Clustering analysis of coronavirus disease 2019 pandemic

Submission: 14-12-2020 Revision: 23-12-2020 Publication: 01-02-2021

Access this article online

Website: http://nepjol.info/index.php/AJMS

DOI: 10.3126/ajms.v12i2.33401

E-ISSN: 2091-0576

P-ISSN: 2467-9100

Sir,

Since the initial reports of the Wuhan outbreak, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic has been spreading around the world at an alarmingly exponential rate.[1-3] As of 10 October 2020, the number of confirmed infections exceeded thirty-five million worldwide, and almost 400,000 cases in Iraq.[2] Complications related to the illness claimed the lives of over one million lives, and almost ten thousand in Iraq.[2] SARS-CoV-2 allocated with to nations from the developed and the developing world as well, including the United States, India, Brazil, Russia, Columbia, Peru, Spain, Mexico, Argentina, South Africa, France, Chile, Iran, UK, Bangladesh, Iraq, Saudi Arabia, Turkey, Italy, and Pakistan.[2] As a top priority for the global health agenda, researchers are working to develop effective vaccines in severalnations of the world, including China, Russia, the United Arab Emirates, the United Kingdom, and the United States.[3] Coronaviruses represent a family of enveloped positive-strand RNA viruses that can infect species of vertebrates, including humans.[1,2] Coronaviruses include thirty-nine species that belong to the family Coronaviridae, suborder Cornidovirineae, order Nidovirales, and the realm of Riboviria.[1] The mutation rates of the genome of RNA viruses occurs faster than those of DNA viruses, which indicates a more efficient adaptation process for survival in Coronaviruses.[1,2] During the past two decades, Coronaviruses caused two epidemics, the severe acute respiratory syndrome (SARS) and the Middle East respiratory syndrome (MERS) in 2002-2003 and 2012, respectively.[1,4]

Our study is one of the first studies in the literature, in connection with the SARS-CoV-2 pandemic, to implement non-Bayesian analytics and unsupervised machine learning. Our primary objective is to provide a cluster analysis of all the countries worldwide, using hierarchical clustering, based on the total cases, total deaths, total recovered, active cases, critical cases, total cases per million of the population (total cases/1m), total deaths/1m, and total tests/1m. Our second objective is to examine the bivariate correlations among the same parameters in addition to other parameters, including the new cases, new deaths, new recovered, and the population count for each country. Our third objective is to scrutinize the case of Iraq, its neighbouring countries, and the Middle East in connection with the initial clustering analysis. Thereby, we are highlighting the noteworthiness of modalities of artificial intelligence, including machine learning, to the international and regional authorities, through which they can deploy precautionary and preventive measures in facing the pandemic.[5-8] The authors carried out the work described in this manuscript following the Code of Ethics of the World Medical Association [Declaration of Helsinki] on medical research involving human subjects, EU Directive (210/63/EU), the uniform requirements for manuscripts submitted to biomedical journals and the ethical principles defined in the Farmington Consensus of 1997.

BIVARIATE CORRELATIONS AND HIERARCHICAL CLUSTERING OF THE PANDEMIC

Several renowned websites are providing a quasi-real-time stream of data on coronavirus disease 2019 (COVID-19) worldwide. These websites are not limited to open access databases available via the World Health Organization, COVID-19 Tracker of Microsoft's Bing search engine, the Worldometer website, and many others.[2] We used the Worldometer website because its data "… is also trusted and used by the UK Government, Johns Hopkins CSSE, the Government of Thailand, the Government of Pakistan, Financial Times, The New York Times, Business Insider,

BBC, and many others." Besides, their data have been "requested by, and provided to Oxford University Press, Wiley, Pearson, CERN, World Wide Web Consortium (W3C),…" .[2] Usinga Python script and a web-scraping tool, we retrieved data on 10 October 2020 for all of the countries of the world.[9]

Using Microsoft Excel and IBM SPSS version 24, we conducted descriptive statistics (Table 1), tests of normality, bivariate correlations, and multiple hierarchical cluster analysis. According to the Shapiro-Wilk test of normality, all the parameters of interest followed a non-normal distribution. Therefore, we ran a nonparametric bivariate correlation, Kendall's tau rank correlation, as a correlation matrix for all of the twelve parameters, as mentioned earlier (Table 2). Almost all correlations were significant except for the new cases versus total cases/1M (τb=0.087, p-value=0.097), total deaths versus tests/1M (τb=0.075,

Table 1: Descriptive statistics

	N	Minimum	Maximum	Mean		Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
Total Cases	216	1	7895738	172108.68	54714.977	8.085	0.166	68.771	0.330
New Cases	216	0	12846	357.89	92.050	5.595	0.166	38.924	0.330
Total Deaths	216	0	218685	4970.71	1423.487	7.446	0.166	63.837	0.330
New Deaths	216	0	411	6.74	2.432	8.431	0.166	83.388	0.330
Total Recovered	216	0	5988822	127051.80	42156.107	7.958	0.166	65.788	0.330
New Recovered	216	0	6781	179.96	49.013	6.008	0.166	42.674	0.330
Active Cases	216	0	2611669	32471.79	13219.339	11.532	0.166	147.660	0.330
Critical Cases	216	0	14777	317.19	94.644	7.555	0.166	65.834	0.330
Total Cases/1M	216	0	45445	6055.72	558.223	2.283	0.166	5.996	0.330
Total Deaths/1M	216	0	1237	125.93	13.829	2.509	0.166	6.967	0.330
Total Tests/1M	216	0	2799996	159530.27	21720.632	4.632	0.166	27.563	0.330
Population	216	0	1439323776	36016655.77	9582895.252	8.856	0.166	84.379	0.330

Table 2: Bivariate Correlations: Kendall's tau-b (τb)

		Total cases	New cases	Total deaths	New deaths	Total recovered	New recovered	Active cases	Critical cases	Total cases /1M	Total deaths /1M	Total tests /1M	Population
Total cases	τb	1.000	.265**	.826**	.331**	.868**	.272**	.703**	.607**	.368**	.466**	.136**	.536**
	Sig.	.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.003	.000
	N	216	216	216	216	216	216	216	216	216	216	216	216
New cases	τb	.265**	1.000	.238**	.752**	.250**	.843**	.253**	.176**	.087	.108*	.113*	.208**
	Sig.	.000	.	.000	.000	.000	.000	.000	.001	.097	.041	.032	.000
	N	216	216	216	216	216	216	216	216	216	216	216	216
Total deaths	τb	.826**	.238**	1.000	.317**	.734**	.244**	.666**	.570**	.301**	.502**	.075	.557**
	Sig.	.000	.000	.	.000	.000	.000	.000	.000	.000	.000	.103	.000
	N	216	216	216	216	216	216	216	216	216	216	216	216
New deaths	τb	.331**	.752**	.317**	1.000	.298**	.744**	.334**	.240**	.144**	.196**	.048	.227**
	Sig.	.000	.000	.000	.	.000	.000	.000	.000	.007	.000	.370	.000
	N	216	216	216	216	216	216	216	216	216	216	216	216
Total recovered	τb	.868**	.250**	.734**	.298**	1.000	.289**	.700**	.518**	.313**	.389**	.088	.487**
	Sig.	.000	.000	.000	.000	.	.000	.000	.000	.000	.000	.055	.000
	N	216	216	216	216	216	216	216	216	216	216	216	216
New recovered	τb	.272**	.843**	.244**	.744**	.289**	1.000	.288**	.192**	.067	.093	.078	.217**
	Sig.	.000	.000	.000	.000	.000	.	.000	.001	.201	.080	.142	.000
	N	216	216	216	216	216	216	216	216	216	216	216	216
Active cases	τb	.703**	.253**	.666**	.334**	.700**	.288**	1.000	.521**	.310**	.411**	.078	.431**
	Sig.	.000	.000	.000	.000	.000	.000	.	.000	.000	.000	.092	.000
	N	216	216	216	216	216	216	216	216	216	216	216	216
Critical cases	τb	.607**	.176**	.570**	.240**	.518**	.192**	.521**	1.000	.399**	.489**	.210**	.322**
	Sig.	.000	.001	.000	.000	.000	.001	.000	.	.000	.000	.000	.000
	N	216	216	216	216	216	216	216	216	216	216	216	216
Total cases /1M	τb	.368**	.087	.301**	.144**	.313**	.067	.310**	.399**	1.000	.674**	.447**	-.088
	Sig.	.000	.097	.000	.007	.000	.201	.000	.000	.	.000	.000	.055
	N	216	216	216	216	216	216	216	216	216	216	216	216
Deaths /1M	τb	.466**	.108*	.502**	.196**	.389**	.093	.411**	.489**	.674**	1.000	.327**	.066
	Sig.	.000	.041	.000	.000	.000	.080	.000	.000	.000	.	.000	.152
	N	216	216	216	216	216	216	216	216	216	216	216	216
Tests /1M	τb	.136**	.113*	.075	.048	.088	.078	.078	.210**	.447**	.327**	1.000	-.184**
	Sig.	.003	.032	.103	.370	.055	.142	.092	.000	.000	.000	.	.000
	N	216	216	216	216	216	216	216	216	216	216	216	216
Population	τb	.536**	.208**	.557**	.227**	.487**	.217**	.431**	.322**	-.088	.066	-.184**	1.000
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.055	.152	.000	.
	N	216	216	216	216	216	216	216	216	216	216	216	216

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

p=0.103), new deaths versus tests/1M (τb=0.048, p=0.370), total recovered versus tests/1M (τb=0.088, p=0.055), new recovered versus total cases/1M (τb=0.067, p=0.201), new recovered versus deaths/1M (τb=0.093, p=0.080), new recovered versus tests/1M (τb=0.078, p=0.142), active case versus tests/1M (τb=0.078, p=0.092), total cases/1M versus population (τb=-0.088, p=0.055), deaths/1M versus population (τb=0.066, p=0.152). All of the correlations were in the positive direction except for the population versus total cases/1M (τb=-0.088, p=0.055), and population versus tests/1M (τb=-0.184, p<0.001). Further, all bivariate correlations had a weak-to-moderate effect size. In contrast, some had a strong effect size, including the total cases versus total deaths (τb=0.826, p<0.001), total cases versus total recovered (τb=0.868, p<0.001), total deaths versus total recovered (τb=0.734, p<0.001), total cases versus active cases (τb=0.703, p<0.001), total recovered versus active cases (τb=0.700, p<0.001), new cases versus new deaths (τb=0.752, p<0.001), new cases versus new recovered (τb=0.843, p<0.001), and new deaths versus new recovered (τb=0.744, p<0.001).

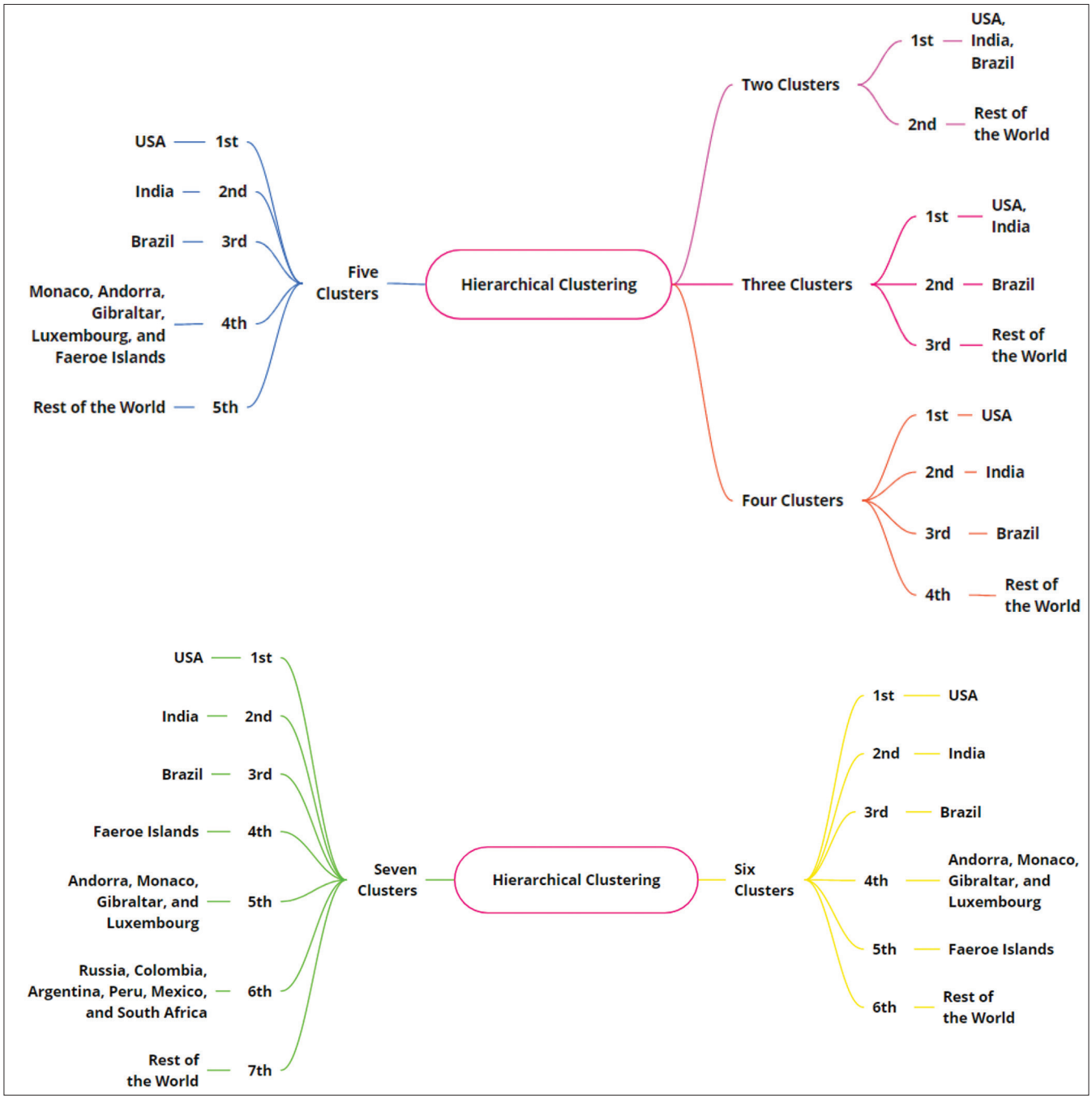

Figure 1: Hierarchical clustering: two, three, four, five, six, and seven clusters' visualization

To summarize, almost all of the correlations were significant and in the positive direction with weak-to-moderate effect size, and some correlations were strongly positive for parameters involving the total cases, total deaths, total recovered, new cases, new deaths, new recovered, and active cases. Besides, the population count for each country had almost no correlation or an inverse correlation of weak effect size with the total cases/1M, deaths/1M, and tests/1M, while having strongly significant correlations (p-value<0.001) with the rests of the parameters, and these were in the positive direction with a moderate effect size.

We conducted a cohort, a total of nine, hierarchial cluster analyses by pre-specifying the number of clusters from two to ten clusters (Figures 1 and 2). The final cluster analysis, with ten clusters, was most conclusive (Figure 2). The first, second, third, fourth, and fifth clusters had the United States, India, Brazil, Faeroe Islands, and Russia, respectively. Spain, France, and the United Kingdom allocated to the sixth cluster, while Andorra, Monaco, Gibraltar, and Luxembourg allocated to the seventh cluster. The eighth cluster had South Africa and some Latin American nations including Columbia, Argentina, Peru, and Mexico. The ninth clusters had two countries from the Middle East, the UAE and Bahrain, as well as Malta, Singapore, Iceland, and Denmark, in addition to small islands, including the Channel Islands, Cayman Islands, Falkland Islands, and Bermuda. The rest of the world, represented by Canada, the vast majority of Africa, Europe, and Australasia allocated to the tenth cluster (Figure 3). These results may have potential links to the "exodus" of archaic humans out of Africa towards Australasia and Europe.[10]

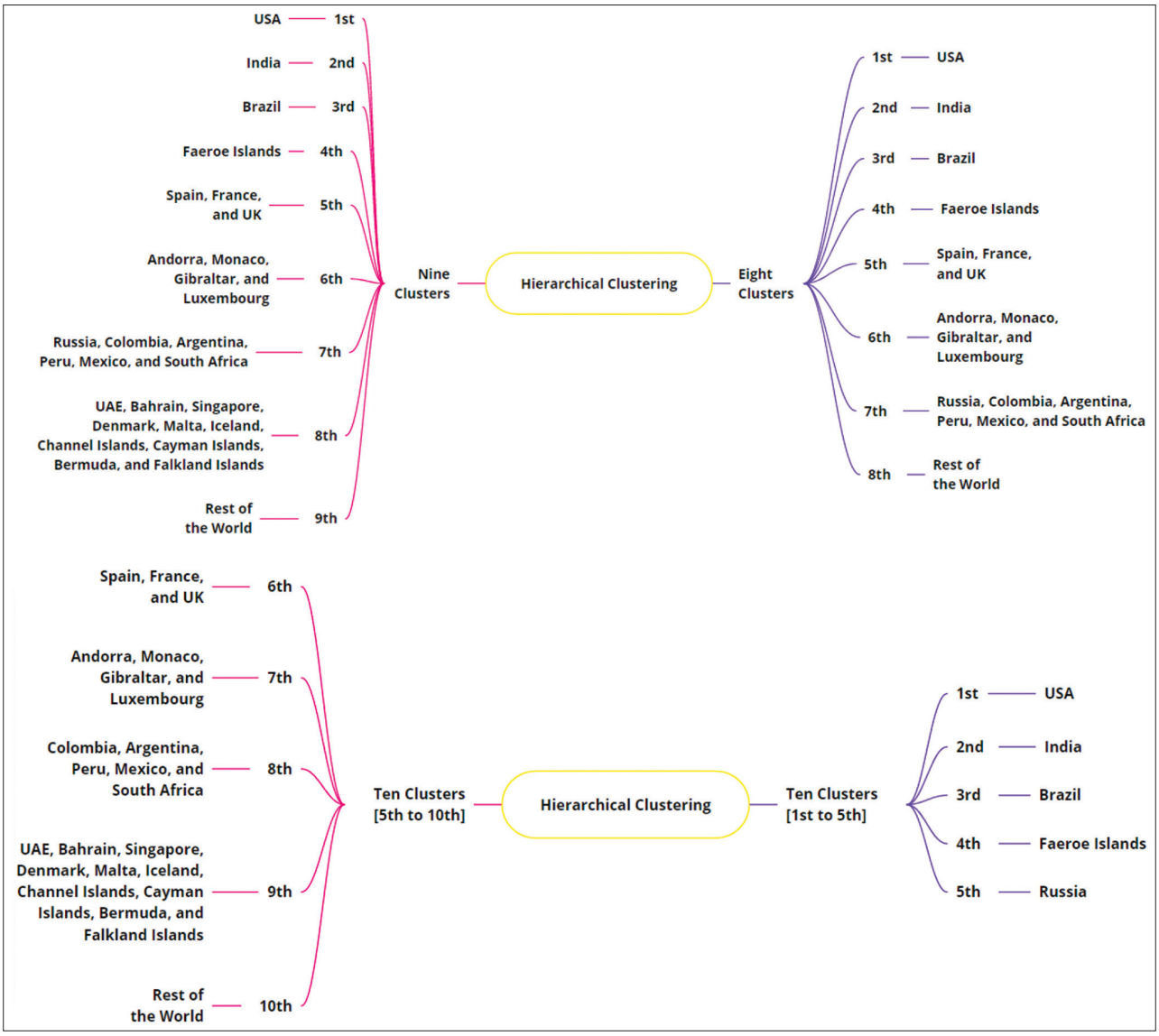

Figure 2: Hierarchical clustering: eight, nine and ten clusters' visualization

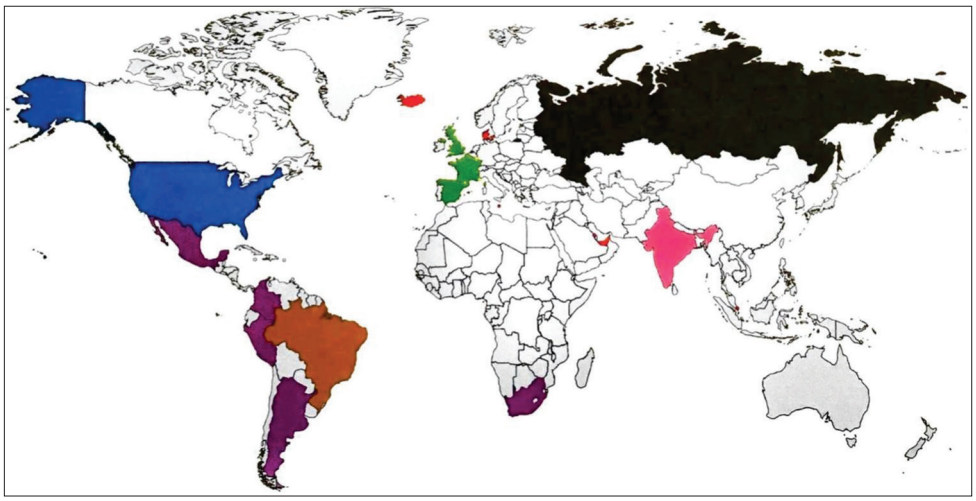

Figure 3: Hierarchical clustering: SARS-CoV-2 world map with ten clusters
*Different colour-coding represent different clusters.
**Some clusters were too small to be visualized on the map, for example, Faeroe Islands.

HIERARCHICAL CLUSTERING: THE CASE OF IRAQ, NEIGHBOURING COUNTRIES, AND THE MIDDLE EAST

According to our summative clustering analysis, Iraq and its neighbours as well as the entire Middle East allocated within the same cluster, together with most of the world countries. However, when we pre-determined the number of clusters into nine and ten, the UAE and Bahrain assigned to a separate clusters together with Malta, Singapore, Denmark, Iceland, Bermuda, Channel Islands, Cayman Islands, and Falkland Islands. As of 5 October 2020, and according to Worldometer website and COVID-19 application (iOS version 0.9.16), the number of daily cases and deaths increased all over the Middle East.[2,11] The neighboring countries of Iraq also witnessed an exponential growth of SARS-CoV-2 infections, including Iran (total cases=475,674, total deaths=27,192, total recoveries=392,293, mortality rate=6.48%, basic reproduction number=1.07), Kuwait (cases=107,592, deaths=628, recoveries=99,549, MR=0.58%, BRN=0.92), Saudi Arabia (cases=336,766, deaths=4,898, recoveries=322,055, MR=1.45%, BRN=0.88), Jordan (cases=17,464, deaths=110, recoveries=5,292, MR=0.63%, BRN=1.81), Syria (cases=4,411, deaths=207, recoveries=1,168, MR=4.69%, BRN=0.99), and Turkey (cases=326,046, deaths=8,498, recoveries=286,370, MR=2.61%, BRN=0.93).[2,10] Unfortunately, Iraq is in the lead of its Arab neighbours in connection with the stats of the pandemic (cases=382,949, deaths=9,464, recoveries=312,158, MR=2.47%, BRN=1.01).[2,11]

CONCLUSION

Our analysis hints territorial, racial and genomic basis of the pandemic, and potential links to ancestral migrations of archaic humans, including Homo ergaster, Homo erectus, and Homo sapiens, out of Africa and across Australasia and Europe.We opine that the current study is of prime importance for epidemiology-based decision making by health officials, and it can provide novel information and insights to predict future changes of the pandemic in developing countries, including countries from the Middle East, that can be conveyed to the Iraqi authorities as well as international authorities for precautionary and preventive measures in facing the pandemic.

AVAILABILITY OF DATA

Our Data, including raw dataset, are available upon request from the corresponding author.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest and that they have self-funded this study.

Key words: Artificial intelligence; cluster analysis; coronaviridae; COVID-19; epidemiology; internet; machine learning; novel coronavirus; SARS-CoV-2

Ahmed Al-Imam[1,2], Usama Khalid[3], Hend Al-Doori[4]

[1]Department of Anatomy and Cellular Biology, College of Medicine, University of Baghdad, Iraq, [2]Barts and The London School of Medicine and Dentistry, Queen Mary University of London, the United Kingdom, [3]Enjaz Limited Liability Company, Baghdad, Iraq, [4]Al-Betool Teaching Hospital, Diyala Health Directorate, Ministry of Health, Iraq.

Address for Correspondence:
Dr. Ahmed Al-Imam, Department of Anatomy and Cellular Biology, College of Medicine, University of Baghdad [Iraq].
Mobile No: +964 (0) 771 433 8199.
E-mail: ahmed.lutfi@uob.edu.iq

REFERENCES

1. Motyka MA, Al-Imam A and Aljarshawi MHA. SARS-CoV-2 pandemic as an anomie. Social Space/PrzestrzeńSpołeczna. 2020; 20.
2. Worldometer - real time world statistics [Internet]. Worldometer. 2020 [cited 29 September 2020]. Available from: https://www.worldometers.info/
3. Amanat F and Krammer F. SARS-CoV-2 Vaccines: Status Report. Immunity. 2020; 52(4): 583-589.
   https://doi.org/10.1016/j.immuni.2020.03.007
4. De Wit E, Van Doremalen N, Falzarano D and Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. Nature Reviews Microbiology. 2016; 14: 523-534.
   https://doi.org/10.1038/nrmicro.2016.81
5. Al-Imam A, Motyka MA and Jędrzejko MZ. Conflicting Opinions in Connection with Digital Superintelligence. IAES International Journal of Artificial Intelligence. 2020; 9(2): 336-348.
   https://doi.org/10.11591/ijai.v9.i2.pp336-348
6. Al-Imam A, Motyka MA, Sahai A and Konuri VK. The "March of Progress": From Cosmic Singularity to Digital Singularity. Current Trends in Information Technology. 2020; 10(1): 1-8.
7. Al-Imam A and Motyka MA. On the Necessity for Paradigm Shift in Psychoactive Substances Research: The Implementation of Machine Learning and Artificial Intelligence. Alcoholism and Drug Addiction/AlkoholizmiNarkomania. 2019; 32(3): 1-6.
   https://doi.org/10.5114/ain.2019.91004
8. Al-Imam A and Al-Lami F. Machine Learning for Potent Dermatology Research and Practice. Journal of Dermatology and Dermatologic Surgery. 2020; 24(1): 1-4.
   https://doi.org/10.4103/jdds.jdds_54_19
9. Web Scraper - The #1 web scraping extension [Internet]. Webscraper.io. 2020 [cited 29 September 2020]. Available from:
   https://www.webscraper.io/
10. Foley R and Lahr M. Beyond "out of Africa": reassessing the origins of Homo sapiens. Journal of Human Evolution. 1992;22(6):523-529.
    https://doi.org/10.1016/0047-2484(92)90085-N
11. COVID-19! [Internet]. App Store. 2020 [cited 6 October 2020]. Available from:
    https://apps.apple.com/us/app/covid-19/id1504906590

Author's Contribution:
Ahmed Al-Imam worked on raw data, conducted data analytics, wrote the first draft of the article, and prepared the manuscript for scholarly submission. Usama Khalid collected data from the surface web using a dedicated python script (code) and web-scraping tools. Hend Al-Doori Contributed to writing the first draft..

Work attributed to:
Department of Anatomy and Cellular Biology, College of Medicine, University of Baghdad, Iraq and Barts and The London School of Medicine and Dentistry, Queen Mary University of London, the United Kingdom.

Orcid ID:
Dr. Ahmed Al-Imam - https://orcid.org/0000-0003-1846-9424

Source of funding: Nil, **Conflict of Interest:** None declared.