The Application of Data Science in Public Health

Dingwei Zhan

University of California, Irvine
dingweiz@uci.edu

ABSTRACT
Since the eruption of COVID-19 pandemic at the end of Dec. 2019, China has relied heavily on health QR code to monitor the movement of people at various locales and regions. Technically, health QR code is a micro-program embedded in Alipay or Wechat that integrates big data and complicated algorithm to allocate a certain color (red, yellow, or green corresponding to different levels of exposure risks) to the code. The placement of health QR code in nearly every smart phone in China has contributed to the establishment of a social barrier to withstand the spread of COVID-19 pandemic. Indeed, before the sweeping of COVID-19 pandemic, QR code was mainly limited to occasions such as electronic payment and seldom applicable to the domain of public health. Gradually, the destiny of QR code has been sublimed to a national necessity and the very techniques that make the code work fall into the category of data science, via which data is collected, processed, integrated, and visually displayed. However, while health QR code contributed to choking of COVID-19 spread in China, the implication of coercive usage of such a system led to concerns that personal information might be misused for unrighteous purposes and the erroneous analytical results might cause negative consequences. Ultimately, this article demonstrates that although the wielding of data science might lead to some undesirable effects, it is inevitable for contemporary human beings to accept and get along with datafication.

Keywords: *Data Science, COVID-19, Health QR Code, Privacy, Public Health.*

1. INTRODUCTION

Data analysis in public health has already been extensively utilized in electronic medical records, remote diagnosis, and AI-driven medical consulting. However, it was not until the outbreak of COVID-19 that data science started to become an indispensable part of Chinese public health sector. Driven by advanced data analytical techniques, a cluster of interwoven black lines reveal where holders of health QR code visited and whether they should be subjected to further quarantine processes. This article will start by enumerating sources of data in the domain public health in China, then expound the main digital technologies, and move on to unveil the mystery of health QR code. The key role health QR code plays in controlling spread of COVID-19 at populous China reiterates the significance of data science in epidemiological domain and will inspire the popularity of data analytical techniques in other fields of public health.

2. VARIOUS SOURCES OF DATA IN PUBLIC HEALTH

For any epidemiological study to be conducted, sufficient amount of data has to be pooled firsthand so that a reliable size of sample can provide statistically significant result for further public health analysis. Consequently, it is necessary to identify several main ways that data regarding public health can be collected.

2.1. Data From Hospital

In China, it is quite rare for people to own private doctors and common citizens will pay a visit to government-funded hospitals even they have only trivial symptoms and minor uncomfortable conditions. Because China has a huge population base and Chinese people tend to visit hospitals frequently, those hospitals will accumulate a lot of health data regarding Chinese citizens' odds to suffer from chronic diseases, their clinical histories, and many other medical records in electronic formats. Moreover, in response to uneven development between urban and rural health care

systems, Chinese government have already enforced a hierarchical medical system to cater exactly to people residing at small counties or geographically located towns, thus further enlarging the sample size of patients[1].

Data From CDC

Centers for Disease Control and Prevention (CDC) shoulder responsibilities to collect citizens' basic health conditions, report injuries caused by occupational exposure, monitor the severity of chronic diseases, supervise the production and injection of vaccinations, and etc [2]. During theses procedures, enormous amount of data will be recorded and many designated platforms have been put in place. For example, under the guidance of WHO (World Health Organization), it has been 10 years since CDC at China launched the Epidemiological Dynamics Data Collection Platform, which incorporated cloud servers and browser/server architecture to substantially reduce the time for data collection and significantly increase the space for data storage[3].

2.2. Data From Large-Scale Data Centers

Aided by Internet of Things, many lines of industry under the tag of public health such as pharmaceutical manufactures, traditional Chinese medicines, basic biological labs have generated enormous amount of data, which are pooled together by government sponsored data centers. For example, Chinese government has assembled the COVID-19 Data Sharing Platform of National Population Health Data Center that stores as much as 50 TB of data and that continues to be enriched by relevant information[4].

3. DATA PROCESSING TECHNIQUES

Data collection is the very first step before reliable data processing techniques help practitioners in public health field to properly interpret data and make best use of them. Commonly used digital technologies are listed as follows:

3.1. Machine Learning

Machine learning is widely used in the establishment of digital epidemiological surveillance system as it can cater to different sizes of sample and disparate characteristics of input. Under the aid of artificial intelligence, machine learning in public health realm is applied to generate predictive models which integrate metabolic parameters, body indexes, and diagnostic figures. The computer has the discretion to glean information from datasets, evaluate the quality of data, eliminate noise data, and ultimately display a statistically reliable result. Moreover, after trials and errors, machine learning has already progressed to the level of deep learning, which means it can not only process textual information, but also conduct imagery analysis[5]. However, machine learning is not a panacea since many key variables in public health are presented in non-standardized forms and the correlation coefficient might not reflect the authenticity of links between modeling results and realities.

3.2. Wearable and Implantable Techniques

Intelligent devices such as Apple Watch can record physiological parameters such as pulse frequency, bone density, blood pressure in real time manner [6]. The most prominent effect brought about by wearable techniques is convenience. Previously, a person had to go to a clinical center to get his or her body examined. But now, even that person is in deep sleep, wearable sensors on his or her body will help regulate his or her body conditions and constantly channel informational flow to medical centers and physicians for further diagnosis. Apart from monitoring physical health, users of wearable devices can independently jot down their stress levels and moods, which are associated with biomarkers such as adrenaline and cortisol [7]. In other words, issues in mental health can also be addressed by wearable and implantable techniques.

Natural Language Processing(NLP)

NLP is a general terminology that delineates the procedures of applying algorithms to discern important connotations in daily language and grasp meanings from seemingly chaotic words written and spoken. It is essentially a branch of computer science. In the domain of public health, it can finish tasks such as summarizing clinical notes from lengthy narratives, maintaining data integrity by extracting data objectively, converting inaccessible information into understandable formats, and recognizing verbal inputs while synchronizing them into texts. However, due to the complicated nature of clinical terms, the semantic meaning of written contexts or speech patterns are easily subjected to misunderstanding [8]. For example, when NLP techniques are applied to reveal the meaning of clinical notes, their accuracy rate might abate because these notes are full of abbreviations and acronyms subjected to high frequency of misinterpretations[8].

4. HOW HEALTH QR CODE WORKS

Unlike previous collection of public health data that happens at designated locations or specific occasions, the health QR code is sampling data anytime and everywhere due to COVID-19's highly contagious nature.

By definition, health QR code is fundamentally an integrated system of data collection, analysis, and visualization. It is a micro-program inscribed in frequently used payment app Alipay and communication

app Wechat since these two apps are almost universally installed on smart phones in China. Facilitated by data analytical techniques, these mobile apps can retrieve users' recent whereabouts, the time interval they spend in potentially contagious environments, and their possible contact with infected people at large who have the disease under incubation. After pooling data from multiple sources, the program can assign a color to the code which can be red, yellow, and green distinguished by exposure risks. Figure 1 below illustrates how the code looks like in different colors. Specifically, green QR code means that the holder of the code has low exposure risks and is free to enter public facilities and take public transportation; yellow QR code means that the holder of the code has medium exposure risks and will have to undergo a 7-day quarantine before the code holder can move at large; red QR code means that the holder of the code has high exposure risk and will have to undergo a 14-day quarantine before the code holder can move about freely.

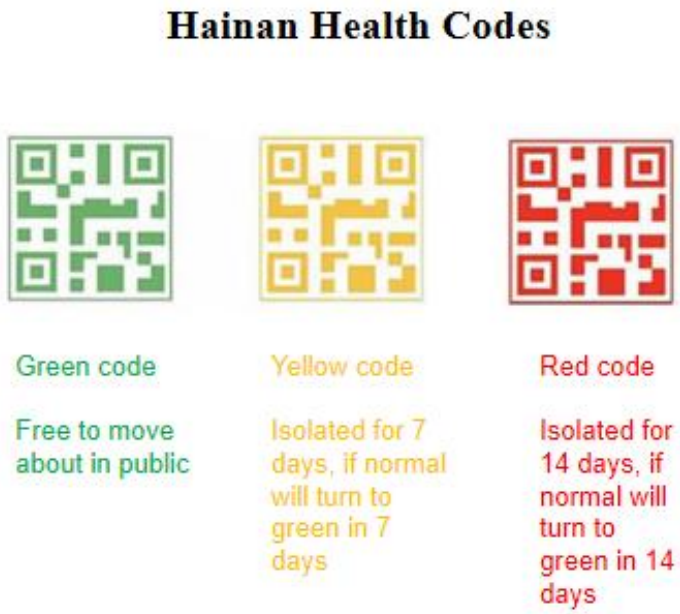

Figure 1. Examples of health QR codes

Techniques involved in maintaining the effectiveness and accuracy of Health QR Code start with Global Positioning System that can give location data and ubiquitous 4G or 5G mobile network that reveals temporal data. Then the combination of spatial and temporal data undergoes algorithm and derives color-based contagion risks. Till the QRof 2020, millions of Chinese have their unique health QR code generated as it is the prerequisite for anyone to commute by public transports and to enter public facilities [9].

5. DATATIFICATION OF PUBLIC HEALTH ACCELERATED BY COVID-19 PANDEMIC

If someone is interrogated by policemen on streets regarding where he or she has been for the last 14 days, that 'suspect' might choose to hide something too private intentionally or omit something too trivial unintentionally. However, the health QR code will objectively reveal all places they go and all people they meet by connecting informational nodes bits by bits. In other words, Chinese society has ushered in the era of datatification as the central government promotes health QR code nationwide[10]. Before the popularity of health QR code, the basic information was input manually by community workers and then transferred to relevant administrative agencies, which was time consuming and subjected to human errors. However, with the introduction of health QR code, all the basic information will be automatically uploaded to cloud end for verification and go through accurate statistical analysis, which can derive reliable result free of human interference.

Essentially, code scanning for cutting off the chain of virus spread is nothing different from code scanning for electronic payment. It takes only seconds for health code holders to present their codes to inspectors at the entrance of public facilities and in the blink of eyes informational streaming without physical barriers occurs and connects streaming of people, streaming of goods, and streaming of virtually everything. Digitization of people's track of movement is a healing process for a society wounded by compulsory social distancing ordinances.

As even tiny bits of information will be recorded to make sure nothing important is omitted, there has to be enough space to store them, which is rendered possible

via data infrastructures. Data infrastructures are cornerstones for data in the realm of public health to function properly. Sufficient storage space, convenient access, and energy-efficient data center are all necessary sections for establishment of data infrastructures, which are still found wanting in the era of COVID-19 pandemic.

6. ETHICAL CONCERNS ACCOMPANYING HEALTH QR CODE

6.1. Potential Leakage of Personal Information

Basically, the generation of the code color is to track the physical movements such as which train or flight the code holder took or whether the code holder had contact with patients contracting COVID-19 virus. However, with the mutation of COVID-19 virus into delta type and the escalating situations of global pandemic, the scope of informational sampling has to be enlarged. For example, in China, the app where health QR code is embedded starts to retrieve data on code holders' history in ibuprofen purchase, their search engine typing record, and many other pieces of information that falls into the category of privacy. It would be a blessing if all the collected information is properly handled for the sake of protection against COVID-19. However, the usual case is that all the information of those diagnosed as COVID-19 patients will be publicized on social media and every bit of their personal information will be magnified just like a microbe under a microscope. For example, at Chengdu, a metropolis located at the southwestern China, a girl who had frequented a night club was diagnosed with COVID-19 and her ID card, family address, even her measurements soon went viral on social media [11,12], which caused a lot of stress for her whole family.

6.2. Accountability of Algorithm

As data has to be processed based on certain rules, it is the complicated algorithm that derives the color of code. To start with, if the color of code can dictate whether its holder should enter a public facility or be quarantined in an individual compartment, its holder has to legal right to know everything about how the color comes into being. There have already been numerous cases that demonstrate the conditions of false alarm, which is to say, the code holder is denied of access to a cinema or a supermarket due to the fact that his or her health QR code is not green, however, further epidemiological study shows that he or she actually has never been anywhere of exposure risk or physically contacted anyone of exposure risk. At this time, if the color holder could understand the underlying algorithm, he or she could locate exactly what went wrong and claim his or her rights.

This is why algorithm accountability matters. Digital trenches between programming engineers and common citizens will lead to disparate levels of recognition of what color reveals [13]. Therefore, if nothing goes wrong, health QR code will play its role in building a barrier against COVID-19 virus; if something goes wrong, the code designers or algorithm developers must be accountable.

6.3. Discrimination Under Healthism

Healthism is supposed to have benign intentions: individuals pursue a way of life that brings about comfortable physical conditions and longevity to themselves [14]. However, amid the global outbreak of COVID-19 pandemic, healthism evolves from seeking for health to evading from disease. Especially in China, anyone who is susceptible to COVID-19 has been prone to be categorized into the domain of "unsanitary". Even from perspectives of public health experts, sanitation has no universally applicable standards [15]. Indeed, health QR code is supposed to build a protective wall against virus, not a discriminating wall against potentially infected persons. Some sociologists have already discerned the side effects caused by health QR code and proposed to pass relevant legislation to restrict the usage of code to only virus prevention.

6.4. More Burdens On Health QR Code

Recently in China, the government has initiated a propaganda campaign that encourages all of its citizens to receive COVID-19 vaccine shots as fast as possible and issued administrative orders that deny unvaccinated people and their family members the access to hospitals, schools and many other public facilities[16]. However, as no one is expected to bring the paper work of vaccination record anytime anywhere, the government has embedded electronic certificate of vaccination into health QR code. Specifically, for holders of health QR code who have received two shots of COVID-19 vaccine, the interface of their QR code will be different from that of those who have not yet finished vaccination. Below is the health QR code at Guangxi Zhuang Autonomous Region and around the code there is a golden perimeter, which will automatically show up after the code holder receives vaccination.

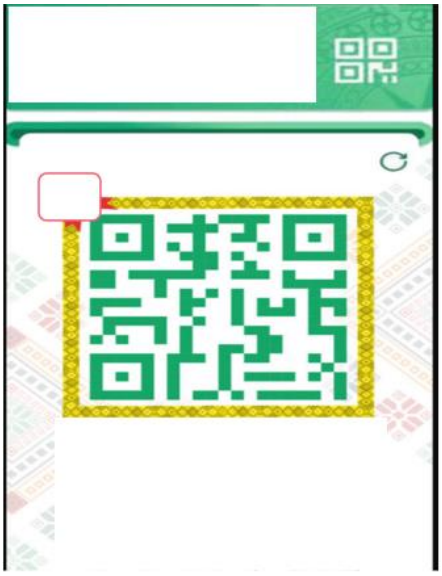

Figure 2. Example of health QR code which demonstrates the completion of vaccine injection

Some human rights activities have voiced against blending of vaccination with health QR code as the code is designed to evaluate the exposure risk of the code holder, which is irrelevant to the action of vaccination. If vaccination history can be integrated into health QR code, perhaps many other modules will be added and eventually a totalitarian, overarching code might come into being that covers everything beyond legitimate boundary.

7. CONCLUSION

COVID-19 prevention, a mission for public health practitioners globally, relies greatly on the deployment of big data solutions as data science evolves thanks to the ongoing construction of data infrastructures and evolution of data analysis tools. The extensive usage of health QR code as a digital tool to track people's physical movement and evaluate their exposure risk contributes a lot to cutting off the contagious chain of COVID-19. Its success relies on sophisticated collection, analysis, and visualization of data. Yet, as it has been gradually permeating into every aspect of people's daily life, it has led to concerns such as leakage of personal information, accountability of analytical errors, and unfair inscription of extra functions. After systematically reviewing data sources and relevant data processing techniques, the future study should lie in the dual missions that on the one hand data science has to be rendered more reliable and on the other hand ethical concerns of application of data science into public health have to be resolved.

REFERENCES

[1] Kok Fong See, Ying Chu Ng, Do hospital reform and ownership matter to Shenzhen hospitals in China? A productivity analysis, Economic Analysis and Policy, Volume 72, 2021, Pages 145-155, ISSN 0313-5926 https://www.sciencedirect.com/science/article/pii/S0313592621000850

[2] Cheng Xi SUN, Bin HE, Di MU, Pei Long LI, Hong Ting ZHAO, Zhi Li LI, Mu Li ZHANG, Lu Zhao FENG, Jian Dong ZHENG, Ying CHENG, Ying CUI, Zhong Jie LI, Public Awareness and Mask Usage during the COVID-19 Epidemic: A Survey by China CDC New Media, Biomedical and Environmental Sciences, Volume 33, Issue 8, 2020, Pages 639-645, ISSN 0895-3988 https://www.sciencedirect.com/science/article/pii/S0895398820301586

[3] Xiaopeng Qi, Nilva Egana, Yujie Meng, Qianqian Chen, Zhiyong Peng, Jiaqi Ma, Description and analysis of design and intended use for Epidemiologic Dynamic Data Collection Platform in China, Volume 204, Investing in E-Health: People, Knowledge and Technology for a Healthy Future, Pages 123-129 https://ebooks.iospress.nl/publication/37242

[4] COVID-19 Data Sharing Platform of Population Health Data Center https://www.ncmi.cn/covid-19/index.html

[5] Bruno Samways dos Santos, Maria Teresinha Arns Steiner, Amanda Trojan Fenerich, Rafael Henrique Palma Lima, Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018, Computers & Industrial Engineering, Volume 138, 2019, 106120, ISSN 0360-8352 https://www.sciencedirect.com/science/article/pii/S0360835219305893

[6] Sri Nuvvula, Eric Y. Ding, Connor Saleeba, Qiming Shi, Ziyue Wang, Alok Kapoor, Jane S. Saczynski, Steven A. Lubitz, Lara C. Kovell, M. Diane McKee, David D. McManus, NExUS-Heart: Novel examinations using smart technologies for heart health—Data sharing from commercial wearable devices and telehealth engagement in participants with or at risk of atrial fibrillation, Cardiovascular Digital Health Journal, 2021, ISSN 2666-6936 https://www.sciencedirect.com/science/article/pii/S266669362100089X

[7] John Torous, Wearable Devices for Mental Health: Knowns and Unknowns, Psychiatric Times, Volume 33, Issue 6, 2016

https://www.psychiatrictimes.com/view/wearable-devices-mental-health-knowns-and-unknowns

[8] Dina Demner-Fushman, James G. Mork, Sonya E. Shooshan, Alan R. Aronson, UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text, Journal of Biomedical Informatics, Volume 43, Issue 4, 2010, Pages 587-594, ISSN 1532-0464 https://www.sciencedirect.com/science/article/pii/S1532046410000201

[9] Jieyu Wu, Xiaowei Xie, Le Yang, Xingyan Xu, Yingying Cai, Tinggui Wang, Xiaoxu Xie, Mobile health technology combats COVID-19 in China, Journal of Infection,Volume 82, Issue 1, 2021, Pages 159-198, ISSN 0163-4453, https://www.sciencedirect.com/science/article/pii/S0163445320305053

[10] Derina Holtzhausen, Datafication: threat or opportunity for communication in the public sphere? Journal of Communication Management, Volume 20, Pages 21-36, ISSN 1478-0852 https://dialnet.unirioja.es/servlet/articulo?codigo=5342209

[11] Huang Zhiling, New local COVID-19 case reported in Chengdu, China Daily Newspaper Online, Dec. 18th, 2020 https://www.chinadaily.com.cn/a/202012/18/WS5fdc53f2a31024ad0ba9cb15.html

[12] A female COVID-10 patient at Chengdu subjected to social media violence, NetEase News Online, Dec. 11th, 2020 https://www.163.com/dy/article/FTI7LHTQ0525KC0U.html

[13] Zhi-peng Wang, Shuai Zhang, Hong-zhao Liu, Yi Qin, Single-intensity-recording optical encryption technique based on phase retrieval algorithm and QR code,Optics Communications, Volume 332, 2014, Pages 36-41, ISSN 0030-4018 https://www.sciencedirect.com/science/article/pii/S0030401814006178

[14] Shiwani Mahajan, Yuan Lu, Erica S. Spatz, Khurram Nasir, Harlan M. Krumholz, Trends and Predictors of Use of Digital Health Technology in the United States, The American Journal of Medicine, Volume 134, Issue 1, 2021, Pages 129-134, ISSN 0002-9343 https://www.sciencedirect.com/science/article/pii/S0002934320306173

[15] Shaomin Guo, Xiaoqin Zhou, Prithvi Simha, Luis Fernando Perez Mercado, Yaping Lv, Zifu Li, Poor awareness and attitudes to sanitation servicing can impede China's Rural Toilet Revolution: Evidence from Western China, Science of The Total Environment, Volume 794, 2021, ISSN 0048-9697. https://www.sciencedirect.com/science/article/pii/S0048969721037323

[16] Ben Westcott and CNN staff, Unvaccinated people in parts of China to be denied access to hospitals, parks and schools, CNN News Online, July 15th, 2021 https://edition.cnn.com/2021/07/15/china/vaccine-china-restrictions-zhejiang-jiangxi-intl-hnk/index.html