Scientific Research Publishing

Open Journal of Statistics, 2021, 11, 259-277
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Predictors of the Aggregate of COVID-19 Cases and Its Case-Fatality: A Global Investigation Involving 120 Countries

Sarah Al-Gahtani[1], Mohamed Shoukri[2]*, Maha Al-Eid[3]

[1]Department of Internal Medicine, King Fahd Medical City, Riyadh, Saudi Arabia
[2]Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London Ontario, Canada
[3]Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
Email: *shoukri.mohamed@gmail.com

How to cite this paper: Al-Gahtani, S., Shoukri, M. and Al-Eid, M. (2021) Predictors of the Aggregate of COVID-19 Cases and Its Case-Fatality: A Global Investigation Involving 120 Countries. *Open Journal of Statistics*, **11**, 259-277.
https://doi.org/10.4236/ojs.2021.112014

Received: February 2, 2021
Accepted: March 6, 2021
Published: March 9, 2021

Abstract

Objective: Since the identification of COVID-19 in December 2019 as a pandemic, over 4500 research papers were published with the term "COVID-19" contained in its title. Many of these reports on the COVID-19 pandemic suggested that the coronavirus was associated with more serious chronic diseases and mortality particularly in patients with chronic diseases regardless of country and age. Therefore, there is a need to understand how common comorbidities and other factors are associated with the risk of death due to COVID-19 infection. Our investigation aims at exploring this relationship. Specifically, our analysis aimed to explore the relationship between the total number of COVID-19 cases and mortality associated with COVID-19 infection accounting for other risk factors. **Methods**: Due to the presence of over dispersion, the Negative Binomial Regression is used to model the aggregate number of COVID-19 cases. Case-fatality associated with this infection is modeled as an outcome variable using machine learning predictive multivariable regression. The data we used are the COVID-19 cases and associated deaths from the start of the pandemic up to December 02-2020, the day Pfizer was granted approval for their new COVID-19 vaccine. **Results**: Our analysis found significant regional variation in case fatality. Moreover, the aggregate number of cases had several risk factors including chronic kidney disease, population density and the percentage of gross domestic product spent on healthcare. **The Conclusions**: There are important regional variations in COVID-19 case fatality. We identified three factors to be significantly correlated with case fatality.

Keywords

Intraclass Correlation Coefficient, Hierarchical Data Structure, Negative Binomial Regression, Data Splitting, Mixed Effects Linear Regression Model

1. Introduction

Coronavirus disease 2019 (COVID-19) is now a considered a pandemic by the World Health Organization. The main objective of this study is to report on the association between regional case fatality of COVID-19, kidney diseases mortality, diabetes, population density and the Gross Domestic Product (GPD). This cross-sectional historical data was constructed by combining data from several sources [1]-[6].

The above sited data sources were accessed on December 2-2020. We included in the study the cumulative number of COVID-19 cases and the associated death counts by country as of December 2-2020. We excluded countries that had cumulative count less than 10,000 cases. The data base has 120 countries, and we divided them into regions according to the classification given in data source number 2, resulting in 15 regions. This classification is shown in **Table 1**. This table has in the first column the names of the countries, the second column is the name of the region they belong to, within brackets, the number of countries in that region, and the third column is a code given to each region.

Basically, we have two outcome variables of interest: 1) The aggregate number of cases per country over the period ending December 2-2020 (AC). 2) The COVID-19 Case Fatality (CF). This is calculated as:

AC = Count of COVID-19 cases analyzed at the regional level.

CF = Count of deaths attributed to COVID-19/(Count of COVID-19 cases) × 100,000.

This paper has three objectives. Firstly, we quantify the degree between regions variation in both AC and CF. The second objective is to identify the important factors associated with AC and CF. We use machine learning algorithm to build a regression models with the entire data set divided into learning set and validation test to quantify the predictive accuracy of the constructed models.

2. Selected Risk Factors

2.1. Chronic Kidney Diseases

Chronic Kidney Disease (CKD) is an important contributor to morbidity and mortality from noncommunicable diseases, and this disease should be actively addressed to meet the UN's Sustainable Development Goal target to reduce premature mortality from non-communicable diseases by a third by 2030 [2]. We extracted the CKD prevalence and the associated mortality from the countries listed in table. The rationale was that there are many published articles

Table 1. Countries and the corresponding Regional classification as given in https://doi.org/10.1016/s0140-6736(20)30045-3.

Countries	Region name	Region Code
Peru, Ecuador, Bolivia	Andean. Latin (3)	10
Kazakhstan, Georgia, Armenia, Azerbaijan, Kyrgyzstan, Uzbekistan, Tajikistan	Central Asia (7)	2
Czechia, Romania, Hungary, Serbia, Bulgaria, Croatia, Slovakia, Bosnia, Slovenia, North-Macedonia, Albania, Montenegro	Central Europe (12)	5
Brazil, Columbia, Mexico, Panama, Costa Rica, Guatemala, Honduras, Venezuela Paraguay, El-Salvador.	Central Latin America (10)	11
Dominican Republic, Puerto Rico, Jamaica	Caribbean (3)	9
Ethiopia, Kenya, Uganda, Zambia, Madagascar. Mozambique, Angola, French Guinea.	CESSA (8)	13
Indonesia, Philippine, China, Myanmar, Malaysia, Sri Lanka, French Polynesia, Maldives.	East Asia (8)	1
Russia, Poland, Ukraine Belarus, Lithuania, Latvia, Estonia	East Europe (7)	6
Japan, Singapore, Republic Korea, Australia.	HIAP (4)	4
Iran, Iraq, Turkey,	MENA (20)	12
Morocco, Saudi Arabia, Israel, Jordan, United Arab, Kuwait, Qatar, Lebanon, Oman, Egypt, Occupied Palestine, Tunisia, Bahrain, Algeria, Libya, Afghanistan, Sudan.		
India, Bangladesh, Pakistan, Nepal.	South Asia (4)	3
Chile, Argentina, Uruguay	South Latin America (3)	8
South Africa, Namibia, Zimbabwe	SSSA (3)	15
USA, France, Spain, UK, Italy, Germany, Belgium, Netherland	West Europe and North America (20)	7
Canada, Switzerland, Portugal, Austria, Sweden, Greece, Denmark, Ireland, Norway, Luxemburg, Finland, Cyprus.		
Mali, Nigeria, Ghana, Cameroon, Ivory Coast	WSSA (8)	14
Senegal, Guinea, Cape Verde		

CSSA = Central Sub-Saharan-Africa, MENA = Middle East and North Africa, HIAP = High Income Asian Pacific, WSSA = Western Sub-Saharan-Africa, SSSA = Southern Sub-Saharan Africa.

highlighting the importance of CKD as a possible risk factor for COVID-19 mortality.

A recent meta-analysis [7] outlined several reasons urging investigators to emphasize the importance of CKD during the COVID-19 infection. It was also noted in [7] that CKD has not attracted enough awareness due to its inconspicuous course, especially in the early stage. Both diabetes and hypertension are the leading causes of CKD in all developed countries and many developing countries, and the long-term or advanced CKD usually increases the risk of cardiovascular diseases. To be noted, these conditions accompanying CKD are all risk factors that exacerbate the COVID-19 patients. The present study highlights the importance of CKD as a risk factor for COVID-19 mortality. Some reports either did not include information on CKD or failed to state the definition of CKD used in the study. By contrast, the study by Williamson *et al.* [8] includes data for three subgroups with CKD. These data also demonstrate that patients with severe forms of CKD have a very high risk of COVID-19 mortality, which is even higher than that of other known high-risk groups, including patients with hypertension, obesity, chronic heart disease or lung disease [9] [10] [11]. The CKD data indicate that these patients deserve special attention with regard to COVID-19.

2.2. Population Density

Two studies [12] [13] utilized data from Japan suggested that the population density, which is somewhat indicative of social distancing, was a significant factor associated with COVID-19 infection. The effect of population density on the morbidity rate was also discussed in a case study of Iran [14]. These studies suggested that several cofactors introduce uncertainty. When discussing the effects of policies, a multi-city analysis representing different countries may be imperative. In multi-country analyses, the number of conducted tests may add uncertainty because this number depends on medical and economic resources for each country.

2.3. Diabetes

A recent population-based study from Italy [15] documented that the presence of comorbidities, including diabetes, were associated with a more severe course of COVID-19 and a higher fatality rate. Other studies from the most affected countries, including China, United States and Italy, seem to indicate that prevalence of diabetes among patients affected by COVID-19 is not higher than that observed in the general population, thus suggesting that diabetes is not a risk factor for SARS-CoV-2 infection. However, a large body of evidence demonstrate that diabetes is a risk factor for disease progression towards critical illness, development of acute respiratory distress syndrome, need for mechanical ventilation or admission to intensive care unit, and ultimately death. The mechanisms underlying the relationship between COVID-19 and diabetes remain to be elucidated. In particular, it is still unresolved whether is diabetes per se, especially if poorly controlled, or rather the various comorbidities/complications associated with it that predispose patients with COVID-19 to a worse prognosis. In fact, conditions that cluster with diabetes in the context of the metabolic syndrome, such as obesity and hypertension, or complicated chronic hyperglycemia, such as cardiovascular disease and chronic kidney disease, have also been associated with poor prognosis in these individuals and the available studies have not consistently shown that diabetes predict disease severity independently of them [15].

The estimated global prevalence was 9.3% in 2019 with an upward trend [16] [17]. In the USA alone, more than 34 million adults had known or undiagnosed diabetes in 2018 [18]. In 2017, diabetes was listed as the underlying or contributing cause of death on 270,702 death certificates, which corresponds to a crude rate of 83.1 per 100,000 persons [18]. Infectious diseases are more frequent and can be associated with worse outcomes in patients with diabetes [19]. Therefore, it is not surprising that diabetes has been considered as a possible risk factor or a predictor for worse outcomes in patients with coronavirus disease 2019 (COVID-19) [20] [21] [22]. COVID-19 rapidly reached the level of a pandemic and has caused more than 850,000 deaths worldwide within a few months despite unprecedented mitigation measures [23]. The strength of the association between diabetes and COVID-19 has been investigated in observational cohorts around the world. We aimed to systematically review and conduct a meta-analysis of the available observational studies reporting the effect of diabetes on mortality among hospitalized patients with COVID-19.

We obtained the data of global prevalence of diabetes by country from [24]. The rationale behind including diabetes in the data base of or study is to explore the strength of the country level diabetes prevalence with the CF outcome variable.

2.4. Gross Domestic Product (GDP) and the Percentage of Spending on Health Care (PHSC)

A recent article [25] analyzed data from 88 countries included as a covariate the percentage of government spending on healthcare. This factor is an indicative of the level of preparedness of the healthcare system to face urgent crisis such as the one caused by the COVID-19 pandemic. The authors showed that there is a significant relationship between an increase in health care spending and reduction in the COVID-19 case fatality. The finding that countries with strong healthcare capacity had fewer deaths per confirmed case was unsurprising. In this sense, the [25] study confirms previous research on the association of overall mortality (all causes) and healthcare funding [26].

3. Data Analysis

We divided the data analysis into two sections. In the first section univariate and descriptive statistics are produced with graphics. We also explore the extent of regional variability for the parameters of interest. This is measured by comparing the between regional variation to the within regional variations. A widely used statistic in this regard is the "Intraclass Correlation Coefficient" (ICCC). In the second stage we first identify the risk factors associated with the outcome of interest. Factors that are significantly correlated with the outcome of interest are used in a multivariate analysis model. The model that we used identifies regions as random effects. Using predictive analytic machine learning regression approach to evaluate the joint effect of the selected covariates on the outcome of interest is done by splitting the data into training data (70% of the entire data) and validation or test set (the remaining 30% of the data).

3.1. Descriptive Measures (Table 2)

The lowest average population density is in region 4, or, HIAP (47267.5 ± 51256.458).

The highest average population density is in region 7, or, West Europe (2628158.14 ± 6383767.55).

The histogram of population density is shown in Figure 1.

In Table 3 we present the summary statistics for the aggregate number of COVID-19 cases.

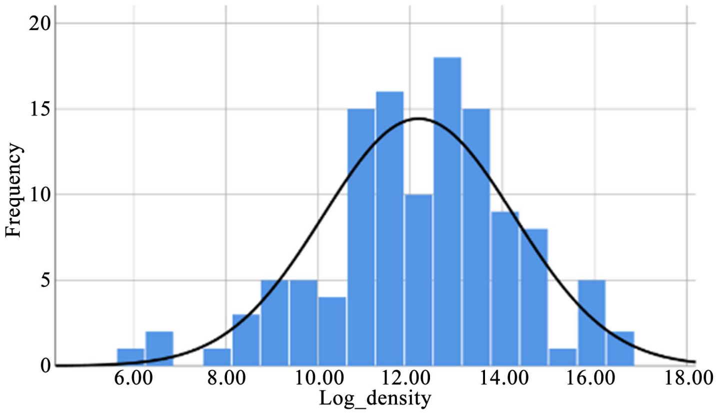

Figure 1. Histogram of log-population density.

Table 2. Summary measures of population density.

region	n	Mean	Std. Deviation
1	7	1,861,024.86	3,518,931.634
2	7	528,770.57	978,243.226
3	4	1,116,063.25	1,488,519.088
4	4	2,042,718.50	3,769,584.918
5	12	71,242.50	61,214.452
6	7	2,628,158.14	6,383,767.547
7	20	1,148,782.75	2,924,033.825
8	2	1,768,251.00	1,431,394.843
9	4	47,267.50	52,156.458
10	3	886,879.33	536,486.935
11	10	1,331,402.60	2,603,264.787
12	20	1,278,450.40	2,620,104.048
13	8	674,711.75	394,170.103
14	8	455,876.25	416,606.224
15	3	812,469.67	415,296.062
Total	119	1,078,510.29	2,576,968.860

Table 3. Summary measures of the aggregate number of COVID-19 cases by region.

region	n	Mean	Std. Deviation
1	7	181,340.86	214,429.947
2	7	105,576.43	55,184.339
3	4	2,650,841.50	4,566,745.811
4	4	68,670.75	57,557.320
5	12	175,681.33	163,726.359
6	7	619,801.86	858,270.813
7	20	1,230,523.65	2,938,619.513
8	2	988,698.50	616,363.061
9	4	55,656.50	62,874.115
10	3	434,431.00	458,961.374
11	10	953,104.00	1,947,275.831
12	20	233,409.90	230,576.529
13	8	36,655.00	38,224.367
14	8	39,255.13	34,628.880
15	3	272,281.33	450,353.611
Total	119	530,003.45	1,619,102.721

From Figure 2 we can see that the distribution of the number of cases is highly skewed to the right and the data in Table 1 shows that the variance is much larger than the mean, a phenomenon known as "over dispersion".

In Table 4 we present the mean, standard deviations of CF as defined above. The largest mean CF is in Andean Latin America (5626.36 ± 1685), while the smallest mean number of CF is in Central Asia (1199.95 ± 402.64).

The large variations in CF can better be depicted from the boxplot of the data as shown by the boxplot as given in Figure 3.

It appears from the above boxplot (Figure 3) that there is considerable variation in the distribution of CF. One way to stabilize this variation is to employ a specific transformation. We selected the logarithmic transformation, so that the dependent variable of interest is Y = Log (CF). The histogram of this new outcome variable Y is given in Figure 4. The Q-Q plot of Y is given in Figure 5 and it shows the closeness of the distribution of Y to that of the normal distribution.

The primary predictor of interest is CKD case fatality (Table 5).

As can be seen from Figure 6, there is great amount of variation among the 15 regions with respect to CKD case fatality. The next covariate we examine is the percent of GDP spending on healthcare. The summary measures are shown in Table 6, with the histogram given in Figure 7.

The above histogram (Figure 7) is quite symmetric apart from an extreme outlier. We shall not employ any transformation on this variable.

In Table 7 we present the summary measures of the diabetes prevalence for each of the 15 regions in the data.

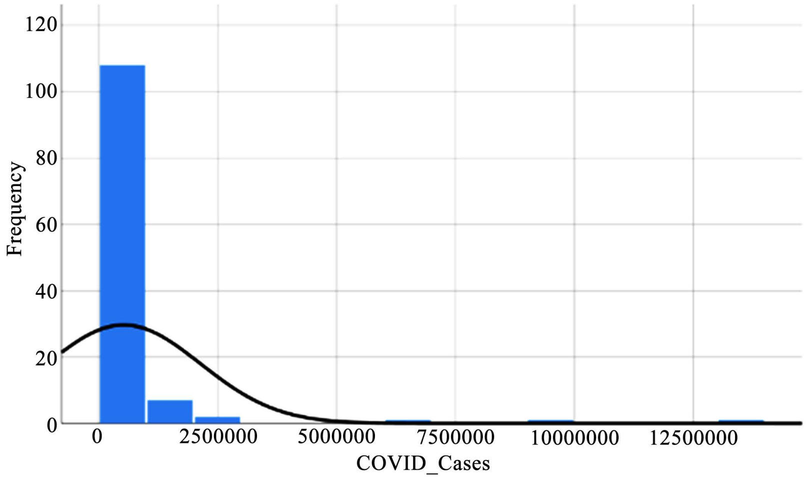

Figure 2. Histogram of the number of COVID-19 cases.

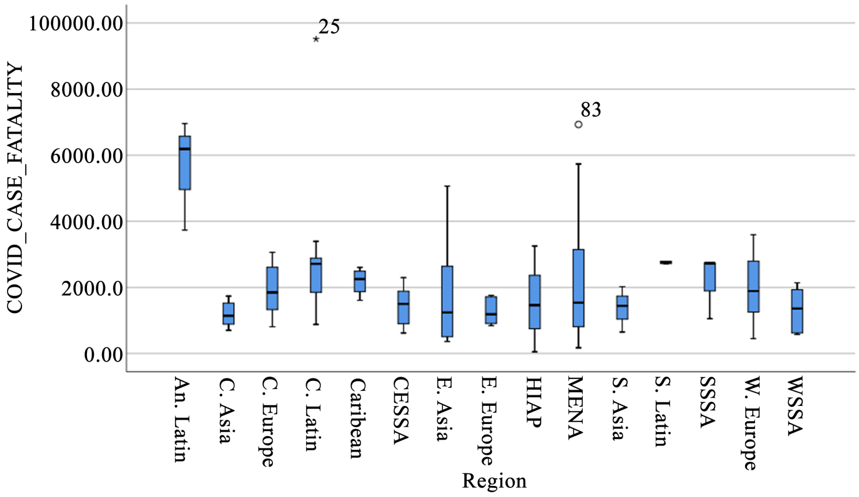

Figure 3. Boxplot of COVID-19 case fatality by region.

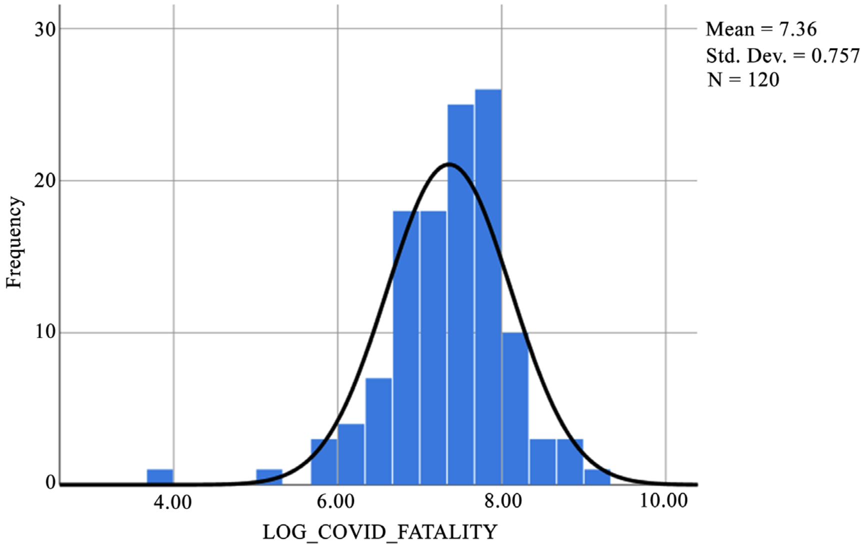

Figure 4. Histogram of Y = Log COVID-19 case fatality.

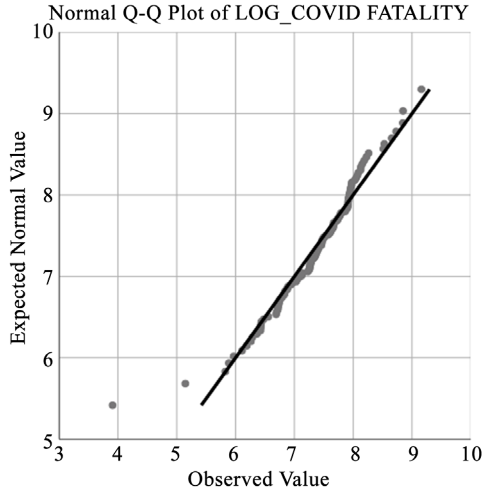

Figure 5. Q-Q plot of the logarithm of COVID-19 case fatality.

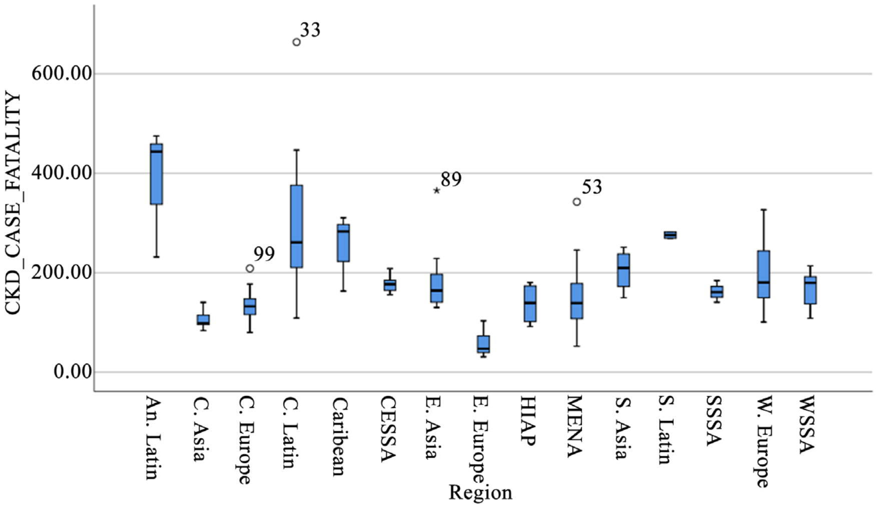

Figure 6. Boxplot for the CKD case fatality.

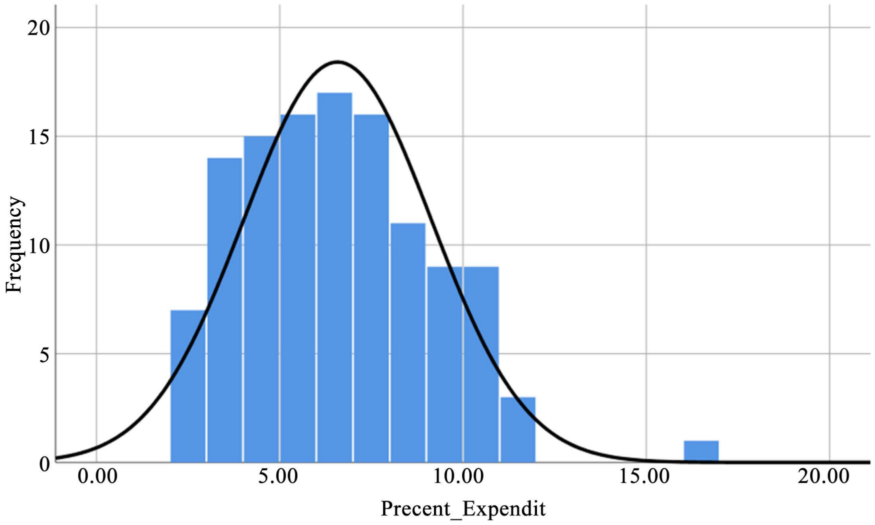

Figure 7. Histogram of the distribution of percentage of GDP on healthcare.

Table 4. COVID_CASE_FATALITY = (COVID death count/COVID Cases) × 100,000.

Region	No. Countries	Mean	Std. Deviation	Std. Error
E. Asia-1	7	1953.2808	1723.40851	651.38719
C.l Asia-2	7	1199.9530	402.63886	152.18318
S. Asia-3	4	1388.4963	562.42803	281.21402
HIAP-4	4	1557.8282	1311.68261	655.84130
C.Euro-5	12	1907.2677	781.13785	225.49507
E.Euro-6	7	1290.2332	426.50378	161.20328
W.Euro7	20	2021.0526	927.78022	207.45796
S.L.A -8	2	2754.8536	51.00766	36.06786
Caribb-9	4	2179.4770	427.92517	213.96259
A. L.A-10	3	5626.3620	1684.96910	972.81736
C. Lat-11	10	3009.9215	2415.83335	763.95358
Mena-12	20	2195.6513	1917.32557	428.72703
Cessa-13	8	1436.8359	589.94954	208.57866
Wssa-14	8	1319.0813	683.02871	241.48712
Sssa-15	3	2173.6344	969.30033	559.62581
Total	119	2015.8405	1466.14430	134.40123

Table 5. CKD_CASE_FATALITY (2017).

Code	Region (n)	Mean	Std. Deviation	Std. Error
1	E. Asia 7	190.6910	83.79775	31.67257
2	C. Asia 7	106.2960	19.44093	7.34798
3	S. Asia 4	205.0174	43.44818	21.72409
4	HIAP 4	137.6728	42.52006	21.26003
5	C. Eur. 12	134.3351	34.96087	10.09233
6	E. Eur. 7	58.2134	26.27258	9.93010
7	W. Eur.20	198.4845	67.26280	15.04042
8	SLA. 2	275.6629	9.04869	6.39839
9	Caribb. 4	259.9511	65.95979	32.97990
10	And. Lat.3	383.4160	132.26184	76.36141
11	CLA. 10	306.2912	159.77184	50.52429
12	MENA 20	149.4680	66.97740	14.97660
13	CESSA 8	177.1631	16.48127	5.82701
14	WSSA 8	167.5075	37.65762	13.31398
15	SSSA 3	161.9803	21.72116	12.54072
Total	119	180.4692	95.11304	8.71900

Table 6. Percent of GDP spending on healthcare (2018).

Code	Region(n)	Mean	Std. Deviation
1	E. Asia 7	4.9057	2.14183
2	C. Asia 7	6.0900	2.42986
3	S. Asia 4	3.7300	1.49457
4	HIAP 4	8.0625	2.77191
5	C. Euro.12	7.2317	1.17092
6	E. Euro. 7	6.3514	0.77802
7	W. Euro 20	9.8350	2.37384
8	SLA 2	9.3800	0.33941
9	Caribb.4	4.3950	1.76847
10	ALA 3	6.5600	1.46738
11	CLA 10	6.7430	1.58911
12	MENA19	5.7611	1.91436
13	CESSA 7	5.0629	1.88558
14	WSSA 8	4.0038	0.60512
15	SSSA 3	6.9867	1.96128
Total	117	6.6104	2.54257

Table 7. Diabetes prevalence worldwide by region (2019).

Code	Region	Mean	Std. Deviation	Std. Error
E. Asia	6	8.9833	4.45260	1.81777
C. Asia	7	6.1143	0.20354	0.07693
S. Asia	4	11.6750	5.63996	2.81998
HAIP	4	5.9000	0.66833	0.33417
C. Eur	12	7.4917	1.45943	0.42130
E. Eur	7	5.8286	2.27502	0.85988
W. Eur	20	6.3700	2.22429	0.49737
SLA	2	7.2500	1.90919	1.35000
Carrib.	4	10.7500	3.09031	1.54515
ALA	3	6.3000	0.70000	0.40415
CLA	10	9.0800	1.96344	0.62090
MENA	20	11.9650	3.95012	0.88327
CESSA	8	5.5750	5.03920	1.78163
WSSA	8	2.9500	1.25584	0.44401
SSSA	3	6.3333	5.67656	3.27736
Total	118	7.7864	3.89396	0.35847

The lowest diabetes prevalence is in WSSA (2.95), while the highest is in MENA region (11.965).

3.2. Detection of Regional Clustering

In this section we shall quantify the degree of clustering for any continuous variable.

We assume that we have k regions and that the individual units within each region are the countries as indicated in Table 1. To articulate this concept, we first assume that the quantity of interest measured in the j^{th} country within the i^{th} *region,* y_{ij} is modelled as

$$y_{ij} = \mu + \tau_i + e_{ij} \tag{3.1}$$

where;

τ_i = random regional effect and represents the collective effects of all regional level unmeasured covariates.

e_{ij} = the country within region deviation from the overall mean of the i^{th} *region*

μ is the grand mean of all measurements in the population. It is assumed that the region effects $\{\tau_i\}$ are normally and identically distributed with mean 0 and variance σ_τ^2, the errors $\{e_{ij}\}$ are normally and identically distributed with mean 0 and variance σ_ϵ^2, and the $\{\tau_i\}$ and $\{e_{ij}\}$ are independent. For this model the ICCC, which may be interpreted as the correlation ρ between any two countries belonging to the same region, is defined as:

$$\rho = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\epsilon^2} \tag{3.2}$$

It is seen by definition that the ICCC is defined as non-negative in this model, a plausible assumption for the application of interest here. We also note that the variance components σ_τ^2, and σ_ϵ^2 can be estimated from the one-way ANOVA mean squares (see; Shoukri, page 47) [27] given in expectation by

$$E(\text{MSB}) = \sigma_\epsilon^2 + n_0\sigma_\tau^2 \tag{3.3}$$

$$E(\text{MSW}) = \sigma_\epsilon^2 \tag{3.4}$$

The ANOVA estimator of ρ is then given by

$$\hat{\rho}_0 = \frac{\text{MSB} - \text{MSW}}{\text{MSB} + (n_0 - 1)\text{MSW}} \tag{3.5}$$

where MSB and MSW are, obtained from the usual ANOVA table, with corresponding sums of squares

$$\text{SSB} = \sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2$$

$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2,$$

where

$$N = \sum_{i=1}^{k} n_i \quad \text{and} \quad n_0 = \frac{1}{k-1}\left[N - \sum_{i=1}^{k} n_i^2 \Big/ N\right].$$

$$\text{MSB} = \text{SSB}/(k-1), \text{ and } \quad \text{MSW} = \text{SSW}/(N-k).$$

Using the delta method, and to the first order of approximation, the variance of $\hat{\rho}_0$ [27] is given by:

$$\operatorname{var}\left(\hat{\rho}_0\right)=\frac{2(1-\rho)^2\left(1+\left(n_0-1\right)\rho\right)^2}{n_0^2\left(k-1\right)\left(1-\frac{k}{N}\right)} \tag{3.6}$$

In our data, $k = 15$, $N = 120$, and $n_0 = 7.47$.

We now start screening for the potential risk factors using bivariate correlation analysis. Table 8 provides the Pearson's correlations and the associated p-values. Significant factors (p-value < 0.05) are candidate for entry into the multivariate model. Table 9 accounts for the regional clustering effect. The clustering parameter (Intra Class Correlation Coefficient) affects the standard errors by a quantity known as the "*Design Effect*" or DEFF. The effect of the clustering is that the standard errors must be multiplied by the corresponding DEFF when we desire to construct confidence limits on the mean of the variable of interest.

It should be noted that the general definition of the "Design Effect" or DEFF is $\left(1+\left(n_0-1\right)\rho\right)^{1/2}$, that is when reporting the standard errors of the means in Tables 3-6 the standard error reported in these tables must be multiplied by the corresponding DEFF. We note that the between regions variations is quite significant for all the parameters in Table 8, except for the COVID-19 cases where the ICCC is quite low for that parameter. This means that there is a large amount of variability between countries within regions with respect to the COVID-19 cases.

Table 8. Bivariate correlations between variables in the data set.

Variable 1	Variable 2	Pearson's correlation	p-value
Log-COVID Case Fatality	Log-CKD Case Fatality (2017)	0.347	0.00001
Log-COVID Case Fatality	Log-Density (2019)	0.567	0.00001
Log-COVID Case Fatality	% of GDP spent on Healthcare (2018)	0.203	0.028
Log-COVID Case Fatality			
Log-COVID Case Fatality	Prevalence of Diabetes (2019)	0.002	0.979
COVID-19 cases	Log (pop. Density)	0.429	0.00001
COVID-19 cases	CKD count	0.508	0.00001
COVID-19 cases	% spending on healthcare	0.331	0.00001

Table 9. Regional clustering of variables of interest. DEFF = (1 + 6.47 * ICCC)$^{1/2}$.

Variable	ICCC	SE	DEFF
1. COVID-19 Cases	0.022	0.060	1.14*
2. COVID Case Fatality	0.190	0.098	1.47
3. CKD Case fatality (2017)	0.461	0.116	2
4. % of GDP spent on Healthcare (2018)	0.480	0.120	1.91
5. Diabetes (2019)	0.410	0.116	2.03

*No regional clustering.

In the next section we develop multiple regression models relating the target outcome variables; COVID-19 cases, and COVID-19 case fatality using the significant predictors.

4. Multivariate Regression Analyses

4.1. Negative Binomial Regression Model (NBRM)

From Table 9, the outcome variable, the number of COVID-19 cases does not exhibit regional clustering. That is one can safely ignore the region as a predictor. However, this variable (which is an integer variable) exhibits large amount of over dispersion, (the variance is much larger than the mean). The commonly used regression models of counts in the presence of over dispersion are constructed using the Negative Binomial Regression Model (NBRM) [28]. For this model we selected number of COVID-19 case as the dependent variable, and the three risk factors that are, in the univariate screening step for potential predictors, significantly correlated with it as shown in Table 8. It is interesting to see that in the NBRM multiple regression model the three covariates are jointly significantly associated the number of COVID-19 cases. The results are shown in Table 10. For this type of models, the "Scaled Deviance" is taken as a measure of goodness of fit of the NBRM. An optimal model should have a scaled deviance very close to unity. The scaled deviance for our model is 1.417. In our opinion this value is not substantially higher than unity, and the results based on the NBRM are quite useful. As the brilliant statistician ***George Box*** once said: *All models are wrong, some of them are useful.*

4.2. Linear Mixed Effects Modeling of COVID-19 Case Fatality

Before we proceed with the data analysis of COVID-19 Case fatality we should emphasize the hierarchical structure of the data. Basically, the data have two levels; the higher level consists of regions and the lower level consists of countries nested within regions. The results of the fitted model are summarized in Table 11 and Table 12. In Table 11 we have the results of the F-statistic to test the significance of all effects. In Table 12, we have the estimated regression coefficients.

Table 10. Results of the negative binomial regression of the aggregate number of COVID-19 cases.

Parameter	Regression coefficient estimate	Standard Error	p-value
Intercept	7.844	0.5027	0.00001
CKD-Count-2017	3.53E−8	9.9415E−9	0.00001
Percent spending on healthcare	0.198	0.0360	0.00001
Log-Population density	0.253	0.0505	0.00001

Dependent variable: COVID cases. Model: (Intercept), CKD_COUNT_2017, Percent-expenditure on healthcare.

Table 11. Dependent variable: LOG_COVID_FATALITY. The type III sums of squares.

Source		Type III Sum of Squares	df	Mean Square	F	p-value.
Intercept	Hypothesis	36.342	1	36.342	127.026	0.000
	Error	31.073	108.611	0.286		
Region	Hypothesis	11.433	14	0.817	2.952	0.001
	Error	27.669	100	0.277		
Log-population density	Hypothesis	17.269	1	17.269	62.411	0.000
	Error	27.669	100	0.277		
GDP_PER_CAPITA	Hypothesis	1.821	1	1.821	6.582	0.012
	Error	27.669	100	0.277		
CKD_CASE_FATALITY	Hypothesis	1.136	1	1.136	4.106	0.045
	Error	27.669	100	0.277		

Table 12. Dependent variable: LOG_COVID_FATALITY.

Parameter	B	Std. Error	t	p-value
Intercept	4.203	0.411	10.226	0.000
[Region = An.Latin]	1.026	0.394	2.603	0.011
[Region = C.Asia]	0.155	0.276	0.562	0.575
[Region = C.Europe]	0.899	0.246	3.655	0.000
[Region = C.Latin]	0.553	0.273	2.024	0.046
[Region = Caribean]	1.044	0.335	3.114	0.002
[Region = CESSA]	−0.048	0.264	−0.181	0.856
[Region = E.Asia]	0.143	0.274	0.522	0.603
[Region = E.Europe]	0.321	0.286	1.125	0.263
[Region = HIAP]	0.267	0.370	0.721	0.473
[Region = MENA]	0.435	0.226	1.923	0.057
[Region = S.Asia]	−0.111	0.324	−0.342	0.733
[Region = S.Latin]	0.258	0.573	0.450	0.654
[Region = SSSA]	0.339	0.358	0.948	0.345
[Region = W.Europe]	0.994	0.302	3.292	0.001
[Region = WSSA]	0^a			
Log_density	0.212	0.027	7.900	0.000
GDP_PER_CAPITA	−1.001E−5	3.904E−6	−2.565	0.012
CKD_CASE_FATALITY	0.001	0.001	2.026	0.045

All risk factors we included in the model are deemed significant, except diabetes which was not significant neither in the univariate screening stage nor in the multiple regression model. In Note that in Table 12, the region WSSA was selected as the reference category when the computer program created dummy variables for the 15 regions.

Now the model-based prediction for the random effects model the correlation between observed and predicted outcome is R = 0.765, which is quite high. Moreover, we found no association between the model-based predictions and the residuals indicating that no model assumptions have been violated.

5. Conclusions

The data that we analyzed are two-levels or hierarchical. We used R software [29] and the commercial software SPSS version 25 [30] for the data analyses. The univariate analyses showed that there is strong regional clustering of CKD, population density and percentage of GDP spending on health care. But the variability between regions in the COVID-19 cases was significantly lower than the variability with regions.

The analyses in this paper targeted two important outcome variables. The first is the aggregate number of COVID-19 cases. The NBRM found that CKD count, population density, and the percentage of GDP spent on healthcare were the most significant risk factors associated with this outcome. The second outcome variable is the COVID-19 case-fatality. The predictive analytic procedure showed that regional effect, log-population density, per-capita GDP, and CKD case fatality are the most important predictors of this outcome. The analysis showed that neither diabetes, nor the percentage of GDP spending on healthcare is significant predictors of COVID-19 case fatality.

Since the analyses in this paper are ecological, the conclusions should not be applied at the within region country level. By contrast, there may be several regional area-level environmental factors that we were not able to explore that may explain the correlation we see between regional level prevalence of health-related risk factors and case fatality of COVID-19.

Since these data did not include within country individual subjects' information on the health status of the case patients, conclusions cannot be drawn about within countries individual risk factors. However, this analysis suggests that there are important regional-level variations in COVID-19 infections that are correlated with variations in other chronic conditions, suggesting that the factors that influence health disparities may also be operating on the distribution of COVID-19.

Furthermore, there presently are limitations with overall infection count data. The analyses were conducted with data only in the first wave of the COVID-19 pandemic in the world. This fact may explain the associations observed with both CKD and population density, rather than indicators of who is being infected, and these data continue to reflect those countries with the most severe outcomes after infection. Finally, the implementation of restrictions on travel and in-person activities may also impact overall rates of COVID-19 during this first wave. It should also be noted that several new strains of the virus have emerged near the end of 2020. Therefore, new and more comprehensive data on the new variant of the COVID virus through the winter of 2021 may clarify the association with other comorbid conditions.

In summary, these analyses found that countries within the fifteen regions estimated CKD case fatality and population density were significant ecologic predictors of case fatality of COVID-19.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] COVID-19 Case Count and Deaths. https://covid19.who.int/table

[2] GBD Chronic Kidney Disease Collaboration (2020) Global, Regional, and National Burden of Chronic Kidney Disease, 1990-2017: A Systematic Analysis for the Global Burden of Disease Study 2017. *The Lancet*, **395**, 709-733.

[3] Population Density and Population Count. https://data.worldbank.org

[4] Global Diabetes Prevalence in 2019. https://www.indexmundi.com/facts/indicators/SH.STA.DIAB.ZS/rankings

[5] National Gross Domestic Product. https://www.worldmeters.info/gdb-per-capita

[6] Percentage of GDP on Health Care Spending. https://www.worldometers.info/gdp/gdp-per-capita/

[7] Zhou, Y.Z., Ren, Q.D., Chen, G., Jin, Q., Cui, Q.X., Luo, H.T., Zheng, K., Qin, Y. and Li, X.M. (2020) Chronic Kidney Diseases and Acute Kidney Injury in Patients with COVID-19: Evidence from a Meta-Analysis. *Frontiers in Medicine*, **7**, Article ID: 588301. https://doi.org/10.3389/fmed.2020.588301

[8] Williamson, E.J., *et al.* (2020) Factors Associated with COVID-19-Related Death Using OpenSAFELY. *Nature*, **584**, 430-436.

[9] Huang, C., *et al.* (2020) Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China. *The Lancet*, **395**, 497-506. https://doi.org/10.1016/S0140-6736(20)30183-5

[10] Guan, W.J., *et al.* (2020) Comorbidity and Its Impact on 1590 Patients with COVID-19 in China: A Nationwide Analysis. *European Respiratory Journal*, **55**, Article ID: 2000547. https://doi.org/10.1183/13993003.01227-2020

[11] Li, X., *et al.* (2020) Impact of Cardiovascular Disease and Cardiac Injury on In-Hospital Mortality in Patients with COVID-19: A Systematic Review and Meta-Analysis. *Heart*, **106**, 1142-1147. https://doi.org/10.1136/heartjnl-2020-317062

[12] Diao, Y.L., Kodera, S., Anzai, D., Gomez-Tames, J. and Rashed, E.A. (2021) Influence of Population Density, Temperature, and Absolute Humidity on Spread and Decay Durations of COVID-19: A Comparative Study of Scenarios in China, England, Germany, and Japan. *One Health*, **12**, Article ID: 100203. https://doi.org/10.1016/j.onehlt.2020.100203

[13] Rashed, E.A., Kodera, S., Gomez-Tames, J. and Hirata, A. (2020) Correlation between COVID-19 Morbidity and Mortality Rates in Japan and Local Population Density, Temperature, and Absolute Humidity. *International Journal of Environmental Research and Public Health*, **17**, 5447. https://doi.org/10.3390/ijerph17155477

[14] Ahmadi, M., Sharifi, A., Dorosti, S., Jafarzadeh Ghoushchi, S. and Ghanbari, N. (2020) Investigation of Effective Climatology Parameters on COVID-19 Outbreak in Iran. *Science of the Total Environment*, **729**, Article ID: 138705. https://doi.org/10.1016/j.scitotenv.2020.138705

[15] Pugliese, G., Vitale, M., Resi, V. and Orsi, E. (2020) Is Diabetes Mellitus a Risk Factor for Corona Virus Disease 19 (COVID-19)? *Acta Diabetologica*, **57**, 1275-1285. https://doi.org/10.1007/s00592-020-01586-6

[16] Palaiodimos, L., Chamorro-Pareja, N., Karamanis, D., Li, W.J., Zavras, P.D., Chang, K.M., Mathias, P. and Kokkinidis, D.G. (2020) Diabetes Is Associated with Increased Risk for In-Hospital Mortality in Patients with COVID-19: A Systematic Review and Meta-Analysis Comprising 18,506 Patients. *Hormones*. https://doi.org/10.1007/s42000-020-00246-2

[17] Zhou, B., Lu, Y., Hajifathalian, K., Bentham, J., Di Cesare, M., Danaei, G., Bixby, H., Cowan, M.J., Ali, M.K., Taddei, C. and Lo, W.C. (2016) Worldwide Trends in Diabetes since 1980: A Pooled Analysis of 751 Population-Based Studies with Million Participants. *The Lancet*, **387**, 1513-1530. https://doi.org/10.1016/S0140-6736(16)00618-8

[18] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A.A., Ogurtsova, K. and Shaw, J.E. (2019) Global and Regional Diabetes Prevalence Estimates for 2019 and Projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes Research and Clinical Practice*, **157**, Article ID: 107843. https://doi.org/10.1016/j.diabres.2019.107843

[19] Centers for Disease Control and Prevention (2020) National Diabetes Statistics Report. US Department of Health and Human Services. https://www.cdc.gov/diabetes/data/statistics-report/index.html

[20] Casqueiro, J., Casqueiro, J. and Alves, C. (2012) Infections in Patients with Diabetes Mellitus: A Review of Pathogenesis. *Indian Journal of Endocrinology and Metabolism*, **16**, S27. https://doi.org/10.4103/2230-8210.94253

[21] Hussain, A., Bhowmik, B. and do Vale Moreira, N.C. (2020) COVID-19 and Diabetes: Knowledge in Progress. *Diabetes Research and Clinical Practice*, **162**, Article ID: 108142. https://doi.org/10.1016/j.diabres.2020.108142

[22] Maddaloni, E. and Buzzetti, R. (2020) COVID-19 and Diabetes Mellitus: Unveiling the Interaction of Two Pandemics. *Diabetes/Metabolism Research and Reviews*, **36**, e33213321. https://doi.org/10.1002/dmrr.3321

[23] Angelidi, A.M., Belanger, M.J. and Mantzoros, C.S. (2020) COVID-19 and Diabetes Mellitus: what We Know, How Our Patients Should Be Treated Now, and What Should Happen Next. *Metabolism*, **107**, Article ID: 154245. https://doi.org/10.1016/j.metabol.2020.154245

[24] Global Diabetes Prevalence. https://www.indexmundi.com/facts/indicators/SH.STA.DIAB.ZS/rankings

[25] Khan, J.R., Awan, N., Islam, M.M. and Muurlink, O. (2020) Healthcare Capacity, Health Expenditure, and Civil Society as Predictors of COVID-19 Case Fatalities: A Global Analysis. *Frontier in Public Health*, **8**, 347.

[26] Korda, R.J. and Butler, J.R.G. (2006) Effect of Healthcare on Mortality: Trends in Avoidable Mortality in Australia and Comparisons with Western Europe. *Public Health*, **120**, 95-105. https://doi.org/10.1016/j.puhe.2005.07.006

[27] Shoukri, M.M. (2018) Analysis of Correlated Data with SAS and R. CRC Press, Boca Raton. https://doi.org/10.1201/9781315277738

[28] Lawless, J.F. (1987) Negative Binomial and Mixed Poisson Regression. *Canadian Journal of Statistics*, **15**, 209-225. https://doi.org/10.2307/3314912

[29] The R Project for Statistical Computing. http://www.r-project.org

[30] SPSS; IBM: Statistical Program for Social Sciences, Version 25.