Original Research MMJ

Assessment of the dissimilarities of totally 186 countries and regions according to COVID-19 indicators at the end of March 2020

Handan Ankarali[1], Unal Uslu[2], Seyit Ankarali[3], Sengul Cangur[4]

1. Department of Biostatistics and Medical Informatics, Faculty of Medicine, Istanbul Medeniyet University, Turkey
2. Department of Histology and Embryology, Faculty of Medicine, Istanbul Medeniyet University, Turkey
3. Department of Physiology, Faculty of Medicine, Istanbul Medeniyet University, Turkey
4. Department of Biostatistics and Medical Informatics, Faculty of Medicine, Duzce University, Turkey

Correspondence: Sengul Cangur (sengulcangur@duzce.edu.tr)

Abstract

Background
This study is aimed at evaluating the relationship between the number of days elapsed since a country's first case(s) of coronavirus disease 2019 (COVID-19), the total number of tests conducted, and outbreak indicators such as the total numbers of cases, deaths, and patients who recovered. The study compares COVID-19 indicators among countries and clusters them according to similarities in the indicators.

Methods
Descriptive statistics of the indicators were computed and the results were presented in figures and tables. A fuzzy *c*-means clustering algorithm was used to cluster/group the countries according to the similarities in the total numbers of patients who recovered, deaths, and active cases.

Results
The highest numbers of COVID-19 cases were found in Gibraltar, Spain, Switzerland, Liechtenstein and Italy were also of that order with about 1500 cases per million population. Spain and Italy had the highest total number of deaths, which were about 140 and 165 per million population, respectively. In Japan, where exposure to the causative virus was longer than in most other countries, the total number of deaths per million population was less than 0.5. According to cluster analysis, the total numbers of deaths, patients who recovered, and active cases were higher in Western countries, especially in central and southern European countries, which had the highest numbers when compared with other countries.

Conclusion
There may be various reasons for the differences between the clusters obtained by fuzzy *c*-means clustering. These include quarantine measures, climatic conditions, economic levels, health policies, and the duration of the fight against the outbreak.

Key Words: COVID-19, total number of cases, total number of deaths, outbreak, clustering

Introduction

Coronavirus disease 2019 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a zoonotic that crossed species to infect human populations and was identified first in Wuhan, China. As for severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS), both of which are human respiratory syndromes, the virus causing COVID-19 also often causes severe respiratory symptoms that can be fatal. The World Health Organization (WHO) first determined that the global risk of a COVID-19 pandemic was "very high" on 28 February 2020, subsequently declaring the outbreak to be a pandemic on 11 March 2020[1]. At that time, COVID-19 had been detected in 81 countries, with 57 countries registering 10 cases or fewer. Around 12 March 2020, the centre of the pandemic moved from China to Europe; subsequently, the number of countries exposed to COVID-19 reached 186 by the end of March[2]. Because the outbreak has affected the world in many respects, a summary of the current situation is of particular importance, and identification of the similarities and differences between countries in terms of the measures being taken is crucial.

The first objective of this study is to define the relationships between the outbreak indicators of 34 countries that had reported the total number of tests conducted by the end of March 2020 and the duration of the fight against the outbreak. The second objective is to cluster totally 186 countries and regions countries according to the outbreak indicators (i.e. the total number of patients who recovered per million population, the total number of deaths per million population, and the total number of active cases per million population) to make it easier to track the outbreak and to evaluate countries' policies related to the pandemic.

Methods

Study population

In this study, the data for the 34 countries that had reported the total number of tests conducted by the end of March 2020 were used for the first objective. The outbreak indicators obtained from totally 186 countries and regions and also two ships were analysed for the second objective.

Study design and data collection technique

This was a cross-sectional study. Data on the total number of tests conducted and the total number of cases in 33 countries were collected between 17 and 20 March 2020. In addition, data from Turkey were collected on 26 March 2020. The data

https://dx.doi.org/10.4314/mmj.v34i2.2.

were analysed according to the following indicators:

- Total number of cases per total number of tests (%):

$$\frac{\text{Total number of cases}}{\text{Total number of tests}} \times 100. \qquad (1)$$

- Total number of cases per million population:

$$\frac{\text{Total number of cases}}{\text{Population of the country}} \times 1000000. \qquad (2)$$

The population size and outbreak indicators, which included the confirmed cases, patients who recovered, deaths, and active cases per day in each country, were mined from open-access public databases on 29 March 2020[3–5]. The ratio of the total number of cases to the total number of tests performed indicates how many people had positive results per 100 tests. In addition, the other indicators used were as follows:

- Daily number of new cases
- Total number of deaths
- Total number of patients who recovered
- Total number of active cases
- Total number of critical cases

The total number of cases was defined as the total number of deaths plus the total number of patients who recovered plus the total number of active cases.

The number of days elapsed between the date of the first reported case and 29 March 2020 was taken into account when we compared countries in terms of outbreak indicators. These days were then divided into ten periods at appropriate intervals, and the effects of these periods on the indicators were re-evaluated from a different perspective. The periods were as follows:

- 31 December 2019 to 15 January 2020
- 16–31 January 2020
- 1–7 February 2020
- 8–15 February 2020
- 16–20 February 2020
- 21–29 February 2020
- 1–7 March 2020
- 8–15 March 2020
- 16–21 March 2020
- 22–29 March 2020

Eligibility criteria

The countries selected for evaluation of the first objective are those that had reported the total number of tests conducted by the end of March 2020.

Data from all countries reporting outbreak indicators published by the end of March 2020 were used to evaluate the second objective.

Ethical considerations

All the data were obtained from open-access public databases; these were Worldometer, the WHO database, and the Johns Hopkins University & Medicine Coronavirus Resource Center database[3–5]. Therefore, ethical approval was not required.

Statistical analysis

The descriptive values, the median value, the 25th and 75th quartiles, the mode, and the minimum and maximum of the outbreak indicators from the countries with outbreaks in the given periods were calculated. All figures were drawn with use of the program Datawrapper[6] for the first objective. The Kruskal–Wallis test followed by the Dunn post hoc test was used for comparison of the ten periods for the four outbreak indicators in Figures 5–8. The fuzzy *c*-means (FCM) clustering algorithm was used to cluster the countries by use of the total number of deaths, total number of patients who recovered, and total number of active cases per million population. All statistical analyses were done with IBM SPSS Statistics for Windows version 25.0 (IBM SPSS, Armonk, NY, USA)[7] and JASP 0.11 (JASP Team, Amsterdam, Netherlands)[8].

FCM clustering

Clustering or cluster analysis is an unsupervised data analysis that is used to partition a set of records or objects into clusters with similar characteristics. Clusters are identified via similarity measures. Clustering involves assigning data points to clusters so that items in the same cluster are as similar as possible, while items belonging to different clusters are as dissimilar as possible. It is a desideratum that the within-cluster variance should be low and the between-cluster variance should be high in clustering.

Fuzzy clustering (also referred to as "soft clustering" or "soft *k*-means") is a form of clustering in which each data point can belong to more than one cluster[9]. Because some countries may be similar to more than one other country in terms of outbreak indicators, fuzzy clustering rather than hard clustering is a more appropriate algorithm. The FCM clustering algorithm is the most widely used partition-based clustering algorithm. FCM clustering with an automatically determined number of clusters could enhance the detection accuracy; it uses the Euclidian distance measure[10]. The FCM clustering algorithm gives the best results for overlapped datasets and is comparatively better than *k*-means and hierarchical clustering algorithms[11].

The algorithm is an iterative clustering method that produces an optimal *c* partition by minimizing the weighted within-group sum of squared error objective function J_{FCM}[11].

$$J_{FCM} = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^q d^2(x_k, v_i).$$

In this study, $X = \{x_1, x_2, \ldots, 186\} \subseteq R^p$ is the dataset in the p (= 3)-dimensional vector space, n is the number of countries, c is the number of clusters with $2 \leq c < n$, u_{ik} is the degree of membership of x_k in the ith cluster, q is a weighting exponent on each fuzzy membership, v_i is the prototype of the centre of cluster i, and $d^2(x_k, v_i)$ is a distance measure between country x_k and cluster centre v_i

J_{FCM} can be obtained via an iterative process, which is performed as follows:

1. Set values for c, q, and ϵ.
2. Initialize the fuzzy partition matrix $U = [u_{ik}]$.
3. Set the loop counter $b = 0$.

https://dx.doi.org/10.4314/mmj.v34i2.2.

4. Calculate the c cluster centres $\left\{v_i^{(b)}\right\}$ with $U^{(b)}$ as follows:

$$v_i^{(b)} = \frac{\sum_{k=1}^{n}(u_{ik}^{(b)})^q x_k}{\sum_{k=1}^{n}(u_{ik}^{(b)})^q},$$

5. Calculate the membership $U^{(b+1)}$. For $k = 1$ to n, calculate the following:

$$I_k = \{i | 1 \leq i \leq c, d_{ik} = \|x_k - v_i\| = 0\}.$$

(a) If $I_k = \phi$ then, $u_{ik}^{(b+1)} = \frac{1}{\sum_{j=1}^{c}(\frac{d_{ik}}{d_{jk}})^{\frac{2}{(q-1)}}}$.

(b) Else next k (k is the number of countries).

6. If $\|U^{(b)} - U^{(b+1)}\| < \epsilon$ (termination criterion between [0, 1]), stop; otherwise set $b = b + 1$ and go to step 4.

A set of cluster validity indices is used to estimate the number of clusters in a set of datasets partitioned by several algorithms. R^2, the Akaike information criterion, the Bayesian information criterion, the within-cluster sum of squares, the Dunn index, the Calinski–Harabasz index, and Silhouette score are used for the validation of the results obtained by the FCM clustering algorithm. These indices are based on internal cluster validity indices. There are a few well-known measures, such as the Silhouette score, the Davies–Bouldin index, the Calinski–Harabasz index, and the Dunn index[12], but these are not enough alone for determining the cluster quality and also the very notion of "good clustering" is a relative concept, based on the point of view and the knowledge of the analyser.

The Dunn index is a ratio-type index where the cohesion is estimated by the nearest-neighbour distance and the separation is estimated by the maximum cluster diameter. Algorithms that produce clusters with a high Dunn index are more desirable.

The Calinski–Harabasz index is the ratio of the sum of between-cluster dispersion and intercluster dispersion for all clusters; the higher the score, the better the performance.

The Silhouette score measures the distance between each data point, the centroid of the cluster it was assigned to, and the closest centroid belonging to another cluster. This index is normalized, and a value close to 1 is always good for whatever clustering one is trying to evaluate. The score is bounded between −1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters.

Results

The relationships between the outbreak indicators and the total number of tests and the duration of the fight against the outbreak

Figure 1 illustrates the relationship between the total number of cases and the total number of tests for the 34 countries that had reported tests performed between 17 and 20 March 2020. In addition, the total number of cases per total number of tests is plotted against the number of days elapsed between the first reported cases and 29 March 2020 for each country in Figure 2; see also Table 1.

The results show that Australia, Russia, Bahrain, Poland, South Africa, South Korea, Taiwan, Vietnam, Hungary, and Thailand performed the highest number of tests per million population and had the lowest number of positive test results (<3%). In eight countries, the rate of positive cases per total number of tests is higher than 10%. Among these countries, Spain, Pakistan, and Italy have the highest proportions.

Figure 3 shows the total number of cases against the number of days after the first case(s).

Gibraltar, Spain, Switzerland, Liechtenstein and Italy had the highest number, about 1500 cases per million population. The number of days elapsed between the first reported cases and 29 March 2020 was 59 in Spain and Italy and 33 in Switzerland. In many countries, however, the number of cases was less than 100 per million population.

The relationships between the total death per million against the number of days after the first case(s) were shown in Figure 4. Spain and Italy had the highest total number of deaths, which were about 140 and 165 per million population, respectively.

On the basis of when the first positive cases were reported in many countries, the number of days elapsed since the outbreak was divided into ten periods, as described in the methods section. Changes in the outbreak indicators in each country according to these periods are presented in Figures 5–10. Specific countries in each period are presented in Tables 2 and 3. The period in which maximum exposure occurred was 8–15 March 2020; 54% of countries saw their first case(s) before 8 March, and approximately 20% of first case(s) occurred after 15 March.

Exposure periods are listed on the *x*-axis in Figures 5–10, from the longest exposure (period 1) to the shortest exposure (period 10). In the periods covering 31 December to 15 January, 8–15 February, and 1–29 March (periods 1, 4, 8, 9, and 10), the median number of active cases per million population was significantly lower than for the other periods (Figure 5 and Table 4; *P*<0.001). Other than that, no significant difference was found. From Figure 5 it can be seen that the highest number of active cases among the countries in the second period was in Italy, and in the sixth period the highest numbers were in Luxembourg and Switzerland.

In Japan, where exposure to the virus was for longer than in most other countries, the total number of deaths is very low (less than 0.5 per million population). The median number of deaths was significantly higher in periods 2, 3, and 6 (16–31 January, 1–7 February, and 21–29 February) than in the other periods (Figure 6 and Table 4; *P*<0.001). Italy and Spain have the highest numbers in period 2 and the Netherlands has the highest number in period 6.

The number of new cases in the countries that have been exposed the longest is quite low. The median number of new cases was significantly lower in periods 1, 4, 8, and 10 (31 December to 15 January, 8–15 February, 8–15 March, and 22–29 March). This was followed by the 16–21 March period, with a significantly higher number of cases than for the other periods (Figure 7 and Table 4; *P*<0.001). Figure 7 shows that among the countries that experienced outbreaks in the 16–31 January period (period 2), Spain, the UK, and Sweden have a significantly higher number of new cases than the other countries. In addition, among the countries exposed in the 1–7 February period, the number of new cases is highest in Belgium.

The number of critical cases is quite low in the countries that have been exposed the longest.

https://dx.doi.org/10.4314/mmj.v34i2.2.

Table 1. Total number of cases per total number of tests in the countries studied

Countries	Total number of cases per total number of tests	Countries	Total number of cases per total number of tests
Australia	0.62%	Panama	9.42%
Armenia	15.01%	Poland	2.72%
Bahrain	1.44%	Philippines	18.12%
Costa Rica	8.37%	Qatar	5.48%
Colombia	3.12%	Romania	4.42%
Denmark	10.73%	Russia	0.14%
Hungary	2.83%	Slovakia	4.40%
Finland	13.33%	S. Africa	2.33%
Indonesia	10.66%	S. Korea	2.73%
Israel	6.23%	Spain	46.35%
Italy	19.83%	Taiwan	0.51%
Japan	6.38%	Thailand	2.99%
Lithuania	3.99%	Turkey	6.04%
Malaysia	6.49%	UK	5.07%
N. Zealand	6.68%	Ukraine	8.23%
Norway	3.55%	USA	13.71%

Table 2. Periods and the number of countries and ships

Period	Country's first COVID-19 outbreak	Number of countries and regions
1	31 Dec 2019–15 Jan 2020	3
2	16–31 Jan 2020	24
3	1–7 Feb 2020	2
4	8–15 Feb 2020	1
5	16–20 Feb 2020	3
6	21–29 Feb 2020	37
7	1–7 Mar 2020	39
8	8–15 Mar 2020	52
9	16–21 Mar 2020	29
10	22–29 Mar 2020	11

Table 3. Countries and ships in the periods determined by considering exposure times

P1	P2	P3	P4	P5	P6	P7	P8		P9	P10
China	Australia	Belgium	Egypt	Iran	Afghanistan	Albania	Antigua	Honduras	Angola	Anguilla
Japan	Cambodia	*Diamond Princess*		Israel	Algeria	Andorra	Aruba	Ivory Coast	Barbados	Belize
Thailand	Canada			Lebanon	Armenia	Argentina	Bahamas	Jamaica	Bermuda	Brit.Virgin Islands
	Finland				Austria	Bangladesh	Benin	Kazakhstan	Cape Verde	Guinea
	France				Azerbaijan	Bhutan	Bolivia	Kenya	Chad	Laos
	Germany				Bahrain	Bosnia and Herzegovina	Brunei	Liberia	Djibouti	Libya
	Hong Kong				Belarus	Bulgaria	Burkina Faso	Mauritania	Dominica	Mali
	India				Brazil	Cameroon	Cayman Islands	Mayotte	El Salvador	MS *Zaandam*
	Italy				Croatia	Chile	Central African Republic	Mongolia	Eritrea	Myanmar
	Macao				Czech Republic	Colombia	Channel Islands	Montenegro	Fiji	Saint Kitts and Nevis
	Malaysia				Denmark	Costa Rica	Cuba	Namibia	Gambia	Turks and Caicos
	Philippines				Ecuador	Gibraltar	Democratic Republic of the Congo	Republic of the Congo	Isle of Man	
	Russia				Estonia	Hungary	Equatorial Guinea	Rwanda	Kyrgyzstan	
	South Korea				Georgia	Indonesia	Eswatini	Saint Lucia	Madagascar	
	Singapore				Greece	Jordan	Ethiopia	Seychelles	Mauritius	

Table 3 Cont...

Spain	Iceland	Latvia	French Polynesia	Somalia	Montserrat
Sri Lanka	Iraq	Liechtenstein	Gabon	Sudan	Mozambique
Sweden	Ireland	Maldives	Ghana	Suriname	New Caledonia
Taiwan	Kuwait	Malta	Greenland	Tanzania	Nicaragua
UK	Lithuania	Martinique	Grenada	Trinidad and Tobago	Niger
United Arab Emirates	Luxembourg	Moldova	Guadeloupe	Turkey	Papua New Guinea
USA	Mexico	Morocco	Guatemala	Uruguay	Sint Maarten
	Netherlands	Peru			
	New Zealand	Poland			
	Nigeria	Portugal			
	North Macedonia	Saudi Arabia			
	Norway	Senegal			
	Oman	Serbia			
	Pakistan	Slovakia			
	Qatar	Slovenia			
	Romania	South Africa			
	Saint Barthélemy	Togo			
	Saint Martin	Tunisia			

P1, 31 December 2019 to 15 January 2020; P2, 16–31 January 2020; P3, 1–7 February 2020; P4, 8–15 February 2020; P5, 16–20 February 2020; P6, 21–29 February 2020; P7, 1–7 March 2020; P8, 8–15 March 2020; P9, 16–21 March 2020; P10, 22–29 March 2020.

Table 4. Descriptive values of the indicators in the ten periods according to the first case(s)

Periods	Statistics	Total cases per million population	Deaths per million population	New cases per million population	Total patients recovered per million population	Active cases per million population	Critical cases per million population	The proportion of Critical cases in active cases (%)
31 Dec–15 Jan	*N*	3	3	3	3	3	3	3
	Mode	13.4	0.02	0	1.39	1.87	0.16	0.86
	Minimum	13.4	0.02	0	1.39	1.87	0.16	0.86
	Maximum	56.6	0.41	2.05	52.4	18.4	0.52	27.6
	Median	19.9	0.10	0.03	3.35	9.6	0.44	4.6
16–31 Jan	*N*	24	24	24	24	24	24	20
	Mode	0.17	0	0	0.03	0.14	0	0.08
	Minimum	0.17	0	0	0.03	0.14	0	0.08
	Maximum	1685.3	165.8	118.9	314.6	1158.8	89.1	72.4
	Median	71.2	0.64	1.95	3.78	77.5	1.02	2.31
16–20 Feb	*N*	3	3	3	3	3	3	3
	Mode	64.2	1.47	3.81	4.40	58.3	0.59	1.01
	Minimum	64.2	1.47	3.81	4.40	58.3	0.59	1.01
	Maximum	456.1	3.14	34.5	147.5	434.5	38.2	13.8
	Median	446.5	1.73	28.4	10.3	277.1	7.63	1.75
21–29 Feb	*N*	37	37	37	37	37	37	28
	Mode	0.47	0	0	0	0.45	0	0.12
	Minimum	0.47	0	0	0	0.45	0	0.12
	Maximum	6601.6	648.4	167.0	334.1	5776.4	471.5	12.1
	Median	143.1	1.46	1.02	3.39	131.1	1.77	2.18
1–7 Mar	*N*	39	39	39	39	39	39	25
	Mode	0.27	0	0	0	0.16	0	0.17
	Minimum	0.27	0	0	0	0.16	0	0.17
	Maximum	7125.9	77.6	336.5	1432.6	7125.9	129.4	153.5
	Median	48.6	0.23	0.48	0.59	45.9	0.18	1.92
8–15 Mar	*N*	52	52	52	52	52	52	15
	Mode	0.11	0	0	0	0	0	0.69
	Minimum	0.11	0	0	0	0	0	0.69
	Maximum	575.2	14.9	17.3	77.7	563.7	10.0	0.65
	Median	10.1	0	0	0	8.77	0	3.63
16–21 Mar	*N*	29	29	29	29	29	29	1
	Mode	0.75	0	0	0	0.11	0	1
	Minimum	0.11	0	0	0	0.11	0	1
	Maximum	1001.8	18.0	58.8	2.6	1001.8	0.79	1
	Median	1.70	0	0	0	1.70	0	1

Table 4 Cont....

22–29 Mar	*N*	11	11	11	11	11	11	0
	Mode	0.15	0	0	0	0.15	0	0
	Minimum	0.15	0	0	0	0.15	0	0
	Maximum	1093.5	0.05	0	0	1093.5	0	0
	Median	5.1	0	0	0	5.05	0	0

As there are only two countries and one ship in the 1–7 February period and only one country in the 8–15 February period, descriptive statistics are not given for these periods.

Table 5. Internal validity criteria and performance of the results

Clusters obtained	*N*	AIC	BIC	Silhouette score	R^2	Dunn index	Calinski–Harabasz index
12	186	85.39	204.13	0.540	0.978	0.004	744.84

The model is optimized with respect to the Bayesian information criterion (BIC).
AIC, Akaike information criterion.

Table 6. Cluster information

	Cluster No											
	1	2	3	4	5	6	7	8	9	10	11	12
Number of countries	3	2	1	15	124	1	1	6	39	6	1	1
Within-cluster sum of squares	6.493	2.172	0.000	1.264	0.272	0.000	0.000	1.408	0.932	0.850	0.000	0.000
Silhouette score	0.071	0.283	0.000	0.267	0.776	0.000	0.000	0.283	0.069	0.335	0.000	0.000
Centroid deaths per million population	0.718	1.417	-0.151	-0.019	-0.145	13.163	-0.149	-0.024	-0.108	-0.004	-0.061	0.025
Centroid total patients recovered per million population	1.394	-0.055	12.458	-0.137	-0.188	1.365	-0.196	1.017	-0.132	-0.090	3.424	2.697
Centroid active cases per million population	1.463	4.939	1.983	0.217	-0.300	6.983	8.682	0.001	-0.146	1.015	1.233	2.992

The between-cluster sum of squares of the 12-cluster model is 587.95. The total sum of squares of the 12-cluster model is 601.34.

Table 7. Clusters obtained from the fuzzy c-means algorithm

Cluster 1 (*n* = 3)	Italy	Spain	Switzerland								
Cluster 2 (*n* = 2)	Andorra	Luxembourg									
Cluster 3 (*n* = 1)	Faeroe Islands										
Cluster 4 (*n* = 15)	Aruba	Channel Islands	Denmark	Estonia	France	Ireland	Isle of Man	Israel	Malta	Netherlands	Saint Barthélemy
	Slovenia	Sweden	UK	USA							
Cluster 5 (*n* = 124)	Afghanistan	Algeria	Angola	Antigua and Barbuda	Argentina	Azerbaijan	Bahamas	Bangladesh	Barbados	Belarus	Belize
	Benin	Bermuda	Bhutan	Bolivia	Brazil	Brit Virgin Islands	Bulgaria	Burkina Faso	Cambodia	Cameroon	Cent African Rep
	Chad	Colombia	Costa Rica	Cuba	Curaçao	DR of the Congo	Djibouti	D. Republic	Egypt	El Salvador	Equ Guinea
	Eritrea	Eswatini	Ethiopia	Fiji	Gabon	Gambia	Georgia	Ghana	Grenada	Guatemala	Guinea
	Guinea-Bissau	Guyana	Haiti	Honduras	Hungary	India	Indonesia	Iraq	Ivory Coast	Jamaica	Japan
	Jordan	Kazakhstan	Kenya	Kuwait	Kyrgyzstan	Laos	Lebanon	Liberia	Libya	Macao	Madagascar
	Maldives	Mali	Mauritania	Mauritius	Mexico	Moldova	Mongolia	Morocco	Mozambique	Myanmar	Namibia
	Nepal	New Caledonia	Nicaragua	Niger	Nigeria	Oman	Pakistan	Palestine	Papua New Guinea	Paraguay	Peru
	Philippines	Poland	Portugal	Rep. of the Congo	Russia	Rwanda	Saint Kitts and Nevis	Saint Lucia	Saint Vincent and the Grenadines	Saudi Arabia	Senegal
	Seychelles	Sint Maarten	Slovakia	Somalia	South Africa	Sri Lanka	Sudan	Suriname	Syria	Taiwan	Tanzania
	Thailand	Timor-Leste	Tobago	Togo	Tunisia	Uganda	Ukraine	United Arab Emirates	Uruguay	Uzbekistan	Venezuela
	Vietnam	Zambia	Zimbabwe								
Cluster 6 (*n* = 1)	San Marino										
Cluster 7 (*n* = 1)	Vatican City										
Cluster 8 (*n* = 6)	Bahrain	Belgium	Brunei	Germany	Iran	South Korea					
Cluster 9 (*n* = 39)	Albania	Anguilla	Armenia	Australia	Bosnia and Herzegovina	Canada	Cape Verde	Cayman Islands	Chile	China	Croatia
	Cyprus	Czech Republic	Dominica	Ecuador	Finland	French Guiana	French Polynesia	Greece	Greenland	Guadeloupe	Hong Kong
	Latvia	Lithuania	Malaysia	Martinique	Mayotte	Montenegro	New Zealand	N. Macedonia	Panama	Qatar	Réunion
	Romania	Saint Martin	Serbia	Singapore	Turkey	Turks and Caicos					
Cluster 10 (*n* = 6)	Austria	Liechtenstein	Monaco	Montserrat	MS *Zaandam*	Norway					
Cluster 11 (*n* = 1)	Gibraltar										
Cluster 12 (*n* = 1)	Iceland										

Table 8. Median values of the indicators in each cluster

Deaths per million population		Total recovered per million population		Active cases per million population	
Cluster no.	**Median**	**Cluster no.**	**Median**	**Cluster no.**	**Median**
3*	0	7*	0	**5**	**7.11**
7*	0	**5**	**0.07**	9	131.6
11*	0	10	0.645	8	242.6
5	**0.01**	4	1.58	4	434.5
9	1.21	9	2.44	10	1023
10	2.305	2	38.42	**1**	**1159**
8	3.05	8	109.2	11*	1246
12*	5.86	6*	176.8	3*	1821
4	6.74	**1**	**204.8**	12*	2649
2	53.21	12*	334.1	2	3532
1	**139.6**	11*	415.5	6*	5776
6*	648.4	3*	1432	7*	7125

*Median values are country values because only one country was found in these clusters.
**: The lowest and highest median values for each outbreak indicator in the columns are bold

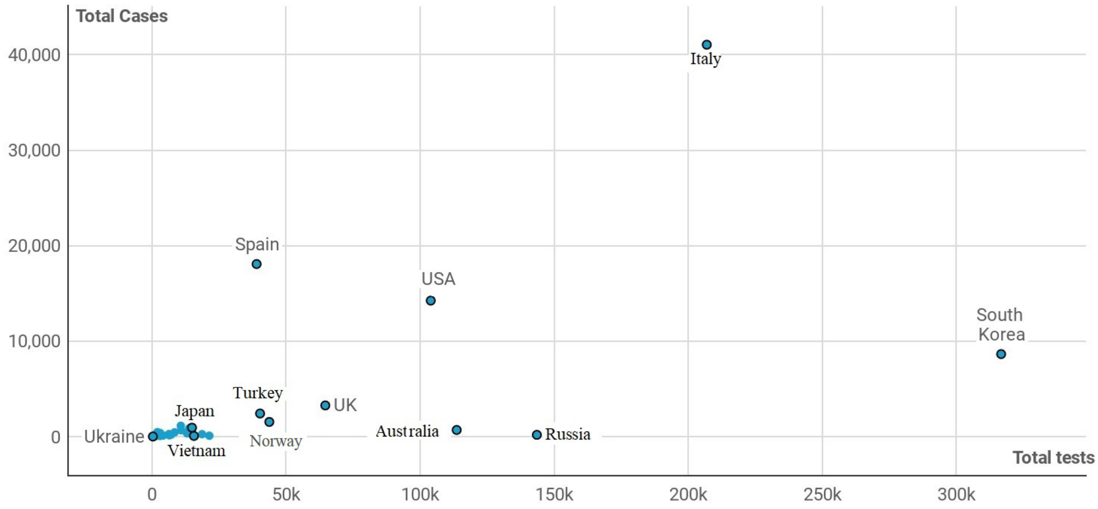

Figure 1 . Relationship between the total number of tests reported and the total number of cases in different countries

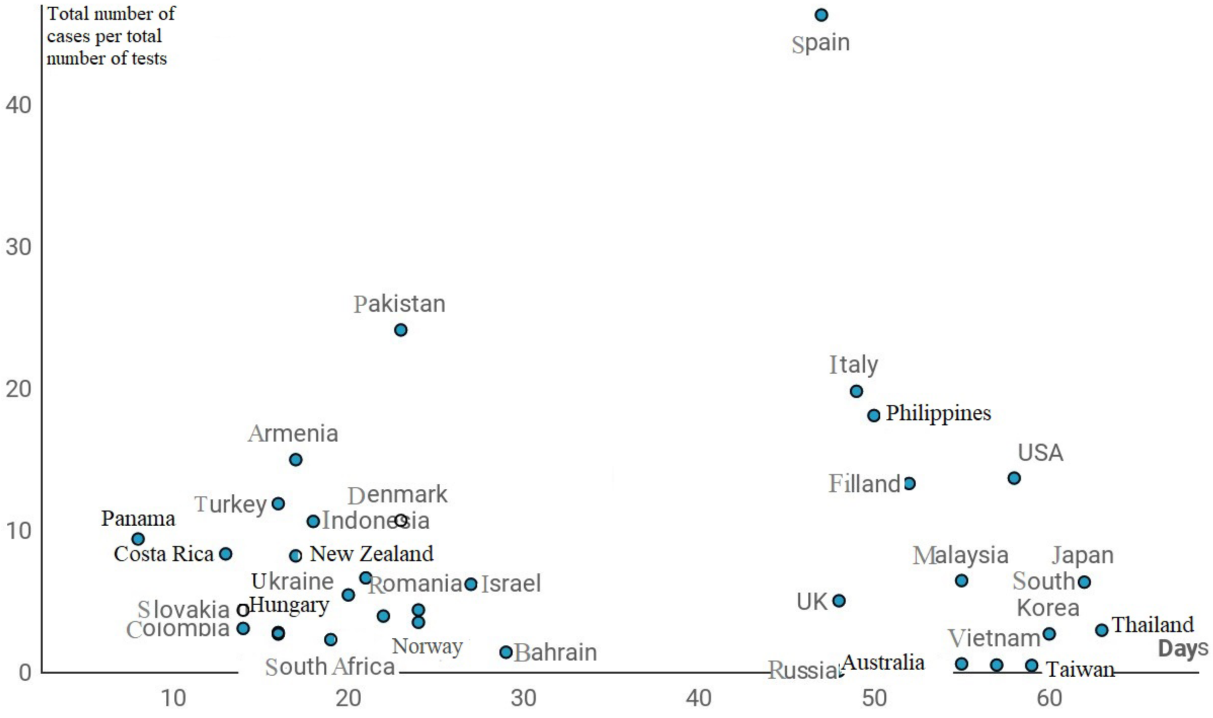

Figure 2 . Total number of cases per million population versus the number of days after the first case(s) in each country.

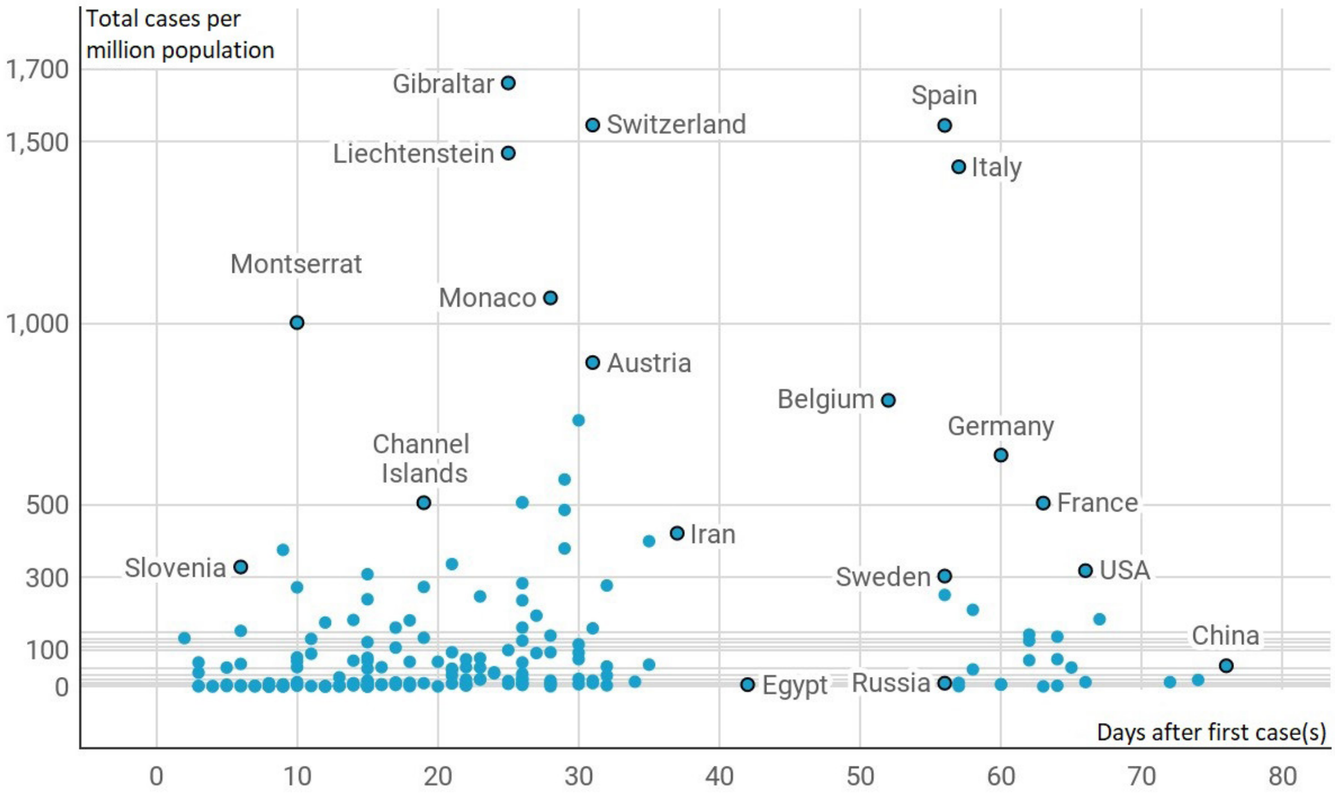

Figure 3 . Total number of cases per million population versus the number of days after the first case(s).

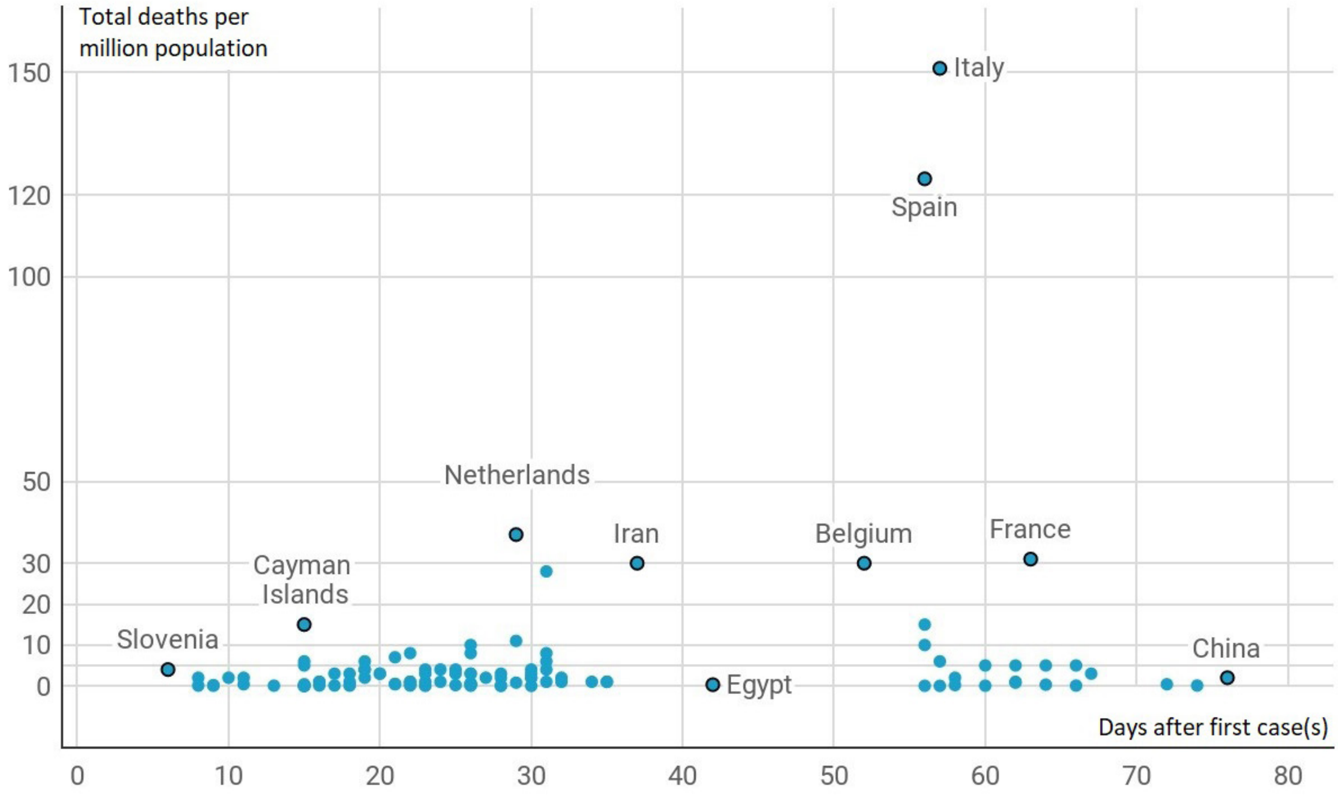

Figure 4 . Total number of deaths per million population versus the number of days after the first case(s).

https://dx.doi.org/10.4314/mmj.v34i2.2.

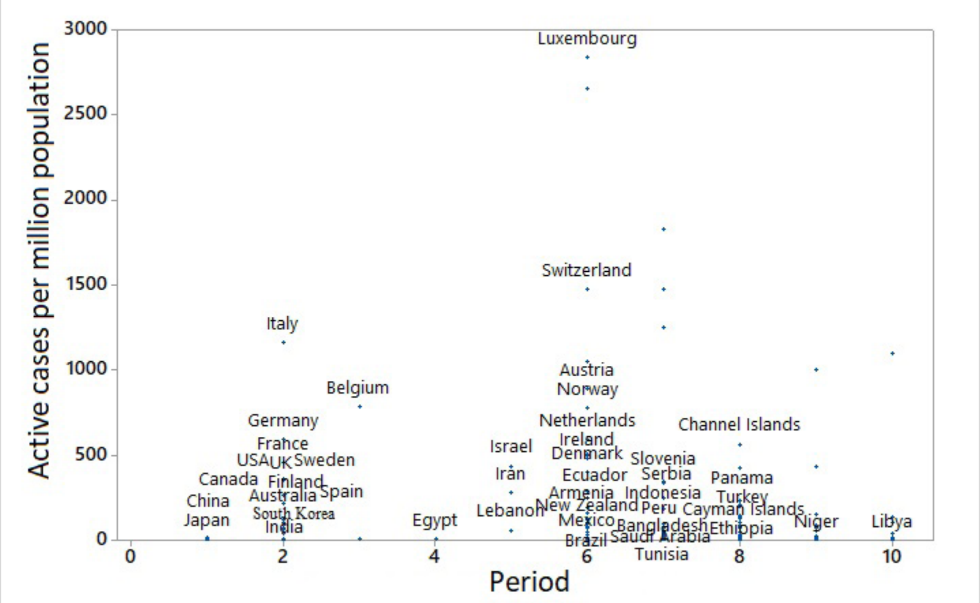

Figure 5 . Total number of active cases per million population according to the period.

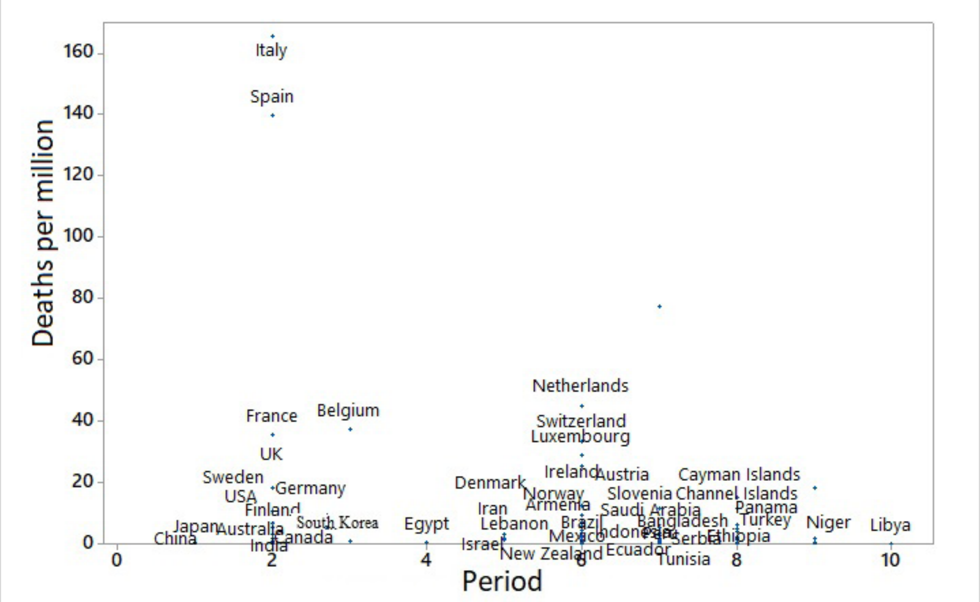

Figure 6. Total number of deaths per million population according to the period.

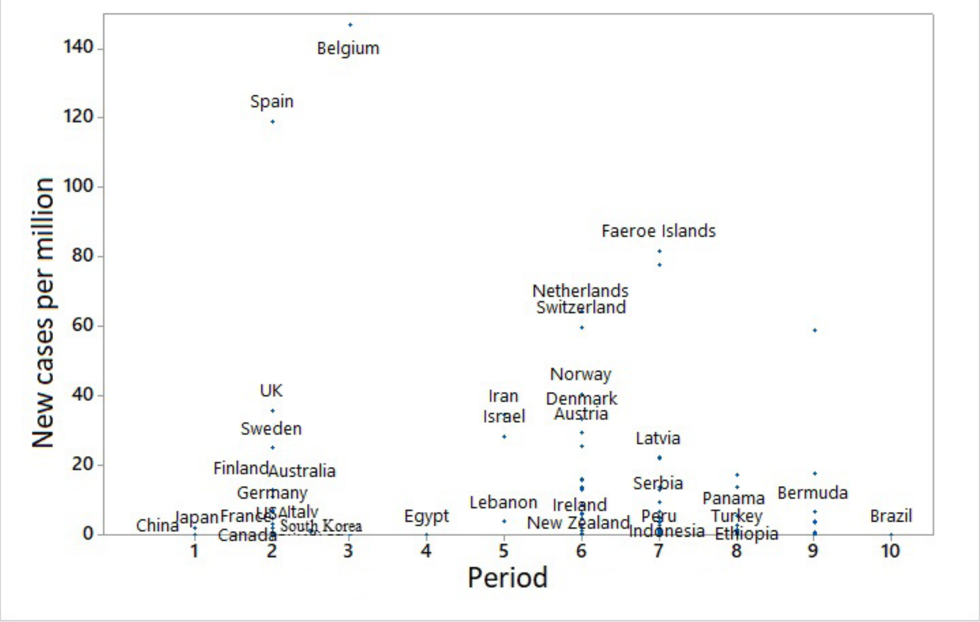

Figure 7. Number of new cases per million population according to the period

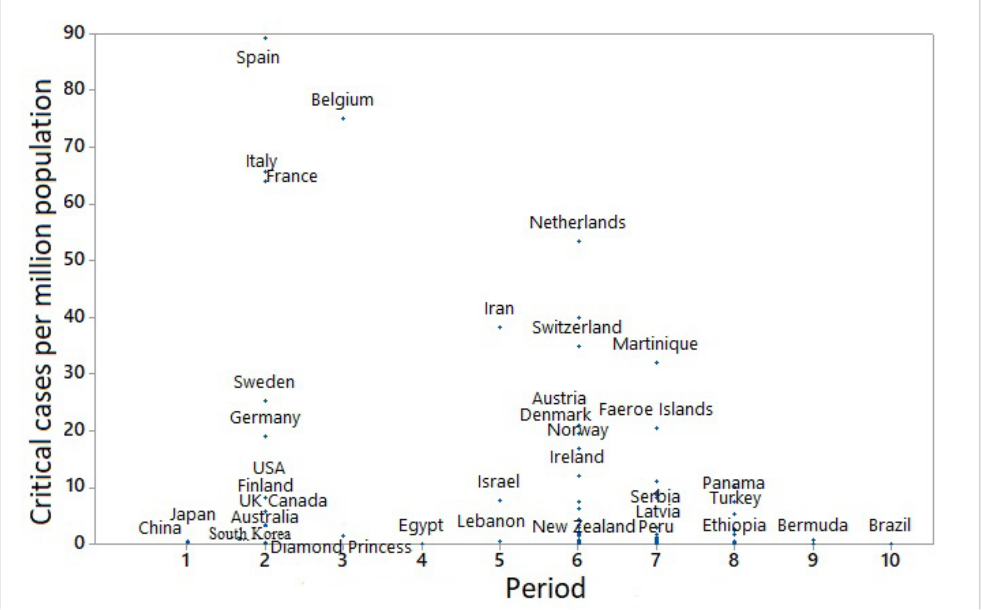

Figure 8. Number of critical cases per million population according to the period.

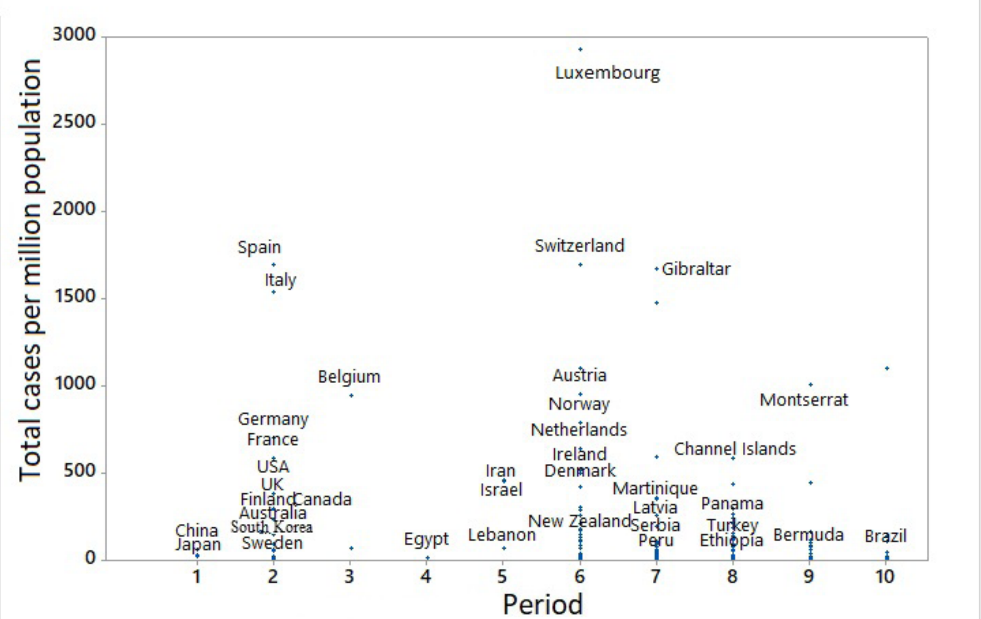

Figure 9: Total number of cases per million population according to the period

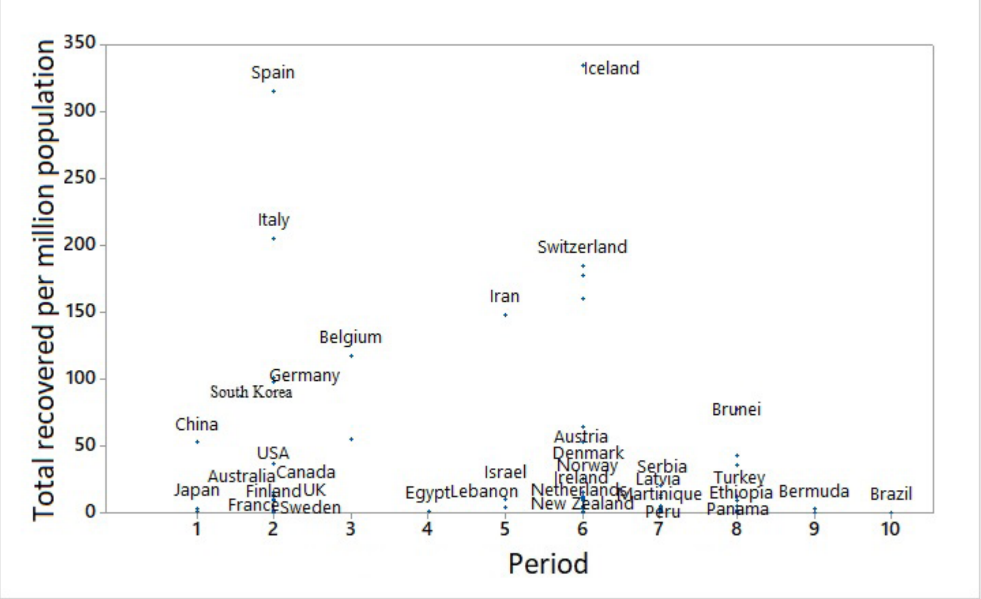

Figure 10: Total number of patients who recovered per million population according to the period

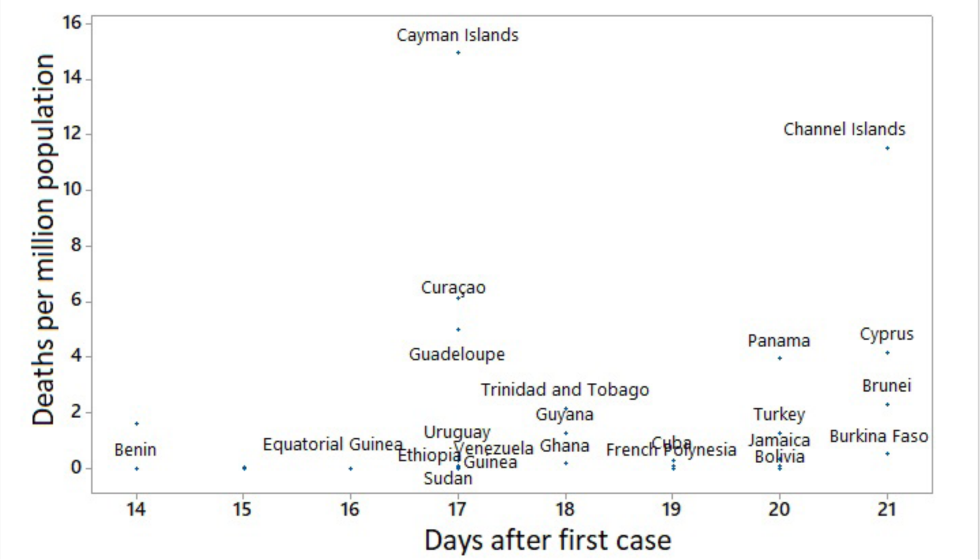

Figure 11: Total number of deaths per million population in different countries (8–15 March 2020).

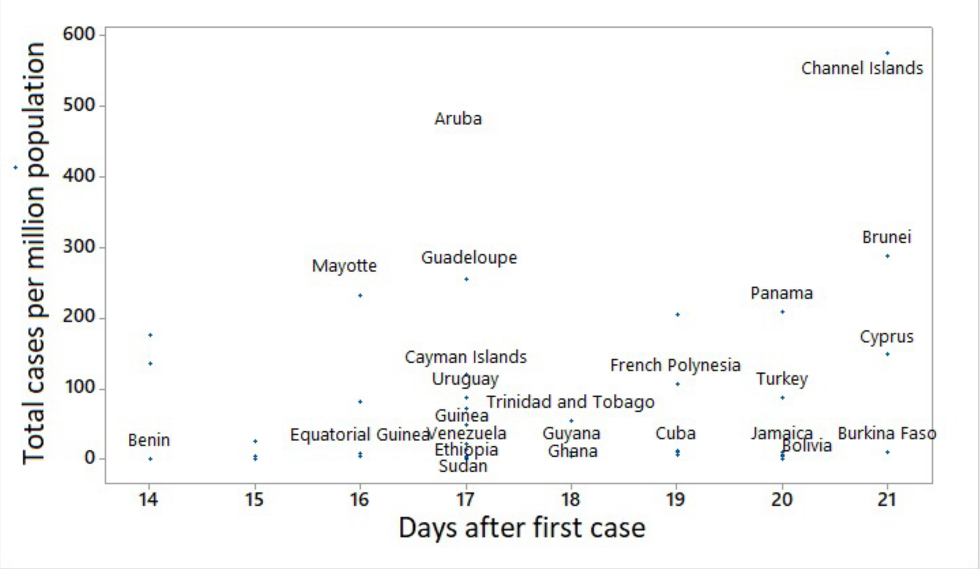

Figure 12: Total number of cases per million population in different countries (8–15 March 2020).

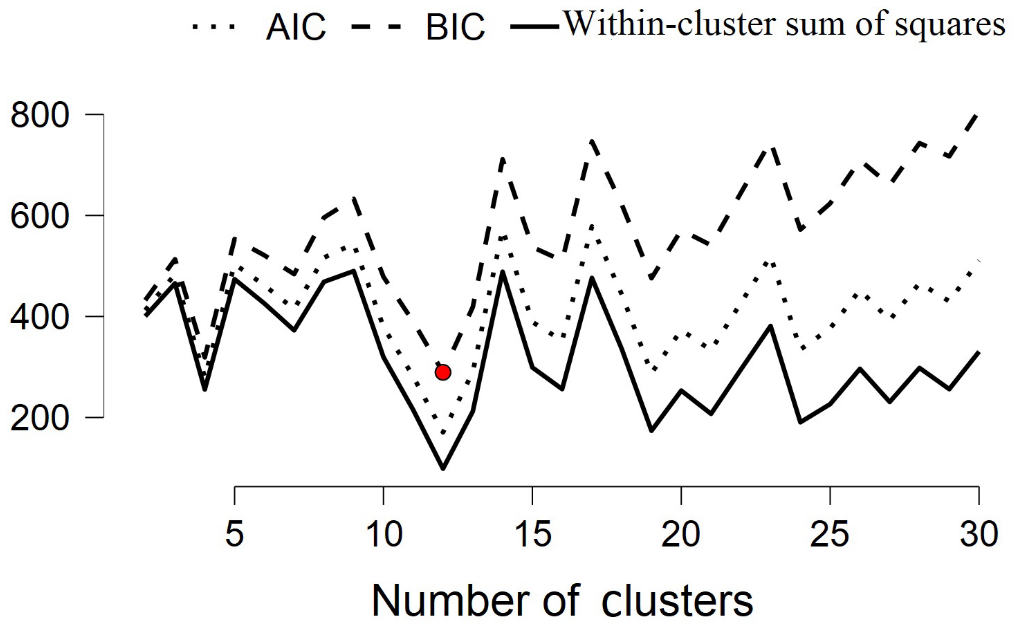

Figure 13 . Optimal number of clusters according to goodness of fit measures. AIC, Akaike information criterion; BIC, Bayesian information criterion

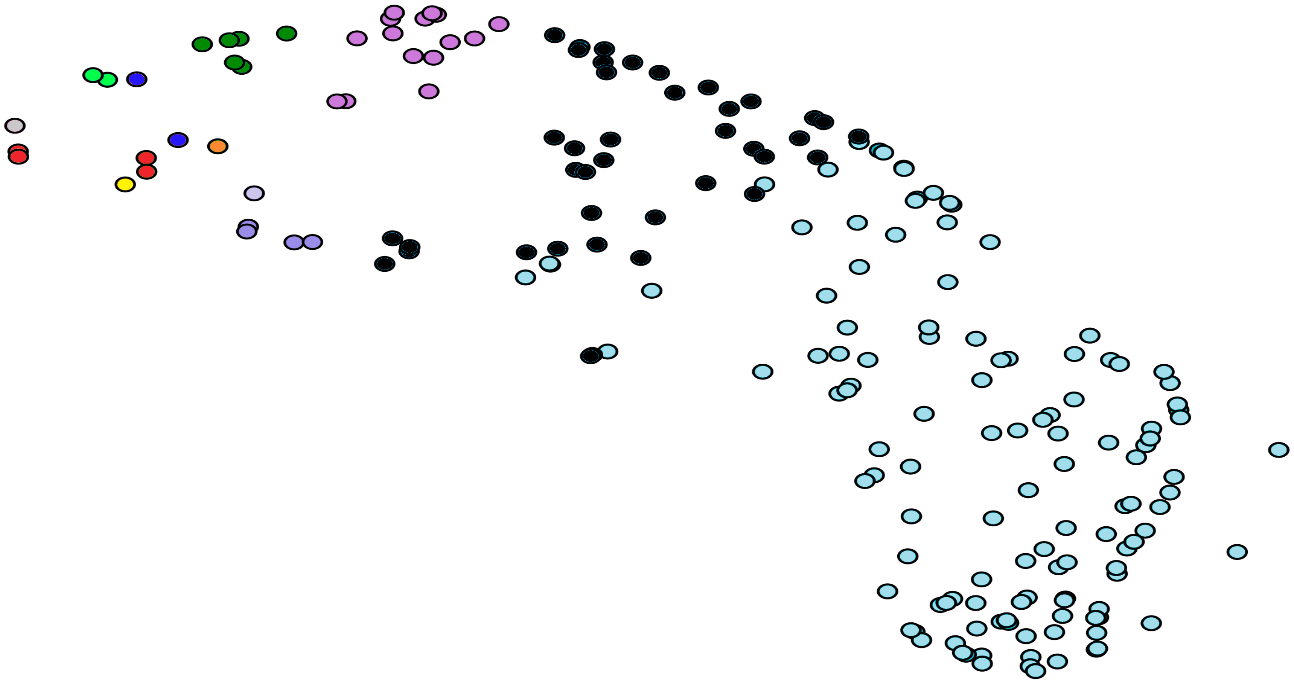

Figure 14. Visualization of fuzzy clustering results by Sammon mapping.

Discussion

The earliest countries to report COVID-19 cases after the outbreak in China were South Korea and Taiwan, but these countries have contained the outbreak with some success[13].

The rapid spread of COVID-19 has led many countries around the world to implement strict measures, and serious problems have started to emerge. To follow the course of the outbreak and to minimize problems, it is of great importance that accurate methods of data analysis should be used. In addition, many indicators and country-specific characteristics should be taken into consideration when one is comparing data from different countries[14]. There are many open-access databases comprising shared data relating to COVID-19 cases that can be used for this purpose[3–5].

In this study, two objectives were achieved. Firstly, the relationship between outbreak indicators (total number of cases, total number of deaths, and total number of patients who recovered) and the number of days after the index case, and also the total number of tests, was clarified. From Figures 2–4, it can be seen that, on the basis of the total number of tests conducted in Italy and Spain, the number of positive cases and the total number of deaths are very high. These numbers have negatively affected the responsiveness of the health systems in those countries. The health systems in Italy, Spain, Belgium, France, the Netherlands, and Iran displayed capacity difficulties. Despite the high numbers of COVID-19 cases in Germany, Austria, Switzerland, and the USA, the health systems in these countries are currently able to respond. Figure 7 leads us to the conclusion that quarantine conditions are not followed adequately in countries with a high number of new cases. Furthermore, it can be seen that the spread of the virus slowed down in period 1 countries, where the virus first spread, whereas the effects of the outbreak in period 2 countries will continue on the current course (Table 3). However, it can also be observed that for countries in periods 6 and 7, the health systems that are struggling to cope with the numbers of COVID-19 patients are likely to see increased numbers of deaths.

Various factors such as demographic structure, geographical structure, economic level, climatic conditions, and measures taken can be affected the pandemic results of the countries. In a study by Violini[15], the importance of exposure times is emphasized in a comparison of 23 countries also. For this reason, the duration of exposure to infection was taken into account in this study as it can affect country differences. The WHO guidelines explained that the pre-epidemic preparations of countries and physician knowledge and skills also affected the rate of positive cases[16].

Secondly, the similarities of countries in terms of outbreak indicators were examined by a multivariate method. Figure 14 summarizes the similarities and differences between the countries studied at the end of March 2020 in terms of the total number of deaths, the total number of patients who recovered, and the total number of active cases. Those with characteristics different from the characteristics of other countries in terms of the effects of the pandemic are generally located in separate clusters. This study determined that the total number of deaths is higher in central and southern European countries, especially Italy, Spain, Switzerland, and Portugal. However, the number of patients who recovered in these countries is also high. Additionally, it was found that the number of active cases is higher in South America, East Asia, and northern European countries such as Italy, Spain, and Switzerland. According to the results of the cluster analysis, countries can make better decisions about the measures to be taken by investigating the reasons for the intra-cluster and inter-cluster differences found.

The clustering of countries according to various indicators is discussed in some studies[17,18]. In the *k*-means cluster analysis conducted by Zoumpekas[17], the total number of cases by country, the daily number of deaths, and the daily number of patients who recovered were considered. For each indicator, data presented in separate time series were used.

Kumar[18] performed a hierarchical cluster analysis to classify Indian states and union territories on the basis of COVID-19 status. He found that it grouped 27 states and five union

https://dx.doi.org/10.4314/mmj.v34i2.2.

territories into six clusters. He found that optimization of monitoring techniques is required to improve government policies and decisions, medical facilities, treatment, etc. to reduce the number of people who die.

Ploner[19] performed two different HDBSCAN cluster analyses. The first included only three features and worked well for countries having only 2.5 weeks of data after the outbreak. In comparison, the second analysis used features from the peak of the curve. For countries with increased numbers of daily cases, the peak moved and, therefore, the results changed. Approximately 60 countries were considered 60. Ploner[19] found higher mortality in Spain, Italy, Belgium, New York, Germany, and Canada than in other countries.

Zarikas et al.[20] presented a novel analysis resulting in the clustering of countries according to active cases, active cases per population, and active cases per population and per area based on Johns Hopkins epidemiological data. They found that after removing Monaco and San Marino, a cluster including Liechtenstein and Andorra and one with Malta and Luxembourg were obtained, while all other countries remained together.

Conclusion

To define and track the progress of the pandemic and its effects, similarities between countries can be examined by considering indicators together. Therefore, better decisions can be made using multivariate analysis techniques such as cluster analysis[21], which is an extremely useful method for finding new relationships and insights[19]. In the event that the pandemic continues, this work offers a basic study that evaluates the measures taken by countries in the periods following outbreaks. In addition, the results of this study will benefit researchers by offering a guide for how to design more comprehensive research. It can be misleading to compare countries one by one in terms of each indicator. In this study, country similarities were investigated by our considering the relationships between outbreak indicators. In conclusion, various features of countries, such as climatic conditions, cultural habits, average age, chronic disease frequency, the epidemic measures taken, and epidemic indicator results, can be related to each other. For this reason, it is recommended to perform data analysis with multivariate models such as cluster analysis, which takes into account the relationships between these features in studies that examine countries comparatively.

Limitations

By the end of March 2020, only 34 of 169 countries, 17 regions and 2 ships struggling with the pandemic had reported the total number of tests. The results can give limited information to show the relationship between outbreak indicators and the total number of tests. Besides, three outbreak indicators were used in the clustering of countries according to their similarities in this study. On the other hand, in addition to the outbreak indicators, more accurate predictions can be made once the similarities of countries are investigated together with many features, such as pandemic measures, economic levels, climatic conditions, and demographic structures.

Conflict of interest

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

HA, UU, SA, and SC conceived the study and HA, UU, and SA designed the study. HA, UU, SA, and SC collected the data, performed the analysis, interpreted the results, drafted the manuscript, revised the manuscript critically for important intellectual content, and approved the final version of the manuscript.

References

1. WHO [Internet]. Geneva, Switzerland: World Health Organization; WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020 [cited 2020 April]. Available from: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

2. ECDC [Internet]. Sweden: European Centre for Disease Prevention and Control; Coronavirus disease 2019 (COVID-19) pandemic: increased transmission in the EU/EEA and the UK – seventh update [update 2020 March 25; cited 2020 April]. Available from: https://www.ecdc.europa.eu/sites/default/files/documents/RRA-seventh-update-Outbreak-of-coronavirus-disease-COVID-19.pdf.

3. Worldometers [Internet]. Dover, Delaware, USA: Worldometers; COVID-19 coronavirus pandemic. 2020 [cited 2020 April]. Available from: https://www.worldometers.info/coronavirus/.

4. WHO [Internet]. Geneva, Switzerland: World Health Organization; WHO Coronavirus Disease (COVID-19) Dashboard [cited 2020 April]. Available from: https://who.sprinklr.com/.

5. coronavirus.jhu.edu [Internet]. Broadway, Baltimore (MD): Johns Hopkins University & Medicine Coronavirus Resource Center [cited 2020 April]. Available from: https://coronavirus.jhu.edu/.

6. Lorenz M, Aisch G, Kokkelink D. Datawrapper: Create charts and maps [software]. 2012 [cited 2020 April]. Available from: https://www.datawrapper.de/.

7. IBM Corp. Released 2017. IBM SPSS Statistics for Windows. Version 25.0 [software]. 2017. Armonk, NY, USA: IBM Corp.

8. JASP Team. JASP. Version 0.11 [software]. 2019. Available from: https://jasp-stats.org

9. Simhachalama B, Ganesan G. Performance comparison of fuzzy and non-fuzzy classification methods. Egypt Inform J. 2016;17(2):183–8.

10. El-Khamy SE, Sadek RA, El-Khoreby MA. An efficient brain mass detection with adaptive clustered based fuzzy C-mean and thresholding. 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA); 2015 Oct 19-21; Kuala Lumpur, Malaysia: IEEE; 2016. p. 429–33. https://doi.org/10.1109/ICSIPA.2015.7412229.

11. Yang Y. Image segmentation by fuzzy c-means clustering algorithm with a novel penalty term. Comput Inform. 2007;26(1):17–31.

12. Arbelaitz O, Gurrutxaga I, Muguerza J, Perez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern Recognit. 2013;46(1):243–56.

13. Cheng SC, Chang YC, Long Y, Chiang F, Chien YC, Cheng M, et al. First case of coronavirus disease 2019 (COVID-19) pneumonia in Taiwan. J Formos Med Assoc. 2020;119(3):747–51.

14. Khafaiea MA, Rahimb F. Cross-country comparison of case fatality rates of COVID-19/SARS-COV-2. Osong Public Health Res Perspect. 2020;11(2):74–80.

15. Violini G. Country comparison of the COVID-19 pandemic. What happens in Latin America?, and why? [cited 2020 April 17]. Available from: https://doi.org/10.13140/RG.2.2.13707.64800.

16. who.int [Internet]. Geneva, Switzerland: World Health Organization; Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19) [cited 2020 April]. Available from: https://www.who.int/

docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf.

17. Zoumpekas T. COVID-19 cluster analysis. Canada: Towards Data Science Inc; 2020 [cited 2020 April]. Available from: https://towardsdatascience.com/covid-19-cluster-analysis-405ebbd10049.

18. Kumar S. Monitoring novel corona virus (COVID-19) infections in India by cluster analysis. Ann Data Sci. 2020;7:417–25. https://doi.org/10.1007/s40745-020-00289-7.

19. Ploner M. Which countries react similar to COVID-19? Machine learning provides the answer. Canada: Towards Data Science Inc; 2020 [cited 2020 April 16]. Available from: https://towardsdatascience.com/which-countries-react-similar-to-covid-19-machine-learning-provides-the-answer-5971ec2f6f31.

20. Zarikas V, Poulopoulos SG, Gareiou Z, Zervas E. Clustering analysis of countries using the COVID-19 cases dataset. Data Brief. 2020;31:105787. https://doi.org/10.1016/j.dib.2020.105787.

21. Pirouz B, Haghshenas SS, Haghshenas SS, Piro P. Investigating a serious challenge in the sustainable development process: analysis of confirmed cases of COVID-19 (new type of coronavirus) through a binary classification using artificial intelligence and regression analysis. Sustainability. 2020;12(6):2427. https://doi.org/10.3390/su12062427.