

Potential conflicts of interest. Author certifies no potential conflicts of interest. The author has submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Salathé M, Bengtsson L, Bodnar TJ, et al. Digital epidemiology. PLoS Comput Biol 2012; 8:e1002616.
2. Salathé M, Freifeld CC, Mekaru SR, Tomasulo AF, Brownstein JS. Influenza A (H7N9) and the importance of digital epidemiology. N Engl J Med 2013; 369:401–4.
3. Ginsberg J, Mohebbi MH, Patel RS, Bramer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009; 457:1012–4.
4. 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:e23610.
5. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol 2013; 9: e1003256.
6. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science 2014; 343:1203–5.
7. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PloS One 2011; 6:e19467.
8. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic. PLoS One 2013; 8:e36872.
9. Bodnar T, Salathé M. Validating models for disease detection using Twitter. In: Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013:697–702.
10. McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput Biol 2014; 10: e1003581.
11. Generous N, Fairchild G, Deshpande A, De Valle SY, Friedlbsky R. Global disease monitoring and forecasting with wikipedia. PLoS Comput Biol 2014; 10:e1003892.
12. Paolotti D, Carnahan A, Colizza V. Web-based participatory surveillance of infectious diseases: the Influenznet participatory surveillance experience. Clin Microbiol Infect 2014; 20:17–21.
13. Chunara R, Goldstein E, Patterson-Lomba O, Brownstein JS. Estimating influenza attack rates in the United States using a participatory cohort. Sci Rep 2015; 5:9540.
14. Salathé M, Kandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 2011; 7:e1002199.
15. Salathe M, Vu DQ, Khandelwal S, Hunter DR. The dynamics of health media: implications for infectious disease dynamics and control. PLoS Comput Biol 2014; 10:e1003892.
16. Office of the Inspector General, Department of Health and Human Services (DHHS). Hospital incident reporting systems do not capture most patient harm. Report OEI-06-09-00091. Washington DC: DHHS, 2012.
17. European Medicines Agency. One-year report on human medicines pharmacovigilance tasks of the European Medicines Agency. Report EMA/171322/2014. London, UK: European Medicines Agency, 2014.
18. Robertson J, Newby DA. Low awareness of adverse drug reaction reporting systems: a consumer survey. Med J Aust 2013; 199:684–6.
19. Fortnum H, Lee AJ, Rupnik B, Avery A. Collaboration YCS. Survey to assess public awareness of patient reporting of adverse drug reactions in Great Britain. J Clin Pharm Ther 2012; 37:161–5.
20. Lopez-Gonzalez E, Herdeiro MT, Figueiras A. Determinants of under-reporting of adverse drug reactions a systematic review. Drug Safety 2009; 32:19–31.

S402 • JID 2016:214 (Suppl 4) • Salathé
21. Vogel L, Sysak T. Physicians should be paid to report adverse drug reactions, experts say. Can Med Assoc J 2012; 184:E409–10.
22. Fox S. The Social Life of Health Information, 2011. 2011.
23. White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. J Am Med Inform Assoc 2013; 20:404–8.
24. White RW, Harpaz R, Shah NH, DuMouchel W, Horvitz E. Toward Enhanced Pharmacovigilance Using Patient-Generated Data on the Internet. Clin Pharmacol Ther 2014; 96:239–46.
25. Freifeld CC, Brownstein JS, Menone CM, Bao W. Digital drug safety surveillance: monitoring pharmaceutical products in Twitter. Drug Safety 2014; 37:555.
26. Adrover C, Bodnar T, Huang Z, Telenti A, Salathe M. Identifying adverse effects of HIV drug treatment and associated sentiments using twitter. JMIR Public Health 2015; 1:e7.
27. Cespedes MS, Aberg JA. Neuropsychiatric complications of antiretroviral therapy. Drug Safety 2006; 29:865–74.
28. Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. PA: Association for Computational Linguistics Stroudsburg. 2010:117–25.
29. Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. AMIA Annu Symp Proc. 2011; 217–26.
30. Vayena E, Salathe M, Madoff LC, Brownstein JS. Ethical challenges of big data in public health. PLoS Comput Biol 2015; 11:e1003904.
31. Santillana M, Nguyen AT, Dredze M, Paul MJ. Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS Comput Biol 2015; 11:e1004513.