Data Science in Pandemic: A review

Krisha Patel(author)[1] Helly Patel(author)[2] Neha Mistry(author)[3] Punit Prajapati(author)[4]
Sejal Thakkar(author)[5]

[1,2,3,4] Student, B.tech IT, Indus University, Ahmedabad, India
[5]Research scholar, Ph.D, CE department, Indus University, Ahmedabad, India

Abstract— The outbreak of the COVID-19 pandemic has underscored the critical role of data science in managing public health crises. This review paper provides an overview of the applications of data science techniques and methodologies in pandemic management. It explores how data science has been instrumental in various aspects of pandemic response, including epidemiological modelling, forecasting, contact tracing, resource allocation, and vaccine distribution. The paper also discusses challenges and ethical considerations associated with the use of data science in pandemic management and highlights future directions for research in this field.

1. Introduction

The utilization of data science techniques in healthcare and public health has been facilitated by the extensive accessibility of large datasets encompassing human movement, tracing of contacts, medical imagery, virological studies, drug screening, bioinformatics, electronic medical records, and scientific publications, alongside the continual advancement in computational capabilities[1–4]. Amidst the ongoing coronavirus disease 2019 (COVID-19) pandemic, the significant dedication of researchers and professionals, coupled with the pressing demand for insights derived from data, has underscored the pivotal role of data science in comprehending and addressing the pandemic with unprecedented efficacy[5]. The global spread of COVID-19, instigated by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has led to significant loss of life, surpassing 3.4 million fatalities as of May 19, 2021[6]. Given its profound impact on both public health and economies worldwide, the COVID-19 pandemic underscores the critical necessity for timely and precise data sources, encompassing individual-level and surveillance and management. In contrast to responses to prior epidemics such as SARS, Ebola, HIV, and MERS, the COVID-19 pandemic has garnered unprecedented attention not only from medical and public health experts but also from professionals in various data and computational science domains, which were previously less central to epidemic responses[7-8].

The onset of the COVID-19 pandemic serves as an opportunity and a valuable reservoir of data for mathematicians, physicists, and engineers to actively engage in understanding the disease through data-driven and computational approaches. While certain datasets were absent during previous epidemics, and others were underutilized, the current pandemic offers a breadth of information that was previously untapped. Despite the effectiveness demonstrated by established public health systems, including those overseen by numerous countries' Centers for Disease Control (CDCs), they were swiftly overwhelmed by the highly transmissible nature of the SARS-CoV-2 virus and the continuous surge in global human mobility.

2. Analysis of a crisis

2.1 Static Visualization-

The data clearly indicated the exponential rise of COVID-19, a typical pattern observed during pandemics. Following the initial reports of confirmed cases, the numbers swiftly escalated, as noted by Stevens (2020). Since the early stages, global attention has been focused on this exponential surge, prompting the World Health Organization (WHO) to emphasize the significance of "flattening the curve," as highlighted by Roberts (2020). Consequently, extensive analysis ensued, resulting in what is poised to become one of the most renowned data visualizations documented in the public discourse surrounding COVID-19. China emerged as the first nation to present this curve and subsequently endeavored to mitigate its upward trajectory (Figure 1). Since the WHO declared the pandemic, every country followed China (Figure 2) [9].

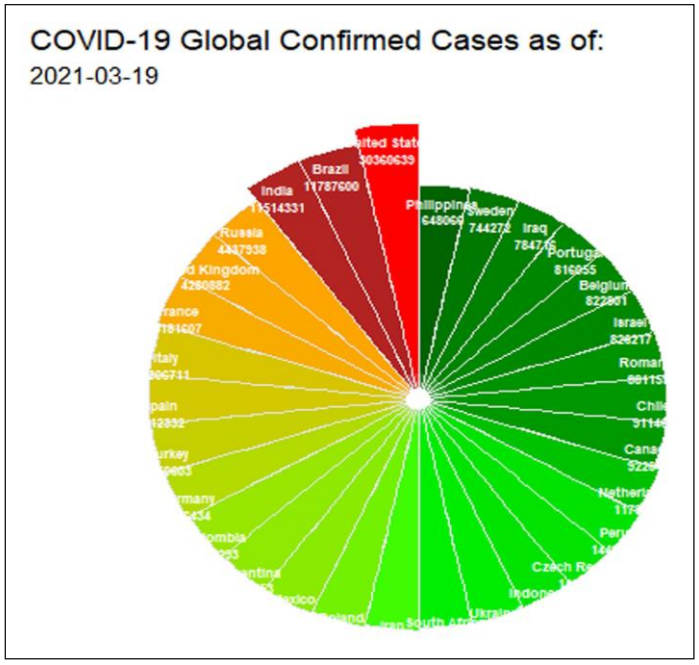

Figure 1: Number of deaths, confirmed cases, and cured cases in China.

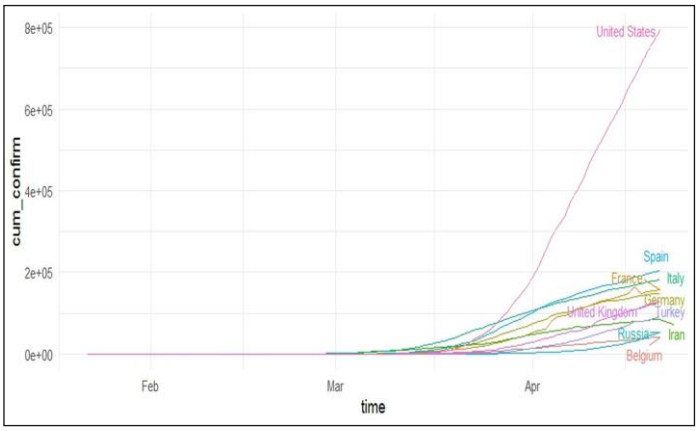

Figure 2: Total number of cases world-wide except China.

2.2 Interactive Visualization –

While static charts can effectively communicate a snapshot of data, researchers, healthcare professionals, policymakers, and decision-makers often require deeper insights into ongoing trends. Interactive visualizations offer the ability to directly engage with real-time data, providing a dynamic perspective. For instance, the accompanying figure showcases two kernel density plots illustrating the number of confirmed cases (figure 3). The left graph, featuring a high bandwidth for density estimation, suggests a relatively normal distribution of cases across all countries. However, this broad bandwidth masks underlying data patterns. Upon adjusting the bandwidth, as depicted in the right graph, the plot unveils a non-normal distribution, revealing multiple waves of cases. Notably, the initial wave corresponds to China, succeeded by Italy and subsequently other countries. Although the United States initially represented a minor wave in March 2020, it later surpassed other nations. This dynamic interaction with the data unveils the evolving landscape of the pandemic[9].

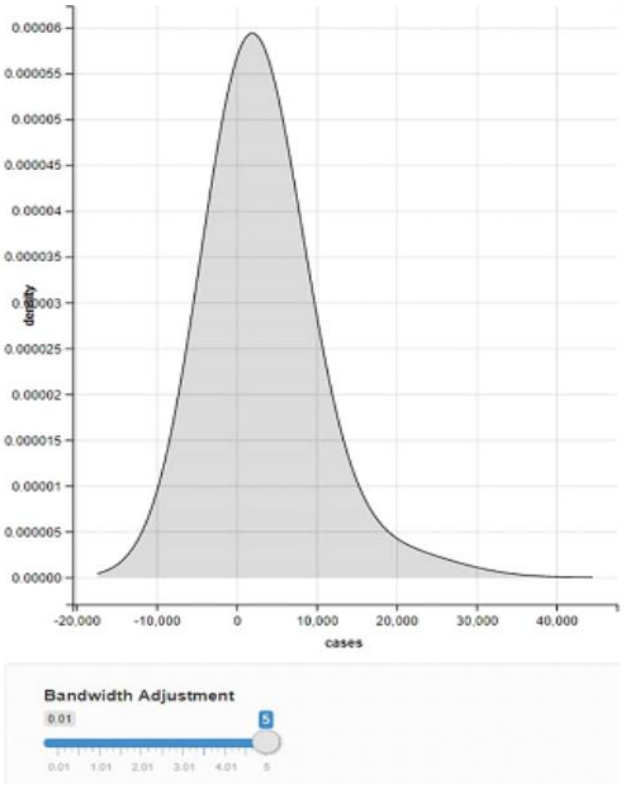

Figure 3: Global confirmed cases by country using an interactive density chart.

3. Modelling Human Mobility

SARS-CoV-2 exhibits a high level of contagiousness among humans, particularly in proximity[6]. Extensive evidence suggests that, akin to other SARS-related coronaviruses, SARS-CoV-2 initially infected a human host through an intermediary host in its natural environment[10,11]. Subsequently, human-to-human transmission emerged as the primary mode of spread. Consequently, the trajectory of the epidemic is heavily influenced by both local and international human movement patterns. This underscores the critical importance of analysing human mobility data for disease surveillance and policy assessment. Fortunately, we now have access to a wealth of human mobility data, encompassing population-level census and survey data that capture general travel patterns, as well as individualized

mobility data sourced from mobile devices, digital transactions, and social media platforms.

Looking back at the initial stages of the outbreak in Wuhan City, China, the rapid spread of the virus resulted in significant underreporting of the situation[12]. This was primarily due to several factors: firstly, many individuals who were infected, including asymptomatic carriers and those with mild symptoms, were unaware of their condition until after recovery; secondly, limited healthcare resources meant that numerous symptomatic individuals couldn't access hospital care. Consequently, the early epidemiological data failed to capture the complete picture of all patients. Initial reports tended to focus on patients with severe illness who were hospitalized, leading to an underrepresentation of those who weren't admitted. This pattern appears to have been observed in other regions globally as well. Consequently, several studies resorted to analyzing human mobility data to estimate key epidemiological parameters such as the basic reproduction number (R0). The movement of individuals out of Wuhan in January and February 2020[10-12] provided a well-documented source of data, enabling researchers to better understand the transmission dynamics of COVID-19 in Wuhan and beyond[13].

4. Manual and Digital Contact Tracing

While privacy concerns often limit the accessibility of such data for research purposes, there have been notable empirical and modeling investigations leveraging it. For instance, Bi et al. conducted a thorough analysis utilizing a comprehensive dataset comprising 391 cases and 1286 close contacts in Shenzhen City, China. This dataset, made available by the Shenzhen CDC, spanned from January 14, 2020, to February 12, 2020. Their findings illustrated the efficacy of contact tracing in substantially reducing the reproduction number, thereby averting a localized outbreak[14]. Zhang et al. conducted an analysis utilizing survey data collected from Wuhan City and Shanghai City, along with comprehensive contact tracing data obtained from Hunan Province, which was made available by the Hunan CDC. They employed this data to develop a transmission model aimed at assessing the effectiveness of Non-Pharmaceutical Interventions (NPIs) on transmission dynamics. Their study revealed that the implementation of NPIs in these regions effectively curtailed the spread of COVID-19 outbreak[15].

Traditional manual contact tracing encounters significant hurdles, including recall bias and delays in obtaining information. However, the widespread usage of smartphones presents an opportunity for innovative digital contact tracing methods to emerge as a viable complement to, if not a substitute for, manual contact tracing efforts[16,17]. Leveraging digital contact tracing holds the potential for heightened effectiveness in curbing the spread of the epidemic, particularly due to the substantial percentage of individuals utilizing smartphones[18].

5. Empirical evaluation of government responses

An exemplary illustration of this is the Oxford Covid-19 Government Response Tracker (OxCGRT, available at: https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker), which systematically gathers data on policy measures implemented by over 180 countries since January 1, 2020. OxCGRT meticulously documents these policies using a scale that denotes the magnitude of governmental intervention, and subsequently generates policy indices based on these scores[19].

Likewise, Porcher introduced Response2covid19 (available at https://response2covid19.org/), a dataset cataloging governments' responses to the COVID-19 pandemic [20] Additionally, Piccoli et al. introduced the Citizenship, Migration, and Mobility in a Pandemic (CMMP, accessible at https://www.cmm-pandemic.com/), another comprehensive global dataset. Another global dataset, the *Citizenship, Migration and Mobility in a Pandemic* (*CMMP*, https://www.cmm-pandemic.com/) was introduced by Piccoli *et al*[21]. Measuring the impact of different Non-Pharmaceutical Interventions (NPIs) represents another crucial challenge. Hsiang et al. gathered data on approximately 1700 NPIs implemented at the local, regional, and national levels across six countries. They then utilized reduced-form econometric techniques to quantitatively assess the effectiveness of these NPIs in flattening the

epidemic curve[22]. Dehning et al. conducted an analysis of data from Germany employing a Bayesian inference model. Their study highlighted the importance of exercising caution when considering the relaxation of Non-Pharmaceutical Interventions (NPIs). They emphasized that the current NPIs implemented had only marginally contained the outbreak, suggesting that any relaxation should be approached with careful consideration[23].

6. Assessing the economic, trade and supply chain impact

Their findings indicated that the magnitude of supply chain losses was contingent upon the number of countries implementing travel restrictions. Conversely, a more prolonged containment strategy, which could effectively control the epidemic, was associated with comparatively smaller losses[24]. The study constructed the global supply chain network by utilizing the Global Trade Analysis Project (GTAP) database[25]. Maliszewska et al. similarly employed GTAP data, which typically requires a subscription fee, along with historical instances of global epidemics. They utilized this data to simulate the effects of the COVID-19 pandemic on both gross domestic product (GDP) and trade. Their analysis yielded comparable conclusions to those of previous studies[26]. In a more recent study, Ye et al. introduced an integrated network model to explore the contagion patterns of shortages in personal protective equipment (PPE) on a global trade network. They sourced their trade network data from the World Customs Organization report. Their analysis revealed that export restrictions on PPE exacerbated shortages and facilitated the rapid spread of shortage contagion, surpassing the rate of contagion observed with the disease itself[27].

7. Social media analytics and web mining.

The World Wide Web and social media platforms have emerged as crucial avenues for the general public to access health-related information. Extensive evidence suggests that individuals' online activities are linked to their health statuses, presenting an opportunity to utilize this data to estimate the prevalence of infectious diseases. Leveraging web and social media data holds promise for providing more timely insights into epidemic trends[28,29].

In their empirical investigation, Bento et al. scrutinized individuals' information-seeking patterns following the initial confirmation of COVID-19 cases in each state of the USA. Their study revealed a significant correlation between the timing of the first confirmed case in a state and the subsequent surge in searches for specific terms[30]. In their correlation analysis, Effenberger et al. identified a relationship between Internet searches, measured through Google Trends, and the prevalence of COVID-19 cases across European countries[31].

8. Discussion

We conducted a bibliographic analysis of the papers mentioned above, visualizing the knowledge transfer from the disciplines of the cited papers to the disciplines of the papers citing them. Figure 5 illustrates this transfer, revealing that Multidisciplinary Sciences emerges as the predominant discipline for both groups of papers. To gain deeper insights, we present bar charts showcasing the disciplines of these papers, excluding Multidisciplinary Sciences.

Several topics, including vaccine prioritization [32,33], vaccine hesitancy [34], screening chatbot [35], crowdsourcing, and the emerging folk science, have not been extensively covered in the literature due to a lack of sufficient publications. However, as the pandemic continues and research efforts intensify, we anticipate that more knowledge will become available in these areas[36-40].

References:

1. Topol EJ. 2019 High-performance medicine: the convergence of human and artificialintelligence. *Nat. Med.* **25**, 44–56. (doi:10.1038/s41591-018-0300-7)
2. Khoury MJ, Ioannidis JPA. 2014 Big data meets public health. *Science* **346**, 1054–1055. (doi:10.1126/science.aaa2709)
3. Wong ZS, Zhou J, Zhang Q. 2019 Artificial intelligence for infectious disease big dataanalytics. *Infect., Dis. Health* **24**, 44–48. (doi:10.1016/j.idh.2018.10.002)
4. Mooney SJ, Pejaver V. 2018 Big data in public health: terminology, machine learning, and privacy. *Annu. Rev. Public Health* **39**, 95–112. (doi:10.1146/annurevpublhealth-040617-014208)
5. Who coronavirus (covid-19) dashboard. https://covid19.who.int/. (accessed 15 May 2021).
6. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. 2020 The architecture of Sars-Cov-2 transcriptome. *Cell* **181**, 914–921.e10. (doi:10.1016/j.cell.2020.04.011)
7. Luengo-Oroz M *et al.* 2020 Artificial intelligence cooperation to support the global response to Covid-19. *Nat. Mach. Intell.* **2**, 295–297. (doi:10.1038/s42256-020-0184-3)
8. Latif S *et al.* 2020 Leveraging data science to combat COVID-19: a comprehensive review. *IEEE Trans. Artif. Intell.* **1**, 85–103. (doi:10.1109/TAI.2020.3020521)
9. Mathaisel, DFX. 2023. Data Science in a Pandemic. *Data Science Journal*, 22: 41, pp. 1–12. DOI: https://doi. org/10.5334/dsj-2023-041
10. Rasmussen AL. 2021 On the origins of SARS-CoV-2. *Nat. Med.* **27**, 9–9. (doi:10.1038/s41591-020-01205-5)
11. WHO. 2021 Who-convened global study of origins of Sars-Cov-2: China part. *World Health Organization*.
12. Tuite AR, Fisman DN. 2020 Reporting, epidemic growth, and reproduction numbersfor the 2019 novel coronavirus (2019-ncov) epidemic. *Ann. Intern. Med.* **172**, 567–568. (doi:10.7326/M20-0358)
13. Hao X, Cheng S, Wu D, Wu T, Lin X, Wang C. 2020 Reconstruction of the full transmissiondynamics of COVID-19 in Wuhan. *Nature* **584**, 420–424. (doi:10.1038/s41586-020-2554-8)
14. Bi Q *et al.* 2020 Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *Lancet Infect. Dis.* **20**, 911–919. (doi:10.1016/S1473-3099(20)30287-5)
15. Zhang J *et al.* 2020 Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science* **368**, 1481–1486. (doi:10.1126/science.abb8001)
16. Bengio Y, Janda R, Yu YW, Ippolito D, Jarvie M, Pilat D, Struck B, Krastev S, Sharma A. 2020 The need for privacy with public digital contact tracing during the COVID-19 pandemic. *Lancet Digit. Health* **2**, e342–e344. (doi:10.1016/S2589-7500(20)30133-3)
17. Kleinman RA, Merkel C. 2020 Digital contact tracing for Covid-19. *CMAJ* **192**, E653–E656. (doi:10.1503/cmaj.200922)
18. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L, Parker M, Bonsall D,Fraser C. 2020 Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368**, eabb6936. (doi:10.1126/science.abb6936)
19. Hale T *et al.* 2021 A global panel database of pandemic policies (oxford Covid-19 government response tracker). *Nat. Hum. Behav.* **5**, 529–538. (doi:10.1038/s41562-021-01079-8)
20. Porcher S. 2020 Response2covid19, a dataset of governments' responses to COVID-19 allaround the world. *Sci. Data* **7**, 1–9. (doi:10.1038/s41597-020-00757-y)
21. Piccoli L,Dzankic J, Ruedin D. 2021 Citizenship, migration and mobility in a pandemic(CMMP): a global dataset of COVID-19 restrictions on human movement. *PLoS ONE* **16**, e0248066. (doi:10.1371/journal.pone.0248066)

22. Hsiang S *et al.* 2020 The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* **584**, 262–267. (doi:10.1038/s41586-020-2404-8)
23. Dehning J, Zierenberg J, Spitzner FP, Wibral M, Neto JP, Wilczek M, Priesemann V. 2020 Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science* **369**, eabb9789. (doi:10.1126/science.abb9789)
24. Guan D *et al.* 2020 Global supply-chain effects of COVID-19 control measures. *Nat. Hum. Behav.* **4**, 577–587. (doi:10.1038/s41562-020-0896-8)
25. Chepeliev M. 2020 GTAP-power data base: Version 10. *J. Global Econ. Anal.* **5**, 110–137. (doi:10.21642/JGEA.050203AF)
26. Maliszewska M, Mattoo A, Van Der. 2020 *The potential impact of COVID-19 on GDP and trade: a preliminary assessment*. World Bank Policy Research Working Paper. Washington, DC: World Bank.
27. Ye Y, Zhang Q, Cao Z, Chen FY, Yan H, Stanley HE, Zeng DD. 2021 Impacts of exportrestrictions on the global personal protective equipment trade network during COVID-19. (http://arxiv.org/abs/2101.12444).
28. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. 2008 Surveillance sans frontieres: internetbased emerging infectious disease intelligence and the healthmap project. *PLoS Med.* **5**, e151. (doi:10.1371/journal.pmed.0050151)
29. Merchant RM, Elmer S, Lurie N. 2011 Integrating social media into emergency-preparednessefforts. *N. Engl. J. Med.* **365**, 289–291. (doi:10.1056/NEJMp1103591)
30. Bento AI, Nguyen T, Wing C, Lozano-Rojas F, Ahn YY, Simon K. 2020 Evidence from internetsearch data shows information-seeking responses to news of local Covid-19 cases. *Proc. Natl Acad. Sci. USA* **117**, 11 220–11 222. (doi:10.1073/pnas.2005335117)
31. Effenberger M, Kronbichler A, Shin JI, Mayer G, Tilg H, Perco P. 2020 Association of theCovid-19 pandemic with internet search volumes: a google trendstm analysis. *Int. J. Infect. Dis.* **95**, 192–197. (doi:10.1016/j.ijid.2020.04.033)
32. Wang LL *et al.* 2020 Cord-19: the covid-19 open research dataset. In *Proc. 1st Workshop on NLP for COVID-19 at ACL 2020*.
33. Wang Q *et al.* 2021 Covid-19 literature knowledge graph construction and drug repurposing report generation. In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*.
34. Reddy RG, Iyer B, Sultan MA, Zhang R, Sil A, Castelli V, Florian R, Roukos S. 2020 Endto-end QA on Covid-19: domain adaptation with synthetic training. (http://arxiv.org/abs/2012.01414)
35. Tang R, Nogueira R, Zhang E, Gupta N, Cam P, Cho K, Lin J. 2020 Rapidly bootstrapping aquestion answering dataset for Covid-19. (http://arxiv.org/abs/2004.11339)
36. Lee J, Yi SS, Jeong M, Sung M, Yoon W, Choi Y, Ko M, Kang J. 2020 Answering questions onCovid-19 in real-time. In *Proc. 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.
37. Chen C, Ebeid IA, Bu Y, Ding Y Coronavirus knowledge graph. a case study. In *Int. Workshop on Knowledge Graph, co-located with Twenty-Sixth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (ACM KDD 2020)*.
38. Zhang E, Gupta N, Nogueira R, Cho K, Lin J. 2020 Rapidly deploying a neural search enginefor the Covid-19 open research dataset. In *Proc. 1st Workshop on NLP for COVID-19 at ACL 2020*.
39. Voorhees E, Alam T, Bedrick S, Demner-Fushman D, Hersh WR, Lo K, Roberts K, Soboroff I,Wang LL. 2021 Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, vol. 54, pp. 1–12. New York, NY: ACM.
40. Su D, Xu Y, Yu T, Siddique FB, Barezi EJ, Fung P. 2020 Caire-covid: a question answering andquery-focused multi-document summarization system for Covid-19 scholarly information management.