Digital Epidemiology: Utilizing Social Networks and Data Science for Health Applications

Dina Gamaleldin Mahmoud

900140088

May 16, 2019

Submitted for the Ahmed Zewail prize, AUC

Digital Epidemiology: Utilizing Social Networks and Data Science for Health Applications

In 2015 around 16,000 children under the age of five died every day, mostly from preventable diseases like pneumonia and malaria (United Nations, 2015). This problem is more significant in disadvantaged areas, which not only have higher under-five mortality rates, but also higher incidence of health problems affecting development like malnutrition. According to the UNICEF Egypt Data Snapshot, "one in five children under five years of age is stunted" (p.1). Furthermore, the number of children, who for their age have low weight for their height, has significantly increased since 2000 (UNICEF, 2018). Information is needed for creating efficient policies to overcome these issues. However, in today's fast-paced world, traditional health surveys are potentially too slow, costly and not representative enough. Consequently, benefiting from advances in fields like data science to create interdisciplinary solutions for finding faster and more accurate ways of discerning information is an essential endeavor.

Epidemiology, literally meaning "the study of what is upon people," is a field in medicine focusing on health and disease amongst humans. It deals with the distribution, incidence rates, origin and manner of spread and disappearance of diseases (Green, Freedman, & Gordis, 2011; Salathé, et al., 2012). It is not an emerging field. With the idea from Aristotle's time, the nineteenth century witnessed increased attention towards epidemiology with relatively large-scale measurements of disease in specific populations (Bonita, Beaglehole, & Kjellström, 2006). For adequate epidemiological studies, data must be gathered to trace patterns of disease spread and to understand their causes. Traditionally, data is gathered by public health agencies through health personnel in hospitals, private practices or elsewhere (Salathé, et al., 2012).

Understanding disease dynamics is vital as it can help prevent future epidemics, decrease incidence rates of illnesses and predict health needs in specific areas. Prediction of health needs might be indispensable in impoverished areas or ones lacking in proper healthcare infrastructure. Since resources are scarce, being able to predict the specific resources more urgently needed would improve the services provided and healthcare spending efficiency. Understanding health and disease dynamics is also essential for economic growth. On the one hand, improvements in the economy and the individual income facilitate enhanced access to healthcare services. This leads to longer life-expectancy. Moreover, since vaccines and treatments for infectious diseases become more available, a shift occurs from infectious diseases to chronic illnesses like diabetes and cardiovascular diseases (Kahn, Yang, & Kahn, 2010).

On the other hand, improvements in the wellbeing and health of the individual increase work productivity, decrease man-hours lost because of illnesses and reduce the needed spending on healthcare. This potentially means an improved economy and more spending in other areas including education. This is crucial in developing countries as they aim towards economic growth, increased productivity and exiting the stage where infectious diseases and chronic diseases are both problematic (Kahn, Yang, & Kahn, 2010). Furthermore, developing countries require more spending in education to enhance and accelerate their development. Improvements in education also feed back into ameliorating the economy and citizen health.

However, traditional epidemiology was made for a different world. Nowadays, diseases can spread to larger geographical areas due to technological advances in transport. In addition, due to vaccines and increased hygiene awareness, chronic and non-communicable diseases constitute a significant portion of diseases in both developing and developed countries, as opposed to times when infectious diseases were the main concern (Salathé, et al., 2012). Therefore, it is a worthy endeavor to investigate ways to leverage advances in connectivity and data science to improve the data used in order to gain better epidemiological insights. According to the 2015 Millennium

Development Goals Report, stronger data production and usage of better data to guide policymakers is a “fundamental means of development.” A “data revolution” is also necessary to improve the availability of high quality data at the needed time to help achieve the development goals, which include reducing child mortality and fighting diseases like malaria and HIV (United Nations, 2015, p. 10). This has been the driving force behind the creation of Digital Epidemiology.

Nowadays, people are constantly using electronic devices connected to the internet to share information relating to their thoughts, whereabouts, spatial movements and health on various online platforms (Salathé, et al., 2012). According to the International Telecommunication Union, mobile cellular subscriptions and fixed-broadband subscriptions have been increasing in all countries, despite larger numbers of subscriptions in developed countries (ITU Telecommunication Development Sector, 2018). The “Digital” in “Digital Epidemiology” comes from using the data from these new sources to extract epidemiologically relevant information (Salathé, et al., 2012).

Potentially useful digital sources include sources outside the public health system like social networks, web search queries and blogs as well as platforms especially designed for health applications (Salathé, et al., 2012; Salathé M. , 2018). The data generated is stored and accessible electronically and can complement traditionally obtained data. In some countries, particularly developing ones, the infrastructure for traditional data collection might still be lacking (e.g. no standard procedure, not enough staff or lack of data computerization) (Kahn, Yang, & Kahn, 2010). In these cases, digital data mining, or attempting to extract useful information from digital data, can be applied (Salathé, et al., 2012).

These digital sources do not only provide insight into locations of disease spread, but also into the behaviors associated with this spread (Salathé, et al., 2012). For example, in Egypt, it might be useful to know which sectors of the population are affected by certain maladies. If an illness is alarmingly prevalent in a specific region or occupation, this might indicate an environmental factor, which must be addressed. Moreover, understanding behaviors or special habits, like smoking, associated with illnesses can instruct the education ministry on topics and behaviors to be taught in schools. This can also guide awareness campaigns in terms of population segments to target, which issues to address and how, as well as give an indication of their effectiveness.

How would this work then? For example, when users search for disease symptoms on a search engine, information can be retrieved regarding the used search queries and their location, which can be known from their IP address (Brownstein, Freifeld, & Madoff, 2009). These data sources, both formal and informal, can provide timely information which can quicken the response of public health officials to disease outbreaks as they occur. Furthermore, since these sources are not strictly governmental, transparency should be better because the information cannot be suppressed. In fact, organizations such as the World Health Organization (WHO) rely on such informal digital sources for their daily monitoring (Brownstein, Freifeld, & Madoff, 2009).

There have been several efforts to utilize these sources for similar purposes even if the term “Digital Epidemiology” was not employed. Earlier examples include the Program for Monitoring Emerging Diseases (ProMED) Mail, created by the International Society for Infectious Diseases in 1994. Its main purpose was sharing information regarding outbreaks by e-mailing and posting case reports, some of which obtained from readers, with experts sharing their comments (Brownstein, Freifeld, & Madoff, 2009). Another example is HealthMap, which utilizes tools to aggregate news and information from various sources, such as ProMED Mail and Google news, as well as to analyze them and visualize their significance, producing a global view of threats of infectious diseases (Brownstein, Freifeld, & Madoff, 2009; HealthMap). Finally, an example, which is no longer functional and hence provides insight into challenges and potential sources of error, is Google Flu Trends (GFT). Launched in 2008, it used search queries regarding symptoms to track influenza and similar illnesses. However, there were many problems, some of them technical, relating to, for example, which terms were used by GFT as indicators for the flu. Others were more general like the private ownership of the underlying algorithms. This meant the impossibility of GFT being verified or improved by external authorities (Salathé M. , 2018).

Despite numerous advantages, many challenges for Digital Epidemiology still present themselves. These are divided into two main categories. The first one encompasses technical challenges which are concerned with how to collect, store and analyze vast amounts of data to obtain relevant and accurate information (Salathé, et al., 2012). This category includes, too, challenges concerned with making data accessible to researchers and to health authorities as some social networks, producing relevant data, restrict access to it. It also deals with development of methods, like machine learning (feeding data to computers so that they can extract patterns, even without getting specific instructions on which patterns to find) to analyze the data. Such techniques require large datasets to be useful (Salathé M. , 2018). This category presents innovation opportunities for computer scientists, electronics engineers, legislators as well as epidemiologists.

Collecting the data requires epidemiological knowledge to be combined with algorithms and tools from the computer science field as well as legal guarantees. This would require lawyers, legislators and computer scientists to share their needs and ideas together to come up with appropriate solutions. This process needs to continue in order to provide improvements based on received feedback. The storage and analysis require electronics engineers and computer scientists to collaborate to design the best hardware for the software techniques used. Economists and business analysts would also need to join forces with them to analyze whether data centers – the dedicated places for the computers processing the data - should be built or rented from service providers. Developing countries would need to consider this cost, versus the disadvantages, cost and benefits of hiring their engineers to build their own data centers. Moreover, researchers from all fields need to cooperate with legislators and business owners to find an ideal solution to make data accessible to enable better innovation.

In addition, since already collected data is needed for machine learning techniques, traditional epidemiology still plays an important role. Formally gathered data can be digitized to serve as a starting point for Digital Epidemiology. This would also be a catalyst for the digitization of governmental data, making it available more quickly to policy makers and researchers. Consequently, an ecosystem for epidemiology will be created encompassing digital sources, which are formal and verified along with informal ones. Thus, a comprehensive picture of the health situation should be always available. Formal sources would verify information from informal ones. The latter would provide earlier indications of potential issues enabling timelier solutions.

The other category of challenges is more ethical. It is concerned with specifying who has access to electronically available data, who can share it and privacy concerns of its owners (Salathé, et al., 2012). This category also deals with who should be considered as the owner of generated data; the individual generating them, or the corporate owning the platform on which they are being generated (Salathé M. , 2018). Policy makers, lawyers and computer scientists can have rich discussions about these challenges in order to come up with privacy preserving laws allowing useful information to be reaped from the data. Furthermore, cryptographers can create algorithms to protect sensitive information within the data while allowing it to be processed to generate useful insights.

Therefore, these challenges should not deter the usage of this potential solution for the improvement of global healthcare. In fact, there have been some solutions presented to the aforementioned challenges. For instance, some work has been done to reconcile the individual's right to privacy and the benefit which society might reap if individuals share their digital data. Improvements have been done to a technique called homomorphic encryption, where data is analyzed while encrypted to preserve the privacy (Salathé M. , 2018). Some researchers have also considered using blockchains (the technology behind bitcoin) for preserving the privacy of personal data and medical records (Esposito, De Santis, Tortora, Chang, & Choo, 2018).

In conclusion, providing quality healthcare for individuals remains an open research area, especially in regions lacking proper infrastructure. However, leveraging the technological revolution happening can improve the situation. This would require experts from various fields to collaborate to create means of using existing technological infrastructure to tackle issues with the health system and present the most effective improvements.

References

Bonita, R., Beaglehole, R., & Kjellström, T. (2006). *Basic Epidemiology* (2nd ed.). Geneva: World Health Organization.

Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital Disease Detection — Harnessing the Web for Public Health Surveillance. *New England Journal of Medicine, 360*(21), 2153-2157. doi:10.1056/NEJMp0900702

Esposito, C., De Santis, A., Tortora, G., Chang, H., & Choo, K.-K. R. (2018). Blockchain: A Panacea for Healthcare Cloud-Based Data Security and Privacy? *IEEE Cloud Computing*, 31-37.

Green, M. D., Freedman, D. M., & Gordis, L. (2011). *Reference Guide on Epidemiology.* Washignton, D.C.: The National Academic Press.

HealthMap. (n.d.). *About HealthMap.* Retrieved from Disease Daily: http://www.diseasedaily.org/about

ITU Telecommunication Development Sector. (2018). *Measuring the Information Society Report Volume 1.* Geneva: International Telecommunication Union. Retrieved April 16, 2019, from https://www.itu.int/en/ITU-D/Statistics/Documents/publications/misr2018/MISR-2018-Vol-1-E.pdf

Kahn, J. G., Yang, J. S., & Kahn, J. S. (2010). 'Mobile' Health Needs And Opportunities In Developing Countries. *Health Affairs*, 252-258. doi:10.1377/hlthaff.2009.0965

Salathé, M. (2018). Digital epidemiology: what is it, and where is it going? *Life Sciences, Society and Policy, 14*(1), 1. doi:10.1186/s40504-017-0065-7

Salathé, M. B., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., Campbell, E. M., . . . Vespignani, A. (2012, July 26). Digital Epidemiology. *PLoS Computational Biology, 8*(7). doi:https://doi.org/10.1371/journal.pcbi.1002616

UNICEF. (2018, October). Child Malnutrition: Unfolding the Situation in Egypt. *UNICEF Egypt Data Snapshot*(1). Retrieved April 25, 2019

United Nations. (2015). *The Millennium Development Goals Report.* New York: United Nations.