OEINSIGHTS
Evidence, Analytics, Decisions

The Promise of Big Data: Conquering COVID-19 with Data and Intelligence

November 7, 2020 | Article No. 27

Contributors

Ayesha Siddiqua MSc PhD

Mohit Bhandari MD FRCSC PhD
Editor-in-Chief, OrthoEvidence

Insights

- Big data has played an integral role in detecting the initial COVID-19 outbreak and subsequently monitoring the scope of the pandemic around the world.
- Analysis based on artificial intelligence can parse through large volumes of data and generate insights in a timely manner.
- Identifying trends in big data can help inform a wide range of strategies in diagnosis, public health, clinical decision making, and therapeutics to curb the impact of the pandemic.
- Several countries have leveraged the power of smart-phone based trackers which collect data at the population level to control the spread of SARS-CoV-2.
- There are still several barriers preventing optimal use of big data, which include lack of availability of linked population level datasets in diverse areas, as well as administrative and legal regulations which limit access to data.
- Our acronym **C.A.P.A.C.I.T.Y.** identifies steps we can take to harness the power of big data to win the battle against COVID-19.

> "Amid the daily news churn, policy makers seem to be facing an impossible choice between saving lives and saving livelihoods."

> "The next step is developing smart prevention capabilities rather than requiring blanket isolation and shutdowns. Our window is short — measured in months — for heading off what Bill Gates has characterized as potentially a once-in-a-century pandemic like the 1918 Spanish Flu, which killed at least 50 million people around the world. We have many technological advantages over those fighting that pandemic a century ago. In many ways, this is our most meaningful Big Data and analytics challenge so far. With will and innovation, we could rapidly forecast the spread of the virus not only at a population level but also, and necessarily, at a hyper-local, neighborhood level."

-Shah & Shah, 2020 (1)

First things first: What is Big Data and Artificial Intelligence?

Before we can appreciate the true potential of big data and artificial intelligence (AI), it is important to review some relevant terminology and concepts. While there have been different definitions for big data and AI over the years, there are now some common definitions that are widely used across industries. Exhibit 1 provides an overview of big data and AI to contextualize our discussion to follow. While there are some differences in opinions regarding the exact size of a dataset that qualifies it as big data, typically big data refers to data storage bigger than one terabyte and can range in size from terabytes to zettabytes (2). Exhibit 2 provides examples of different sizes of big datasets that are currently used in diverse industries.

1 kilobyte (KB): 1000 bytes
1 megabyte (MB): 1000 kilobytes
1 gigabyte (GB): 1000 megabytes
1 terabyte (TB): 1000 gigabytes
1 petabyte (PB): 1000 terabytes
1 exabyte (EB): 1000 petabytes
1 zettabyte (ZB): 1000 exabytes

Exhibit 1: Descriptions and Key Characteristics of Big Data and Artificial Intelligence (2-7)

Big Data	
Definition	Big data is used to describe large data sets whose size or type is beyond the capabilities of traditional relational databases to capture, manage, and process data with low latency. Big data also requires advanced analysis methods to analyze different types and sizes of data that is not always possible with traditional analysis methods.
Importance	The value of big data does not depend on how much data there is, but rather what is done with the data. Big data can be analyzed to generate insights for a wide range of industries.
Key Characteristics	
Volume	Data can be collected from a wide range of sources at organizations. Historically, storage for a large volume of data has been a challenge, however, newer and cheaper platforms such as data lakes provide effective solutions.
Velocity	With the growth of the Internet, data is now streamed into organizations at unprecedented speed which should be handled and managed in a timely manner.
Variety	Data can come in a wide range of formats, such as structured numeric data or unstructured text files, emails, videos, and audios, as well as semi-structured data.
Variability	Data flows can be unpredictable – they can change often and vary greatly. This requires tools that can detect changes in data patterns quickly.
Veracity	Since data can come from many different sources, they can be difficult to link, match, clean, and transform given the differences between those sources. Effective strategies for data linkages are needed to manage data.
Artificial Intelligence	
Definition	This refers to machines that respond to stimulation consistent with traditional responses from humans, with human capacity for contemplation, judgement, and intention. These include software systems that make decisions requiring human level of expertise and have strong predictive capabilities.
Importance	AI allows machines to learn from experience, adjust to new inputs, as well as perform human-like tasks. It is particularly helpful when AI can be used to train computers to process large amounts of data and recognize patterns in the data.
Key Characteristics	
Intentionality	AI algorithms are designed to make decisions using big data that is often collected in real time. They are more advanced than passive machines that can only produce mechanical or predetermined responses. AI uses sensors, digital data, or remote inputs and combines information from a wide range of sources to instantly analyze data and generate insights.
Intelligence	AI methods mimic human decision-making processes by using machine learning based data analytics. Machine learning is used to look for underlying trends in data using algorithms that can identify patterns. Deep learning is a subset of machine learning that uses artificial neural networks and algorithms inspired by the human brain to learn from large amounts of data. Just as humans learn from experience, deep learning algorithms perform a task repeatedly, while refining performance each time to improve the outcome.
Adaptability	Systems that use AI have the ability to learn and adapt as they continue to make decisions without human development. For example, semi-autonomous vehicles using AI can let drivers know of upcoming traffic congestion or construction.

Exhibit 2: Examples of Big Data of Different Sizes (8)

Dataset	Description
Social media	This is continually growing in size. E.g. 12+ terabytes of tweets everyday.
British Library UK Website Crawler	~110 terabytes per domain crawl to be delivered
Large Hadron Collider (world's largest and highest-energy particle collider and the largest machine in the world)	13-15 petabytes (2010)
Internet Communications (Cisco)	667 exabytes (2013)
Digital universe	7.9 zettabytes (2015)

> "This immediate burst of cases and their health data have created a vital source of information and knowledge. There is an immediate requirement to store such a large amount of data of these cases, using different data storage technologies. These data are used to undertake research and development about the virus, pandemic and measures to fight this virus and its after-effects. Big data is an innovative technology which can digitally store a large amount of data of these patients."
>
> **Haleem et al. 2020 (9)**

Application of Big Data: Our Only Hope to Win the Battle Against COVID-19?

Since the beginning of the COVID-19 pandemic, both developing and developed countries have faced an unprecedented volume of challenges that have debilitated many of their healthcare systems and economies. In order to rapidly generate mitigation strategies to minimize the impact of the pandemic, many countries have turned to insights generated from big data to inform their planning. Thus far, the applications of big data have been diverse. Big data is currently being used to store population level information on different types of cases affected by COVID-19 – including those infected, recovered, and expired (9). Several countries have implemented tracking systems to anticipate potential disease outbreaks as well as identify different risk factors for the disease (10). As we progress further through the pandemic, the utility of big data has become more apparent to not only help identify a vaccine and

treatment for COVID-19, but also thoroughly study the broader socioeconomic impact of this disease. It is increasingly recognized that **AI** can play a critical role for revealing patterns, trends, associations, and differences in big data – and when this is done in tandem with big data collected in real time – it can provide a powerful tool for making evidence-based decisions in a timely manner (9). Despite the growing momentum and the enthusiasm for harnessing the advantages of big data, its current applications have taught many important lessons that can pave the way for more refined and meaningful uses of this valuable resource beyond the pandemic. Reflecting on the history of threats to population health, there is now consensus that future global health crises are only inevitable. In order to save millions of lives around the world and prevent economies from collapsing again, the time to master strategies to maximize benefits of big data as well as build capacity and infrastructure for its use is now.

> Machines that respond to stimulation consistent with traditional responses from humans, with human capacity for contemplation, judgement, and intention. These include software systems that make decisions requiring human level of expertise and have strong predictive capabilities.

"This is, in essence, a big data problem. We're trying to track the spread of a disease around the world."

"Between recognizing signs and symptoms, tracking the virus, and monitoring the availability of hospital resources, researchers are dealing with enormous amounts of information – too much for humans to comprehend and analyze on their own. It's a situation that is seemingly tailor-made for advanced analytics technologies."

"There are several big data components to this pandemic where artificial intelligence can play a big role."

Dr. James Hendler, the Tetherless World Professor of Computer, Web,
and Cognitive Science at Rensselaer Polytechnic Institute (RPI) and
Director of the Rensselaer Institute for Data Exploration and Applications (IDEA) (11)

> "The start-up (BlueDot), which grew out of Dr. Kamran Khan's research at the University of Toronto, combines natural language processing and machine learning to gather insights on the emergence and spread of infectious diseases around the globe. BlueDot was the first to warn the world of a potentially dangerous new illness – now known as COVID-19 – ringing the alarm before the U.S. Centers for Disease Control and Prevention did on Jan. 6 and before the World Health Organization followed suit three days later."

> "BlueDot will provide the (Canadian) federal government with insights and intelligence to help combat the virus – in part by using anonymous location data from hundreds of millions of mobile devices to see how the public health response is working."

Vendeville, 2020 (12)

Big Data and COVID-19: Progress Thus Far

Big data has already played a tremendous role in curbing the impact of the COVID-19 pandemic, including tracking, controlling, research, and prevention of the disease (9). Around the world, there are increasing initiatives to collect data at the population level or aggregate large data sets that can be parsed using AI (9). In the US, the National COVID Cohort Collaborative (N3C) database was launched to collect data from patients' electronic health records if they have been tested for COVID-19 or if they reported symptoms of this disease (13). This large data repository, which includes over 1.2 million patients, is now available for researchers to use (10). There is also the COVID Symptom Tracker app in the US, which currently has over 4 million users and continues to collect information in real time for large populations (10). While the data from this app are not available to researchers yet, the makers of this app have formed partnerships with those conducting clinical trials and longitudinal studies, highlighting the power of big data collected in real time to inform a wide range of strategies to overcome the pandemic (10).

During this global public health crisis, AI has played an integral role in generating insights from big data. In fact, AI was used to identify the COVID-19 outbreak in the first place. BlueDot, a Canadian company that specializes in infectious disease forecasting, uses an AI engine to continuously gather data on different diseases from a wide range of data sources around the world. Using their AI engine, BlueDot predicted the COVID-19 outbreak and alerted their users even before the World Health Organization (14).

Given the incredible time constraints, finding the right solutions for COVID-19 in a speedy manner is of utmost importance. AI has the power to eliminate false tracks and identify potential targets for solutions faster, instead of trying out 100 or 1000 different options (11). This can expedite the process for finding a vaccine and treatment for COVID-19. With other disease outbreaks, such as the 2009 SARS (H1N1 flu) pandemic, there was no pressing need to identify the effectiveness of social distancing, which is crucial to limit the spread of SARS-CoV-2. It is now recognized that different social distancing techniques may impact the spread of this virus differently in different places as a function of a multifactorial process (11). AI is well suited to distill this process given its ability to learn and analyze a large volume of multifactorial data.

Indeed, since the beginning of the pandemic, AI has been used in a wide range of areas, including diagnosis, public health, clinical decision making, and therapeutics (15). There are already several data-driven AI based tools and models that demonstrated the need for customized COVID-19 interventions given the geographic region, as well as proposed the predicted outcomes of different intervention strategies (16,17). Exhibit 3 provides examples of some AI based tools that are currently being used to combat COVID-19.

Exhibit 3: Examples of AI based Tool for COVID-19 (15)

Diagnosis	**COVNet (18)** ● Fully automatic 3D deep-learning framework developed to detect COVID-19. ● Can extract relevant information from 2D and 3D images gathered from a CT scan to determine a probability score to distinguish patients with COVID-19 from patients with non-COVID-19 community acquired pneumonia. **InferRead CT (19)** ● Deep-learning based diagnostic system that can identify the features of SARS-CoV-2 infection in CT scans of patients with false-negative reverse-transcription polymerase chain reaction (RT-PCR) results.
Public health	**BlueDot's AI Engine** ● Predicted outbreak of COVID-19 and currently monitoring its progress worldwide. AI-based prediction model for epidemiological trends of COVID-19 by Yang et al. (2020) (20) ● This model is trained based on the 2003 SARS epidemic data from China and includes epidemiological parameters of COVID-19, as well as public health interventions in Hubei, China. ● This model was used to predict when COVID-19 cases would peak in Hubei.

Clinical decision making	**AI-augmented systems for early identification of at-risk patients** ● These systems have been used for plain chest radiographs to track and predict the pulmonary progression of hospitalized patients with COVID-19, which helped identify those requiring critical care (21). ● There is evidence that AI models can predict deterioration at initial presentation, with greater accuracy compared to traditional logistic regression (22).
Therapeutics	AI based technology has been used to screen 1.3 billion compounds from the ZINC15 library to determine if any drugs can be repurposed (23). AI based technology also has been used to predict binding affinity values between currently available antiviral drugs and target proteins on SARS-CoV-2 (24).

Several countries have already extensively leveraged the power of big data from diverse sources and used AI-based strategies to contain the COVID-19 outbreak. In particular, both China and South Korea have implemented smart-phone based tracking programs that have critically informed the implementation of intervention strategies. Health Barcode (from China) and COVID-19 SMS (from South Korea) are both smart-phone based trackers that have been made possible as a result of public and private partnership (25). These trackers facilitate COVID-19 risk management using individual self-reported health status and travel history in combination with big data from aviation, railway and ground transportation systems, social media, COVID-19 database, mobile GPS, as well as payment records to retrace the movement of an individual (25). AI and machine learning algorithms are then applied to retrace the movement of an individual with the disease and all people they came in contact with, while determining risk of infection in these people (25). By identifying hotspots where lots of people have high risk of infection, public health interventions are applied to limit the spread of SARS-CoV-2 (25).

"While AI and other analytics technologies appear to be the best possible tools for assessing and mitigating a global pandemic, researchers can't always access what they need to build these models."

"The ideal data is hospital data that would tell us who is experiencing certain impacts from the virus…For example, one project we'd love to do would be to correlate environmental or genomic factors to the people who are getting advanced respiratory problems, which is what's killing most people with this disease. Is there a genetic component to that? Is it something where environmental factors are some kind of comorbidity? But we can't get that kind of data because of HIPAA restrictions."

Dr. James Hendler, the Tetherless World Professor of Computer, Web, and Cognitive Science at Rensselaer Polytechnic Institute (RPI) and Director of the Rensselaer Institute for Data Exploration and Applications (IDEA) (11)

Lessons from Using Big Data: We Still Have a Long Way to Go

Despite the many successes with big data during the COVID-19 pandemic, many challenges and opportunities for improvements have been identified along the process. One of the biggest takeaway lessons from this global health crisis is the need for population level linked databases that collect a wide range of data on a continuous basis. This will allow continuous AI-based analysis to generate insights, foresee any potential disease outbreak in the near future, and prepare accordingly. Planning and capacity building ahead of time, particularly when the collection of individual level data in real time is involved, is of utmost importance. Specifically, the success of smart-phone based trackers depends on a large uptake along with strong public health enforcement, as both contact tracing and isolation are individualized endeavors (25). For example, for the UK's 'Test, Track and Trace' strategy to be effective in controlling the spread of the virus, it has been estimated that at least 60% of the residents need to use the contact tracing app (25). However, it is difficult to achieve high levels of uptake of any technology within a short period of time. This was demonstrated by many examples such as the Canadian COVID Alert App - although there were 2.2 million downloads within the first month of its launch, this number represents only 15 percent of the population of Ontario, Canada (26). This highlights the importance of being prepared to implement population level data collection strategies as soon as there are alerts of a potential disease outbreak, and way before the occurrence of a pandemic.

To compound the challenge of lack of data, researchers also experience many administrative and legal barriers to access available data which significantly limits the comprehensiveness of the models that are built using AI. Modified data access regulations and open data sourcing which protects individual privacy through anonymity and other related measures can maximize the benefits gained from existing big data. AI-based tracking and prediction strategies are already implemented in many industries (e.g. logistics and transportation companies) – collaborating with them can increase the efficiency and effectiveness of the interventions the public health sector deploys during a disease outbreak. The importance of public and private partnerships is emphasized by the case observed in the UK – where a COVID-19 tracker app by the government had glitches and was replaced in favour of the apps developed by Apply and Google (27). If they formed partnerships earlier and worked in tandem, they easily could have saved precious time during a global public health crisis and arrive at a solution faster.

Big data and analytics provide powerful means for winning the battle against the current COVID-19 and future pandemics. Our acronym **C.A.P.A.C.I.T.Y.** serves as a reminder of steps we can take to harness the power of these resources.

Exhibit 4: Steps to Harness the Power of Big Data

C	Collaborations and partnerships between public and private stakeholders across industries
A	Adaptive public health intervention planning based on emerging insights from big data
P	Prioritizing quality at every stage, including data collection, management, analysis, as well as developing evidence-based solutions
A	Assessing resource requirements for different case scenarios generated by data driven models and planning solutions accordingly
C	Capacity building and infrastructure development for linked population level datasets in diverse areas
I	Integrating perspectives of stakeholders from different disciplines in planning and implementing big data projects, as well as designing solutions
T	Time is critical – planning early and not compromising on safeguards to ensure methodological rigour of research and utility of solutions
Y	You. It starts with your willingness and commitment to learn about and champion big data and its applications in your work.

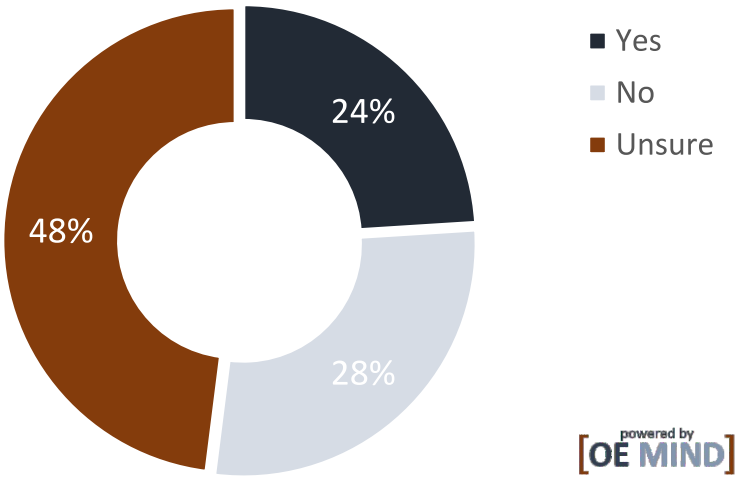
[powered by OE MIND]

OE INSIGHTS
www.myorthoevidence.com

OE Community Perspectives on Big Data

We conducted a poll within the OE community to gain their perspectives on big data analytics. Overall, close to half of the participants voted that they are unsure whether they trust results from big data analytics, whereas 24% of the participants indicated they do trust these results (Exhibit 5).

Exhibit 5: OE Poll: Do you trust results from "Big Data Analytics"?



[powered by OE MIND]

Contributors

Ayesha Siddiqua MSc, PhD
Ayesha Siddiqua has a Masters and a PhD from the Health Research Methodology Program in the Department of Health Research Methods, Evidence, and Impact at McMaster University.

Mohit Bhandari, MD, PhD
Dr. Mohit Bhandari is a Professor of Surgery and University Scholar at McMaster University, Canada. He holds a Canada Research Chair in Evidence-Based Orthopaedic Surgery and serves as the Editor-in-Chief of OrthoEvidence.

References

1.Shah J & Shah N (2020). Fighting Coronavirus with Big Data. Retrieved from https://hbr.org/2020/04/fighting-coronavirus-with-big-data

2.IBM (2020). Big data analytics. Retrieved from https://www.ibm.com/analytics/hadoop/big-data-analytics

3.SAS Institute Inc. (2020). Big Data: What it is and why it matters. Retrieved from https://www.sas.com/en_ca/insights/big-data/what-is-big-data.html

4.Shubhendu S & SJ Vijay (2013). Applicability of Artificial Intelligence in Different Fields of Life. IJSER 1(1): 28-35.

5.SAS Institute Inc. (2020). Artificial Intelligence: What it is and why it matters. Retrieved from https://www.sas.com/en_ca/insights/analytics/what-is-artificial-intelligence.html

6.West DM & Allen JR (2018). How artificial intelligence is changing the world. Retrieved from https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/

7.Marr B (2018). What Is Deep Learning AI? A Simple Guide With 8 Practical Examples. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/10/01/what-is-deep-learning-ai-a-simple-guide-with-8-practical-examples/#5191a2c78d4b

8.Neves PC & Bernardino J (2015). Big Data in the Cloud: A Survey. Open Journal of Big Data 1(2): 1-18.

9.Haleem A et al (2020). Significant Applications of Big Data in COVID-19 Pandemic. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7204193/

10.Yasinski A (2020). Big Data and Collaboration Seek to Fight COVID-19. Retrieved from https://www.the-scientist.com/news-opinion/big-data-and-collaboration-seek-to-fight-covid-19-67759

11.Kent J (2020). Understanding the COVID-19 Pandemic as a Big Data Analytics Issue. Retrieved from https://healthitanalytics.com/news/understanding-the-covid-19-pandemic-as-a-big-data-analytics-issue

12.Vendeville G (2020). U of T infectious disease expert's AI firm now part of Canada's COVID-19 arsenal. Retrieved from https://www.utoronto.ca/news/u-t-infectious-disease-expert-s-ai-firm-now-part-canada-s-covid-19-arsenal

13.National Center for Advancing Translational Sciences (2020). National COVID Cohort Collaborative (N3C). Retrieved from https://ncats.nih.gov/n3c

14.Bowles J (2020). How Canadian AI start-up BlueDot spotted Coronavirus before anyone else had a clue. Diginomica. Retrieved from https://diginomica.com/how-canadian-ai-start-bluedot-spotted-coronavirus-anyone-else-had-clue

15.Chen J & See KC (2020). Artificial Intelligence for COVID-19: Rapid Review. JMIR 22(10): e21476 https://www.jmir.org/2020/10/e21476

16.Kent J (2020). AI Model Shows Stricter Interventions Needed to Manage COVID-19 in TX. Retrieved from https://healthitanalytics.com/news/ai-model-shows-stricter-interventions-needed-to-manage-covid-19-in-tx

17.Kent J (2020). Stanford Launches Data-Driven Model Evaluating COVID-19 Interventions. Retrieved from https://healthitanalytics.com/news/stanford-launches-data-driven-model-evaluating-covid-19-interventions

18.Li L et al (2020). Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy. Radiology 296(2):E65-E71.

19.Li Y & Xia L (2020). Coronavirus Disease 2019 (COVID-19): Role of Chest CT in Diagnosis and Management. American Journal of Roentgenology 214(6):1280-1286.

20.Yang Z et al (2020). Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis 12(3):165-174.

21.Hurt B et al (2020). Deep Learning Localization of Pneumonia: 2019 Coronavirus (COVID-19) Outbreak. J Thorac Imaging 35(3):W87-W89.

22.Jiang X et al (2020). Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity. Tech Science Press 63(1):537-551.

23.Ton A et al (2020). Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. Mol Inform 39(8):e2000028.

24.Beck BR et al (2020). Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Comput Struct Biotechnol J 2020;18:784-790.

25.Lin L & Hou Z (2020). Combat COVID-19 with artificial intelligence and big data. J of Travel Med 27(5):1-4.

26.Forani J (2020). Low uptake and disclosure numbers plague feds' COVID Alert app. Retrieved from https://www.ctvnews.ca/health/coronavirus/low-uptake-and-disclosure-numbers-plague-feds-covid-alert-app-1.5089379

27.Murphy S et al (2020). Piloted in May, ditched in June: the failure of England's Covid-19 app. Retrieved from https://www.theguardian.com/world/2020/jun/18/piloted-in-may-ditched-in-june-the-failure-of-englands-covid-19-app