Springer Proceedings in Business and Economics

Hui Yang
Robin Qiu
Weiwei Chen *Editors*

AI and Analytics for Public Health

Proceedings of the 2020 INFORMS International Conference on Service Science

Springer

Springer Proceedings in Business and Economics

Springer Proceedings in Business and Economics brings the most current research presented at conferences and workshops to a global readership. The series features volumes (in electronic and print formats) of selected contributions from conferences in all areas of economics, business, management, and finance. In addition to an overall evaluation by the publisher of the topical interest, scientific quality, and timeliness of each volume, each contribution is refereed to standards comparable to those of leading journals, resulting in authoritative contributions to the respective fields. Springer's production and distribution infrastructure ensures rapid publication and wide circulation of the latest developments in the most compelling and promising areas of research today.

The editorial development of volumes may be managed using Springer's innovative Online Conference Service (OCS), a proven online manuscript management and review system. This system is designed to ensure an efficient timeline for your publication, making Springer Proceedings in Business and Economics the premier series to publish your workshop or conference volume.

More information about this series at http://www.springer.com/series/11960

Hui Yang • Robin Qiu • Weiwei Chen
Editors

AI and Analytics for Public Health

Proceedings of the 2020 INFORMS International Conference on Service Science

Springer

Editors
Hui Yang
Department of Industrial Engineering
Pennsylvania State University
University Park, PA, USA

Robin Qiu
Division of Engineering & Info Sci
Pennsylvania State University
Malvern, PA, USA

Weiwei Chen
Department of Supply Chain Management
Rutgers, The State University of New Jer
Piscataway, NJ, USA

ISSN 2198-7246 ISSN 2198-7254 (electronic)
Springer Proceedings in Business and Economics
ISBN 978-3-030-75165-4 ISBN 978-3-030-75166-1 (eBook)
https://doi.org/10.1007/978-3-030-75166-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This proceeding publishes the papers submitted, peer-reviewed, and presented at the 2020 INFORMS Conference Service Science (ICSS 2020), held in the all-live and virtual format on Dec. 19–21, 2020. This conference provided an excellent opportunity for scholars and practitioners to present their service science related research and practice work, to learn about the emerging technologies and applications, and to network with each other for further collaborative opportunities.

2020 was a difficult and challenging year for the world. The COVID-19 pandemic was unprecedented. Containing the pandemic was and still is challenging to humanity. Contributing to combating the unprecedented COVID-19 crisis, the ICSS 2020 conference theme was *AI and Analytics for Public Health*, aimed at promoting and facilitating the development of healthy and strong communities where we live, work, learn, and play, and uncovering solutions to protect the health of people and the communities, nationally and internationally. This conference attracted scholars and practitioners around the world to come together virtually to share what had been found, helping each other by timely sharing solutions and stimulating new ideas, which further helped enhance the needed solutions and extend them to uncharted territories. We are confident that in the fight against any virus, humanity will and must prevail.

This year we had over 120 submissions from around the world. All full/short paper submissions were carefully peer reviewed. After the rigorous review and revision process, 37 papers were finally accepted to be included in this proceeding. The major areas covered at the conference and included in this proceeding include:

- Public Health Service, Policy, Administration, Response, and Systems
- Service Management, Operations, Engineering, Design, Innovations, and Marketing
- Smart Cities, Sustainable Systems, IT and Service System Analytics, and Self-service Systems
- Smart and Intelligent Service, Healthcare Analytics, FinTech, Learning Analytics, and Others

- Big Data, Machine Learning, Artificial Intelligence, and Data-Driven Decision Making
- Systems Modeling, Management, and Simulation in Manufacturing, Supply Chain, Logistics, and Others
- AI, Data Analytics, and Data-driven Applications in Health, Energy, Finance, Transportation, Sport, and Governmental/Public Services

In addition to the accepted research papers, ICSS 2020 provided an opportunity for scholar and practitioners to share their ongoing studies. We invited six well-known service science experts to deliver plenary speeches:

- Dr. Jim Spohrer, director of IBM Cognitive Opentech Group, presented "Future of AI and Post-Pandemic Society: A Service Science Perspective."
- Prof. Paul Maglio, University of California at Merced, former EIC of INFORMS *Service Science,* discussed "What is Service Science?"
- Prof. Weiwei Chen, Rutgers University, delivered "Improving Service Designs and Operations Using Analytics."
- Prof. Saif Benjaafar, Distinguished McKnight University Professor, University of Minnesota, EIC of INFORMS *Service Science*, articulated "Dimensioning On-Demand Vehicle Sharing Systems."
- Prof. Dmitry Ivanov, Berlin School of Economics and Law, explained "Supply Chain Resilience Theory and COVID-19 Pandemic: What We Know, Where We Failed, and How to Progress."
- Prof. Victor Chan, Tsinghua-Berkley Shenzhen Institute, reviewed "Recent Mathematical and Computational Studies of COVID-19."

The conference had 16 parallel sessions, including 77 presentations. ICSS 2020 also had successfully organized the best conference paper competition and the best student paper competition. We would like to thank all authors, speakers, track chairs, session chairs, reviewers, and participants.

Finally, we would like to thank all authors for submitting their high-quality works in the field of service science, and the conference organizing and program committee members, listed on the following pages, for their tireless efforts and time spent on reviewing submissions. We are very grateful to Springer's editors, Neil Levine and Faith Su, and the production editor, Shobha Karuppiah, who have contributed tremendously to the success of the ICSS 2020 conference proceedings. We would also like to acknowledge the NSF I/UCRC Center for Healthcare Organization Transformation (CHOT), NSF I/UCRC award IIP-1624727, for sponsoring the conference.

Co-Editors – Proceedings of 2020 INFORMS Conference on Service Science

Malvern, PA, USA — Robin Qiu

University Park, PA, USA — Hui Yang

Piscataway, NJ, USA — Weiwei Chen

ICSS 2020 Committees

Program Committee

- Ralph Badinelli, Virginia Tech, USA
- Victor Chan, Tsinghua University, China
- Ozgur Araz, University of Nebraska-Lincoln, USA
- Jenny Chen, Dalhousie University, Canada
- Weiwei Chen, Rutgers University, USA
- Hongyan Dai, Central University of Finance and Economics, China
- David Ding, Rutgers University, USA
- Qiang Duan, Penn State, USA
- Yucong Duan, Hainan University, China
- Tijun Fan, East China University of Science and Technology, China
- Siyang Gao, City University of Hong Kong, China
- Yan Gao, University of Shanghai for Science and Technology, China
- Dmitry Ivanov, Berlin School of Economics and Law, Germany
- Hai Jiang, Tsinghua University, China
- Zhibin Jiang, Shanghai Jiaotong University, China
- Haitao Li, University of Missouri–St. Louis, USA
- Zhenyuan Liu, Huazhong University of Science and Technology, China
- Kelly Lyons, University of Toronto, Canada
- Rym M'Hallah, Kuwait University, Kuwait
- Juan Ma, iHeartMedia, USA
- Xin Ma, Texas A&M University, USA
- Paul Maglio, UC Merced, USA
- Aly Megahed, IBM, USA
- Paul Messinger, University of Alberta, Canada
- Chuanmin Mi, Nanjing University of Aeronautics & Astronautics, China
- Ran Mo, Central China Normal University, China
- Ashkan Negahban, Penn State, USA
- Kai Pan, Hong Kong Polytechnic University, China

- Patrick Qiang, Penn State, USA
- Robin Qiu, Penn State, USA
- Lun Ran, Beijing Institute of Technology, China
- Tina Wang, University of Oxford, UK
- Hui Xiao, Southwestern University of Finance and Economics, China
- Xiaolei Xie, Tsinghua University, China
- Hui Yang, Penn State, USA
- Ming Yu, Tsinghua University, China
- Canrong Zhang, Tsinghua University, China

Conference Organizing Committee

- Conference Co-chair(s): Prof. Robin Qiu and Prof. Hui Yang
- Program Co-chair(s): Prof. Weiwei Chen
- Invited Tracks:
  - Special Sessions (Prof. Qiang Duan and Prof. Xiaolei Xie)
  - Sharing Economy (Prof. Ashkan Negahban)
  - Healthcare Service and Analytics (Prof. David Ding and Prof. Xiaolei Xie)
  - Service Design, Operations, and Analytics (Prof. Victor Chan and Prof. Canrong Zhang)
  - Service Economy in the Emerging Market (Prof. Qiang Qiang)

Best Student Paper Award Committee

- Laura Anderson, IBM, USA
- Clara Bassano, University of Salerno, Italy
- Chiehyeon Lim, UNIST, South Korea
- Kelly Lyons, University of Toronto, Canada (Chair)
- Eleni Stroulia, University of Alberta, Canada

Best Conference Paper Award Committee

- Weiwei Chen, Rutgers University, USA
- Dmitry Ivanov, Berlin School of Economics and Law, Germany
- Yingdong Lu, IBM, USA (Chair)
- Ashkan Nagahban, Penn State, USA
- Jie Song, Peking University, China

Contents

Epidemic Informatics and Control: A Review from System Informatics to Epidemic Response and Risk Management in Public Health

Hui Yang, Siqi Zhang, Runsang Liu, Alexander Krall, Yidan Wang, Marta Ventura, and Chris Deflitch

1 Introduction

Epidemic outbreaks impact the health of our society and bring significant disruptions to the US and the world. For example, Coronavirus Disease 2019 (COVID-19) is currently ravaging multiple countries and was declared as a global pandemic by the World Health Organization (WHO) in March 2020. COVID-19 has caused a total of approximately 7.82 million infected cases and 432 K deaths worldwide, as well as 2.17 million infected cases and 118 K deaths in the US by June 16, 2020 (CDC, 2019). The abrupt increase of cases quickly exceeds the capacity of health systems and highlights the shortages of workers, beds, medical supplies and equipment. Many governments have taken a variety of actions (e.g., lockdown, large-scale testing, stay-at-home) to flatten the curve and avoid overwhelming health systems, but these reactionary policies have resulted in great economic losses. The US unemployment rate has skyrocketed from 3.5% in February 2020 to 14.7% in April 2020 (The Bureau of Labor Statistics, n.d.). The number of unemployed persons has increased to 23.1 million, which is even worse than the Great Depression in 1930s. The economic uncertainty has caused US stock markets to trigger the circuit breakers to halt trading for a historical 4 times in the week of March 9–16, 2020 (Zhang et al., 2020). The US GDP shrunk 4.8% in the first quarter of 2020.

H. Yang (✉)
Department of Industrial Engineering, Pennsylvania State University, University Park, PA, USA
e-mail: huy25@psu.edu

S. Zhang · R. Liu · A. Krall · Y. Wang · M. Ventura
Center for Health Organization Transformation, The Pennsylvania State University, University Park, PA, USA

C. Deflitch
Department of Emergency Medicine, Penn State Health Milton S. Hershey Medical Center, Hershey, PA, USA

H. Yang et al. (eds.), *AI and Analytics for Public Health*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-030-75166-1_1

When the COVID-19 epidemic emerged, it was not uncommon to encounter a misperception or misinformation that coronavirus is like the seasonal influenza (flu). Although there are similarities (e.g., causing respiratory illness) between coronavirus and flu virus, they are significantly different. COVID-19 or severe acute respiratory syndrome (SARS) is caused by the family of coronavirus, which is not the same as the flu virus. There are three major types of flu viruses – Types A, B and C. Type A flu virus caused many epidemics in the past 100 years (e.g., 1918 Spanish Flu (Trilla et al., 2008), 1968 H3N2 epidemic (Alling et al., 1981), and 2009 H1N1 epidemic (Sullivan et al., 2010)). It is worth mentioning that Type A flu virus infects a wide variety of animals (e.g., poultry, swine, aquatic birds) and easily evolves and mutates genes. Once transported and adapted to humans, it can evolve into an epidemic. Types B and C flu viruses infect only humans as the typical seasonal flu and has rarely been the cause of past epidemics (Taubenberger et al., 2005). It is estimated by Center for Disease Control and Prevention (CDC) that seasonal flu causes approximately 140,000–810,000 hospitalizations and 12,000–61,000 deaths annually since 2010 (Disease Burden of Influenza, n.d.). However, the death toll of 1918 Spanish Flu is about 50 million worldwide and 675,000 in the US.

Historically, epidemics are inevitable and recur at more or less near-periodic cycles. It is difficult to predict when a new virus will emerge and cause an epidemic. The infection rate of a virus is commonly measured by the basic reproduction number R_0, which characterizes how many people on average can be infected by one infected individual in a susceptible population. For COVID-19, R_0 is estimated to range from 1.4 to 6.49, with a mean of 3.28 (Liu et al., 2020). The potential transmission pathway can be either through air droplets, which are generated when infected individuals talk, cough, or sneeze, or through contact with an infected person or surface that is contaminated with the virus. At the start of an outbreak, antivirals and vaccines are often not available. People can only resort to non-pharmaceutical interventions (NPIs) for the control and containment of virus spread (Davies et al., 2020). Traditional NPI methods include the practice of good personal hygiene, the use of disinfectants, the isolation and quarantine of infected individuals, and the limitation of public gatherings. From 1918 Spanish flu epidemic to COVID-19, this situation does not change much although health systems become more advanced and medical resources are richer than before.

However, one thing that does change is the faster and augmented capability of medical testing and diagnostics, thanks to rapid advances of gene/DNA, microbiology, and imaging technologies (Ravi et al., 2020). As such, large amounts of data are collected in the evolving process of epidemic outbreaks. The availability of data calls upon the development of analytical methods and tools to gain a better understanding of virus spreading dynamics, optimize the design of healthcare policies for epidemic control, and improve the resilience of health systems. Therefore, this paper presents a review of the system informatics approach of **D**efine**, M**easure, **A**nalyze, **I**mprove, and **C**ontrol (DMAIC) for epidemic management through the intensive use of data, statistics and optimization. Despite the sustained successes of DMAIC in a variety of established industries such as manufacturing, logistics, services and

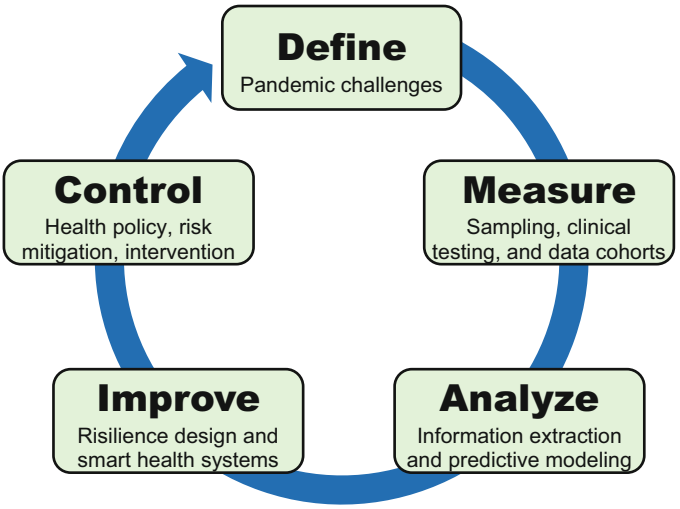

Fig. 1 The flowchart of system informatics for epidemic response and risk management

beyond (Yang et al., 2021; Knowles et al., 2005; Kumar et al., 2007), there is a dearth of concentrated review and application of the data-driven DMAIC approach in the context of epidemic outbreaks. As shown in Fig. 1, The DMAIC methodology consists of five phases: **(1) Define**: outline the societal challenges posed by the epidemic; **(2) Measure**: collect data about key variables in the epidemic process; **(3) Analyze**: extract useful information pertinent to the spread of epidemic; **(4) Improve**: design solutions and methods to improve the resilience of health systems; **(5) Control**: develop health policies, management plans, and intervention methods to control the spread of infectious diseases. The goal of this paper is to catalyze more in-depth investigations and multi-disciplinary research efforts to accelerate the application of system informatics methods and tools in epidemic response and risk management.

The rest of the paper is organized as follows: Section 2 discusses specific societal challenges arising from large-scale outbreaks of infectious diseases. Section 3 reviews the sampling and testing strategies to increase information visibility for epidemic management. Then, we present a review of analytical methods and tools for the extraction of useful information in Sect. 4. Continuous improvements and re-design to improve the resilience of health systems are discussed in Sects. 5 and 6 presents the health policies and intervention strategies for the control of virus spread. Section 7 discusses the system informatics approach for epidemic management and concludes this paper.

2 Epidemic Challenges to Our Society

2.1 Health System Challenges

Epidemic outreak calls upon the execution of large amounts of clinical testing to examine the prevalence of a virus in the population. No doubt, such a large demand poses significant challenges on the manufacturing and supply chain systems. Fortunately, advanced medical technology (e.g., gene/DNA, microbiology) enables the provision of viral and/or antibody testing kits to the US population. For example, as of June 19, 2020, there are a total of 26,781,666 viral tests performed to determine whether an individual is currently infected by the coronavirus (CDC, 2019). Approximately 10% of the test results are positive. Among a sample of 1,934,566 individuals with COVID-19, most of them are within 18–44 and 45–64 age groups (41.4% and 32.8%, respectively). For the rest, 5.1% and 9.5% are aged 0–17 and 65–74, respectively, and 11% of them are above 75 (CDC, 2019). In general, when the age of patients increases, the hospitalization rate also becomes higher. Hospitalization rate is the ratio between the number of individuals who are hospitalized within 14 days after a positive viral test and the total population in a spatial region. As shown in Table 1, the overall cumulative hospitalization rate is 94.5 per million (CDC, 2019). For people aged 50–64 and above 65, the rates increase to 143 and 286.9 per million, respectively. However, for people aged 0–4 and 5–17, the rates declined to 7.4 and 3.5, respectively.

The upsurge of positive cases poses significant challenges on the hospital capacity. As shown in Table 2, as of June 18, 2020, 70% of inpatient beds are occupied, in which 5% is used for COVID-19 patients. Also, nearly 63% of intensive care units (ICU) beds are occupied (CDC, 2019). In addition, the shortages of medical supplies (e.g., personal protection equipment (PPE)) become more and more prevalent in the health systems with a rising number of coronavirus cases and hospitalizations. In the era of globalization, US medical supplies are heavily dependent on importation, nearly 72% of active pharmaceutical ingredients (APIs) are imported from other countries. Specifically, approximately 13% of medical products are from China, and 18% of pharmaceutical imports are provided by India (COVID-19: Impact on Global Pharmaceutical and Medical Product Supply Chain Constraints U.S. Production, 2019). Also, generic drugs imported from these two countries account for about 90% of medicine supplies in the US. However,

Table 1 A summary of cumulative hospitalization rate for each age group

Age Group	Hospitalization rate per million
Overall	94.5
0–4 years	7.4
5–17 years	3.5
18–49 years	56.5
50–64 years	143.0
65+ years	286.9

Table 2 National estimates of hospital bed occupancy in the COVID-19 in the US

Estimates for June 18	Number (95% CI)	Percentage (95% CI)
Inpatient Beds Occupied (all Patients)	524,610 (500,844–548,376)	65% (64–66%)
Inpatient Beds Occupied (COVID-19)	40,112 (37,682–42,541)	5% (5–5%)
ICU Beds Occupied (all Patients)	77,029 (72,135–81,922)	63% (61–64%)

the COVID-19 outbreak in January shuts down almost all manufacturing facilities and non-essential businesses in China. Even though manufacturing activities were resumed in late February, the average capacity utilization at top 500 manufacturing enterprises in China was only 58.98% (Fernandes, 2020; ISM Report on Business, 2019). As such, a disrupted supply chain causes serious shortages of medical products in the US, which endangers the healthcare workers in the front line.

Indeed, healthcare workers are among the most vulnerable group of people who face a higher probability to get infected during the epidemic outbreak. The higher risk is due to their closer contact with patients, the shortage of PPEs, the delay of testing program in the early stage, and the high infection rate in the hospital. As the COVID-19 proliferates, healthcare workers suffer from occupational burnout and fatigue. The key factors include occupational hazards, emergence responses, process inefficiencies, and financial instability (Sasangohar et al., 2020; Shechter et al., 2020; Greenberg et al., 2020). During the period of February 12–April 9, 2020, approximately 19% of COVID-19 patients are healthcare workers. Therefore, this fact further exacerbates the shortage of staffing in the hospital. To avoid secondary infection in the hospital, screening and masks are required for all people upon entry into the hospital (Bartoszko et al., 2020). Patients with suspected or confirmed COVID-19 are placed in a single-occupancy room with a closed door and a separated bathroom. Also, all healthcare workers should wear PPE, isolation gowns and non-sterile gloves upon entering these patients' room. When transporting patients out of the room, both patients and healthcare workers should wear PPE. Moreover, hospitals conduct routine cleaning and disinfection procedures. Enhanced environmental cleaning and disinfection are preferred for rooms used by patients with suspected or confirmed COVID-19, and for areas used by healthcare workers who care for such patients (Chirico et al., 2020).

2.2 Economic Challenges

The COVID-19 epidemic made the nation shut down non-essential businesses, schools and instituted travel bans, which have greatly impacted the U.S. economy. The shocks to supply chain bring significant disruptions to manufacturing. Small and medium manufacturing enterprises faced unprecedented challenges, while some have to shut down entirely to mitigate the virus spread. With social distancing measures in place, many workers can only work from home. The production

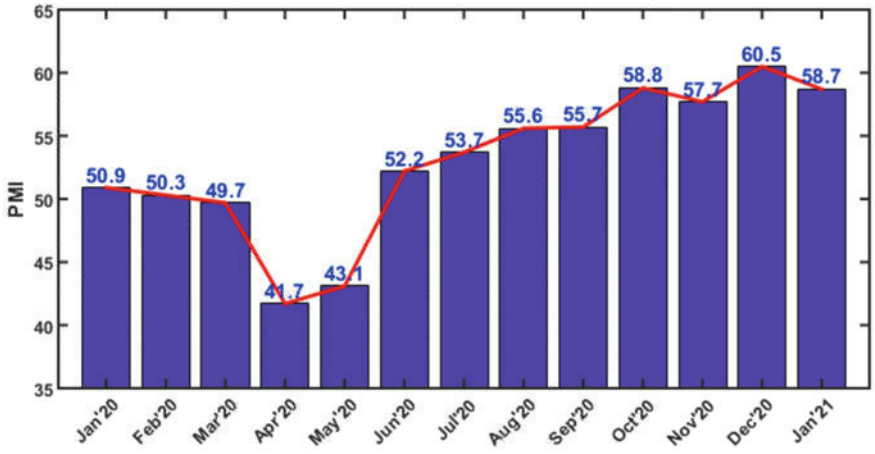

Fig. 2 The variations of Purchasing Manager's Index (PMI) from January to May 2020

lead time has doubled due to shortages of workers and materials (ISM Report on Business, 2019). Also, a limited number of products can be distributed worldwide by air or ocean because of trade wars, hiking tariffs, and importation restrictions. All these impacts of COVID-19 make companies question the just-in-time strategy and reconsider the design of supply chain. In March 2020, there was a 6.3% drop in manufacturing production, which was the largest 1-month drop since 1946 (ISM Report on Business, 2019; Bonaccorsi et al., 2020). The drop was even larger for April 2020. Note that the Purchasing Manager's Index (PMI) shows the impacts of COVID-19 on the economy. PMI is a composite index, ranging from 0 to 100, of economic activities including new orders, inventory levels, production, supplier deliveries, and employment. If the PMI is above 50, the manufacturing sector is generally expanding. If PMI is below 50, it is generally contracting. As shown in Fig. 2, US economic growth is strong in January 2020 with PMI 50.9, but decreases from January to April 2020 (ISM Report on Business, 2019; Bonaccorsi et al., 2020). When the COVID-19 outbreak occurred in March 2020, the PMI fell below 50, further dropped to 41.7 in April 2020, and then remained low through May 2020. From March to May 2020, COVID-19 poses significant challenges on the US economic activities due to unexpected outbreaks, lockdowns, and non-pharmaceutical interventions. After June 2020, the US economical activities recover with the rollout of stimulus plans, increasing manufacturing productions, and new modes for businesses such as teleconferencing, e-commerce and online learning.

A worse impact on the manufacturing industry during the epidemic would be caused by decreased spending because of job loss or reduced incomes. The disruption in the manufacturing industry and the tremendous drop in demand led to the layoff of workers. As of May 2020, the unemployment rate in the manufacturing industry increased to 11.6%. Table 3 summarizes the number of employees in the manufacturing sector as issued by the U.S. Bureau of Labor Statistics, for both

the non-seasonally adjusted case and the seasonally adjusted case (Manufacturing: NAICS 31-33, n.d.). As shown in Table 3, when it is not seasonally adjusted, the number of employees in the manufacturing sector decreased by 1.32 million from March 2020 to April 2020, with about 0.90 million in durable goods manufacturing and 0.42 million in non-durable goods manufacturing. Meanwhile, there were about 1.13 million fewer jobs in May 2020, compared to May 2019. When it is seasonally adjusted, the U.S. manufacturing lost about 1.29 million jobs from March 2020 to April 2020. About 69% (0.91 million) of the job loss was in the durable good manufacturing, while the rest 31% (0.38 million) was in the non-durable good manufacturing. Compared to May 2019, there were 1.12 million fewer jobs in May 2020 (Manufacturing: NAICS 31-33, n.d.).

Schools and universities across the country have also been disrupted. In March 2020, most schools started to switch from in-person instruction to online-only instruction, which gave rise to the concerns about instruction quality (Crawford et al., 2020). Meanwhile, it is not uncommon that many universities faced financial challenges. As students moved out of on-campus housing, universities issued pro-rated refunds to them, which was a substantial amount of unexpected expenses. Also, universities needed to allocate additional funds for dorm cleaning and technology essentials for online classes. Moreover, due to the cancellation of college entrance exams worldwide and limitation on travel, the enrollment for the fall 2021 semester is likely to drop, which will also cause financial issues to universities.

These paramount challenges posed by epidemics call upon multiple scientific disciplines to design and develop new enabling methods and technological innovations for rapid response and management. For example, a complete picture of the new virus is urgently needed from the community of medical scientists. The manufacturing community should be agile to innovate the design and increase the production of personal protective equipment (PPE). In this paper, we propose a system informatics approach for data-driven epidemic response and operational management, thereby mitigating the risks and controlling the virus spread. In the following sections, "**Measure**" provides statistical methods for optimal sampling and testing of the population for the presence of virus, as well as a review of data management and data visualization methods. "**Analyze**" focuses on the handling and analysis of heterogeneous and interconnected datasets (e.g., from CDC, Census Bureau, Food and Drug Administration, state and federal health departments) that are collected during the epidemic lifecycle. "**Improve**" exploits data-driven knowledge to improve the resilience design of health systems, including healthcare capacity, resources, workflows, and operations. Further, "**Control**" focuses on the learning and optimization of health policies and action strategies for controlling the spread of virus. The system informatics methods and tools will complement medical, clinical and pharmaceutical research efforts, helping safeguard the population from infectious diseases and make health systems more resilient to overwhelming epidemic events.

Table 3 Employees on nonfarm payrolls in manufacturing (in thousands) (Manufacturing: NAICS 31-33, n.d.)

	Not seasonally adjusted				Seasonally adjusted				
	May 2019	Mar. 2020	Apr. 2020	May 2020	May 2019	Mar. 2020	Apr. 2020	May 2020	Change Apr.–May 2020
Manufacturing	12,810	12,747	11,427	11,677	12,829	12,806	11,482	11,707	225
Durable goods	8052	8013	7109	7234	8056	8031	7124	7243	119
Nondurable goods	4758	4734	4318	4443	4773	4775	4358	4464	106

3 Measure the Epidemic Dynamics

The "measure" step is directly aimed at testing the population for the prevalence of virus, which is critical to monitoring the temporal evolution of an epidemic in a spatial region. Rapid advances of gene, microbiology and imaging technologies have greatly improved the design and development of testing methods (e.g., speed and accuracy) of coronavirus and influenza. As discussed in Sect. 2, an epidemic poses paramount challenges on the health and economy of our society. The prevalence of a virus in a large population often incurs large amounts of testing, which leads to spatially-temporally big data. This provides an opportunity for the "analyze" step to develop an in-depth understanding of dynamically evolving statuses of an epidemic. Here, data could be collected in disparate efforts by private companies, research centers, universities, and government agencies, thereby leading to the formation of data cohorts to address issues of data management. Epidemic data can then be visualized in various ways to provide comprehensible information about the spatiotemporal variations of an epidemic. An effective visualization further helps the "analyze" step to estimate and extract salient features for the prediction of future trajectory or the monitoring of transmission risks.

3.1 Testing and Sampling

Clinical testing is a critical first step to stopping the spread, which consists of viral testing (i.e., examine whether an individual is currently infected or not) (Esbin et al., 2020) and antibody testing (i.e., check whether an individual was infected before and currently has the presence of antibodies in the blood) (Lipsitch et al., 2020). In the case of COVID-19, specimens are often collected through swabs in the nose or throat for the viral testing. If specimens show the existence of a virus's ribonucleic acid (RNA) or proteins, the test will be positive. The antibody testing is typically done by collecting a sample of blood serum and then examining the presence of antibodies. In order to monitor the prevalence of virus, testing can be performed in three different ways as follows:

- **100% testing**: Population is the entire collection of individuals of interests in a region of interest (e.g., university, city, county, or state). If the cost is not a concern, 100% testing makes sure everyone is tested and then all the infected individuals can be isolated and quarantined. This is an effective approach to stop the spread, but often encounters practical limitations such as inadequate supply of testing kits, prohibitive cost, and population instability due to mobility and immigration.
- **Acceptance sampling**: Sample is a representative subset of the population that can be tested for statistical inference. Acceptance sampling, also called Lot Quality Assurance Sampling (LQAS) (Hedt et al., 2012), is a middle ground between 0% and 100% testing and requires a small sample size for population

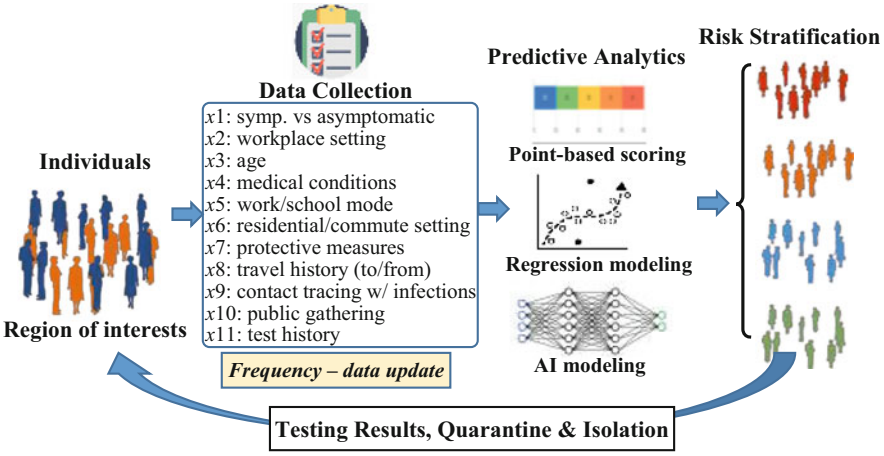

Fig. 3 Data-driven risk scoring systems for categorized sampling and testing

surveys. The population can be stratified into sub-groups (or lots), and each lot can be sampled for clinical testing so as to "accept" or "reject" the lot according to the risk tolerance levels. Also, these samples can be aggregated to establish the confidence interval of infected proportion for testing the hypothesis on the prevalence of an epidemic virus.

- **0% testing**: This means that no testing will be done for the individuals in a specific region. In the onset of an epidemic, few tests are performed because the new virus is just emerging and has not caught enough attention from the public. Once the epidemic virus is captured (e.g., genome sequenced and shared), testing kits can then be designed and developed.

Figure 3 shows that mobile or web-based applications can be used for data collection from individuals in a spatial region of interests, if the testing capacity is constrained and 100% testing cannot be implemented. Examples of the predictors may include x_1: symp. vs asymptomatic; x_2: workplace setting; $x3$: age; $x4$: medical/comorbidity conditions; $x5$: work/school mode; $x6$: residential/commute setting; $x7$: protective measures; $x8$: travel history (to/from); $x9$: contact tracing with infections; $x10$: public gathering; $x11$: test history; The response variable will be the risk probability of infection (range from 0 to 1). The data-driven decision support system helps stratify the individuals into groups (or lots) and then optimize the testing decisions. The risk scoring system categorizes the population into different groups with various levels of risk probability. For example, four groups can be stratified based on the risk probability, which helps further optimize the allocation of testing resources and identify the infected individuals for isolation and quarantine.

As shown in Fig. 3, risk scoring systems can be established in three different ways, namely point-based systems, regression modeling, or AI-based modeling. Such scoring systems help categorize the acuity levels of patients and then improve

the quality of healthcare services (e.g., surgical procedures, medication usages, care guidelines, treatment plans, and resource allocations) (Chen & Yang, 2014; Imani et al., 2019). Point-based scoring systems use the simple points or weights, and can be easily implemented in questionnaire form. The points or weights can be adjusted for different predictors (or factors). For example, if the symptom is weighted more than other predictors, it may be assigned with a larger point (or weight). In clinical practice, point-based scoring systems are widely used to stratify the patients, e.g., Acute Physiology and Chronic Health Evaluation (APACHE) (Zimmerman et al., 2006), Sequential Organ Failure Assessment (SOFA) (Raith et al., 2017), Simplified Acute Physiology Score (SAPS) (Metnitz et al., 2005; Moreno et al., 2005), and Mini-mental state examination (MMSE) (Galasko et al., 1990). Figure 3 shows an example of risk factors for the design of point-based scoring systems, which also helps reduce the number of variables to compile into a short survey. An increasing score indicates a higher risk of infection. In addition, the infection risk can be derived using a multivariate logistic regression model as: $\log\left(\frac{risk}{1-risk}\right) = a + \sum_i b_i x_i$, where *Risk* is the risk of death, $\left(\frac{risk}{1-risk}\right)$ is the odds ratio, a is the intercept, b_i is the coefficients and $x_i^{'}s$ are independent predictors. Here, training data or medical domain knowledge can be used to adjust the regression coefficients for different predictors (or factors). Finally, it is not uncommon that AI modeling (e.g., neural networks) are utilized to learn from complex-structured data for risk stratification. AI models, however, need large amounts of data for training and learning the weights, and are difficult to implement for testing and sampling in an epidemic.

Statistical sampling is a cost-effective approach to survey the groups (or lots) of individuals when the testing capacity and supply chain are constrained. First, the confidence interval for the proportion of infections p can be estimated from testing data. If there are c infected individuals for a random sample of size n, then an approximate $100(1-\alpha)\%$ confidence interval for p is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} \tag{1}$$

where $\hat{p}$ is c/n, and $z_{\alpha/2}$ is the z value with an upper tail area of $\alpha/2$. This estimation tends to be more reliable when the number of confirmed individuals c is greater than 6 in the sample, and is also applicable in the case of hypergeometric distribution when the sample size n is small. Here, the choice of sample size is dependent on the significant level α and the margin of error (MOE), i.e., $z_{\alpha/2}\sqrt{\hat{p}\left(1-\hat{p}\right)/n}$. If a specific MOE value e is desired, then the sample size n is approximately $z_{\alpha/2}^2\hat{p}\left(1-\hat{p}\right)/e^2$. Note that the function $\hat{p}\left(1-\hat{p}\right)$ reaches the maximum 1/4 when $\hat{p} = 1/2$. Hence, the MOE is guaranteed not to exceed e if the sample size is chosen to be $z_{\alpha/2}^2/4e^2$. For example, it is 95% confident that the MOE will not exceed 0.02 when the sample size is $1.96^2/(4 \times 0.02^2) = 2401$.

Acceptance sampling is useful to help the decision-making process on whether or not to lockdown or reopen a region (or "lot") for regular businesses. As shown in Fig. 4a, the operating characteristic (OC) curve describes an acceptance sampling plan in terms of the probability of reopening versus the proportion infected. For example, the probability of reopening is $1 - \alpha$ if the region meets the acceptance risk level (ARL) p_{ARL}. The probability of reopening is β if the region is on the rejection risk level (RRL) p_{RRL}. Assuming a binomial distribution, the sample size n and acceptable number a can be obtained as:

$$1 - \alpha = \sum_{c=0}^{a} \frac{n!}{c!\,(n-c)!} p_{ARL}^{c} (1 - p_{ARL})^{n-c} \tag{2}$$

$$\beta = \sum_{c=0}^{a} \frac{n!}{c!\,(n-c)!} p_{RRL}^{c} (1 - p_{RRL})^{n-c} \tag{3}$$

Then, for this acceptance sampling plan, if there are more than a infections in the random sample of size n from the region, lockdown will be implemented. If there are less than or equal to a infections, the risk is below the ARL level and the region can be reopened. For example, Fig. 4b shows the acceptance sampling plans with $n = 2000$ and a is ranging from 15 to 95. When the acceptance number a increases, this does not significantly change the slope, but rather move the OC curves to the right. If the acceptance number a is small, the risk tolerance levels tend to be low. For larger values of a, both ARL and RRL levels are higher. If a region is above the RRL, NPIs such as lockdown and stay-at-home should be implemented. On the other hand, rectification testing programs can further screen individuals in the rejected region. Often, 100% testing can be performed to identify all the infected individuals, then isolate and quarantine them.

In the practice of clinical testing, acceptance sampling may have the following limitations. First, if the sample size is finite, then the distribution tends to be hypergeometric instead of binomial. However, binomial approximation of hypergeometric is valid if the ratio between sample size and lot size is less than 1/10. Second, acceptance sampling assumes the selection of samples at random from each region. Although clinical testing is prioritized for symptomatic cases or traced contacts of infected individuals, it can however assume that the infection of an individual is at random. Then, clinical testing can be assumed to be implemented on individuals who are infected at random, albeit with the introduction of bias to some extent. Third, individuals are assumed to be homogeneous in a region. In other words, homogeneity refers to the fact that the probability to get infected is approximately the same if in contact with pathogens. This is a reasonable assumption for a susceptible population, although there may be slight differences in the infection probabilities for uncontrollable factors such as age groups and blood types. These limitations and assumptions should be considered during the practice of acceptance sampling for clinical testing.

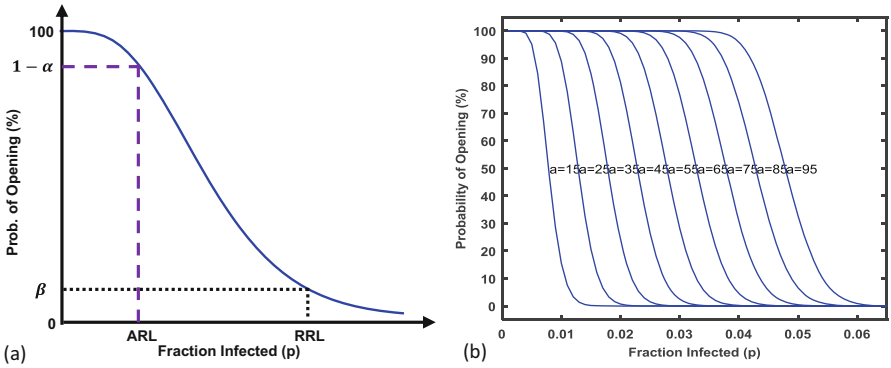

Fig. 4 (**a**) An illustration of operating characteristic (OC) curve, (**b**) OC curves of acceptance sampling plans with the sample size $n = 2000$ and the acceptance number a is ranging from 15 to 95

3.2 Spatiotemporal Surveillance of Epidemic Processes

Clinical testing brings significant amount of data pertinent to the evolution of an epidemic. The epidemic data may include total cumulative cases (or per capita), daily new cases, total deaths for multiple spatial regions (or lots) of interest and are dynamically changing over time. Therefore, the epidemic evolution is a spatiotemporal process, i.e., varying in both space and time. The availability of data provides a great opportunity to design monitoring charts and develop epidemiology surveillance programs. Statistical monitoring methods help health systems leverage sequentially observed data to trigger the alarms and identify the outbreak region. However, raw data are often not normalized and cannot be directly used to develop monitoring charts. For example, spatial regions often have different population sizes. Total cases should be adjusted for the population in a region. As such, features need to be extracted from the data to describe the epidemic characteristics in a region. Examples of features may include cases per million, the incidence rate, or transmission risk index that are characterized with data-driven models.

If the monitoring objective is to detect abnormal changes of incidence rates x_1, x_2, ..., x_k over k regions, then the feature vector will be $\boldsymbol{x} = [x_1, x_2, \ldots, x_k]^T$. The statistical test is aimed at setting up the null and alternative hypotheses, then seeking data-driven evidence to determine whether an anomaly is present in any dimension (i.e., a region) of the feature vector or not. Under the null hypothesis H_0, the incidence rates over k regions do not change over time. As such, the feature vector $\boldsymbol{x}$ is assumed to follow a multivariate normal distribution with population mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, i.e.,

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{4}$$

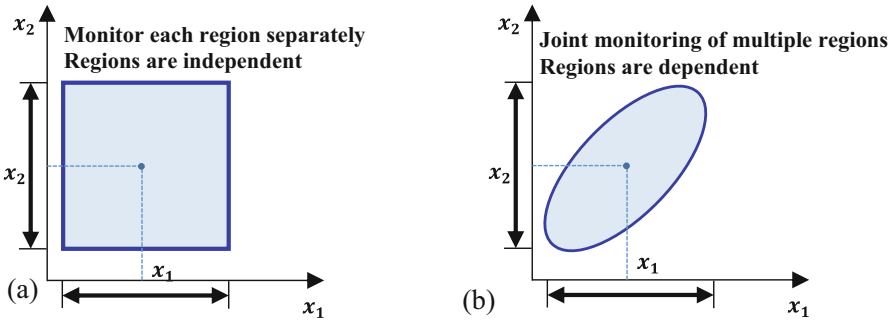

Fig. 5 Multivariate monitoring schemes for epidemic surveillance: (**a**) Monitor each region separately and regions are independent, (**b**) Joint monitoring of multiple regions and regions are dependent

If an outbreak occurs in one region or multiple adjacent regions, then the assumption of multivariate normal distribution is no longer valid. The alternative hypothesis H_1 that the joint distribution of multivariate features is non-normal will tend to hold. The hypothesis test accepts or rejects the null hypothesis H_0 at a significance level α. Although the assumption of multivariate normality is required to formally establish confidence limits in the statistical test, a slight deviation will not severely impact the results (Chen & Yang, 2016a). Here, multivariate normal probability plotting can be used to evaluate whether the extracted features of incidence rates are approximately normally distributed for multiple regions of interests.

As shown in Fig. 5a, most of traditional monitoring schemes assume that k regions are independent. Therefore, a common approach is to monitor each feature independently in the literature. In the bivariate case, control limits will form a rectangular region. If the pair of observations fall within this rectangular region, then the null hypothesis H_0 holds. If the pair of observations reside outside this region, then the null hypothesis H_0 is rejected. However, this monitoring scheme has limited applications due to the "curse of dimensionality". For example, if the probability of Type I error is α for each feature, then Type I error for monitoring k features independently is $1 - (1 - \alpha)^k$. The probability that all k observations fall within the confidence limits is $(1 - \alpha)^k$ if all the k regions are in control (Yang & Chen, 2014; Chen & Yang, 2015). Hence, the error is significant when the dimensionality of the feature vector increases. It may also be noted that k features are oftentimes not independent because adjacent regions tend to be correlated with each other in an epidemic situation.

Therefore, multivariate statistical methods that consider spatial correlations and jointly monitor these regions (or features) are urgently needed. As shown in Fig. 5b, due to the correlation among adjacent regions, the pair of observations now resides in the elliptical region for the bivariate case. Under the null hypothesis H_0, k regions will follow the multivariate normal distribution with the population covariance matrix $\boldsymbol{\Sigma}$. As such, the test statistic $\chi^2 = (\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ follows

a chi-square distribution with k degrees of freedom. The joint distribution changes in the presence of regional anomalies. If there are shifts in at least one out of k regions, then χ^2 values will be above the upper control limit $UCL = \chi^2_{\alpha,p}$, where α is the significance level. If χ^2 values are below the upper control limit, then the null hypothesis H_0 holds and there will be no significant evidence of anomalies. The control ellipse of bivariate case in Fig. 5b is due to region-to-region correlations. Because off-diagonal elements are no longer zero in covariance matrix $\boldsymbol{\Sigma}$, the principal axes of the ellipse are not parallel to the $\overline{x}_1$, $\overline{x}_2$ axes any more.

In the real world, population mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ are often unknown and need to be estimated from the data. If the sample mean $\overline{\boldsymbol{x}}$ and covariance matrix $\boldsymbol{S}$ are used instead, then the test statistic becomes $T^2 = (\boldsymbol{x} - \overline{\boldsymbol{x}})' \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}})$, which is commonly called as the Hotelling T^2 statistic (Mason et al., 1997; Li et al., 2008). The new UCL for the Hotelling T^2 statistic is:

$$UCL = \frac{p\,(N+1)\,(N-1)}{N^2 - Nk} F_{\alpha,k,N-k} \tag{5}$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N$ are N sequentially observed samples of epidemic data from k regions, $F_{\alpha,\,k,\,N-k}$ is the upper $100\alpha\%$ critical point of F distribution with k and $N - k$ degrees of freedom. Note that control limits are established in Phase I with in-control datasets (i.e., without the presence of anomalies). For Phase II monitoring, the control chart plots control limits and the test statistic $T^2(i)$, $i = 1, 2, \cdots N$ for each sample. When a new sample arrives, we will then compute the test statistic and check the conformance in the control chart. Note that it is not feasible to graphically construct the control ellipse for more than two regions as shown in Fig. 5b. The composite index (i.e., Hotelling T^2 statistic) helps characterize the multivariate distribution of k features (or regions), and further establish the control chart to effectively detect whether there are shifts in at least one out of k regions (i.e., multivariate epidemic monitoring and surveillance).

3.3 Data Management and Visualization

As the epidemic progresses, large amounts of data are organized in the form of data cohorts or lakes. Medical scientists collect pertinent data about the clinical picture of a new virus for the development of effective intervention methods, such as antivirals and vaccines. Epidemiologists and engineers leverage the public health data to develop analytical models for the prediction of virus spread dynamics. Real-time data of epidemic situations is critical to understand the spread, trace the contacts, and control the propagation. Data management is indispensable to integrate disparate data efforts from government agencies, universities, and private companies. Here, data cohort connects various organizations to manage the data using the defining characteristics, which help researchers save tremendous amount of time in finding, analyzing, evaluating and validating relevant data for useful

information and insights to stop the epidemic. Nonetheless, data lake is a repository of unorganized data in the raw format. Data cohort may include necessary data from on-going and completed research, as well as contact tracing data. This type of data could contain the patient location, sociodemographic information, and the list of contacts during the elicitation window and where the patient has visited. When the number of infections become prevalent, data management gets increasingly complex. This is partly due to the large number of cases, as well as the long list of traced contacts of each positive case. Data management depends on the use of database systems to support such many-to-many relational tables and provide a higher level of flexibility of routine data storage, update, security, reporting, and On-Line Analytical Processing (OLAP).

Note that the epidemic data is varying in both space and time. Table 4 provides examples of data repositories and cohorts developed by government agencies, institutions, and private companies. These data cohorts are open access to the public or limited access by applications. The UN data lab, US CDC and European Centers for Diseases Control (ECDC) organize and publish the real-time position data of virus spread in either country level or county level. Such information can be used to study and track the spread of the disease. US National Science Foundation (NSF) supported a research project to develop the COVID Information Commons, which is an open website to promote data and knowledge sharing across different COVID research efforts. National Institute of Health (NIH) initiated an National COVID Cohort Collaborative (N3C) project for collaboration on data collection, sharing, and analytics, which also provides the open access to research literature about COVID-19 genomics, virus structures, and clinical studies.

Also, academic institutions such as John Hopkins University (JHU) and the University of Washington provides the organized COVID-19 data and popular dashboards for data visualization. This, in turn, greatly facilitates the general public in visualizing the spread and trend of epidemic, thereby promoting situational awareness. In addition, there are data cohorts from private companies and foundations that provide targeted information about the disease. For example, the COVID-19 tracking project assembles the testing data, hospitalization rates, treatment outcomes, race and ethnicity data for researchers to investigate the outbreak scale, the mortality rate, and regional effects of the disease. COVID-19 Open Research Dataset (CORD19) provides an application programming interface (API) to retrieve the infection data, research feed, and COVID related texts. This API can help researchers query data in a fast manner. Surgo Foundation provides the community vulnerability index, social distance tracking, and nurse sentiment data to help develop analytical methods and tools for epidemic response.

Large amounts of data are readily available from different sources. The next step is to visualize and represent the data so that useful information and salient features can be easily comprehensible by the audience. Data visualization focuses on compact representations of trends and patterns in the data with graphical methods and tools such as time series charts, density graphs, and heat maps. The human brain can perceive information in graphics and images better than pale texts or data tables. An effective visualization helps condense a thousand words in one picture.

Table 4 Examples of COVID-19 data repository/cohort and features

Data cohorts and repositories	Descriptions and features
Center for Disease Control and Prevention (CDC) https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/	US infection data with cases, race, ethnicity, testing, hospital capacity and other data streams at local, state, and national levels
World Health Organization https://www.who.int/	Global case updates with total confirmed cases and deaths, new cases and deaths, and transmission classifications
European CDC https://www.ecdc.europa.eu/en/covid-19-Epidemic	COVID-19 situation updates, case counts and distributions for the EU/EEA, UK, and worldwide.
National Institutes of Health https://datascience.nih.gov/covid-19-open-access-resources	COVID-19 data and resources such as official data, related studies, and high-performance computing consortium
National COVID Cohort Collaborative (N3C) https://cd2h.org/	A very large patient-level COVID-19 clinical dataset shared by CTSA, CD2H and other distributed clinical data networks
Clinicaltrials.gov https://clinicaltrials.gov/ct2/results?cond=COVID-19	Detailed information about active and recruiting clinical trials such as intervention and phase
Johns Hopkins University https://github.com/CSSEGISandData/COVID-19	Global and US daily situation update at country and state level, along with time-series summary
NSF COVID Information Commons https://covid-info-commons.site.drupaldisttest.cc.columbia.edu/	Open website to facilitate knowledge sharing and collaboration focused on NSF funded COVID rapid response research projects
New York Times https://github.com/nytimes/covid-19-data	US state level and county level situation updates, with historical and live data
Twitter Dataset https://github.com/thepanacealab/COVID-19_twitter	Tweets and retweets data acquired from Twitter stream related to COVID-19 chatter with all languages
The COVID Tracking Project https://covidtracking.com/data	US infection data with cases, tests, hospitalized, severity (in ICU, on ventilator, etc.), and outcomes
CORD-19 https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge	A dataset of over 167,000 scholarly articles about COVID-19, SARS-CoV-2 and related coronavirus
Ding Xiang Yuan https://ncov.dxy.cn/	Global case updates with active, confirmed, recovered. China regional case updates with city level native/imported counts
OPENICPSR https://www.openicpsr.org/openicpsr/search/COVID-19/studies	Data cohort which contains links to US state policy database, government response dataset, and COVID-19 impact survey