Quantifying the Regional Disproportionality of COVID-19 Spread

Kenji Sasaki [1]*, Yoichi Ikeda [1] and Takashi Nakano [1,2]

1. Center for Infectious Disease Education and Research, Osaka University, Osaka 565-0871, Japan

2. Research Center for Nuclear Physics, Osaka University, Osaka 567-0047, Japan

*** Correspondence: kenjis@cider.osaka-u.ac.jp**

Abstract

Background:
The COVID-19 pandemic has caused serious health problems and has had major economic and social consequences worldwide. Understanding how infectious diseases spread can help mitigating the social and economic impact.
Objective:
The study focuses to capture the degrees of disproportionality in prevalence rates of infectious disease across different regions over time.
Methods:
We analyze the numbers of daily COVID-19 confirmed cases in the United States collected by Johns Hopkins University over 1100 days since the first reported case in January 2020 in order to assess quantitatively the disproportionality of the confirmed cases using the Theil index, a measure of imbalance used in economics.
Results:
Our results reveal a dynamic pattern of interregional disproportionality in the confirmed cases by monitoring variations in regional contributions to the Theil index as the pandemic progresses.
Conclusions:
The combined monitoring of this indicator and the confirmed cases is crucial for understanding regional differences in infectious diseases and for effective planning of response and resource allocation.

Keywords: Infectious disease; COVID-19; Epidemiology; Pandemic; Inequality measure; Information theory; Kullback-Leibler divergence

Introduction

The COVID-19 pandemic has caused serious health problems and has had major economic and social consequences worldwide. A number of indicators and models have been proposed to address the problem, and mechanisms for the spread of the

infection and intervention measures to control the pandemic have been studied [1–7].

Several studies have investigated regional differences in COVID-19 prevalence [8–11]. Differences in prevalence rates between regions underscore the importance of understanding regional imbalances in pandemic response strategies. Effectively addressing these imbalances requires accurate quantification and understanding of the regional disproportionalities of daily COVID-19 confirmed cases.

In the field of economics, various indicators have been developed to measure resource and income inequality, including one proposed by Theil incorporating information theory [12]. In reference [13], the authors demonstrate the value of using inequality indices to monitor changes in geographic inequality, and the Theil index was used to track geographic inequality over time in the COVID-19 pandemic, providing important insights to inform public health policy.

The aim of this paper is to quantify the interregional disproportionality in numbers of the confirmed cases using the Theil index, which corresponds to the Kullback-Leibler (KL) divergence in information theory [14]. It is an effective method of measuring the degree of disproportionality and objectively assessing the bias in the interregional distribution of infected individuals.

Methods

The Theil index is commonly applied in various fields including economics, sociology, and information theory. The index quantifies the relative differences between various components of a dataset. In the context of regional analysis of the confirmed cases, the Theil index can be employed to evaluate the distribution of infected individuals across different regions. In this study, we utilize the Theil index to identify regions with disproportionate numbers of the confirmed cases relative to their population size by comparing the distribution of the confirmed cases with the overall population distribution.

The discrete form of Theil index is expressed as

$$T = \sum_{i=1}^{N} p_i ln \frac{p_i}{q_i}$$

where N is the total number of considering regions and ln is the natural logarithm. The p_i which is described as the discrete probability distributions in region i is the ratio of daily confirmed cases in a region and it in whole region per a day and, similarly, the q_i is the ratio of the population in a region and it in whole region.

The Theil index, which shares the same property as the KL divergence, is a non-symmetric metric that measures the relative entropy or difference in information represented by two distributions. It is sensitive to the interregional distribution of the confirmed cases, with its maximum value attained when the confirmed cases are concentrated in areas with the smallest population proportion. Consequently, the index tends to exhibit higher values when a small number of regions account for a large share of the confirmed cases, and conversely, lower values when the confirmed cases are more evenly distributed across regions. Notably, it remains non-negative and reaches a minimum value of 0 only when the two distributions are identical. Therefore, applying the Theil index to the timeline data of the confirmed cases, changes in the index over time can be used to quantify the degree of spread of infectious diseases and to assess whether a disproportionate concentration of infected individuals relative to the population is occurring.

Results

We analyze the time trend of daily COVID-19 confirmed cases in the United States over 1100 days since the first reported case on January 21, 2020 [15]. Data are taken from the COVID-19 data repository at the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [16]. Population data by U.S. state were obtained from [17]. Note that population changes due to migration, births, and deaths were ignored throughout the analysis.

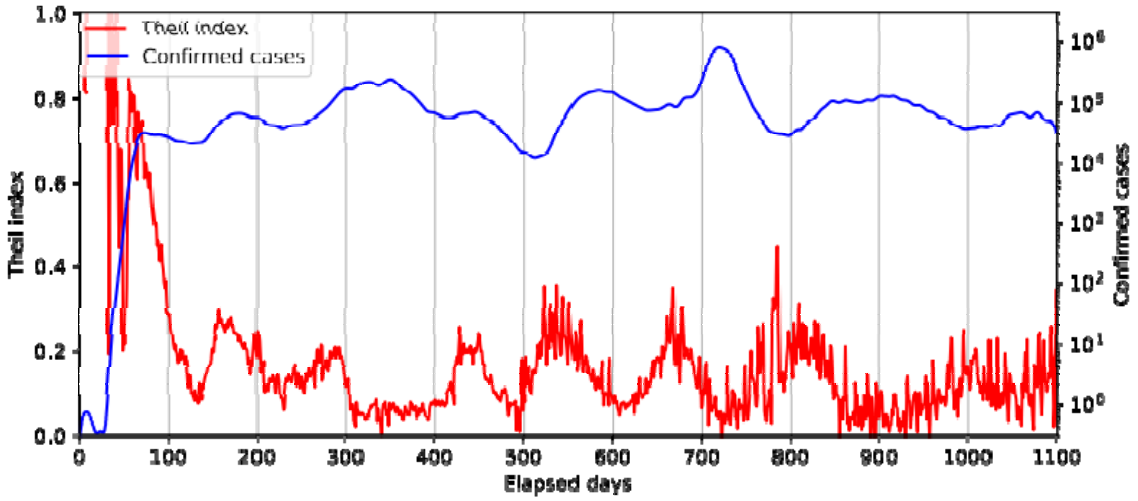

Figure 1. Time trends of the Theil index on the left axis and the number of 7-day averages of the confirmed cases on the right axis on a logarithmic scale are shown in the red and blue curves, respectively. The horizontal axis is the number of days elapsed since January 21, 2020.

Before showing the results, if we use the confirmed cases as it is, the Theil index will fluctuate greatly due to the way the data is aggregated in holidays differs depending on regions, so we use the 7-day average instead of raw data.

The Fig. 1 is a two-axis graph showing the time trends of the Theil index on the left and the number of the confirmed cases on the right, on a logarithmic scale. The horizontal axis is the number of days, denoted by t in the text, elapsed since the date when the first case was reported in the U.S..

It is important to note that increases and decreases in the Theil index simply indicate the degree of disproportionality in the confirmed cases and do not correspond to increases or decreases in the number of infected individuals. In other words, this indicator is effective when monitored in conjunction with actual trends in the number of the confirmed cases.

In Fig. 1, there are eight notable surges of the confirmed cases, occurring at around t =80(1st), 180(2nd), 350(3rd), 450(4th), 580(5th), 720(6th), 900(7th), and 1080(8th) respectively. Before the first peak, the number of the confirmed cases is quite low and the Theil index fluctuates unstably. As t increases near the 1st peak, the Theil index appears to gradually decrease, reaching a local minimum around t = 120. This implies that a fairly localized epidemic at the beginning of the COVID-19 pandemic spreads rather equally throughout the U.S.. For the other peaks in the confirmed cases, the similar phenomena can be confirmed, namely the increase of the Theil index occurs to some extent before the peak and then the Theil index decreases. This can be seen as a precursor to a surge of infected individuals as discussed in ref. [5].

It is interesting to check some examples. Practically, when the index value is high and the number of the confirmed cases is low (t = 60, 550, etc.), it indicates that the infection is occurring locally and spreading to various regions. On the other hand, when the index is low and the number of the confirmed cases is high (t = 750, etc.), it indicates that there is no clear epicenter of infection and that the number of the confirmed cases is increasing equally in different regions.

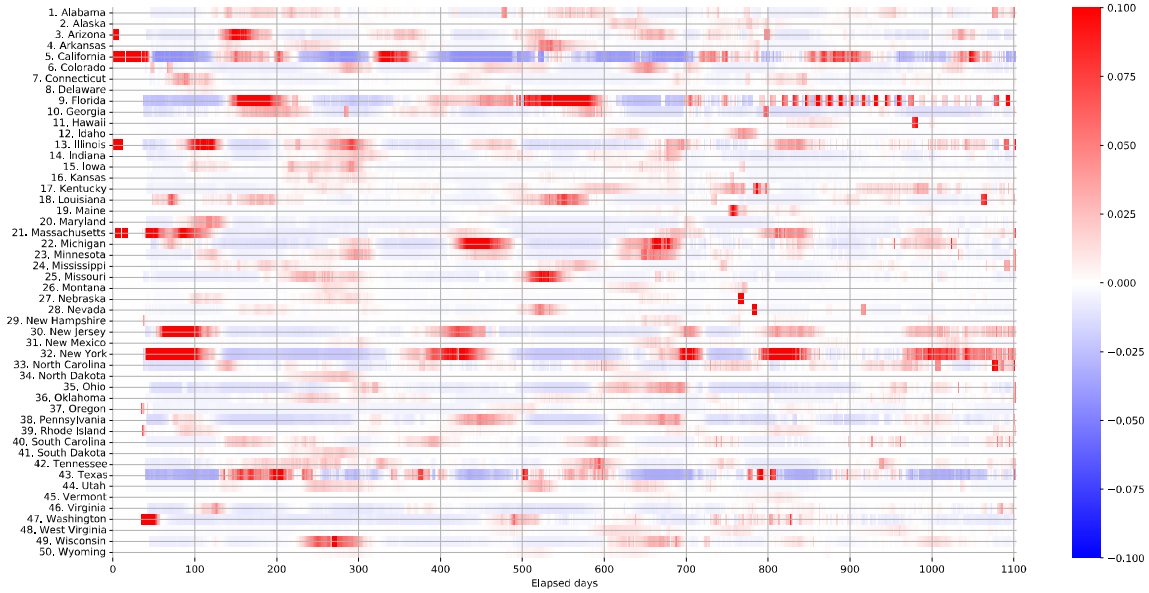

Figure 2. Contributions to the Theil index from each region are shown by a heat-map over the time. The positive (high concentration of prevalence) and negative (low concentration) contributions to the Theil index correspond to the deep red and blue color, respectively.

The contributions to the Theil index from each region over time are visualized by the heat-map as shown in Fig. 2 where regions with a high concentration of the confirmed cases relative to the population are colored in red, while blue regions indicate lower concentrations. There is a somewhat long interval between the deep red patches in some regions such as California, Florida and New York. In other words, the periods of intense infection represented by the deep red patches were not repeated at short intervals. This phenomenon is of great importance in infectious disease responses. Once a major epidemic in an area has subsided, the interval between subsequent outbreaks provides an opportunity to rebuild the healthcare system and implement preventive measures before the next epidemic occurs.

Based on the observations from Fig. 2, it is apparent that the epicenter of infectious diseases, indicated by the red patch, alternates among New York, California, and Florida. This insight is crucial for understanding the spread mechanism of future infectious diseases. Furthermore, after t = 750, both the red and blue colors fade over time, suggesting the absence of a clear epicenter and indicating a widespread outbreak of COVID-19. This suggests the ineffectiveness of countermeasures against the spread of infectious diseases under these circumstances.

Discussion

Regional disproportionality is a critical factor influencing strategy formulation in the fight against the COVID-19 pandemic. A detailed analysis of infection patterns in different regions facilitates the development of more targeted and efficient region-specific interventions. Furthermore, if the distribution of infectious diseases is highly imbalanced, one could consider that there is an opportunity to rearrange the allocation of health care resources.

The combined monitoring of Fig. 1 and 2 allows us to show the epicenter of infectious diseases and their concentration at any time. As indicated in the results section, it was also found that once an infectious disease concentration decreases, there is some interval before the next concentration occurs. Therefore, while monitoring, it is necessary to concentrate countermeasures in areas where there is a concentration of infections and prepare for the next concentration of infections in other areas.

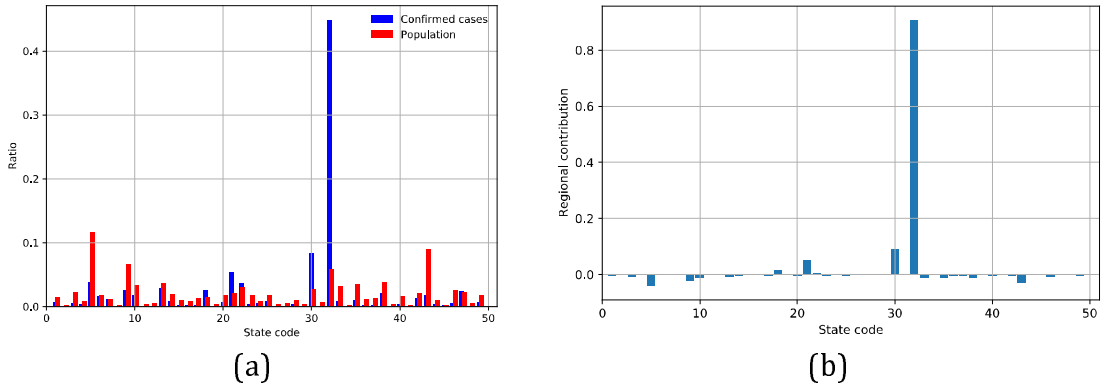

Figure 3. Contributions to the Theil index from each region at t = 60. The horizontal axis shows the state code given in Tab. 1. (a) Contributions to the Theil index from each region. The vertical axis shows the strength of the contribution to the Theil index. (b) Comparison of the distribution of the confirmed cases and population. The vertical axis shows the ratio of a part to the whole region for populations and for the confirmed cases.

Table 1. The list of state code used in this paper.

1: Alabama	2: Alaska	3: Arizona	4: Arkansas	5: California
6: Colorado	7: Connecticut	8: Delaware	9: Florida	10: Georgia
11: Hawaii	12: Idaho	13: Illinois	14: Indiana	15: Iowa
16: Kansas	17: Kentucky	18: Louisiana	19: Maine	20: Maryland
21: Massachusetts	22: Michigan	23: Minnesota	24: Mississippi	25: Missouri
26: Montana	27: Nebraska	28: Nevada	29: New Hampshire	30: New Jersey
31: New Mexico	32: New York	33: North Carolina	34: North Dakota	35: Ohio
36: Oklahoma	37: Oregon	38: Pennsylvania	39: Rhode Island	40: South Carolina
41: South Dakota	42: Tennessee	43: Texas	44: Utah	45: Vermont
46: Virginia	47: Washington	48: West Virginia	49: Wisconsin	50: Wyoming

As it is mentioned in the previous section, enhancement of the Theil index can be seen as a precursor of a surge of the confirmed cases. In fact, just before the 1st, 4th, 6th and 7th surges, concentration of the confirmed cases is occurred in New York, and before the 2nd, 5th surges occurred in Florida.

Fig. 3(a) shows the contributions to the Theil index by region at t = 60. The horizontal axis in the figure shows the state code given in Tab. 1. There is a notable contribution to the Theil index from New York State which is prominent relative to the other regions. We also find that New Jersey and Massachusetts have a relatively large positive contribution to the Theil index due to the concentration of the number of the confirmed cases relative to the population and are better treated as the same group as New York due to their geographic proximity to each other. It can be clearly seen in Fig. 3(b) that the confirmed cases are quite localized in these regions.

There is a relatively large negative contribution to the Theil index from California, Florida, and Texas, which are regions of high population ratio. It is interesting to note that these are regions where there was little risk of infection at that time, but

where the number of infected individuals rapidly increased after the concentration of the confirmed cases in New York took place.

The concentration of infections in New York at t = 60 as shown in Fig. 3(a) and (b) cannot be overlooked when considering infection control. The lockdown was implemented in New York city at the time when the contribution to the Theil index was concentrated in New York state. While it is impossible to assess the impact of the lockdown measures on the Theil index alone, the concentration of the confirmed cases in New York state suggests that the lockdown was implemented at the appropriate time. However, given that the contribution to the Theil index from New Jersey and Massachusetts which can be treated as the same group as New York was larger in positive values than the other regions, if lockdown measures are an appropriate response to contain COVID-19 infection, it may have been necessary to implement strong measures in these regions simultaneously to prevent the spread of COVID-19 throughout the country.

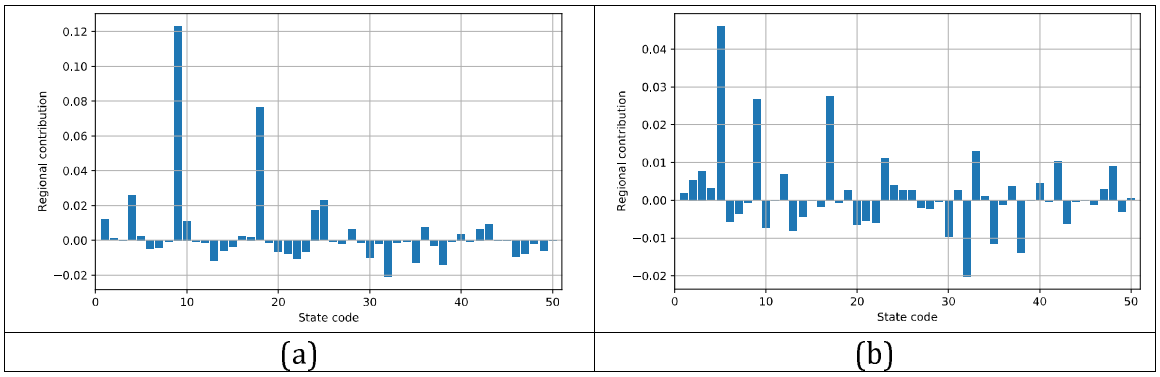

Figure 4. Contributions to the Theil index from each region at a specific date. The vertical axis shows the strength of contribution to the Theil index. The horizontal axis shows the state code given in Tab. 1. (a) Contributions of the Theil index at t = 550. (b) Contributions of the Theil index at t = 750.

The Fig. 4 shows the contributions of the Theil index from each region at t = 550 and t = 750. At t = 550 shown in Fig. 4(a), the Theil index is a peak and trend of the confirmed cases is increasing, suggesting that the new epidemic occurred mainly in Florida and Louisiana. However, their contributions are fairly smaller than that from New York at t = 60 in Fig. 3. This indicates that the regional imbalance is much less than in the early stage of the COVID-19 pandemic. It is also interesting to look at the case at t = 750 shown in Fig. 4(b), when the confirmed cases in the US are at their maximum. In this figure, although there are several regions with large contributions to the Theil index, the epicenter of COVID-19 is no longer clear, meaning that COVID-19 is evenly distributed across the regions.

Conclusions

In conclusion, this study demonstrates the effectiveness of Theil index in quantifying regional disproportionalities in the confirmed cases and monitoring their evolution over time. By analyzing data of the confirmed cases in the United States, we have

clarified patterns of disproportionalities in the confirmed cases, specifying epicenters and occurring localized outbreaks.

Continued monitoring and analysis of regional differences in COVID-19 transmission remains essential, especially in light of emerging variants and evolving public health responses. Our findings highlight the importance of understanding regional dynamics of infected individuals for responses of pandemic. Metrics such as Theil index provide valuable tools for policymakers and public health officials to allocate resources effectively and tailor interventions to specific regional needs.

Incorporating evidence from this study will enable policymakers to refine strategies and address the different needs of different regions, ultimately increasing the effectiveness of pandemic response efforts and mitigating the impact of future health crises.

Lastly, the decomposability of the Theil index makes it possible to quantify and compare disproportionality in groups with specific characteristics, such as age, vaccination coverage, and accessibility of health care, and identifying these disproportionalities will provide important insights for future pandemic responses.

Acknowledgements

We thank the members of the Division of Scientific Information and Public 184 Policy (SiPP) at the Center for Infectious Disease Education and Research (CiDER) Osaka University for useful discussions. This research was supported by The Nippon Foundation—Osaka University Project for Infectious Disease Prevention.

Conflicts of Interest

None declared.

Abbreviations

COVID-19 : COrona Virus Infectious Disease, emerged in 2019
KL : Kullback-Leibler

References

1. Perkins, T.A.; España, G. Optimal Control of the COVID-19 Pandemic with Non-pharmaceutical Interventions. *Bull. Math. Biol.* 82, 118 (**2020**). https://doi.org/10.1007/s11538-020-00795-y
2. Nakano, T.; Ikeda, Y. Novel Indicator to Ascertain the Status and Trend of COVID-19 Spread: Modeling Study. *J. Med. Internet Res.* **2020**, *22*, e20144. https://www.jmir.org/2020/11/e20144.
3. Buchy, P.; Buisson, Y.; Cintra, O.; Dwyer, D.E.; Nissen, M.; Ortiz de Lejarazu, R.; Petersen, E. COVID-19 pandemic: lessons learned from more than a century of pandemics and current vaccine development for pandemic control. *Int J Infect Dis.* **2021** Nov;112:300-317. doi: 10.1016/j.ijid.2021.09.045. Epub 2021 Sep 23. PMID: 34563707; PMCID: PMC8459551.
4. Stenseth, N.C.; Dharmarajan, G.; Li, R.; Shi, ZL.; Yang, R.; Gao, G.F. Lessons Learnt From the COVID-19 Pandemic. *Front Public Health.* **2021** Aug 2;9:694705. doi: 10.3389/fpubh.2021.694705. PMID: 34409008; PMCID: PMC8365337.

5. Ikeda, Y.; Sasaki, K.; Nakano, T. A New Compartment Model of COVID-19 Transmission: The Broken-Link Model. *Int. J. Environ. Res. Public Health* **2022**, *19*, 6864. https://doi.org/10.3390/ijerph19116864.
6. Nature Outlook; Pandemic preparedness https://www.nature.com/collections/jaacfgeief (accessed on: 14 March 2024).
7. Sasaki, K.; Ikeda, Y.; Nakano, T. The Effects of Behavioral Restrictions on the Spread of COVID-19. Reports 2022, 5, 37. https://doi.org/10.3390/reports5040037
8. Jinjarak, Y.; Ahmed, R.; Nair-Desai, S.; Xin, W.; Aizenman, J. Accounting for Global COVID-19 Diffusion Patterns, January-April 2020. *Econ Disaster Clim Chang.* **2020**;4(3):515-559. doi: 10.1007/s41885-020-00071-2. Epub 2020 Sep 4. PMID: 32901228; PMCID: PMC7471593.
9. Mollalo, A.; Vahedi, B.; Rivera, K.M. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Sci Total Environ.* **2020** Aug 1;728:138884. doi: 10.1016/j.scitotenv.2020.138884. Epub 2020 Apr 22. PMID: 32335404; PMCID: PMC7175907.
10. Villanustre, F.; Chala, A.; Dev, R.; Xu, L.; LexisNexis, J.S.; Furht, B.; Khoshgoftaar, T. Modeling and tracking Covid-19 cases using Big Data analytics on HPCC system platformm. *J Big Data.* **2021**;8(1):33. doi: 10.1186/s40537-021-00423-z. Epub 2021 Feb 15. PMID: 33614394; PMCID: PMC7883950.
11. Yue, H.; Hu, T.; Geographical Detector-Based Spatial Modeling of the COVID-19 Mortality Rate in the Continental United States. *Int. J. Environ. Res. Public Health* **2021**, *18*, 6832. https://doi.org/10.3390/ijerph18136832
12. Theil, H *Economics and Information Theory*, North-Holland Publishing Company: Amsterdam, North-Holland, 1967.
13. Manz, K.M.;, Mansmann, U. Inequality indices to monitor geographic differences in incidence, mortality and fatality rates over time during the COVID-19 pandemic. *PLoS ONE* 16(5): (**2021**) e0251366. https://doi.org/10.1371/journal.pone.0251366
14. Kullback, S.; Leibler, R. A. On information and sufficiency, *Annals of Mathematical Statistics* 22: (**1951**) 79-86.
15. Gini, C.; Concentration and dependency ratios (in Italian). English translation in *Rivista di Politica Economica*, 87 (**1997**), 769–789.
16. Gini, C. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[Fasc. I.]* (in Italian). Tipogr. di P. Cuppini: Bologna, Italy. 1912.
17. Our World in Data; Measuring inequality: What is the Gini coefficient? https://ourworldindata.org/what-is-the-gini-coefficient (accessed on: 14 March 2024).
18. Holshue, M.L.; DeBolt, C.; Lindquist, S.; Lofy, K.H.; Wiesman, J.; Bruce, H.; Spitters, C.; Ericson, K.; Wilkerson, S.; Tural, A.; Diaz, G.; Cohn, A.; Fox, L.; Patel, A.; Gerber, S.I.; Kim, L.; Tong, S.; Lu, X.; Lindstrom, S.; Pallansch, M.A.; Weldon, W.C.; Biggs, H.M.; Uyeki, T.M.; Pillai, S.K. Washington State 2019-nCoV Case Investigation Team. First Case of 2019 Novel Coronavirus in the United States. *N. Engl. J. Med.* **2020** Mar 5;382(10):929-936. doi: 10.1056/NEJMoa2001191. Epub 2020 Jan 31. PMID: 32004427; PMCID: PMC7092802.
19. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. https://github.com/CSSEGISandData/COVID-19 (accessed on: 14 March 2024).
20. U.S. Census Bureau; National Population Totals and Components of Change: 2020-2023. https://www.census.gov/data/datasets/time-series/demo/popest/2020s-national-total.html (accessed on: 14 March 2024).
21. Shehzad, K.; Bilgili, F.; Koçak, E.; et al. COVID-19 outbreak, lockdown, and air quality: fresh insights from New York City. *Environ Sci Pollut Res* 28, 41149–41161 (**2021**). https://doi.org/10.1007/s11356-021-13556-8
22. Huang, Y.; Li, R. The lockdown, mobility, and spatial health disparities in COVID-19 pandemic: A case study of New York City. *Cities.* **2022** Mar;122:103549. doi: 10.1016/j.cities.2021.103549. Epub 2022 Jan 3. PMID: 35125596; PMCID: PMC8806179.