Journal of Medical Engineering & Technology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ijmt20

A region-specific clustering approach to investigate risk-factors in mortality rate during COVID-19: comprehensive statistical analysis from 208 countries

Poojita Garg & Deepak Joshi

To cite this article: Poojita Garg & Deepak Joshi (2021): A region-specific clustering approach to investigate risk-factors in mortality rate during COVID-19: comprehensive statistical analysis from 208 countries, Journal of Medical Engineering & Technology, DOI: 10.1080/03091902.2021.1893398

To link to this article: https://doi.org/10.1080/03091902.2021.1893398

Published online: 22 Mar 2021.

JOURNAL OF MEDICAL ENGINEERING & TECHNOLOGY
https://doi.org/10.1080/03091902.2021.1893398

Taylor & Francis
Taylor & Francis Group

INNOVATION

A region-specific clustering approach to investigate risk-factors in mortality rate during COVID-19: comprehensive statistical analysis from 208 countries

Poojita Garg[a] and Deepak Joshi[b,c]

[a]Computer Science and Engineering, University Institute of Engineering and Technology, Panjab University, Chandigarh, India; [b]Centre for Biomedical Engineering, IIT Delhi, New Delhi, India; [c]Department of Biomedical Engineering, AIIMS Delhi, New Delhi, India

ABSTRACT

Since the outbreak of the novel coronavirus, COVID-19 has continuously spread across the globe briskly. However, since its existence, the symptoms of the disease have been varying widely; thus, developing an urgent need to stratify high-risk categories of people who show more propensity to be affected by this deadly virus will be beneficial for health care. Using the open-access data and machine learning algorithms, this paper aims to cluster countries in groups with similar profiles with respect to the country level pre COVID-19 pandemic parameters. The purpose of performing the data analysis is to measure the extent to which these major risk factors determine the mortality rate due to the coronavirus disease 2019. An unsupervised machine learning model (*k*-means) was employed for two hundred and eight countries to define data-driven clusters based on thirteen country-level parameters. After performing the one-way ANOVA for comparing the clusters in terms of total cases, total deaths, total cases per population, total deaths per population, and death rate, the paradigm with four and seven clusters showed the best ability to stratify the countries according to total cases per population and death rate with *p*-values of less than 0.05 and 0.001, respectively. However, the model could not stratify countries in total deaths/cases and total deaths per population.

ARTICLE HISTORY
Received 1 February 2021
Accepted 14 February 2021

KEYWORDS
K-means; machine learning; COVID-19; risk-factors

1. Introduction

The global pandemic caused by the novel coronavirus has pigeonholed the entire population on the basis of geographical, socio-economic, age, and health factors in terms of risk associated with them. Soon the novel virus started firming its grip after being first sighted at Wuhan, China. In late December 2019, researchers have been keen in finding the pertinent factors affecting both its spread as well as to identify the high-risk categories of the population. With over 32 million cases and about a million deaths worldwide, the outbreak has served no less than a catastrophe that has been able to throw both our world and our lives into turmoil in no time. Along with the burgeoning rate of fatalities, the widespread signs of stigma and depression [1] have also proven to be major concerns worldwide. Till any effective vaccine is available for COVID-19 worldwide, countries have been empowering their communities to adhere to the guidelines such as appropriate social distancing, wearing masks, and frequent handwashing as issued by the world health organisation (WHO) to prevent the spread of this infectious respiratory syndrome. The current research highlights a variety of probable COVID-19 risk-factors. A worldwide view of country-level data would aid us to see the universal trends as well as serve as a benchmark to predict the future behaviour of the countries based on the past trends of similar countries in terms of the risk factor values. The pervasive nature of the coronavirus and its ability to affect humans at large irrespective of geography has initiated a need to investigate the risk factors globally. The research is an extension to a previous study [2], which overcomes some of its shortcomings by including a larger number of countries as well as newly recognised risk factors. Thus, in this research, we aim to test and analyse a few of the major risk factors such as population, GDP, extreme poverty levels, count of hospital beds, human development indicators, age, smoking, cardiovascular diseases, diabetes, and life expectancy associated with COVID-19 on global data in order to present a quantitative analysis of the

CONTACT Deepak Joshi joshid@iitd.ac.in Centre for Biomedical Engineering, IIT Delhi, New Delhi, India; Department of Biomedical Engineering, AIIMS Delhi, New Delhi, India

trends shown by the countries with similar values of the risk factors by employing unsupervised machine learning algorithms.

2. Methodology

2.1. Data source

The data for the total cases, deaths, and risk factors were obtained from the Our World in Data website [3]. All the data used for analysis in the study is until 16 September 2020. A total of 13 parameters as described in Table 1 have been selected for 208 countries and discussed in the following subsection. These 13 predictors were used as inputs to cluster the countries by employing unsupervised machine learning algorithms where the analysis unit was a country.

2.1.1. Population

COVID-19 being an infectious respiratory syndrome that spreads through cough and sneezing, has led to social distancing, is an important preventive measure in order to limit the spread of coronavirus [4]. Thus, countries with high population and population density pose a greater challenge to maintain social distance thus are likely to have an increasing trend in fatalities caused due to COVID-19. We use the population as one of the determinants for our cluster analysis.

2.1.2. Age

As per the study [5], there is a general consensus about the coronavirus disease posing a greater risk on the elderly population, especially those above the age of 65. The probable death of patients is as high as 15% for the people above 80 years of age as well as the people in their 50's is expected to be three times more prone to death due to the COVID-9 than the ones in their 40's [6]. In order to incorporate age as a factor in the clustering analysis, we have three predictors indicating the median age of the population of the country, the share of population above 65 years of age, and the share of population above 70 years of age.

2.1.3. Socio-economic factors

The major socio-economic factors for a country include GDP per capita, extreme poverty statistics, life expectancy, hospital beds' availability per 1000 people, and human development index, which is a composite measure of the key factors of human development such as literacy, the standard of living and life expectancy. These predictors have a significant effect on the pandemic spread and represent the basis for future preventive measures to effectively challenge another COVID-19 wave or any future pandemics [7].

2.1.4. Physiological factors

According to the Centres for Disease Control and Prevention, certain medical conditions place people of any age group at a higher risk of severe illness due to COVID-19 [8]. Some of these common diseases have been considered in the study, including the death rate from cardiovascular diseases and the percentage of the population having diabetes. Also, as smoking is

Table 1. Description of COVID-19 parameters and risk factors variables.

Concept	Variables	Description
COVID-19 parameters	Total cases	Total confirmed cases of COVID-19(as of 16/09/2020)
	Total deaths	Total confirmed deaths of COVID-19(as of 16/09/2020)
	Total cases per population	Total confirmed cases per population(as of 16/09/2020)
	Total deaths per population	Total confirmed deaths per population(as of 16/09/2020)
	Death rate[a]	Total confirmed deaths per total confirmed cases(as of 16/09/2020)
Population	Population	Population in 2020
Age	Median age	Median age of the population, UN projection for 2020
	Aged 65 years older	Share of the population that is 65 years and older, most recent year available
	Aged 70 years older	Share of the population that is 70 years and older in 2015
Socio-economic factors	GDP per capita	Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available
	Extreme poverty	Share of the population living in extreme poverty, most recent year available since 2010
	Hospital beds per thousand	Hospital beds per 1000 people, most recent year available since 2010
	Life expectancy	Life expectancy at birth in 2019
	Human development index	Summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living
Physiological factors	Diabetes prevalence	Diabetes prevalence (% of population aged 20–79) in 2017
	Cardiovascular death rate	Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)
	Female smokers	Share of women who smoke, most recent year available
	Male smokers	Share of men who smoke, most recent year available

[a]As per equation 1.

one of the major reasons for weak lungs and other respiratory ailments, we have considered two predictors for the share of women who smoke as well as that for men for the purpose of our analysis.

2.2. Data preprocessing

A preprocessing step was performed on the available data before analysis. The data obtained from the source had several missing values. We used a machine learning model based imputer to handle the missing values so that minimal loss of information occurs. The model used for the purpose was a k-nearest neighbour (kNN) with *k* equal to 10 and 'nan-euclidian' as the distance metric from the Scikit-Learn library. The missing values of one feature are imputed by the average of those from its k nearest neighbours, which are calculated using the remaining features of the dataset. The principle followed by the kNN imputer is more robust and sensitive than any other value estimation method. Apart from missing value imputation, the data was normalised for accurate analysis using the MinMax scaler from the Scikit-Learn library. All the preprocessing and data analysis is done using Python 3.

2.1. Clustering using K-means

For the purpose of clustering the countries into groups, a popular unsupervised machine learning algorithm, namely *k*-means from Scikit Learn was used. The above mentioned 13 common predictors for each country were used as input to our model. The elbow test depicted in Figure 1 was used to predict the number of categories best suited for our data. Though the data is substantially limited in number, thus in such cases, the values near the predicted number of categories resulted after the elbow test provide similar information, thus necessitating the need for visual inspection. So, for the purpose of visual inspection, t-Distributed Stochastic Neighbour Embedding (t-SNE), which is a method for visualisation of high-dimensional datasets by reducing the dimensions well suited for low dimensional datasets, was used. It is a similar method like principal component analysis(PCA) but contrary to the fact that PCA is a mathematical technique whereas t-SNE being a probabilistic one. The principle followed by t-SNE is to minimise the divergence between the distribution that measures pairwise similarities of the input objects and that of which measures the pairwise similarities of the respective low dimensional data points. A silhouette

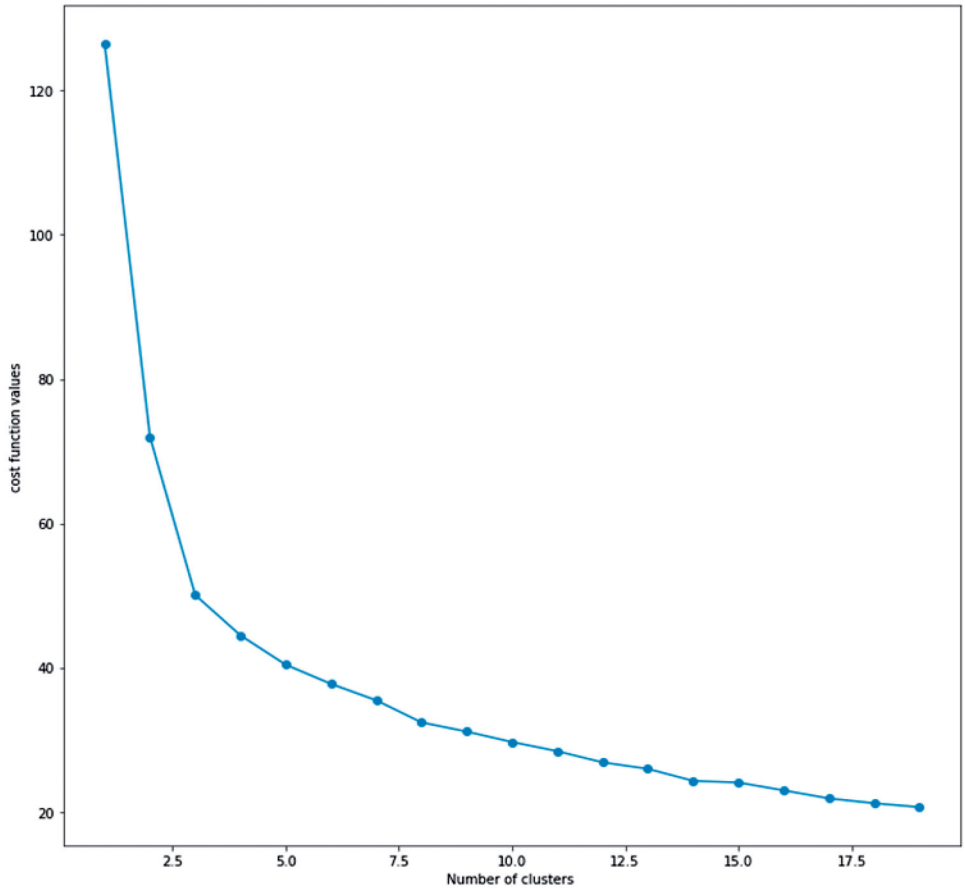

Figure 1. Elbow test for the K-means analysis.

Table 2. *p*-Values obtained after ANOVA test for different number of clustered countries with their respective parameters with the text highlighted in bold representing values strictly less than threshold and the text in bold italics representing marginally significant values.

Cluster #	Death rate	Total cases	Total deaths	Total deaths/pop	Total cases/pop
3	>0.05	>0.05	>0.05	>0.05	***0.059***
4	>0.05	>0.05	>0.05	>0.05	**0.038**
5	>0.05	>0.05	>0.05	>0.05	>0.05
6	>0.05	>0.05	>0.05	>0.05	***0.066***
7	**0.00074**	>0.05	>0.05	>0.05	>0.05

Table 3. *p*-Values corresponding to the *t*-tests on seven clusters with the text highlighted in bold representing values strictly less than the threshold.

Clusters #	0	1	2	3	4	5
1	0.064					
2	0.66	0.149				
3	0.82	0.59	0.185			
4	0.077	0.908	0.172	0.654		
5	**0.00010**	0.008	**0.0007**	0.0083	0.0075	
6	0.358	0.59	0.521	0.95	0.658	0.0172

score was used to interpret the consistency within clusters.

2.2. Statistical analysis

The statistical analysis was performed on five features of COVID-19, such as total confirmed cases, total confirmed deaths, death rate given in Equation 1, total confirmed deaths per population, and total deaths per population.

$$\text{Death rate} = \text{Total confirmed deaths}/\text{Total confirmed cases} \quad (1)$$

The clusters were statistically examined by employing a one-way ANOVA test with the commonly accepted threshold being 0.05. The clusters were pairwise compared using the *t*-tests along with employing the Bonferroni correction for multiple tests. The analysis was performed using the statistical functions of the SciPy library.

2. Results

The clustering and the *post hoc* analysis was done on 208 countries. The consistency within the clusters was measured using the silhouette score. The silhouette scores for different categories are 2(0.35), 3(0.33), 4(0.26), 5(0.24), 6(0.23), 7(0.23), 8(0.21), 9(0.21), 10(0.21); for all other categories ranging from 2 to 10 the silhouette score was less than 0.21. Results after employing *k*-means with *k* being 3 showed the maximum silhouette score but with a high concentration of data points in one category, thus leading to an imbalance within categories. The post- hoc analysis was done on all of the above clusters. The p-values obtained after performing the ANOVA test on the different number of clusters are shown in Table 2.

Table 4. *p*-Values corresponding to the *t*-tests on three clusters with.

Clusters #	0	1
1	0.029	
2	0.155	0.1832

Table 5. *p*-Values corresponding to the *t*-tests on four clusters.

Clusters #	0	1	2
1	0.044		
2	0.020	0.497	
3	0.102	0.748	0.345

Table 6. *p*-Values corresponding to the *t*-tests on six clusters.

Clusters #	0	1	2	3	4
1	0.788				
2	0.0326	0.025			
3	0.18	0.150	0.455		
4	0.05	0.038	0.6828	0.660	
5	0.6579	0.539	0.082	0.3580	0.1335

After observing and comparing the resulting p-values with the threshold value of 0.05, the 7 clusters showed a significant value of 0.00074, and when tested for the parameter of death rate along with 4 clusters, which gave a *p*-value of 0.038 for the parameter of total cases per population. Furthermore, clusters 3 and 6 showed marginally significant *p*-values of 0.059 and 0.066, respectively. Additionally, to further investigate the dissimilarities between the clusters, these categories were chosen. Upon applying t-tests on each pair of categories for specific numbers of clusters, the results for that of 7 clusters for the parameter of death rate are depicted in Table 3. The threshold was modified by applying the Bonferroni correction for multiple tests, which came out to be 0.0023. The tests revealed a significant dissimilarity between category 5 with respect to categories 0 and 2 with *p*-values of 0.00010 and 0.0007.

Similarly, Tables 4–6 depict the p-values obtained after performing the pairwise t-tests on clusters 3,4, and 6 with respect to the parameter total cases per population with the modified threshold after correction of 0.016, 0.008, 0.003. It can be evidently stated that no

Table 7. Average input values across seven clusters.

Clusters #	0	1	2	3	4	5	6
Population	0.014	0.117	0.071	0.011	0.040	0.017	0.008
Median age	0.684	0.115	0.413	0.873	0.386	0.831	0.530
Aged 65 older	0.463	0.071	0.195	0.658	0.211	0.722	0.219
Aged 70 older	0.427	0.067	0.180	0.626	0.196	0.701	0.199
GDP per capita	0.283	0.019	0.074	0.202	0.093	0.422	0.320
Extreme poverty	0.011	0.509	0.074	0.014	0.078	0.004	0.011
Cardiovascular disease death rate	0.179	0.418	0.631	0.447	0.253	0.076	0.247
Diabetes	0.403	0.173	0.355	0.281	0.318	0.225	0.682
Female smokers	0.349	0.053	0.122	0.664	0.130	0.562	0.139
Male smokers	0.384	0.224	0.587	0.502	0.295	0.261	0.360
Hospital beds per 1000	0.281	0.066	0.242	0.483	0.135	0.335	0.172
Life expectancy	0.836	0.278	0.572	0.751	0.618	0.907	0.711
Human development index	0.817	0.210	0.548	0.789	0.567	0.935	0.735

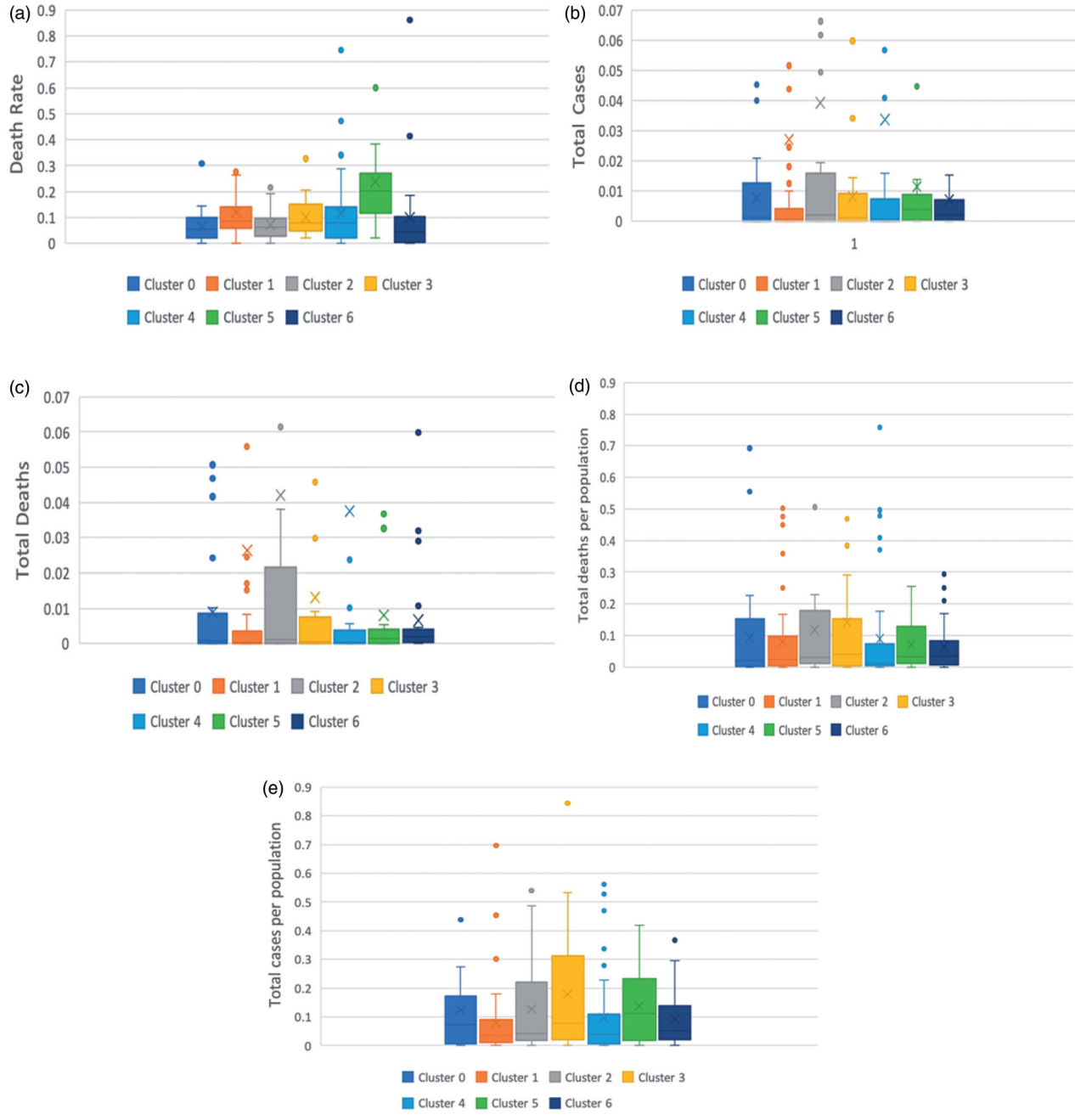

Figure 2. Boxplots showing the distribution of COVID-19 parameters across categories of seven clusters.

pairwise comparison gave a significant p-value. However, a p-value of 0.029 and 0.020 obtained after performing the *t*-tests of between categories 0 and 1 with clusters 3 and categories 0 and 1, and 0 and 2 with clusters 4 showed the closest results but not enough to be considered statistically significant.

3. Discussion

The study was performed to assess the significance of various popular COVID-19 risk-factors with regard to the trend of COVID-19 cases and deaths worldwide. The study assumes its application in predicting the countries' future behaviour with respect to how significant the impact of the risk-factors would be. The global pandemic has brought the whole world together to fight this serious threat to mankind; however, the socio-economic, geographic, and epidemiological differences between the countries have served as the prime factors that could explain the varying impact of this epidemic on different countries.

The clusters were made to group the countries with similar risk-factor values, which would foster those countries to learn from the shortcomings as well as guide them to adopt similar strategies as that of the other countries of the same cluster. The input parameters could potentially provide an explanation for the cluster configuration. As per the average values of the input parameters shown in Table 7, in the model with 7 clusters with respect to the death rate, category 6 showed the highest death rate than other clusters, implying a high proportion of people dying if once affected by the virus. Figure 2 depicts the median and interquartile range of the COVID-19 parameters for different categories of 7 clusters.

The possible reason could be having a high median age as well as a high proportion of people over 65 years of age. This is in compliance with the past studies depicting the vulnerability of the elderly to COVID-19 [8]. With a high GDP per capita, low population, highest female smoker population, considerable amount of hospital beds per 1000, and human development index as compared to other categories of the same model also concludes that the biological immunity and respiratory health are the major factors impacting the deaths caused by this acute respiratory disease. In contrast, the category 1 and 3 showed the comparatively lowest death rate amongst the seven categories. Few of the possible reasons explaining this contrast could be explained by the fact of having a lower proportion of elderly population and female smokers in the clustered countries [9].

4. Conclusion

In this paper, an unsupervised machine learning algorithm has been developed which groups the countries with similar profiles with respect to the country-level risk factors and evaluates them based on the metrics of total deaths/cases, total deaths per population, total cases per population, and death rate. The work provides a paradigm to help identify the countries more vulnerable to the ongoing pandemic as well as present the clusters of similar profile countries and whose past trends could be an important consideration while developing their future mitigation plans for the ongoing pandemic.

Acknowledgements

The authors are grateful to the Our World in Data Website [3] for providing the open data for our research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Rajkumar RP. COVID-19 and mental health: a review of the existing literature. Asian J Psychiatr. 2020;52:102066.

[2] Carrillo-Larco RM, Castillo-Cara M. Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: an unsupervised machine learning approach. Wellcome Open Res. 2020;5:56.

[3] https://ourworldindata.org/coronavirus. [cited 2020 Dec 15].

[4] Rajan K, Dhana K, Barnes LL, et al. Strong effects of population density and social characteristics on distribution of COVID-19 infections in the United States. medRxiv. 2020.

[5] Novel CP. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi. 2020;41(2):145.

[6] https://www.businessinsider.com/coronavirus-death-age-older-people-higher-risk-2020-2?r=US&IR=T. [cited 2020 Dec 15]

[7] Gangemi S, Billeci L, Tonacci A. Rich at risk: socio-economic drivers of COVID-19 pandemic spread. Clin Mol Allergy. 2020;18(1):1–3.

[8] https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fneed-extra-precautions%2Fgroups-at-higher-risk.html. [cited 2020 Dec 15].

[9] Mahase E. Covid-19: why are age and obesity risk factors for serious disease? BMJ 2020;371:m4130.