Received: 23 June 2021 | Revised: 20 July 2021 | Accepted: 27 July 2021

DOI: 10.1002/jmv.27242

RESEARCH ARTICLE

JOURNAL OF MEDICAL VIROLOGY WILEY

Statistical analysis of the Hungarian COVID-19 victims

Elnaz Gholipour[1] | Béla Vizvári[1] | Tareq Babaqi[1] | Szabolcs Takács[2]

[1]Department of Industrial Engineering, Eastern Mediterranean University, Famagusta, Turkey

[2]Department of Psychology, Karoly Gaspar University, Budapest, Hungary

Correspondence
Elnaz Gholipour, Department of Industrial Engineering, Eastern Mediterranean University, Famagusta—North Cyprus, P.O. Box: 99628, Mersin 10, Turkey.
Email: elnaz.gholipour@emu.edu.tr

Abstract

With the wide spread of Coronavirus, most people who infected with the COVID-19, will recover without requiring special treatment. Whereas, elders and those with underlying medical problems are more likely to have serious illnesses, even be threatened with death. Many more disciplines try to find solutions and drive master plan to this global trouble. Consequently, by taking one particular population, Hungary, this study aims to explore a pattern of COVID-19 victims, who suffered from some underlying conditions. Age, gender, and underlying medical problems form the structure of the clustering. K-Means and two step clustering methods were applied for age-based and age-independent analysis. Grouping of the deaths in the form of two different scenarios may highlight some concepts of this deadly disease for public health professionals. Our result for clustering can forecast similar cases which are assigned to any cluster that it will be a serious cautious for the population.

KEYWORDS

clustering, coronavirus disease, hungary, statistical analysis

1 | INTRODUCTION

The almost whole world today has been infected by the phenomenon of a deadly virus that is very speedily transmitted to humans.[1] There is a worldwide endeavor to figure out the medical, economic and sociological effects of the pandemic.[2] Many more global studies about the impact of human mobility networks and rapid spread of COVID-19 have been surveyed.[3-6] All efforts are biased to keep different aspects of the society safe and human health is placed as the first priority. The risk factors for severe COVID-19 has been updated every day and age is the strongest risk factor for coronavirus outcomes (Underlying Medical Conditions for Clinicians, 2021).[7,8] Older age, being black or African American and being male were associated with immense odds of death in variate analysis.[9,10] A research surveyed age and sex patterns of mortality, based on reported deaths from Western Europe and the United States, which underscores a high correlation between demographic vulnerability to coronavirus death rates.[11] Severity and mortality rate in COVID-19 reached from sex-difference comorbidities and behaviours, marks the necessity to gather sex and age-disaggregated data to understand disease pathology and guide clinical care deeply,[12] as we carried out in our study.

Subjects older than 60 years and any case with chronic medical problems struggling with coronavirus condition showed no improvement and considerable mortality rate.[13] In another study, male gender and old age with some underlying condition like hypertension, diabetes and heart disease assigned with fatal outcome.[14] The age distribution of COVID-19 victims, moved toward younger age groups from May through August 2020.[15] Different study revealed the similar age distribution of mortality in Italy, Japan, and Spain, even though the total deaths are entirely different among them.[16] Out of severe medical conditions, diabetes and asthma for greater age that associated with COVID-19 resulted in death.[10] It has been reported some underlying medical conditions that put human beings at increased risk for potential severe and life-threatening outcomes from COVID-19 infections.

Cancer, chronic lung disease, COPD, serious heart disease, obesity, Type 2 diabetes, chronic kidney disease, HIV infection or those with weakened immune systems, pregnancy down syndrome, Asthma, Hypertension (high blood pressure), neurologic conditions, such as Dementia, liver disease, pulmonary fibrosis (having damaged or scarred lung tissue), thalassemia (a type of blood disorder), Type 1 diabetes.[8,17,18]

The full review of 54 papers observed that diabetes, hypertension and cholesterol levels have detectable relation to COVID-19 severity. Cancer, kidney diseases and stoke need further research to be explored a connection to the virus.[19] A nationwide study in Turkey, concluded male gender, diabetes, heart failure and dementia as significant factors with high mortality rate.[20] This research declares, chronic kidney disease, hypertension and cancer are the risky underlying medical conditions for the patients aged between 60 and 79. To compare between severe and nonsevere patients, hypertension, diabetes and cardiovascular disease are risky factors.[21] One survey from New York City shows,[22] regardless of age, individuals with diabetes and immune disorders are at high risk, and the youngest age groups are the least likely to exhibit symptoms. The aim of the research is to explore the special characteristics of the deceased. In spite of the previous related studies, we desired to have three main factors; age, gender and 19 high-risk underlying medical conditions all together in the form of different clusters to group the population of COVID-19 victims of Hungary to report the similar risky cases statistically and recognize the dangerous medical problems for unlike gender and range groups. Through this investigation, positive and negative correlation between diseases may shed light on the particular cases for further professional studies. The Hungarian government publishes all cases of death by four types of data as follows: number, age, gender and underlying conditions[23] i.e., containing the existence/nonexistence of 19 types of diseases. Latter ones can be coded as binary, i.e., 0 or 1 parameters. Since, the classification and clustering problems are admired topics of research in the area of pattern identification,[24] it is taken into consideration the current research. Clustering is one of the most applicable approach in COVID-19 screening that particularly is used to identify the transmission risks assessment that recently has been engaged with multiple articles,[25-27] besides, clustering is a method to develop a map of COVID-19 occurrences to achieve the optimal handling of the pandemic that the provinces were clustered by K-Means method[28] to inform the population by raising their awareness of the diseases spread. Numerous studies classified the geographical regions based on COVID-19 patients as local clustering approach.[29-33]

On the report of more studies about COVID-19, age as a dominant variable, structures two scenarios for additional analysis of the present article. Age-Dependent and Age-Independent. When age is taken from more examination, the number and attributes of the clustering will be changed. Another contribution of this study is to compare the statistical consequences of Hungarian victims while age is included versus it is excluded from the survey.

2 | METHODS AND MATERIALS

All 7238 cases were collected from the population of Hungary who died till 15th of December 2020 because of COVID-19. Related information of three attributes of the sample composed of age, gender and underlying conditions, i.e., containing the existence/nonexistence of 19 types of diseases, obtained from medical Hungarian website.

TABLE 1 The frequencies of the 19 diseases

Disease	# Of cases	Disease	# Of cases
High blood pressure (Hypertension)	4710	Stroke	219
Tumor	919	Kidney failure	854
Chronic lung disease	85	Asthma	190
Cardiac arrhythmia	479	COPD	37
Ischemic heart disease and heart attack	858	Parkinson disease	203
Pneumonia	158	Vasoconstriction	66
Dementia	728	Alzheimer disease	144
Atrial fibrillation	269	Reflux	119
Diabetes	2044	Schizophrenia	36
Obesity	220		

There are 3517 female and 3720 male patients among them. The youngest person was 18 years old and the eldest was 103 years old. The frequencies of the diseases are in Table 1. The complete table of the correlations among the groups of diseases can be found in the appendix.

There are six groups such that the correlation of any two diseases is positive. All six disease groups are related to the vascular system and blood oxygen levels. They are as follows: hypertension, chronic lung disease, COPD, arterial fibrillation, heart disease, and vasoconstriction. Four groups of diseases, such as Alzheimer's disease, dementia, Parkinson's disease, schizophrenia, have a negative correlation with each of the following diseases: (a) hypertension, (b) chronic lung disease, (c) COPD, (d) tumor and cancer, (e) asthma, (f) renal failure, (g) obesity. The strongest negative correlation is between hypertension and tumor/cancer. Its value is −0.105. Clustering is a task of dividing the population into a number of classes such that data points in the same classes are more similar to each other than those in other classes. It helps create and segregate the particular classes, whereas, each cluster defines its own attributes based on the factors. The continues variable is age. If it is not scaled, i.e., it keeps its original value, then it is the dominant variable as its absolute value is greater than the sum of the rest of the variables. Two different types of clustering were applied as two scenarios: *age-dependent*, i.e., without changing the value of the age and *age-independent* when the age value is divided by 100. In the latter case, this value is transformed between 0 and 1 and it gets to the same range as the Boolean variables. Generally, there are three different types of clustering as: (a) Hierarchical, (b) nonhierarchical; K-Means and (c) Mix Method; Two-step clustering. For this research, we used two-step clustering in the case of age-dependent analysis to identify the number of the optimal clustering and K-Means clustering to classify the dataset for both scenarios. The SPSS statistical program package was used for this purpose.

2.1 | First scenario; age-based analysis

In the initial part of the study, age is a basis of the cluster description. Classifying the dataset carried out for different ages to reveal natural groupings. Out of 21 variables, (1 continues, 1 categorical, and 19 logical factors), 8 factors with frequency less than 1.5% of the studied population as nonsignificant variables were removed. Consequently, similarity of 13 variables for different age groups constructs distinct clusters. It is a higher quality, when the ratio size of the clusters is 2 or less. It means no cluster in the cluster set is more than two times as large as any other clusters, then there is no huge differences between the size of the clusters and the distribution of the cases between clusters is quite well. In our case, by trail and error, two-step clustering results showed that 10 clusters (K = 10) seem reasonably suitable as we reached the ratio of cluster sizes at almost 2.0, which is an acceptable spread of the cases (Figure 1). Whereas, it was 3.53 for cluster size of 8 and 2.99 for cluster size of 12. Therefore, according to Figure 1, the size of the largest cluster divided by the size of the smallest one is 2.04.

2.2 | Second scenario; age-independent analysis

For the second part of the research, Age factor was divided by 100 to have the same range of the values for all categorical, continues and binary variables. It causes to decline the impact of the age in clustering and grouping the data mostly based on the other factors. The further cluster arrangement is driven with binary variables which are diverse underlying conditions for different genders. To uncover the properties of clusters as age-independent inspection, 21 variables categorized into the heterogeneous classes by K-Means clustering. Technically the K-Means method has been defined by a distance function. It initializes clustering by k centers and assigning all objects to the nearest center and then moving a center to the mean of its members. Reassigning the objects, moving the centers and iterate them adequately is to have no objects changed membership. In this section of our study, clusters describe the similarity of the Hungarian COVID-19 victims exclusive of the age. The small value of the distance function means that the two victims share many common properties. K = 8, 9, and 10 defined patterns of the clustering appropriately to manifest genders with varied underlying conditions based on 7238 victims' attribute.

3 | RESULTS

3.1 | Consequences of the statistical analysis

The descriptive statistics of 19 diseases registered besides COVID-19 can be found in Table 2. The number of the occurrences of the diseases are considered random variable depending on the ages of the patients. The hypothesis that the number of occurrences has a normal distribution is rejected based on skewness and/or kurtosis in the cases signed by yellow background. Hypertension has the highest number of cases, 4710 out of 7238 giving 65.08%. The second most frequent disease is diabetes with 2044 cases, 28.24%. There are two diseases having the number of cases close to 1000; cancer/tumor (919, 12.7%), and kidney failure (854, 11.8%). The number of cases of any other disease is less than 500 (6.9%). COPD, and schizophrenia have the least number of cases, 37, and 36, respectively. The relation of the age and the deceased can be seen also from Table 2. However, Table 3 gives it in an ordered way. Obesity has the lowest value which is 64.5 years. Schizophrenia is the other one having an average under 70. It is 69.69. There are again three diseases having an average above 80. It is not a surprise that Alzheimer has the highest value at 80.85. The other ones are Dementia and Atrial fibrillation with 80.33 and 80.07 correspondingly. It is also easy to accept that the fourth place is occupied by Parkinson disease with 79.78. The high value of heart attack is noteworthy.

If the ages of the individuals are considered random variables, then it is possible to check if the random variable has a normal distribution. Based on skewness, i.e., third moment, and on kurtosis, i.e., the fourth moment, the normality can be rejected in the case of diseases as follows: Atrial fibrillation, Pneumonia, Vasoconstriction, Kidney failure, Arrhythmia, and cancer/tumor. It means that additional statistical test can be made with the data of these diseases if the theory of the statistical test does not assume normality. The Kolmogorov–Smirnov test gives the same result. The skewness of every disease is negative. The reason is that every disease has early incidences. It is true even for Alzheimer where the youngest patient is 62 years old.

The fitted normal distribution and the histogram of obesity is in Figure 2. The hypothesis of the normal distribution is accepted.

It can be observed that the histogram of tumor has two peaks (Figure 3). The hypothesis of the normal distribution is rejected because of the two peaks.

3.2 | Output of age-dependent clustering

SPSS uses a function based on the absolute value of the attributes. This property gave a high weight to the age.

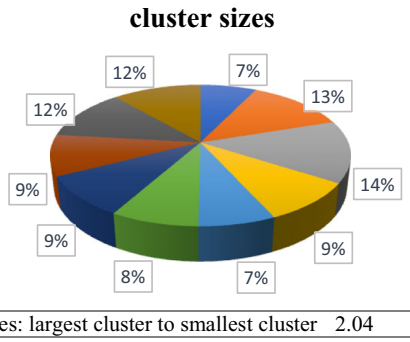

FIGURE 1 The relative sizes of the clusters in the case of 10 clusters

TABLE 2 Basic statistics of the registered diseases

	N	*Min age*	*Max age*	*Mean*		*Std.*	*Skewness*		*Kurtosis*	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error	Statistic	Std. Error
Alzheimer	144	62	99	80,85	0,556	6,671	-0,129	0,202	0,059	0,401
Arrhythmia	479	32	102	78,46	0,475	10,396	-0,863	0,112	1,477	0,223
Asthma	190	34	96	72,56	0,878	12,104	-0,545	0,176	0,119	0,351
Fibrillation	269	46	100	80,07	0,556	9,126	-0,795	0,149	0,804	0,296
COPD	37	52	101	76,08	1,893	11,512	0,151	0,388	-0,401	0,759
Diabetes	2044	26	101	74,64	0,241	10,900	-0,698	0,054	0,654	0,108
Heart Attack	858	39	103	79,01	0,346	10,137	-0,667	0,083	0,192	0,167
Hypertension	4710	25	102	76,89	0,157	10,786	-0,721	0,036	0,682	0,071
Kidney failure	854	20	101	76,65	0,400	11,700	-1,064	0,084	1,876	0,167
Lung	85	57	90	75,65	0,940	8,671	-0,079	0,261	-0,946	0,517
Obesity	220	20	95	64,50	0,937	13,894	-0,544	0,164	0,185	0,327
Parkinson	203	57	97	79,78	0,509	7,258	-0,208	0,171	-0,105	0,340
Pneumonia	158	32	100	76,22	0,931	11,705	-0,844	0,193	1,025	0,384
Reflux	119	43	95	75,91	1,060	11,559	-0,491	0,222	-0,327	0,440
Schizophrenia	36	47	93	69,69	1,990	11,942	-0,085	0,393	-0,494	0,768
Stroke	219	44	99	76,43	0,676	10,005	-0,565	0,164	0,460	0,327
Tumor/Cancer	919	18	100	73,93	0,372	11,275	-0,783	0,081	1,569	0,161
Vasoconstriction	66	46	96	77,68	1,165	9,464	-0,414	0,295	0,838	0,582
Dementia	728	20	101	80.33	0.397	10.72	-1.62	0.089	4.61	0.181

Note: In the case of diseases signed by yellow background, the assumption of normality is rejected.

TABLE 3 Average age of victims as depending on the disease increasing order

Disease	Obesity	Schizophrenia	Asthma	Tumor	Diabetes	Lung
Avg. age	64,50	69,69	72,56	73,93	74,64	75,65
Disease	Reflux	COPD	Pneumonia	Stroke	Kidney failure	Hypertension
Avg. age	75,91	76,08	76,22	76,43	**76,65**	76,89
Disease	Vasoconstriction	Arrhythmia	Heart attack	Parkinson	Atrial fibrillation	Alzheimer
Avg. age	77,68	78,46	79,01	79,78	80,07	80,85
Disease	Dementia					
Avg. age	80,33					

We marked the results on the report of unbiased observations for both scenarios separately. The final result of clustering for 10 clusters by consideration of age, gender and disease has been shown in Table 4.

The number of the cases in each cluster, helps us to find the riskiest clusters with more deaths. Moreover, the feature of each cluster may warn the population about high probability of death if any new infected person with similar attributes assigns to the particular pattern of the cluster. Additionally, the output of the clustering, shows some positive and negative correlation between underlying medical conditions, that it may be valuable for medicine. Sorting the clusters from the highest risk to the lowest owing to the number of the cases involved will be as follows: 7-6-1-8-10-9-2-5-4-3.

From the sorted clusters, the first four clusters have the bulk proportion of the reported cases (7, 6, 1, 8). An age group of 80–87 of Females with medical problems of high blood pressure, lung disease, Arrhythmia, heart disease, dementia, atrial fibrillation and kidney failures

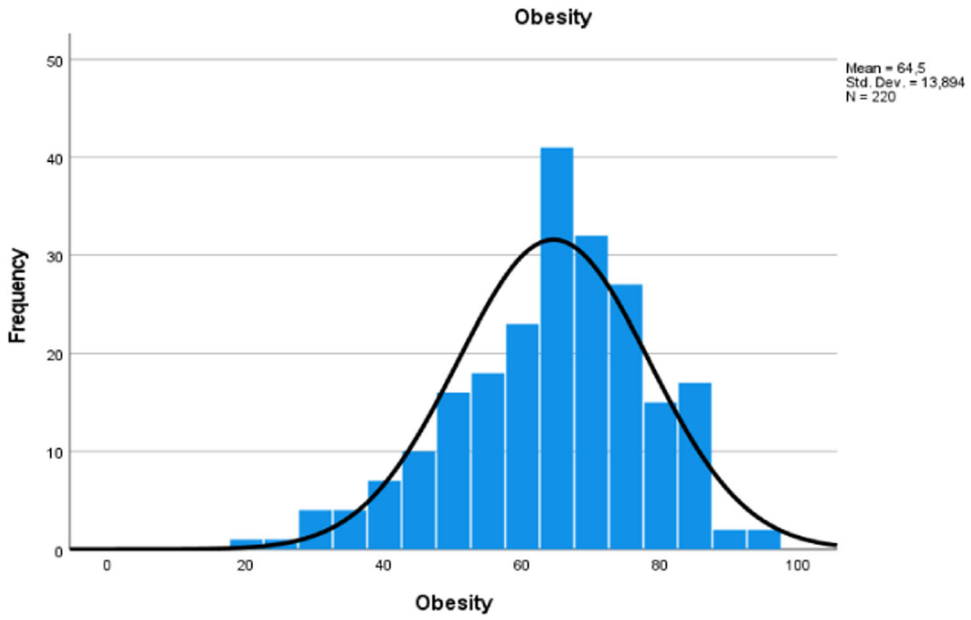

FIGURE 2 The histogram and fitted normal distribution of obesity

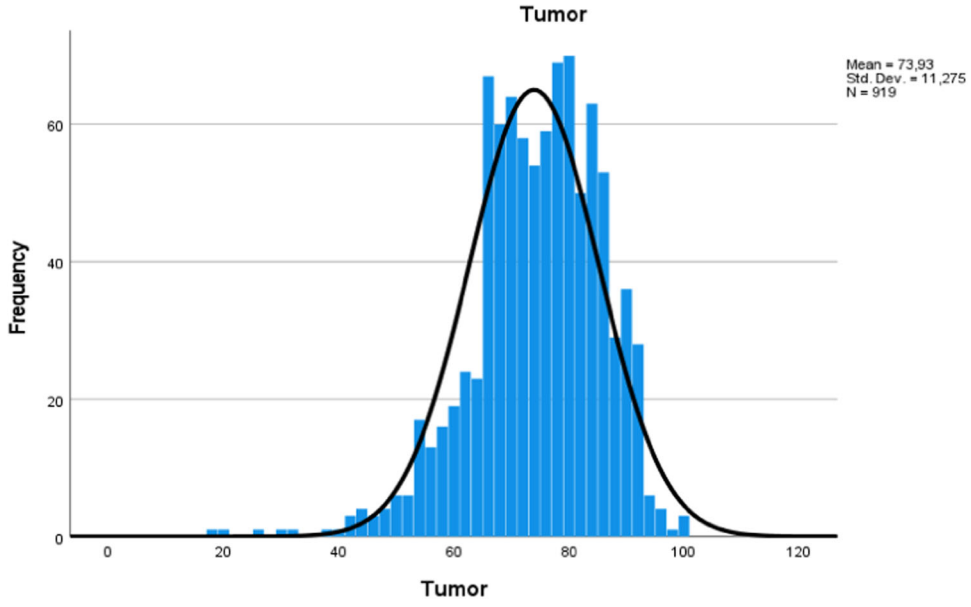

FIGURE 3 The histogram and fitted normal distribution of cancer/tumor

as the riskiest class will be in the serious trouble if they infected by COVID-19. The second murderous class is for males in the age group of 67–74 with underlying conditions of high blood pressure, tumor/cancer, lung disease, Arrhythmia, heart disease, diabetes, asthma. We observed that, lung disease in any cluster with infected cases, carries serious risk. It means, everyone by lung disease problem, in the case of infection by coronavirus may die with high probability (Table 5). White color of fonts shows the significant cases whose values are more than average.

Asthma as a big threat will be a killer for males if COVID-19 hits them. Most of the female victims are classified in the clusters by mean age of 80 and over, whereas the majority of male victims are grouped in whole ranges from 24 to below 74 (Table 6).

3.3 | Output of age-independent clustering

In this scenario, because of the separated reports of the health care system for Arrhythmia and heart disease, despite their similarities, we found 59 different cases between them. Accordingly, we distinguish both separately. It is more likely to realize some essential conclusions through

TABLE 4 The center points in the age-dependent clustering when K = 10

Factors	Clusters 1	2	3	4	5	6	7	8	9	10
Age	87.2	52.9	24.1	34.7	44.6	74.0	80.9	67.1	60.4	93.2
High blood pressure	0.7	0.5	0.12	0.29	0.38	0.65	0.7	0.63	0.57	0.67
Tumor/cancer	0.09	0.15	0.18	0.08	0.11	0.14	0.12	0.18	0.13	0.08
Chronic-lung-disease	0.01	0	0	0	0	0.02	0.01	0.02	0.01	0
Arrhythmia	0.08	0.03	0	0.04	0.03	0.08	0.07	0.05	0.04	0.09
Heart disease	0.15	0.08	0	0.02	0.02	0.11	0.13	0.09	0.09	0.18
Dementia	0.15	0.06	0.12	0.12	0.03	0.09	0.11	0.05	0.03	0.18
Atrial-fibrillation	0.06	0.01	0	0	0.01	0.03	0.05	0.02	0.01	0.05
Diabetes	0.24	0.31	0.12	0.14	0.24	0.33	0.29	0.3	0.34	0.16
Obesity	0.01	0.1	0.18	0.18	0.13	0.03	0.01	0.05	0.05	0.01
Kidney-Failure	0.13	0.1	0.12	0.14	0.1	0.12	0.13	0.1	0.1	0.14
Asthma	0.02	0.4	0	0.04	0.04	0.03	0.02	0.03	0.05	0.01
Gender	1.36	1.64	1.71	1.59	1.67	1.55	1.47	1.67	1.67	1.3
The number of members of each cluster	**1279**	**286**	**17**	**51**	**131**	**1470**	**1792**	**1161**	**522**	**528**

TABLE 5 The important properties of the age-dependent clusters

Factors	3	4	5	9	8	2	6	7	1	10	**Average**
Gender	1.71	1.59	1.67	1.67	1.67	1.64	1.55	1.47	1.36	1.3	
Age	24.1	34.7	44.6	60.4	67.1	52.9	74.0	80.9	87.2	93.2	
Hypertension	0.12	0.29	0.38	0.57	0.63	0.5	0.66	0.7	0.7	0.67	0.522
Tumor/Cancer	0.18	0.08	0.11	0.13	0.18	0.15	0.14	0.12	0.09	0.08	0.126
Lung-disease	0	0	0	0.01	0.02	0	0.02	0.01	0.01	0	0.007
Arrhythmia	0	0.04	0.03	0.04	0.05	0.03	0.08	0.07	0.08	0.09	0.051
heart disease	0	0.02	0.02	0.09	0.09	0.08	0.11	0.13	0.15	0.18	0.087
Dementia	0.12	0.12	0.03	0.03	0.05	0.06	0.09	0.11	0.15	0.18	0.094
Atrial fibrillation	0	0	0.01	0.01	0.02	0.01	0.03	0.05	0.06	0.05	0.024
Diabetes	0.12	0.14	0.24	0.34	0.3	0.31	0.33	0.29	0.24	0.16	0.247
Obesity	0.18	0.18	0.13	0.05	0.05	0.1	0.03	0.01	0.01	0.01	0.075
kidney failure	0.12	0.14	0.1	0.1	0.1	0.1	0.12	0.13	0.13	0.14	0.118
Asthma	0	0.04	0.04	0.05	0.03	0.04	0.03	0.02	0.02	0.01	0.028
#Cases of each cluster	17	51	131	522	1161	286	1470	1792	1279	528	
Young age group											
Middle age group											
Old age group											

three different cluster-running of K = 8, 9, 10 when age is not taken in the interpretation (Appendix).

As a result of 8 clusters running, the bulk part of the population came out into the following clusters in order; 1-5-4-2-7. The population of the clusters 3,6,8 is low, then they may not represent some kind of important risk. In this running, Reflux, Vasoconstriction, Parkinson, COPD, Stroke, Arrhythmia, are nonsignificant from analysis of variance (ANOVA) table, then they were removed from further analysis. Lung and heart diseases and Schizophrenia, follow similar and slight movements through all clusters, then there are not remarkable differences between clusters in the case of them. On the report of 9 clusters running, interpretation of the ANOVA table about nonsignificant factors, Arrhythmia, COPD, Vasoconstriction and Reflux has been withdrawn from additional study. The prominent part of the infected cases contained in the upcoming clusters; 5-9-3-6-8-4. Schizophrenia, equally distributed among clusters then there is no meaningful impact of that on any particular cluster. Clusters 5 and 9

TABLE 6 The most important diseases for some selected clusters

Underlying conditions	Cluster 1	Cluster 5	Cluster 4	Cluster 2	Cluster 7
High blood pressure	√		√		√
Asthma	√				
Pneumonia				√	
Dementia				√	
Diabetes					√
Tumor/cancer		√			
Atrial fibrillation			√		
Kidney failure					√

Note: Describes the most considerable underlying medical disease of the clusters for the K = 8 running.

seeing that the most crowded places, every member has a high blood pressure problem. To compare to other clusters, the frequency of the lung disease is the highest amount in clusters 5 and 1 in 0.03 which is allocated for men. Cause of the considerable percentage of obesity in the clusters 1,4,6, it seems more dangerous for males. Every member by 100% probability of Kidney problem is classified in the cluster 1. Further to this running, Kidney Failure, dramatically is a serious problem for men that it may need deeper studies. For 10 clusters running, everyone in cluster 3,5,6,8 and almost 9 suffers from Hypertension. Substantial medical problems for cluster 1 are Diabetes and Kidney failure. Second noticeable cluster for Kidney failure is 8 with 43% in which next serious cluster in the case of Dementia placed at 65%. Subject to cluster 4, Dementia in the highest probability (100%) accompanied by Parkinson, Schizophrenia and Alzheimer at the rate of three times more than average. Essential frequency of Obesity, was observed in cluster 9 nearly equal distribution between genders, tended to be biased towards females. Abundance of Arrhythmia happens in this cluster also at 61% (six times more than average). Diabetes is a serious problem for all members of the clusters 1,5,7,9 that interestingly, the spread of the different gender is equal in all. Stroke, COPD and Vasoconstriction's cases, without any remarkable mean differences, categorized as nonsignificant factors for this running. The riskiest cluster is 6 that except Hypertension that exist in all corresponding members, the other underlying medical conditions placed below average in this cluster, consequently, Hypertension (high blood pressure) is the most crucial disorder made an unsafe condition for the high percentage of Hungarian victims.

4 | DISCUSSION

4.1 | Age-dependent scenario

4.1.1 | Analysis of clusters between different genders

High blood pressure and Kidney failure sicknesses distributed equally between genders. It demonstrates that the risk of these two underlying conditions is same for male and female. Murderess of tumor-cancer is biased towards males significantly. Also, Arrhythmia and Atrial fibrillation are significant underlying conditions lead to death for infected females more than males. Asthma is a quite serious underlying problem for male substantially than female in dealing with COVID-19. Despite, almost equal distribution of heart disease and dementia between genders, higher proportion of these diseases for female in clusters, signifies that they caused deaths of females more than males in Hungary. Interestingly, diabetes and obesity accelerate the probability of death for males who contaminated by coronavirus.

4.1.2 | Examination of clusters within age groups

Further to the final clustering result, if we divide the age range to three main groups; (24–44), (45–65), (66–94) based on the clusters, we may observe that obesity is the most dangerous underlying reason for young people who died from coronavirus. Dementia, Asthma, Kidney failure, and Tumor/Cancer are placed into the following levels of risk, respectively, for young population. Since Dementia is caused by damage to brain cells by aging, mostly affects older adults. Then, existence the considerable number of it among young-aged groups was drawn the attention to detect its distribution on Figure 4.

In regard to the Figure 4, Dementia trend follows the expectation to go up by getting old starting almost from 60 years old. Middle infected age group; (45–65), is threatened to pass away by tumor/cancer, asthma and diabetes with the most probability. Furthermore, high blood pressure, lung and heart disease are listed in the second level of importance for this age group. For elders, heart disease is the most noticeable problem that depends on the absolute value remarks trouble. The lower level of death possibility for old people as underlying illnesses is coming out by dementia, Arrhythmia and Atrial fibrillation. Kidney problems and high blood pressure categorized as the next extent of the seriousness.

4.2 | Age-independent scenario

4.2.1 | Cluster K = 8

While we go through special cluster, further to high frequency of the factors, we can find some diseases as more threatening cases,

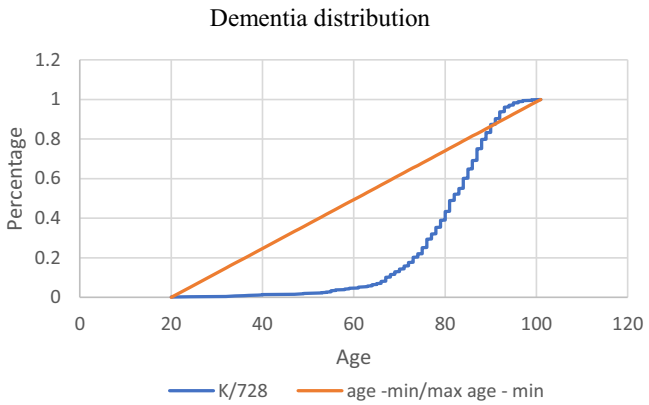

FIGURE 4 The distribution of Dementia as the function of age

since there are substantial differences between their quantities with other items in any singular cluster. For example, in cluster 1, amongst 13 significant diseases, high blood pressure and asthma are notable, as everybody has high blood pressure and in the case of asthma, is double compared to other clusters proportions. The largest ratio of cluster 5, as a second important cluster in this running, is assigned to tumor/cancer as a greatest risk in this cluster. High blood pressure, by covering all the population of cluster 4, showing the most crucial medical problem that it is accompanied by Atrial Fibrillation at its highest quantity, 0.05. In comparison with other clusters, the largest frequency percentage is specified to Pneumonia and dementia at 0.05 and 0.17, respectively. There is a positive association between high blood pressure, diabetes and kidney failure as the most outstanding cases in cluster 7. The biggest quantity of Kidney Failure is allocated to cluster 8 at 0.64 to may confirm the existence of the positive correlation with diabetes at 1.0 again. In the case of tumor/cancer, every infected person has this disease in the cluster of 3 and 6, but due to a few cases contained in these clusters, it may not validate the noticeable danger, that it is observed only in the half of the clusters. Also, male victims who suffered from tumor/cancer outnumbered females regard to this running. Surprisingly, high blood pressure and tumor/cancer in clusters 1,4,5,7 are excluded each other. Indeed, in any cluster that we have complete percentage of them, we do not observe another one at all. Pneumonia, dementia and asthma carry out high risks for females. Meanwhile, obesity, heart and lung diseases have almost similar effect between genders. Also, there can be a positive correlation between High blood pressure and tumor/cancer when there is another disease involved like diabetes.

Exceptionally, some sequences of the particular sicknesses, according to clusters 6 and 7, may exclude schizophrenia as below;

(high blood pressure, tumor/cancer) and (kidney failure, diabetes, high blood pressure).

As a last observation, Kidney failure is a serious risk for infected females by most probability.

4.2.2 | Cluster K = 9

The consequences of 9-cluster running show that Hungarian women who died by COVID-19 had suffered from Pneumonia, dementia, asthma, Parkinson, and Alzheimer more than men. In contrast, males outnumbered females with higher death probability in stroke and obesity. Amongst all the diseases, High blood pressure and diabetes are the remarkable risks to the infected population.

4.2.3 | Cluster K = 10

Since in our study, we combined all types of the cancers and tumor, therefore it is obvious to see some slight positive correlation between high blood pressure and cancer, but in clusters 2 and 10 of this running, both are excluding each other. Depends on the type of the cancer, we may have low or high blood pressure.

The frequency of different gender's underlying medical conditions has been presented in Table 7.

The impressive phenomenon of Table 7 is similar extent of some diseases among female and male like; Arrhythmia, COPD, Parkinson, Vasoconstriction, Schizophrenia. In contrast, big difference between genders in the case of Dementia, Diabetes, Tumor/Cancer is noticeable.

TABLE 7 The number of male and female cases of the registered diseases

Underlying conditions	The number of female	The number of male
Hypertension	2369	2341
Tumor/cancer	420	499
Lung disease	30	55
Arrhythmia	240	239
Heart disease	424	434
Pneumonia	88	70
Dementia	463	265
Atrial fibrillation	115	154
Diabetes	952	1092
Obesity	107	113
Stroke	97	122
Kidney failure	453	401
Asthma	109	81
COPD	19	18
Parkinson	102	101
Vasoconstriction	34	32
Alzheimer	86	58
Reflux	69	50
Schizophrenia	17	19

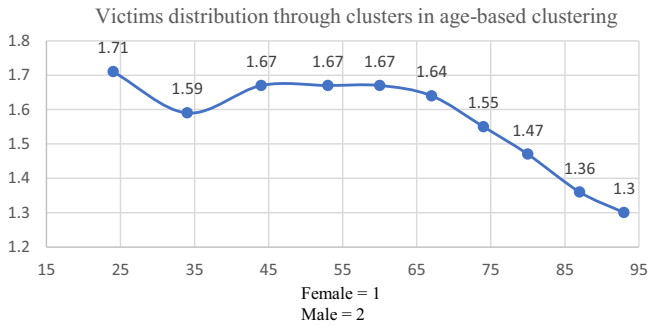

FIGURE 5 Victims distribution through clusters between genders

The correlation between underlying medical problems has been analyzed by SPSS and the result summarized in two tables; Table 8 part "a" and "b" (Appendix). The most noteworthy items are positive correlation between Hypertension with Diabetes and Alzheimer with Dementia. Contrary, negative association between Tumor/Cancer with Hypertension case noticed, that it may originate from chemotherapy or biological drugs which Cancer patients use results in hypotension or low blood pressure.

5 | CONCLUSION

Our result for clustering can forecast similar cases which are assigned to any cluster that it will be a serious cautious for the population. The nonsignificant logical factors have been eliminated from the analysis. Most of the researches emphasized on the importance of *Age* factor in COVID-19 effectiveness. To study deeper, we split the statistical analysis into two parts; when *Age* is taken into the consideration with gender and underlying medical problems, on the other hand, *Age* is excluded from inspection to examine the trend and the correlation of the other factors while *Age* is out of the observation. For both parts, factors with proportion more than cluster center points, interpreted.

At age-based analysis, male victims' numbers in young and middle age exceeds female's. Figure 5 describes that most of the victims in old-age group are women. Indeed, the tendency is that the high ratio of males decreases. Particularly, after 70 years old, the diagram starts falling down towards 1 which valued female gender. Whereas, for younger age groups, it is biased towards 2 as a value of male. As a result of this section, Hypertension, Arrhythmia, Dementia, heart disease, and Kidney failure commonly observed in the older age group of the population. High percentage of Tumor/Cancer and Diabetes assigned to the middle age group (52–67). Obesity was a serious medical problem for young victims of Hungary. To conclude age-independent study, we surveyed *Gender* and *Underlying conditions'* tendency and correlation with three varied clusters running K = 8, 9, 10. K = 8 with equal gender distribution throughout all clusters is a well-structured running. Indeed, on average, the number of the clusters occupied by different sex is equal, four clusters for male and four for female. The riskiest disease for cluster 1 with high number of victims, are Hypertension and Asthma. In cluster 5 of this running, regardless of being second dangerous cluster, there is no any main underlying disease which may be placed up to the cluster center point, they mostly stand on the average or even less. Cluster 4 defines third unsafe cluster, which sheds light on Hypertension, Atrial Fibrillation, and Lung disease as significant medical problems for the members. In K = 9 running, cluster 9, Hypertension with the highest risk for infected population, Dementia and Asthma, in the following level of the risky disease, recognized due to the big quantity of the cases. All members of cluster 5, as second hazardous class, suffered from Hypertension. Also, remarkable weight of Lung-disease detected there. Cluster 3 in spite of having a large number of the victims after clusters 9 and 5, does not have any dominant underlying disorders. Output of clusters 2 and 7 reveals that Tumor/Cancer and Heart disease were riskier for females, in contrast, Kidney failure in males. However, Hypertension and Diabetes spread impartially between various genders. Many numerous victims in the K = 10 running, assigned to cluster 6 with enormous underlying conditions of Hypertension and Lung-diseases. Cluster 3 is a specific class in which every member experienced Hypertension. Also, to compare to cluster center point, the rate of Asthma is double. Clusters 2 and 5, as the following high-risk clusters of this running, showed duplicated Tumor/Cancer and a perfect positive correlation between Hypertension and Diabetes. Gender-based study seems to be important in the second scenario, since significant pure numbers (exactly 1 or 2) detected for genders. Overall, it may be said, in the case of Hungarian victims of COVID-19 till December of 2020, Hypertension, Diabetes, Tumor/Cancer, Heart disease and Kidney failure are the outstanding underlying conditions with the frequency of 3517 and 3720 for female and male respectively. The consequences of the study may guide to be taken precedence of riskier *Underlying medical problems*, *Gender* and *Age* in hospitalization and vaccination.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

AUTHOR CONTRIBUTIONS

Elnaz Gholipour: Idea of the research; literature review; methodology; analysis and conclusion. **Bela Vizvari:** Interpretation of the result and adding some extra part for methodology. **Tareq Babaqi:** Data collection and adding some points for discussion. **Szabolcs Takács:** Advisor of the medical results and conclusions.

DATA AVAILABILITY STATEMENT

"All data collected from Hungarian government website: https://koronavirus.gov.hu/elhunytak."

ORCID

Elnaz Gholipour http://orcid.org/0000-0002-5278-7381
Béla Vizvári https://orcid.org/0000-0002-1349-1035
Tareq Babaqi http://orcid.org/0000-0003-2232-6370

REFERENCES

1. Hoseinpour Dehkordi A, Alizadeh M, Derakhshan P, Babazadeh P, Jahandideh A. Understanding epidemic data and statistics: a case study of COVID-19. *J Med Virol*. 2020;92(7):868-882. https://doi.org/10.1002/jmv.25885
2. 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
3. Hâncean M-G, Slavinec M, Perc M. The impact of human mobility networks on the global spread of COVID-19. *Journal of Complex Networks*. 2020;8(6):1-14. https://doi.org/10.1093/COMNET/CNAA041
4. Hâncean M-G, Perc M, Lerner J. Early spread of COVID-19 in Romania: imported cases from Italy and human-to-human transmission networks. *R Soc Open Sci*. 2020;7(7):200780. https://doi.org/10.1098/RSOS.200780
5. Priesemann V, Balling R, Brinkmann MM, et al. An action plan for pan-European defence against new SARS-CoV-2 variants. *Lancet (London, England)*. 2021;397(10273):469-470. https://doi.org/10.1016/S0140-6736(21)00150-1
6. Priesemann V, Brinkmann MM, Ciesek S, et al. Calling for pan-European commitment for rapid and sustained reduction in SARS-CoV-2 infections. *The Lancet*. 2021;397(10269):92-93. https://doi.org/10.1016/S0140-6736(20)32625-8
7. Underlying Medical Conditions for Clinicians (2021). Available at: https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/underlyingconditions.html. Accessed 6 April 2021.
8. Rosenthal N, Cao Z, Gundrum J, Sianis J, Safo S. Risk factors associated with in-hospital mortality in a US national sample of patients with COVID-19. *JAMA network open*. 2020;3(12):e2029058. https://doi.org/10.1001/jamanetworkopen.2020.29058
9. Harrison SL, Fazio-Eynullayeva E, Lane DA, Underhill P, Lip G. "Comorbidities associated with mortality in 31,461 adults with COVID-19 in the United States: A federated electronic medical record analysis". *PLoS Med*. 2020;17(9):1003321. https://doi.org/10.1371/JOURNAL.PMED.1003321
10. Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. 2020;584(7821):430-436. https://doi.org/10.1038/s41586-020-2521-4
11. Guilmoto CZ. COVID-19 death rates by age and sex and the resulting mortality vulnerability of countries and regions in the world. *medRxiv*. *medRxiv*. 2020. https://doi.org/10.1101/2020.05.17.20097410
12. Alwani M, Yassin A, Al-Zoubi RM, et al. Sex-based differences in severity and mortality in COVID-19. *Rev Med Virol*. 2021;31(2):1-11. https://doi.org/10.1002/rmv.2223
13. Zhang J, Wang X, Jia X, et al. Risk factors for disease severity, unimprovement, and mortality in COVID-19 patients in Wuhan, China. *Clin Microbiol Infect*. 2020;26(6):767-772. https://doi.org/10.1016/j.cmi.2020.04.012
14. Li X, Xu S, Yu M, et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *J Allergy Clin Immunol*. 2020;146(1):110-118. https://doi.org/10.1016/j.jaci.2020.04.006
15. Rossen LM, Branum AM, Ahmad FB, Sutton P, Anderson RN. Excess Deaths Associated with COVID-19, by age and race and ethnicity — United States, January 26–October 3, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(42):1522-1527. https://doi.org/10.15585/mmwr.mm6942e2
16. Omori R, Matsuyama R, Nakata Y. The age distribution of mortality from novel coronavirus disease (COVID-19) suggests no large difference of susceptibility by age. *Sci Rep*. 2020;10(1):16642. https://doi.org/10.1038/s41598-020-73777-8
17. Underlying conditions in confirmed cases of COVID-19 in Ireland (2020). Available at: www.hpsc.ie. Accessed 7 April 2021.
18. COVID-19 : Prevention and Groups at Higher Risk - NYC Health (no date). Available at: https://www1.nyc.gov/site/doh/covid/covid-19-prevention-and-care.page. Accessed 7 April 2021.
19. Zaki N, Alashwal H, Ibrahim S. Association of hypertension, diabetes, stroke, cancer, kidney disease, and high-cholesterol with COVID-19 disease severity and fatality: a systematic review. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*. 2020;14(5):1133-1142. https://doi.org/10.1016/j.dsx.2020.07.005
20. Esme M, Koca M, Dikmeer A, et al. Older adults with coronavirus disease 2019: a nationwide study in Turkey. *J Gerontol A Biol Sci Med Sci*. 2021;76(3):e68-e75. https://doi.org/10.1093/gerona/glaa219
21. Yang J, et al. 'Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis'. 2020. https://doi.org/10.1016/j.ijid.2020.03.017
22. Cincotta R. Population age structure: the hidden factor in COVID-19 mortality. The blog of the Wilson Center's Environmental Change and Security Program, (December). 2020. Available at https://www.newsecuritybeat.org/2020/05/population-age-structure-hidden-factor-covid-19-mortality/
23. Koronavírus. (2020). Available at: https://koronavirus.gov.hu/elhunytak. Accessed 7 April 2021.
24. Sonbhadra SK, Agarwal S, Nagabhushan P. Target specific mining of COVID-19 scholarly articles using one-class approach. *Chaos Solitons Fractals*. 2020;140:110155. https://doi.org/10.1016/j.chaos.2020.110155
25. Sagar PV, Pavan T, Krishna G, Nageswara M. COVID-19 transmission risks assessment using agent-based weighted clustering approach. *Int J Adv Comput Sci Appl*. 2020;11(11):532-537. https://doi.org/10.14569/IJACSA.2020.0111167
26. Mao S, Huang T, Yuan H, et al. "Epidemiological analysis of 67 local COVID-19 clusters in Sichuan Province, China". *BMC Public Health*. 2020;20(1):1525. https://doi.org/10.1186/s12889-020-09606-4
27. Pung R, Chiew CJ, Young BE, et al. Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures. *The Lancet*. 2020;395(10229):1039-1046. https://doi.org/10.1016/S0140-6736(20)30528-6
28. Virgantari F, Faridhan YE. K-Means Clustering of COVID-19 Cases in Indonesia's Provinces. 2020. Available at: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-atthe-mission-briefing-on-covid-19-12-. Accessed 7 April 2021.
29. Hutagalung J, et al. COVID-19 cases and deaths in Southeast Asia Clustering using K-Means Algorithm. Journal of Physics: Conference Series. IOP Publishing Ltd. 2021. https://doi.org/10.1088/1742-6596/1783/1/012027
30. Bhunia GS, Roy S, Shit PK. Spatio-temporal analysis of COVID-19 in India—a geostatistical approach. *Spat Inf Res*. 2021;27:1-12. https://doi.org/10.1007/s41324-020-00376-0
31. Maugeri A, Barchitta M, Agodi A. A clustering approach to classify italian regions and provinces based on prevalence and trend of sars-cov-2 cases. *Int J Environ Res Public Health*. 2020;17(15):1-14. https://doi.org/10.3390/ijerph17155286
32. Adam DC, Wu P, Wong JY, et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nature Med*. 2020;26(11):1714-1719. https://doi.org/10.1038/s41591-020-1092-0
33. Azarafza M, Azarafza M, Akgün H. Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran.

medRxiv. *medRxiv*. 2020. 2020.05.22.20109942. https://doi.org/10.1101/2020.05.22.20109942

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Gholipour E, Vizvári B, Babaqi T, Takács S. Statistical analysis of the Hungarian COVID-19 victims. *J Med Virol*. 2021;93:6660-6670. https://doi.org/10.1002/jmv.27242