Journal of Health and Social Sciences Advance Publication Online

Published Online September 30, 2020 doi 10.19204/2020/crrl6

The Italian Journal for Interdisciplinary Health and Social Development

ORIGINAL ARTICLE IN PUBLIC HEALTH

Correlation between vitamin D levels, individual and socio-demographic characteristics and COVID-19 infection and death rates in 20 European countries: A modelling study

John C. DEARDEN[1], Philip H. ROWE[2]

Affiliations:

[1] Ph.D., Emeritus Professor, School of Pharmacy & Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, U.K.
[2] Ph.D., Visiting Research Fellow, School of Pharmacy & Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, U.K.

Corresponding Author:

John C. Dearden, Emeritus Professor, School of Pharmacy & Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, U.K. E-mail: j.c.dearden@ljmu.ac.uk

Abstract

Introduction: Numerous potentially controlling demographic factors such as age, poverty, obesity, and cardiovascular and respiratory co-morbidities have been suggested, and high vitamin D levels have been found, to be associated with lower levels of COVID-19 infection. This study aimed to explore the correlation between vitamin D levels and socio-demographic characteristics with COVID-19 cases and deaths in 20 European countries.

Methods: A quantitative ecological study was designed. Multiple linear regression analysis was used to examine which of vitamin D levels and 20 demographic factors correlated well with COVID-19 cases and deaths up to 9 May 2020 in 20 European countries. Data distributions were normalised by the Box and Cox approach and the Minitab routine 'Best Subsets' was used to select the best descriptor sets for each quantitative model of cases and deaths.

Results: Cases were best modelled by vitamin D levels, stroke deaths, respiratory deaths, smoking, and human development levels. Deaths were best modelled by the number of cases, stroke deaths,

proportion of African/Afro-Caribbean people, proportion of over 65-year-olds, population density, and levels of physical inactivity. Good correlations were obtained for each model, with coefficients of determination (r^2) being around 0.7 or greater. The correlation of cases with vitamin D levels, stroke deaths, and respiratory deaths was $r^2 = 0.712$, while the correlation of deaths with population density, levels of physical inactivity, and stroke deaths was $r^2 = 0.745$.

Discussion and Conclusions: These results help to explain the variability of COVID-19 cases and deaths, and offer guidance in planning against future coronavirus pandemics. Controlling for a wide range of factors reduces the risk that the apparent protective effect of vitamin D might be confounded.

KEY WORDS: COVID-19; correlation; socio-demographic factors; European countries; modelling study; multiple linear regression; vitamin D.

Riassunto

Introduzione: Potenzialmente numerosi fattori demografici come l'età, la povertà, l'obesità e le co-morbilità cardiovascolari e respiratorie sono stati suggeriti ed alti livelli di vitamina D sono stati associati a minori livelli di infezione. L'obiettivo di questo studio è stato quello di esplorare la correlazione tra i livelli di vitamina D e le caratteristiche socio-demografiche con il numero di casi e di morti di COVID-19 in 20 Paesi europei.

Metodi: E' stato adottato uno studio ecologico di tipo quantitativo. La regressione lineare multipla è stata usata per esaminare quale tra i livelli di vitamina D e 20 fattori demografici correlavano bene con i casi e le morti per infezione da COVID-19 fino al 9 Maggio 2020 in 20 Paesi europei. La distribuzione dei dati è stata normalizzata con l'approccio Box e Cox e la Minitab di routine "Best Subsets" è stata usata per selezionare il miglior set di descrittori per ciascun modello quantitativo di casi e di morti.

Risultati: I casi erano modellati meglio dai livelli di vitamina D, dalle morti per stroke e per insufficienza respiratoria, dal fumo e dai livelli di sviluppo umano. Le morti erano meglio modellate dal numero dei casi, dalle morti per stroke, dalla proporzione di Africani/Afrocaraibici nella popolazione, dalla proporzione di anziani (> 65 anni) e dalla densità della popolazione, cosi come dai

livelli di inattività fisica. Buone correlazioni sono state ottenute per ciascun modello, con coefficienti di determinazione (r^2) che erano intorno a 0.7 o più grandi. La correlazione dei casi con i livelli di vitamina D, le morti per stroke e per insufficienza respiratoria (r^2) è risultata pari a 0.712, mentre la correlazione (r^2) delle morti con la densità di popolazione, i livelli di inattività e le morti per stroke è risultata pari a 0.745.

Discussione e Conclusioni: Questi risultati aiutano a spiegare la variabilità dei casi e delle morti per COVID-19 ed offrono una guida per la pianificazione contro future pandemie da coronavirus. Controllare un'ampia gamma di fattori riduce il rischio che l'effetto protettivo apparente della vitamina D possa essere confuso.

TAKE-HOME MESSAGE: COVID-19 infections are associated with low vitamin D levels, and are related to stroke death rates, respiratory death rates, levels of smoking, and human development levels; COVID-19 deaths are related to stroke death rates, proportions of African/Afro-Caribbean people and elderly people, population density, and physical inactivity levels. This knowledge could help in planning against such infections.

Competing interests: None declared

Cite this article as: Dearden JC, Rowe PH. Correlation between vitamin D levels, individual and socio-demographic characteristics and COVID-19 infection and deaths rates in 20 European countries: a modelling study [published online ahead of print September 30, 2020]. J Health Soc Sci. doi 10.19204/2020/crrl6.

Received: 1 Aug 2020 **Accepted:** 13 Sept 2020 **Published Online:** 30 Sept 2020

INTRODUCTION

The recent and ongoing outbreak of the viral infection COVID-19 was declared a pandemic by the World Health Organisation on 12 March 2020 [1]. Many thousands of deaths have been reported, and it is possible that lasting morbidities will persist in some recovered patients. Whilst some similarities with other coronavirus infections have been observed, some COVID-19 symptoms are worryingly different. For example, it appears to be extremely contagious [2], it affects mainly elderly people [3],

it can attack different organs in different people [4], and Black, Asian and Minority Ethnic (BAME) groups appear to be more susceptible to it [5]. There is, therefore, an urgent need to find and examine what factors control infection rates and deaths [6]. The level of immunity to viral infection is clearly important, and vitamin D3 (1,25-dihydroxyvitamin D, or calcitriol) is known to enhance immune response [7]. Daneshkhah et al [8] found that the risk of severe COVID-19 illness was 18.5% greater in patients with severe vitamin D deficiency, whilst two research groups [3,9] found negative correlations between mean vitamin D levels in European countries and COVID-19 cases/deaths in those countries. Latitude is a possible factor here, as sunlight is less intense in higher latitudes. However, vitamin D levels tend to be lower at lower latitudes, owing to deliberate decreased exposure to the sun in hot climates [3]. Zinc is also known to improve immune function, especially in the elderly [10]. As pointed out above, there are a number of demographic factors that can potentially have a role in COVID-19 infections and deaths. Older people are known to be more susceptible [8], but that could be because they are more likely to have underlying health conditions, especially cardiovascular and respiratory diseases, and also because they generally have lower levels of vitamin D [3]. Numerous other potentially controlling demographic factors such as poverty, obesity, and cardiovascular and respiratory co-morbidities, have been suggested [11–13]. Therefore, this study aims to explore the relationships between some health conditions, including vitamin D levels and socio-demographic characteristics, with COVID-19 cases and deaths in European countries.

METHODS

This quantitative ecological study was designed to see how well vitamin D levels and 20 individual and socio-demographic parameters could model COVID-19 cases and deaths, in order to subject vitamin D to the widest possible challenge, and to allow the other parameters to compete against vitamin D, to see if any could displace the vitamin from our models. Some parameters, such as the proportion aged over 65 years, were included as they were reasonable candidates for a direct causal link to the disease. Others, such as number of dementia or stroke deaths, while unlikely to be directly

causal, might act as a proxy for some factor more directly linked to the disease. The COVID-19 data used were cases and deaths up to 9 May 2020. The methodology used was that widely used in QSAR (quantitative structure-activity relationships) [14].

Data and study variables

COVID-19 cases and deaths were obtained [15] (per million inhabitants) for the twenty European countries selected by Ilie et al [3]. A number of possible factors affecting COVID-19 have been mentioned in recent media reports, and data relating to these and other factors have been obtained from various sources. Data for Turkey are for that part of the country that is in Europe, so far as could be ascertained. No data could be found for zinc levels in European populations. The possible factors investigated were: Vitamin D levels (ng/L) [3]; cardiovascular disease deaths (cvd) [16] (per 100,000 population in 2016); coronary heart disease deaths (chd) [17] (per 100,000 population, latest); stroke deaths [18] (per 100,000 population, latest); respiratory disease deaths [19] (per 100,000 population in 2016); dementia deaths [20] (per 100,000 population, latest); diabetes prevalence [21] (% of population aged 20-79 with type 1 or type 2 diabetes in 2019); obesity [22] (% of population in 2016); smokers [23] (% of population in 2020); poverty [24] (% of population with income less than half the median household income in 2019 or latest); physical activity [25] (%) of population spending at least 2½ hours per week of leisure time on physical activities in 2014; inactivity [26] (% of population not meeting a predefined level of activity (at least 150 minutes of moderate physical activity or at least 75 minutes of vigorous physical activity per week) in 2001 to 2016); vegetarianism [27] (% of population who were vegetarians in ca. 2017); alcohol consumption [28] (equivalent litres of pure alcohol per person in 2019); over 65-year-olds [29] (% of population over 65 years of age in 2016; African/Afro-Caribbean people [30] (% of population who were of African/Afro-Caribbean origin (no data were found for the Czech Republic, Estonia, Hungary and Slovakia; these countries are likely to have very small African/Afro-Caribbean populations, so an arbitrary figure of 0.01% was ascribed; for this reason the baseline was displaced by +1 to aid the normality of distribution)); population density in

2017 [31]; Human Development Index [32] (HDI is a composite index of life expectancy, education and per capita income indicators); latitude of capital cities [33] (for Turkey, the latitude of Istanbul was used rather than that of Ankara, which is in Asia, although the two cities have similar latitudes); pollution index and exponential pollution index [34] (the latter gives more weighting to cities). It should be noted that World Life Expectancy figures are compiled from the latest data from the World Health Organisation [35], the World Bank [36], UNESCO [37], and individual country databases. In addition to the above 21 variables, it would have been useful to have the proportion of Asian people in each country's population, but these data do not appear to be publicly available.

Statistical analysis

The data were analysed with Minitab statistical software, version 19.2 [38], using multiple linear regression. This technique allows a dependent variable (in this case, the number of COVID-19 cases and deaths that occurred up to 9 May 2020) to be modelled by more than one 'independent' variable. It is very rare that any pair of variables will be totally independent of each other, so QSAR modellers use a cut-off in the coefficient of determination (r^2), above which the pair are deemed too similar for both to be used in the model. There is no definitive cut-off value, but an r value of ± 0.8 ($r^2 = 0.64$) is widely used. Among the data sets investigated, several were strongly positively or negatively skewed (Table 2), with data sets containing either high or low outlier values that could have inappropriately strong influences on regression relationships. Because there were both positively and negatively skewed parameters, no single transformation was appropriate in all cases. The authors wished to avoid making individual decisions as to how to handle each variable, as this could allow subjective biases to creep in. Consequently, all the data sets were subjected to the method of Box and Cox [39] as implemented in Minitab. The Minitab implementation of Box-Cox raises each data set to a power of plus or minus 0.5, 1, 2, 3 or 5, or where appropriate uses a log transform. The power used is selected to bring a data set as close as possible to a normal distribution. In this case, transformations were restricted to nothing greater than plus or minus 3 as raising data to the fifth power seemed

excessive. The transformed data sets were analysed using the Normality Test procedure in Minitab and all closely approximated normal distributions and had Anderson-Darling test p-values well in excess of 0.05. The best subsets routine in Minitab was used to select the independent variables that best modelled the COVID-19 cases and deaths. For 20 objects (in this case countries), the *Topliss* and *Costello* rule [40] states that no more than four variables should be used, to minimise the risk of chance correlations. Whilst a good r^2 value is desirable, it has to be recognised that almost all types of data, especially *in vivo* results, include error; it is generally accepted that a model involving *in vivo* data should have an r^2 value no greater than about 0.8, otherwise it could be modelling error.

RESULTS

Table 1 shows the collected data for all the variables.

Table 1. COVID-19 cases and deaths up to 9 May 2020, and potential governing factors.

Country	Cases	Deaths	Vit D	CVD	CHD	Stroke	Resp-iratory	Dem-entia	Diabetes	Obesity	Smoking
1. Belgium	4509	741	49.3	268.8	47.3	22.0	100.6	22.0	4.6	22.1	23.3
2. Czech Rep.	754	25	62.5	569.9	124.8	35.6	80.5	11.7	7.0	26.0	33.2
3. Denmark	1755	90	65.0	248.3	38.6	23.9	116.9	24.6	8.3	19.7	17.0
4. Estonia	1299	42	51.0	381.1	161.6	29.7	43.2	4.5	4.2	21.2	33.1
5. Finland	1038	47	67.7	360.2	68.9	27.3	38.2	50.8	5.6	22.2	20.9
6. France	2054	386	60.0	197.2	31.0	16.6	57.0	19.3	4.8	21.6	27.7
7. Germany	2040	89	50.1	381.1	73.5	22.2	71.2	15.8	10.4	22.3	30.4
8. Hungary	325	40	60.6	737.5	181.8	48.7	79.4	14.8	6.9	26.4	28.4
9. Iceland	4919	27	57.0	315.1	62.6	20.0	78.5	29.4	5.8	21.9	16.1
10. Ireland	4548	285	56.4	309.0	60.6	21.3	134.0	26.8	3.2	25.3	22.2
11. Italy	3605	501	50.0	296.2	51.3	26.8	62.0	13.8	5.0	19.9	24.0
12. Netherlands	2410	307	59.5	264.4	42.9	22.9	80.7	32.9	5.4	20.4	25.1
13. Norway	1501	41	65.0	247.5	46.6	20.9	98.9	24.4	5.3	23.1	22.3
14. Portugal	2653	108	39.0	296.7	38.4	38.3	122.7	16.1	9.8	20.8	22.6
15. Slovakia	267	5	81.5	620.2	133.7	46.6	78.5	22.3	6.5	20.5	28.7
16. Spain	4732	558	42.5	237.3	38.9	19.4	92.8	21.5	6.9	23.8	29.2
17. Sweden	2444	307	73.5	318.6	58.3	21.7	63.0	28.2	4.8	20.6	20.6
18. Switzerland	3511	177	46.0	263.0	47.5	15.3	52.2	25.5	5.7	19.5	23.3
19. Turkey (Eur)	1630	44	51.8	523.7	122.6	42.2	155.2	57.6	11.1	32.1	26.0
20. U.K.	3181	470	47.4	253.3	47.6	21.6	136.4	37.6	3.9	27.8	19.2

	Poverty	Inactive	Exer-cise	Veget-arian	Alco-hol	≥65	%Afro	Popn Density	HDI	Lat N	Pollution Index	Exp. Index
1.	10.2	42.7	24.0	7	15.9	18.2	6.99	337	91.9	50.85	52.94	90.05

2.	5.6	25.0	28.4	2	19.1	18.3	0.01	130	89.1	50.08	39.99	65.48
3	5.8	35.1	54.6	5	13.9	18.8	0.05	125	93.0	55.68	21.33	32.87
4	15.8	17.2	23.2	1.8	15.9	19.0	0.01	28	88.2	59.37	19.81	33.12
5.	6.3	37.8	54.6	4	14.8	20.5	1.09	16	92.5	60.25	11.55	19.66
6.	8.1	32.5	25.0	5	16.7	18.8	7.45	104	89.1	48.83	43.56	72.66
7.	10.4	28.0	48.3	10	16.9	21.1	0.72	233	93.9	52.50	29.03	46.68
8.	8.0	26.0	28.6	1	17.1	18.3	0.01	108	84.5	47.48	48.29	83.58
9.	5.4	18.2	60.8	3	12.8	13.9	0.98	3	93.8	64.17	16.21	28.12
10.	9.0	53.2	29.1	4.3	16.0	13.2	1.42	65	94.2	53.35	33.99	56.48
11.	13.9	54.7	18.2	8.6	12.0	22.0	1.99	192	88.3	41.90	55.63	94.51
12.	8.3	18.2	25.0	5	12.0	18.2	5.17	393	93.3	52.38	27.41	44.83
13.	8.4	44.2	56.8	4	9.4	16.4	2.47	13	95.4	59.92	20.35	32.66
14.	10.7	51.0	18.4	1.2	17.8	20.7	1.46	109	85.0	38.70	30.89	49.68
15.	8.5	22.2	29.4	2	16.6	14.4	0.01	111	85.7	48.17	39.66	66.56
16.	14.8	50.2	34.0	1.5	14.6	18.7	1.18	92	89.3	40.42	39.99	65.48
17.	9.3	44.2	54.1	10	12.5	19.8	1.84	20	93.7	59.33	18.09	28.24
18.	9.1	23.7	21.8	14	14.2	18.0	1.76	207	94.6	46.95	22.39	36.03
19.	17.2	56.0	4.7	4	2.4	9.1	0.10	102	80.6	41.02	67.35	117.28
20.	11.9	63.3	36.7	7	15.6	17.9	5.27	267	92.0	51.60	40.56	67.05

The best transformations found using the Box and Cox approach are shown in Table 2.

Table 2. Transformation of cases, deaths and predictor variables using Box and Cox approach.

Variable (X)	**Skewness**	**Transformation from X**	**Transformation code in eqns.**
Cases/million	0.29	Square root(X)	Tr1(X)
Deaths/million	1.02	Log(X)	Tr2(X)
Vitamin D level	0.51	Log(X)	Tr2(X)
CVD deaths/100,000	1.49	1/X	Tr3(X)
CHD deaths/100,000	1.33	1/X	Tr3(X)
Stroke deaths/100,000	1.10	1/X	Tr3(X)
Respiratory deaths/100,000	0.53	Log(X)	Tr2(X)
Dementia deaths/100,000	1.14	Square root(X)	Tr1(X)
% Diabetic	0.99	1/Square root(X)	Tr4(X)
% Obese	1.49	$1/X^3$	Tr5(X)
% Smokers	0.12	Square root(X)	Tr1(X)
% In poverty	0.78	1/Square root(X)	Tr4(X)

% Inactive	0.16	X	Tr0(X)
% Exercising	0.36	Square root(X)	Tr1(X)
% Vegetarian	1.06	Log(X)	Tr2(X)
% ≥ 65 years	-1.36	X^3	Tr6(X)
% Afro + 1	1.39	1/Square root(X)	Tr4(X)
Alcohol (litres/year)	-1.94	X^2	Tr7(X)
Population density	0.98	Square root(X)	Tr1(X)
HDI	-0.88	X^3	Tr6(X)
Latitude	-0.05	X	Tr0(X)
Pollution index	0.51	Log(X)	Tr2(X)
Exponential pollution index	0.63	Log(X)	Tr2(X)

Correlations between the COVID-19 cases/deaths and each of the 21 independent variables give an indication of the contribution that each can make; the r^2 values are shown in Table 3. The r values are also shown, so that the sign of the contribution that each makes can be seen. It should be noted that r^2 represents the fraction of the variation in the dependent variable that is explained by that variable; for example, for SqRtCOVID-19 cases vs. log vitamin D levels, the r^2 value of 0.349 means that transformed vitamin D levels explain 34.9% of the variation in the transformed number of COVID-19 cases.

Table 3. Correlation of COVID-19 SqRt (Cases) and Log (Deaths) with single transformed independent variables.

	SqRt Cases		**Log Deaths**	
Variable	**r^2**	**r**	**r^2**	**r**
SqRt (Cases)	1.000	1.000	0.521	0.722
Log (Deaths)	0.521	0.722	1.000	1.000
Log (Vitamin D levels)	0.349	-0.570	0.248	-0.498
1/(Cardiovascular deaths)	0.405	0.636	0.495	0.704

1/(Coronary heart disease deaths)	0.326	0.570	0.453	0.673
1/(Stroke deaths)	0.438	0.662	0.344	0.586
Log (Respiratory deaths)	0.052	0.229	0.013	0.115
SqRt (Dementia deaths)	0.018	0.134	0.0029	0.054
1/SqRt (% Diabetic)	0.113	0.336	0.162	0.402
$1/(\%\ \text{Obese})^3$	0.013	0.113	0.0082	0.091
SqRt (% Smokers)	0.199	-0.446	0.040	-0.199
1/SqRt (% Poverty)	0.049	-0.221	0.140	-0.374
% Inactive	0.150	0.387	0.256	0.505
SqRt (% Exercising)	0.0002	0.014	0.016	-0.126
Log (% Vegetarians)	0.166	0.407	0.230	0.480
$(\% \geq 65\ \text{years})^3$	0.0001	0.010	0.130	0.361
1/SqRt (% Afro + 1)	0.379	-0.616	0.569	-0.754
$(\text{Alcohol consumption})^2$	0.046	-0.214	0.0063	-0.079
SqRt (Population density)	0.021	0.146	0.206	0.454
$(\text{HDI})^3$	0.184	0.429	0.088	0.296
Latitude	0.0033	-0.057	0.058	-0.241
Log (Pollution index)	0.0001	0.010	0.055	0.234
Log (Exponential pollution index)	0.000	0.000	0.044	0.209

For COVID-19 cases, the best models were found using three variables:

Tr1(Cases) = 261.6(54.8) – 105.0(25.4) Tr2(Vitamin D) + 581(174) Tr3(Stroke deaths)

– 11.08(4.28) Tr1(%Smokers) (1)

n = 20 $r^2 = 0.758$ $q^2 = 0.630$ s = 8.630 $F = 16.7$ All p values ≤ 0.02

where numbers in brackets are the standard errors on each of the coefficients, n = number of countries, q^2 = cross-validated r^2 (a measure of internal validation of the model), s = standard error of the model

prediction, *F* = variance ratio or Fisher coefficient (a measure of goodness of fit), and p = probability that random sampling error might lead to an apparent effect in the absence of any real relationship; a p value of 0.02 means a 2% risk of such an event.

It is important to stress that more than one good model can sometimes be found from the same data. Note that 'good' means 'with good statistics'. In the present case, best subsets analysis also found two other good three-variable models for COVID-19 cases:

$$\text{Tr1(Cases)} = 108.1(65.7) - 80.7(28.6)\ \text{Tr2(Vitamin D)} + 817(184)\ \text{Tr3(Stroke deaths)} + 24.3(13.8)\ \text{Tr2(Respiratory deaths)} \quad (2)$$

$n = 20 \quad r^2 = 0.712 \quad q^2 = 0.584 \quad s = 9.408 \quad F = 13.2 \quad$ All p values ≤ 0.097

$$\text{Tr1(Cases)} = 285.0(59.8) - 133.5(27.8)\ \text{Tr2(Vitamin D)} - 10.42(5.05)\ \text{Tr1(\% Smokers)} + 0.000063(0.000026)\ \text{Tr6(HDI)}^3 \quad (3)$$

$n = 20 \quad r^2 = 0.698 \quad q^2 = 0.533 \quad s = 9.631 \quad F = 12.3 \quad$ All p values ≤ 0.056

For COVID-19 deaths, best subsets analysis found three good, three-term models, with very similar statistics.

$$\text{Tr2 (Deaths)} = 1.599(0.583) + 0.0167(0.0058)\ \text{Tr1(Cases)} - 1.050(0.420)\ \text{Tr4(\%Afro + 1)} + 0.000064(0.000030)\ \text{Tr6(\% } \geq 65 \text{ years)}^3 \quad (4)$$

$n = 20 \quad r^2 = 0.745 \quad q^2 = 0.543 \quad s = 0.319 \quad F = 15.6 \quad$ All p values ≤ 0.053

$$\text{Tr2(Deaths)} = -0.206(0.338) + 0.0453(0.0148)\ \text{Tr1(Population density)} + 0.0190(0.0051)\ \text{Tr0(\% Inactive)} + 26.15(5.92)\ \text{Tr3(Stroke deaths)} \quad (5)$$

$n = 20 \quad r^2 = 0.745 \quad q^2 = 0.573 \quad s = 0.319 \quad F = 15.6 \quad$ All p values ≤ 0.007

Tr2 (Deaths) = 1.692(0.580)+ 0.0157(0.0059) Tr1(Cases)

+ 0.0303(0.0158) Tr1(Population density) – 1.019(0.434) Tr4(Afro + 1) (6)

n = 20 $r^2 = 0.736$ $q^2 = 0.493$ s = 0.324 $F = 14.9$ All p values ≤ 0.074

In each of the above equations, correlations among the independent variables are very low ($r^2 \leq 0.379$), meaning that there is no significant collinearity between them in each equation.

DISCUSSION

It can be seen from Table 3 that the strength of correlations between cases/deaths and individual potential contributory factors varies widely. Some factors that have been postulated, for example in the media, to be significant, such as obesity, have only weak correlations. That could be interpreted as indicating that obesity is a proxy for physical inactivity, diabetes, and/or cardiovascular diseases. Some factors, such as the proportion of elderly, and physical inactivity, correlate more strongly with deaths than with cases, which is to be expected.

Correlation of cases with potential contributory factors

The log (vitamin D) term in equation 1 has a negative sign, meaning that COVID-19 cases are lower in countries with high vitamin D levels, consistent with the known relationship between vitamin D and COVID-19 infection levels (2-4). The sign on the 1/(Stroke deaths) term is positive, indicating that stroke deaths are inversely correlated with COVID-19 cases. It was not anticipated that a factor such as the proportion with a previous stroke would be causally linked to COVID-19, so its emergence as a significant predictor, may arise because it acts as a proxy for some other societal factor. It is noticeable that infections are relatively low in eastern European countries such as Slovakia and Hungary, whilst stroke deaths are high, which could explain the inverse correlation. Table 3 shows that the proportion of smokers in a population correlates positively with both cases and deaths, which would be expected intuitively. However, when other factors are controlled for, as in linear regression equations 1 and 3, the Tr1(%Smokers) term has a negative sign, which is surprising. Nevertheless, the finding is supported by recent work by Hippisley-Cox et al [41], who reported that smokers were less likely to

contract COVID-19, and had 88% lower risk of being taken into Intensive Care Units, compared with non-smokers. A similar finding was also reported by Williamson et al [5].

The sign on the log (Respiratory deaths) term in equation 2 is positive, indicating that people with respiratory problems are more likely to succumb to COVID-19 infection, which is itself primarily a respiratory disease [29]. The positive sign on the $(HDI)^3$ term indicates that COVID-19 infections increase with a high HDI (in effect a measure of standard of living), whereas it might be expected that countries with a high HDI would have fewer COVID-19 infections. The range of HDI values in this study is very narrow (80.6 – 95.4; see Table 1), which means that the prediction error for equation 3 could cause a reversal of sign of the $(HDI)^3$ term. To check this, the correlation between SqRt (Cases) on 27 July 2020 [42] and $(HDI)^3$ [43] for 126 countries, with an HDI range of 45.2 – 95.3, was determined. The correlation was still positive, with $r^2 = 0.219$ [(very close to that of 0.184 for the 20 European countries (see Table 2)], indicating that the effect was real. An examination of Table 1 shows that European countries with lower HDI values generally had lower COVID-19 cases up to 9 May 2020. It is important to note that all three of the best fitting, three-term models for cases retained vitamin D as a predictor. No other factor, or combination of factors, was able to displace it.

Correlation of deaths with potential contributory factors

Clearly a large number of cases would be expected to lead to a large number of deaths, and the correlation between SqRt (Cases) and log (Deaths) is the highest of all SqRt (Cases) correlations (Table 3). The inclusion of the 1/SqRt (% Afro + 1) term is to be expected, given the widely known susceptibility of BAME people to the coronavirus [5], although it is perhaps surprising that this term did not appear in the SqRt (Cases) correlations. The well-known vulnerability of older people to death from COVID-19 [44] is confirmed by the selection of the $(\% \geq 65 \text{ years})^3$ term in equation 4.

The selection of SqRt (Population density) in equation 5 indicates that proximity of people may be a key factor in COVID-19 deaths. It could have been expected that it should be important in causing COVID-19 cases also, but Table 3 shows that it correlates much more strongly with deaths. That may

reflect the fact that people living in crowded conditions have more co-morbidities [45]. Inactivity is another key factor in COVID-19 deaths, and as with population density, it correlates more strongly with deaths than with cases (see Table 3), again perhaps reflecting more co-morbidities. The inclusion of 1/(Stroke deaths) in equation 5 is perhaps surprising, since it correlates less strongly with log(Deaths) than do 1/(Cardiovascular deaths) and 1/(Coronary heart disease deaths). Nevertheless, replacement of the stroke deaths term by either the cardiovascular disease deaths or the coronary heart disease deaths term reduces the goodness of fit of equation 5. The positive sign on the 1/(Stroke deaths) term can be explained in the same way as was done for equations 1 and 2.

The significance of each of the selected variables in equation 6 has already been explored above.

Despite obesity being considered a risk for COVID-19 infection and death [46], obesity does not appear as a significant variable in any of the equations, and the correlation of obesity with either COVID-19 cases or deaths is very weak (Table 3). That is surprising, but may be because obesity *per se* may not be a significant factor, but could be in effect a proxy for inactivity and cardiovascular and respiratory problems.

Study limitations

Although great care was taken in the design and undertaking of the study, the very newness of COVID-19 has given rise to some potential problems. A particular focus of the current study was the risk that there might be no direct link between vitamin D levels and COVID-19; that is, there could be confounding. Those nations with low vitamin D levels might also have some other characteristic that was more directly related to the disease. Secondly, there are differences between the ways that different countries record the numbers of COVID-19 cases and deaths [47]. There are also possible differences between countries in the measurement of the socio-demographic factors that constitute the independent variables used in this study, for example in the determination of causes of death [48]. Although Asian people have been reported as having higher susceptibility to the coronavirus [49], figures were not available for the proportions of Asian people in many of the 20 European countries studied here, so

their susceptibility to the coronavirus could not be determined. Finally, although a large number of socio-demographic factors were examined in this study, there may be unknown confounding factors that could change the results reported here.

CONCLUSIONS

Analysis of a large number of factors that could potentially affect both the infection rate of, and mortality from, COVID-19 has quantitatively highlighted several that are of high significance. For infection levels, best subsets analysis identified vitamin D levels, the Human Development Index, level of stroke deaths, level of respiratory disease deaths, and proportions of smokers as important. For COVID-19 deaths, the number of COVID-19 cases, proportion of African and Afro-Caribbean people, proportion of over-65-year-olds, population density, level of inactivity, and level of stroke deaths were the key factors selected. The key variables identified in this study as affecting infection by and deaths from COVID-19 are, on the whole, broad-brush factors. Nonetheless they offer some important pointers regarding the determination of the spread and lethality of the disease. The study thus throws light on the behaviour of the coronavirus, and so should help in the battle to control and reduce its effects. The authors recognise that community-level epidemiological studies of this type could never demonstrate that low vitamin D levels directly cause a greater risk of COVID-19 infection, but the fact that this factor has withstood competition from such a wide range of other parameters usefully reduces the risk that the apparent relationship is mere confounding. Vitamin D levels appear to carry unique predictive information that none of the other factors can provide.

Acknowledgments

The authors thank Ms. Ruth Bibby for assistance with data collection.

References

1. WHO declares Covid-19 a pandemic. 2020 [cited 2020 Aug 31]. Available from:https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-a-pandemic.

2. Van Doremalen N, Bushmaker T, Morris DH, Holbrook MG, Gamble A, Williamson BN, et al. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. N Engl J Med. 2020;382:1564–1567. Available from: https://doi.org/10.1056/NEJMe2004973.
3. Ilie PC, Stefanescu S, Smith L. The role of vitamin D in the prevention of coronavirus disease-2019 infection and mortality. Aging Clin Exp Res. 2020;32(7):1195–1198. https://doi.org/10.1007/s40520-020-01570-8.
4. Lukiw W, Pogue A, Hill J. SARS-CoV-2 infectivity and neurological targets in the brain. Cell Mol Neurobiol. 2020. https://link.springer.com/article/10.1007/s10571-020-00947-7.
5. Williamson E, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. OpenSAFELY: factors associated with COVID-19-related hospital death in the linked electronic hospital records of 17 million adult NHS patients. medRχiv. 2020. http://doi.org/10.1101/2020.05.06.20092999.
6. Chirico F, Magnavita N. Covid-19 infection in Italy: An occupational injury. S Afr Med J. 2020; published online 8 May 2020. https://doi.org/10.7196/SAMJ.2020.v110i6.14855.
7. Aranow C. Vitamin D and the immune system. J Investig Med. 2011;59(6):881–886. https://doi.org/10.231/JIM.0b013e31821b8755.
8. Daneshkhah A, Eshein A, Subramanian H, Roy HK, Backman V. The role of vitamin D in suppressing cytokine storm in COVID-19 patients and associated mortality. medRχiv. 2020. https://doi.org/10.1101/2020.04.08.20058578.
9. Laird E, Rhodes J, Kenny RA. Vitamin D and inflammation: potential implications for severity of Covid-19. Ir Med J. 2020;113(5):81–88.
10. Andriollo-Sanchez M, Hininger-Favier I, Meunier N, Toti E, Zaccaria M, Brandolini-Bundon M, et al. Zinc intake and status in middle-aged and older European subjects: the ZENITH study. Eur J Clin Nutrition. 2005;59:Suppl 2,S37–S41. https://doi.org/10.1038/sj.ejcn.1602296.

11. Liu T, Liang W, Zhong H, He J, Chen Z, He G, et al. Risk factors associated with Covid-19 infection: a retrospective cohort study based on contacts tracing. Emerg Microbes Infect. 2020;9(1):1546–1553. https://doi.org/10.1080/22221751.2020.1787799.
12. Li AY, Hannah CT, Durbin J, Dreher N, McAuley FM, Marayati NF, et al. Multivariate analysis of black race and environmental temperature on COVID-19 in the US. Am J Med Sci. 2020;Jun 20;S0002-9629(20)30257-3. https://doi:10.1016/j.amjms.2020.06.015.
13. Menzak D, Menzak A, Analysis of environmental, economic, and demographic factors affecting COVID-19 transmission and associated deaths in the U.S.A. Social Science Research Network. 6 July 2020 [cited 2020 Aug 31]. Available from: http://dx.doi.org/10.2139/ssrn.3644677.
14. Dearden JC, Cronin MTD, Kaiser KLE. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). SAR QSAR Environ Res. 2009;20(3-4):241–266. http://dx.doi.org/10.1080/10629360902949567.
15. COVID-19 cases and deaths (Pandemic by Country and Territory – Wikipedia) [cited 2020 May 10]. Available from: https://en.wikipedia.org/wiki/COVID-19_pandemic_by_country_and_territory.
16. Townsend N, Wilson L, Bhatnagar P, Wickramsinghe K, Rayner M, Nichols M. Cardiovascular disease in Europe: epidemiological update 2016. Eur Heart J. 2016;37(42):3232–3245. https://doi.org/10.1093/eurheart.ehw334.
17. Coronary heart disease deaths [cited 2020 May 20]. Available from: https://worldlifeexpectancy.com/cause-of-death/coronary-heart-disease/by-country/.
18. Stroke deaths [cited 2020 May 20]. Available from: https://www.worldlifeexpectancy.com/cause-of-death/stroke/by-country/.
19. Respiratory disease deaths [cited 2020 May 15]. Available from: https://ec.europa.ec/eurostat/statistics-explained/index.php?title=File:Causes-of-

death_diseases_of_the_respiratory_system_residents_2016_Health2019.png. Accessed 15 May 2020.

20. Dementia deaths [cited 2020 May 20]. Available from: https://www.worldlifeexpectancy.com/cause-of-death/alzheimers-dementia/by-country/.
21. Diabetes prevalence [cited 2020 May 12] Available from: https://www.indexmundi.com/facts/indicators/SH.STA.DIAB.ZS/rankings.
22. Obesity [cited 2020 May 14] Available from: https://who.int/gho/ncd/risk-factors/overweight-obesity/obesity_adults/en/.
23. Smokers [cited 2020 May 14]. Available from: https://worldpopulationreview.com/countries/smoking-rates-by-country/.
24. Poverty [cited 2020 May 14]. Available from: https://data.oecd.org/inequality/poverty-rate.htm.
25. Physical activity [cited 2020 May 15]. Available from: https://ec.europa.eu/eurostat/web/products_eurostat-news/-/DDN-20170402-1.
26. Guthold R, Stevens GA, Riley M, Bull FC. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1.9 million participants. Lancet Global Health. 2018;6(10):e1077–e1086. http://dx.doi.org/10.1016/S2214-109X(18)30357-7.
27. Vegetarianism [cited 2020 May 14]. Available from: https://en.wikipedia.org/wiki/Vegetarianism-by-country.
28. Alcohol consumption [cited 2020 May 16]. Available from: https://worldpopulationreview.com/country-rankings/alcohol-consumption-by-country.
29. Over-65 year-olds [cited 2020 May 14]. Available from: https://ec.europa.eu/eurostat/cache/infographs/elderly/index.html.

30. African/Afro-Caribbean people [cited 2020 May 18]. Available from: https://en.wikipedia.org/wiki/Afro-European.
31. Population density [cited 2020 May 14]. Available from: https://www.worldatlas.com/articles/european-countries-by-population-density.html.
32. Human Development Index. United Nations Development Programme: Human Development Report 2019 [cited 2020 May 14]. Available from: https://hdr.undp.org/en/content/2019-human-development-index-ranking.
33. Latitude [cited 2020 May 18]. Available from: https://www.csgnetwork.com/llinotable.html.
34. Pollution index [cited 2020 May 18]. Available from: https://www.numbeo.com/pollution_rankings_by_country.jsp.
35. WHO COVID-19 data [cited 2020 Sep 04]. Available from: https://covid19.who,int/.
36. World Bank COVID-19 data [cited 2020 Sep 04]. Available from: www.datatopics.worldbank.org/universal-health-coverage/coronavirus/.
37. UNESCO COVID-19 data [cited 2020 Sep 04]. Available from: www.en.unesco.org/covid19/communicationinformation response/opensolutions.
38. Minitab [cited 2020 Sep 04]. Available from:https://minitab.com.
39. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc Ser B Stat Methodol. 1964;26(2):211–252. https://doi.org/10.2307/2984418.
40. Topliss JG, Costello RJ. Chance correlations in structure-activity studies using multiple linear regression. J Med Chem. 1972;15(10):1066–1068. https://doi.org/10.1021/jm00280a017.
41. Hippisley-Cox J, Young D, Coupland C, Channon KM, Tan PS, Harrison DA, et al. Risk of severe COVID-19 disease with ACE inhibitors and angiotensin receptor blockers: cohort study including 8.3 million people. Heart. 2020;0:1–9. https://doi.org/10.1136/heartjnl-2020-317393.

42. Johns Hopkins University Coronavirus Tracker [cited 2020 Jul 27]. Available from: https://www.covidtracker.com.

43. World Population Review: Human Development Index by Country 2020 [cited 2020 Jul 27]. Available from: https://worldpopulationreview/country-rankings/hdi-by-country.

44. Mueller AL, McNamara MS, Sinclair DA. Why does COVID-19 disproportionately affect older people? Aging 2020;12(10):9959–9981. https://doi.org/10.18632/aging.103344.

45. Housing for Health [cited 2020 Sep 04]. Available from: www.housingforhealth.com/the-guide/health housing/reducing-the-negative-impacts-of-crowding/.

46. Anderson MR, Gelens J, Anderson DR, Zucker J, Nobel YR, Freedberg D, et al. Body mass index and risk for intubation or death in SARS-CoV-2 infection. Ann Intern Med. 2020: published on-line 29 July. https://doi.org/10.7326/M-20-3214.

47. WHO Commentary: Estimating mortality from COVID-19 [cited 2020 Sep 02]. Available from: https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19.

48. Rampatige R, Mikkelsen L, Hernandez B, Riley I, Lopez AD. Systematic review of statistics on causes of deaths in hospitals: strengthening the evidence for policy-makers. Bull World Health Organ. 2014;92(11):807–816. https://doi.org/10.2471/BLT.14.137935.

49. Public Health England. COVID-19: understanding the impact on BAME communities. Published 16 June 2020 [cited 2020 Sep 04] Available from: https://www.gov.uk/government/publications/covid-19-understanding-the-impact-on-bame-communities.