PATHOGENS AND GLOBAL HEALTH
2022, VOL. 116, NO. 3, 146–177
https://doi.org/10.1080/20477724.2021.1993676

Taylor & Francis
Taylor & Francis Group

COVID-19 transmission risk factors

Alessio Notari[a] and Giorgio Torrieri[b]

[a]Departament de Física Quàntica i Astrofisíca & Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona, Barcelona, Spain; [b]Instituto de Fisica Gleb Wataghin (IFGW), Campinas, SP, Brazil

ABSTRACT

We analyze risk factors correlated with the initial transmission growth rate of the recent COVID-19 pandemic in different countries. The number of cases follows in its early stages an almost exponential expansion; we chose as a starting point in each country the first day d_i with 30 cases and we fitted for 12 days, capturing thus the early exponential growth. We looked then for linear correlations of the exponents a with other variables, for a sample of 126 countries. We find a positive correlation, *i.e. faster spread of COVID-19*, with high confidence level with the following variables, with respective p-value: low Temperature ($4 \cdot 10^{-7}$), high ratio of old vs. working-age people ($3 \cdot 10^{-6}$), life expectancy ($8 \cdot 10^{-6}$), number of international tourists ($1 \cdot 10^{-5}$), earlier epidemic starting date d_i ($2 \cdot 10^{-5}$), high level of physical contact in greeting habits ($6 \cdot 10^{-5}$), lung cancer prevalence ($6 \cdot 10^{-5}$), obesity in males ($1 \cdot 10^{-4}$), share of population in urban areas ($2 \cdot 10^{-4}$), cancer prevalence ($3 \cdot 10^{-4}$), alcohol consumption (0.0019), daily smoking prevalence (0.0036), and UV index (0.004, 73 countries). We also find a correlation with low Vitamin D serum levels ($0.002 - 0.006$), but on a smaller sample, $\sim$50 countries, to be confirmed on a larger sample. There is highly significant correlation also with blood types: positive correlation with types RH- ($3 \cdot 10^{-5}$) and A+ ($3 \cdot 10^{-3}$), negative correlation with B+ ($2 \cdot 10^{-4}$). We also find positive correlation with moderate confidence level (p-value of 0.02$\sim$0.03) with: CO_2/SO emissions, type-1 diabetes in children, low vaccination coverage for Tuberculosis (BCG). Several of the above variables are correlated with each other, and so they are likely to have common interpretations. We thus performed a Principal Component Analysis, to find the significant independent linear combinations of such variables. The variables with loadings of at least 0.3 on the significant PCA are: greeting habits, urbanization, epidemic starting date, number of international tourists, temperature, lung cancer, smoking, and obesity in males. We also analyzed the possible existence of a bias: countries with low GDP-per capita might have less intense testing, and we discuss correlation with the above variables.

KEYWORDS

COVID-19; statistical data; epidemiology; risk factors

I. Introduction

The recent coronavirus (COVID-19) pandemic is now spreading essentially everywhere in our planet. The growth rate of the contagion has, however, a very high variability among different countries, even in its very early stages, when government intervention is still almost negligible. Any factor contributing to a faster or slower spread needs to be identified and understood with the highest degree of scrutiny. In [1] the early, growth rate of the contagion has been found to be correlated at high significance with temperature T. In this work, we extend a similar analysis to many other variables. This correlational study could help further investigation to establish causal factors, and it can help policy makers in their decisions. See also [2–4] for relevant control measures on the epidemic.

Some factors are intuitive and have been found in other studies, such as temperature [1,5–11] (see also [12] for a different conclusion) and air travel [12,13]; we aim here at being more exhaustive and at finding also factors which are not 'obvious' and have a potential biological origin, or correlation with one. As discussed in detail in the Conclusions section, a few of the main driving factors that we will discuss have been examined in the literature previously, including by mechanistic models.

The paper is organized as follows. In section II, we explain our methods; in section III, we show our main results; in section IV, we show the detailed results for each individual variable of our analysis; in section V, we discuss correlations among variables; and in section VI, we draw our conclusions.

II. Method

As in [1], we use the empirical observation that the number of COVID-19 positive cases follows a common pattern in the majority of countries: once the number of confirmed cases reaches order 10 there is a very rapid growth, which is typically well approximated for a few weeks by an exponential. Subsequently, the exponential growth typically gradually slows down, probably due to

CONTACT Alessio Notari notari@fqa.ub.edu Departament de Física Quàntica i Astrofisíca & Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona, Martí i Franquès 1, Barcelona 08028, Spain

other effects, such as: lockdown policies from governments, a higher degree of awareness in the population or the tracking, and isolation of the positive cases. The growth is then typically stopped and reaches a peak in countries with a strong lockdown/tracking policy.

Our aim is to find which factors correlate with the speed of contagion, in its first stage of *free* propagation. For this purpose we analyzed a datasets of 126 countries taken from [14] on 15 April 2020. We have chosen our sample using the following rules:

- We start analyzing data from the first day d_i in which the number of cases in a given country reaches a reference number N_i, which we choose to be $N_i = 30$[1].
- We include only countries with at least 12 days of data, after this starting point.
- We excluded countries with too small total population (less than 300 thousands inhabitants).

We then fit the data for each country with a simple exponential curve $N(t) = N_0 e^{\alpha t}$, with 2 parameters, N_0 and α; here t is in units of days.

Note that the statistical errors on the exponents α are typically only a few percent of the spread of the values of α among the various countries. For this reason, we disregarded statistical errors on α. The analysis was done using the software *Mathematica*, from Wolfram Research, Inc.

III. Main results

We first look for correlations with several individual variables. Most variables are taken from [15], while for a few of them have been collected from other sources, as commented below.

1. Non-significant variables

We find *no* significant correlation of the COVID-19 transmission in our set of countries with many variables, including the following ones:

(1) Number of inhabitants;
(2) Asthma-prevalence;
(3) Participation time in leisure, social and associative life per day;
(4) Population density;
(5) Average precipitation per year;
(6) Vaccinations coverage for: Polio, Diphteria, Tetanus, Pertussis, Hepatitis B;
(7) Share of men with high-blood-pressure;
(8) Diabetes prevalence (type 1 and 2, together); and
(9) Air pollution ('Suspended particulate matter (SPM), in micrograms per cubic metre').

2. Significant variables, strong evidence

We find *strong* evidence for correlation with:

(1) Temperature (negative correlation, p-value $4.4 \cdot 10^{-7}$);
(2) Old-age dependency ratio: ratio of the number of people older than 64 relative to the number of people in the working-age (15–64 years) (positive correlation, p-value $3.3 \cdot 10^{-6}$);
(3) Life expectancy (positive correlation p-value $8.1 \cdot 10^{-6}$);
(4) International tourism: number of arrivals (positive correlation p-value $9.6 \cdot 10^{-6}$);
(5) Starting day d_i of the epidemic (negative correlation, p-value $1.7 \cdot 10^{-5}$);
(6) Amount of contact in greeting habits (positive correlation, p-value $5.0 \cdot 10^{-5}$);
(7) Lung cancer death rates (positive correlation, p-value $6.3 \cdot 10^{-5}$);
(8) Obesity in males (positive correlation, p-value $1.2 \cdot 10^{-4}$);
(9) Share of population in urban areas (positive correlation, p-value $1.7 \cdot 10^{-4}$);
(10) Share of population with cancer (positive correlation, p-value $2.8 \cdot 10^{-4}$);
(11) Alcohol consumption (positive correlation, p-value 0.0019);
(12) Daily smoking prevalence (positive correlation, p-value 0.0036);
(13) UV index (negative correlation, p-value 0.004; smaller sample, 73 countries); and
(14) Vitamin D serum levels (negative correlation, annual values p-value 0.006, seasonal values 0.002; smaller sample, ~50 countries).

3. Blood types, strong evidence

We also find strong evidence for correlation with blood types:

(1) RH + blood group system (negative correlation, p-value $3 \cdot 10^{-5}$);
(2) A+ (positive correlation, p-value $3 \cdot 10^{-3}$);
(3) B + (negative correlation, p-value $2 \cdot 10^{-4}$);
(4) A– (positive correlation, p-value $3 \cdot 10^{-5}$);
(5) 0- (positive correlation, p-value $8 \cdot 10^{-4}$); and
(6) AB– (positive correlation, p-value 0.028).

We find moderate evidence for correlation with:

(1) B– (positive correlation, p-value 0.013).

[1] In practice, we choose, as the first day, the one in which the number of cases N_i is closest to 30. In some countries, such a number N_i is repeated for several days; in such cases we choose the last of such days as the starting point. For the particular case of China, we started from January 16th, with 59 cases, since the number before that day was essentially frozen.

We find no significant correlation with:

(1) 0+;
(2) AB+.

4. Significant variables, moderate evidence

We find moderate evidence for correlation with:

(1) CO_2 (and SO) emissions (positive correlation, p-value 0.015);
(2) Type-1 diabetes in children (positive correlation, p-value 0.0008); and
(3) Vaccination coverage for Tuberculosis (BCG) (negative correlation, p-value 0.028).

5. Significant variables, counterintuitive

Counterintuitively we also find correlations in a direction opposite to a naive expectation:

(1) Death-rate-from-air-pollution (negative correlation, p-value $3.5 \cdot 10^{-5}$);
(2) Prevalence of anemia, adults and children, (negative correlation, p-value $1.4 \cdot 10^{-4}$ and $7. \times 10^{-6}$, respectively);
(3) Share of women with high-blood-pressure (negative correlation, p-value $1.6 \cdot 10^{-4}$);
(4) Incidence of Hepatitis B (negative correlation, p-value $2.4 \cdot 10^{-4}$); and
(5) PM2.5 air pollution (negative correlation, p-value 0.029).

We performed also non-linear fits with a quadratic function of α for each variable and used the Akaike Information Criterion (AIC) for model comparison, $\Delta\text{AIC} \equiv 2\Delta k - 2\Delta \ln(\mathcal{L})$, where Δk is the increase in the number of parameters, compared to the simple linear model and $\Delta \ln(\mathcal{L})$ is the change in the maximum log-likelihood between the two models. Only for a few variables, the ΔAIC is slightly in favor (i.e. ΔAIC less than -2) of the quadratic model: $\Delta\text{AIC} = -4, -2.7, -2.6, -3.8, -2.4, -3.5, -2.4$ for life expectancy, number of tourist arrivals, starting date of the epidemic, obesity in males, urbanization, death rate from air pollution, Hepatitis B. None of these is large enough to strongly rule out the linear relation. We will therefore omit quadratic fits from now on.

A. Bias due to GDP: lack of testing?

We also find a correlation with GDP per capita, which should be an indicator of lack of testing capabilities. Note, however, that GDP per capita is also quite highly correlated with another important variable, life expectancy, as we will show in section V: high GDP per capita is related to an older population, which is correlated with faster contagion.

Note also that correlation of contagion with GDP disappears when excluding very poor countries, approximately below 5 thousand $ GDP per capita: this is likely due to the fact that only below a given threshold the capability of testing becomes insufficient. Figure 1

We performed 2-variables fits, including GDP and each of the above significant variables, to check if they remain still significant. In section IV, we will show the results of such fits, and also the result of individual one variable fits excluding countries below the threshold of 5 thousand $ GDP per capita. We list here below the variables that are still significant even when fitting together with GDP.

1. Significant variables, strong evidence

In a 2-variable fit, including GDP per capita, we find strong evidence for correlation with:

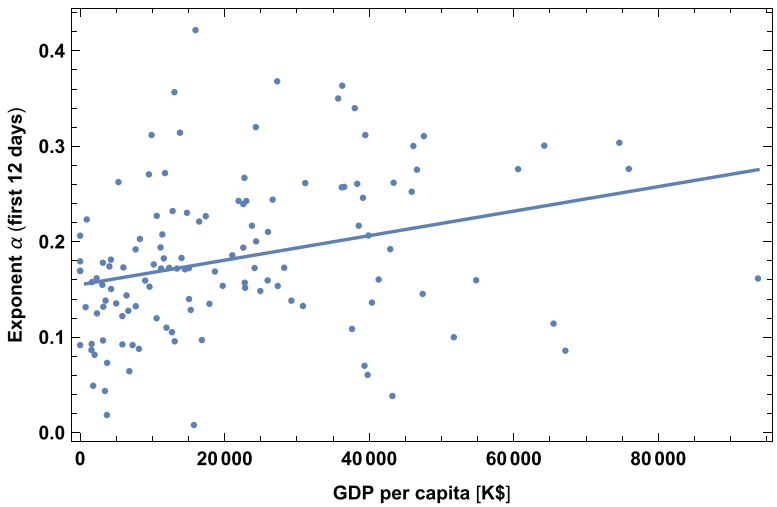

Figure 1. Exponent α for each country vs. GDP per capita. We show the data points and the best-fit for the linear interpolation.

Table 1. In the left panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with GDP per capita (GDP). In the right panel: R^2 for the best-estimate and number of countries N.

	Estimate	Standard Error	t-Statistic	p-value		
1	0.155	0.0111	14.	6.79×10^{-27}	R^2	0.087
GDP	1.28×10^{-6}	3.81×10^{-7}	3.37	0.001	N	121

(1) Amount of contact in greeting habits (positive correlation, p-value $1.5 \cdot 10^{-5}$);
(2) Temperature (negative correlation, p-value $2.3 \cdot 10^{-5}$);
(3) International tourism: number of arrivals (positive correlation, p-value $2.6 \cdot 10^{-4}$);
(4) Old-age dependency ratio: ratio of the number of people older than 64 relative to the number of people in the working-age (15–64 years) (positive correlation, p-value $5.5 \cdot 10^{-4}$);
(5) Vitamin D serum levels (negative correlation, annual values p-value 0.0032, seasonal values 0.0024; smaller sample, ~50 countries). To be confirmed on a larger sample.
(6) Starting day of the epidemic (negative correlation, p-value 0.0037);
(7) Lung cancer death rates (positive correlation, p-value 0.0039); and
(8) Life expectancy (positive correlation, p-value 0.0048);

2. Blood types, strong evidence

We still find strong evidence for correlation with blood types:

(1) RH + blood group system (negative correlation, p-value $1 \cdot 10^{-3}$);
(2) B + (negative correlation, p-value $2 \cdot 10^{-3}$);
(3) A– (positive correlation, p-value $1 \cdot 10^{-3}$); and
(4) 0- (positive correlation, p-value $3 \cdot 10^{-3}$);

We find moderate evidence for correlation with:

(1) A+ (positive correlation, p-value 0.028);
(2) B– (positive correlation, p-value 0.039); and
(3) AB– (positive correlation, p-value 0.012).

3. Significant variables, moderate evidence

We find moderate evidence for:

1. UV index (negative correlation, p-value 0.01; smaller sample, 73 countries);
2. Vaccination coverage for Tuberculosis (BCG) (negative correlation, p-value 0.023);
3. Obesity in males (positive correlation, p-value 0.02);
4. CO_2 emissions (positive correlation, p-value 0.02);
5. Alcohol consumption (positive correlation, p-value 0.03);
6. Daily smoking prevalence (positive correlation, p-value 0.03); and
7. Share of population in urban areas (positive correlation, p-value 0.04).

4. Significant variables, counterintuitive

Counterintuitively we still find correlations with:

(1) Death rate from air pollution (negative correlation, p-value 0.002);
(2) Prevalence of anemia, adults and children, (negative correlation, p-value 0.023 and 0.005);
(3) Incidence of Hepatitis B (negative correlation, p-value 0.01); and
(4) Share of women with high-blood-pressure (negative correlation, p-value 0.03).

In the next section, we analyze in more detail the significant variables, one by one (except for those which are not significant anymore after taking into account of GDP per capita, but including Type 1 Diabetes). In section V, we will analyze cross-correlations among such variables, and this will also give a plausible interpretation for the existence of the 'counterintuitive' variables.

IV. Results for each variable

We first analyze each individual variable, then we analyze variables that have a 'counterintuitive' correlation, and finally, we separately analyze blood types.

1. Temperature

The average temperature T has been collected for the relevant period of time, ranging from January to mid April, weighted among the main cities of a given country, see [1] for details. Results are shown in Figure 2 and Table 2.

We also found that another variable, the absolute value of latitude, has a similar amount of correlation as for the case of T. However, the two variables have a very high correlation (about 0.91) and we do not show results for latitude here. Another variable which is also very highly correlated is UV index and it is shown later.

2. Old-age dependency ratio

This is the ratio of the number of people older than 64 relative to the number of people in the working-age (15–64 years). Data are shown as the proportion of dependents per 100 working-age population, for the year 2017. Results are shown in Figure 3 and Table 3. This is an interesting finding, since it may suggest that

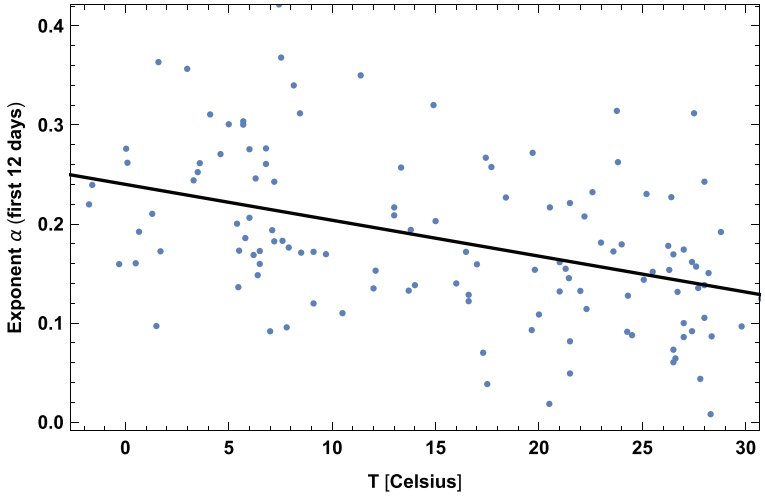

Figure 2. Exponent a for each country vs. average temperature T, for the relevant period of time, as defined in [1]. We show the data points and the best-fit for the linear interpolation.

Table 2. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of a with temperature T. We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of a with temperature T and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.239	0.0124	19.2	4.64×10^{-39}		R^2	0.186
T	−0.00359	0.000676	−5.32	4.73×10^{-7}	7.1×10^{-5}	N	126
	Estimate	Standard Error	t-Statistic	p-value			
1	0.223	0.0188	11.9	7.45×10^{-22}		R^2	0.212
GDP	5.62×10^{-7}	3.93×10^{-7}	1.43	0.155		N	121
T	−0.0033	0.000761	−4.33	0.0000311		Cross-correlation	0.425

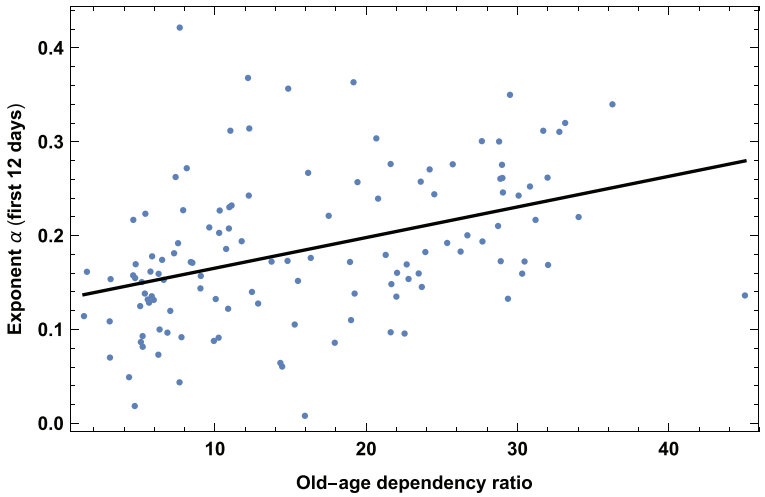

Figure 3. Exponent a for each country vs. old-age dependency ratio, as defined in the text. We show the data points and the best-fit for the linear interpolation.

old people are not only subject to higher mortality, but also more likely to be contagious. This could be either because they are more likely to become sick, or because their state of sickness is longer and more contagious, or because many of them live together in nursing homes, or all such reasons together.

Note that a similar variable is life expectancy (which we analyze later); other variables are also highly correlated, such as median age and child dependency ratio, which we do not show here. In analogy to the previous interpretation, data indicate that a younger population, including countries with high percentage of

Table 3. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with old-age dependency ratio, OLD. We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of α with OLD and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.132	0.0126	10.5	9.03×10^{-19}		R^2	0.164
OLD	0.00326	0.000669	4.87	3.37×10^{-6}	0.0025	N	123
	Estimate	Standard Error	t-Statistic	p-value			
1	0.126	0.0132	9.56	2.42×10^{-16}		R^2	0.188
GDP	7.18×10^{-7}	4.04×10^{-7}	1.78	0.0778		N	120
OLD	0.0027	0.00076	3.55	0.000557		Cross-correlation	−0.4561

children, is more immune to COVID-19, or less contagious. See also [16] for a subsequent analysis of the effects of the age structure on the transmission dynamics of the disease.

3. Life expectancy

This dataset is for year 2016. It has high correlation with old-age dependency ratio. It also has high correlations with other datasets in [15] that we do not show here, such as median age and child dependency ratio (the ratio between under 19-year-olds and 20- to 69-year-olds). Results are shown in Figure 4 and Table 4.

4. International tourism: number of arrivals

The dataset is for year 2016. Results are shown in Figure 5 and Table 5. As expected, more tourists correlate with higher speed of contagion. This is in agreement with [12,13], that found air travel to be an important factor, which will appear here as the number of tourists as well as a correlation with GDP.

5. Starting date of the epidemic

This refers to the day d_i chosen as a starting point, counted from 31 December 2019. Results are shown in Figure 6 and Table 6, which shows that earlier contagion is correlated with faster contagion. One possible

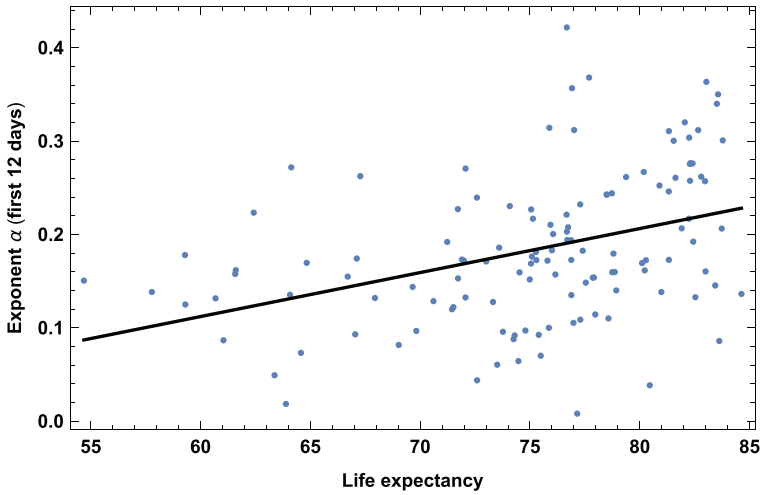

Figure 4. Exponent α for each country vs. life expectancy. We show the data points and the best-fit for the linear interpolation.

Table 4. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with life expectancy, LIFE. We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of α with LIFE and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	−0.147	0.0716	−2.05	0.0424		R^2	0.151
LIFE	0.00446	0.00096	4.65	8.56×10^{-6}	0.0041	N	125
	Estimate	Standard Error	t-Statistic	p-value			
1	−0.11	0.0929	−1.19	0.237		R^2	0.160
LIFE	0.00386	0.00134	2.87	0.00485		N	120
GDP	3.99×10^{-7}	4.98×10^{-7}	0.801	0.424		Cross-correlation	−0.679

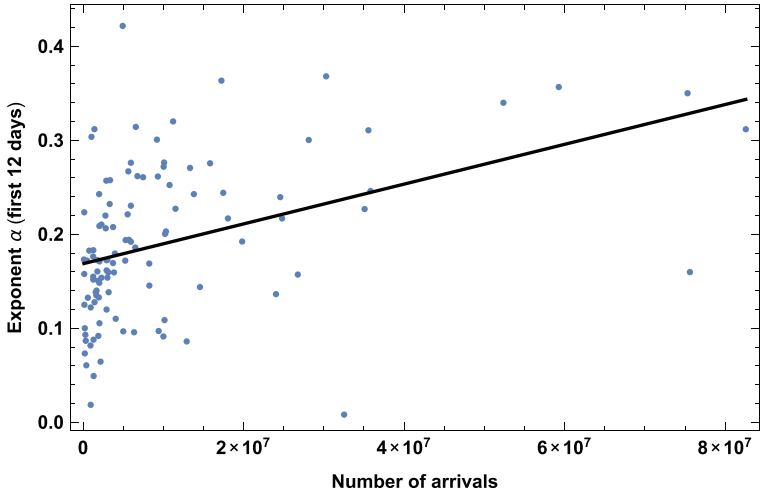

Figure 5. Exponent a for each country vs. number of tourist arrivals. We show the data points and the best-fit for the linear interpolation.

Table 5. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of a with number of tourist arrivals, ARR. We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of a with ARR and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.168	0.00846	19.9	7.38×10^{-38}		R^2	0.169
ARR	2.13×10^{-9}	4.55×10^{-10}	4.69	8.12×10^{-6}	0.0002	N	110
	Estimate	Standard Error	t-Statistic	p-value			
1	0.146	0.0116	12.6	1.48×10^{-22}		R^2	0.226
GDP	1.11×10^{-6}	4.02×10^{-7}	2.76	0.00691		N	107
ARR	1.77×10^{-9}	4.68×10^{-10}	3.78	0.000265		Cross-correlation	−0.288

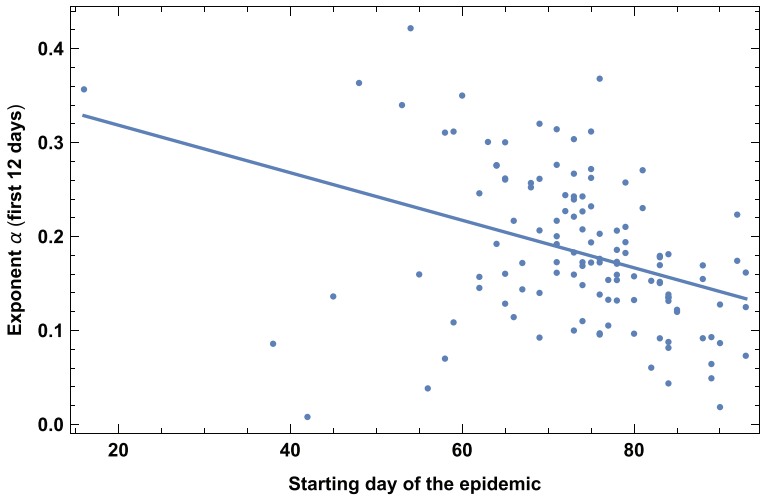

Figure 6. Exponent a for each country vs. starting date of the analysis of the epidemic, DATE, defined as the day when the positive cases reached $N = 30$. Days are counted from 31 December 2019. We show the data points and the best-fit for the linear interpolation.

interpretation is that countries which are affected later are already more aware of the pandemic and therefore have a larger amount of social distancing, which makes the growth rate smaller. Another possible interpretation is that there is some other underlying factor that protects against contagion, and therefore, epidemics spreads both later *and* slower.

6. Greeting habits

A relevant variable is the level of contact in greeting habits in each country. We have subdivided the countries in groups according to the physical contact in greeting habits; information has been taken from [17].

Table 6. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of a with vs. starting date of the analysis of the epidemic, DATE, as defined in the text. We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of a with DATE and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.369	0.0421	8.76	1.21×10^{-14}		R^2	0.139
DATE	−0.00253	0.000565	−4.48	0.0000168	0.011	N	126
	Estimate	Standard Error	t-Statistic	p-value			
1	0.32	0.0568	5.64	1.18×10^{-7}		R^2	0.151
GDP	5.85×10^{-7}	4.38×10^{-7}	1.33	0.185		N	121
DATE	−0.00204	0.000689	−2.96	0.00367		Cross-correlation	0.539

(1) No or little physical contact, bowing. In this group, we have: Bangladesh, Cambodia, Japan, Korea South, Sri Lanka, and Thailand.
(2) Handshaking between man-man and woman-woman. No or little contact man-woman. In this group we have: India, Indonesia, Niger, Senegal, Singapore, Togo, Vietnam, and Zambia.
(3) Handshaking. In this group, we have: Australia, Austria, Bulgaria, Burkina Faso, Canada, China, Estonia, Finland, Germany, Ghana, Madagascar, Malaysia, Mali, Malta, New Zealand, Norway, Philippines, Rwanda, Sweden, Taiwan, Uganda, United Kingdom, and United States.
(4) Handshaking, plus kissing among friends and relatives, but only man-man and woman-woman. No or little contact man-woman. In this group, we have: Afghanistan, Azerbaijan, Bahrain, Belarus, Brunei, Egypt, Guinea, Jordan, Kuwait, Kyrgyzstan, Oman, Pakistan, Qatar, Arabia Saudi, Arab Emirates United, and Uzbekistan.
(5) Handshaking, plus kissing among friends and relatives. In this group, we have: Albania, Algeria, Argentina, Armenia, Belgium, Bolivia, and Bosnia Herzegovina, Cameroon, Chile, Colombia, Costa Rica, Côte d'Ivoire, Croatia, Cuba, Cyprus, Czech Republic, Denmark, Dominican Republic, Ecuador, El Salvador, France, Georgia, Greece, Guatemala, Honduras, Hungary, Iran, Iraq, Ireland, Israel, Italy, Jamaica, Kazakhstan, Kenya, Kosovo, Latvia, Lebanon, Lithuania, Luxembourg, Macedonia, Mauritius, Mexico, Moldova, Montenegro, Morocco, Netherlands, Panama, Paraguay, Peru, Poland, Portugal, Puerto Rico, Romania, Russia, Serbia, Slovakia, Slovenia, Africa South, Switzerland, Trinidad and Tobago, Tunisia, Turkey, Ukraine, Uruguay, and Venezuela.
(6) Handshaking and kissing. In this group, we have: Andorra, Brazil, and Spain.

We have arbitrarily assigned a variable, named GRE, from 0 to 1 to each group, namely GRE $= 0, 0.25, 0.5, 0.4, 0.8$ and 1, respectively. We have chosen a ratio of 2 between group 2 and 3 and between group 4 and 5, based on the fact that the only difference is that about half of the possible interactions (men-women) are without contact. Results are shown in Figure 7 and Table 7.

Note also that an outlier is visible in the plot, which corresponds to South Korea. The early outbreak of the disease in this particular case was strongly affected by the Shincheonji Church, which included mass prayer and worship sessions. By excluding South Korea from the dataset one finds an even larger significance, p-value $= 6.4 \cdot 10^{-8}$ and $R^2 \approx 0.23$.

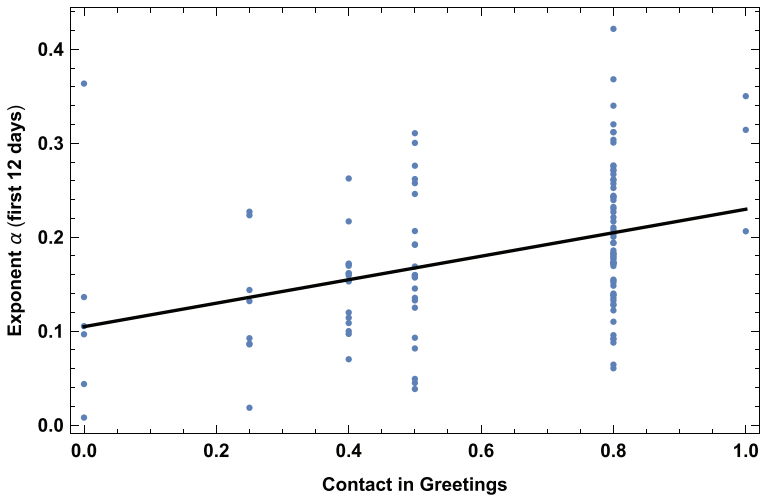

Figure 7. Exponent a for each country vs. level of contact in greeting habits, *GRE*, as defined in the text. We show the data points and the best-fit for the linear interpolation.

Table 7. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with level of contact in greeting habits, GRE, as defined in the text. We also show the p-value, excluding countries below 5 thousand \$ GDP per capita. In the left bottom panel: same quantities for correlation of α with GRE and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K\$		
1	0.111	0.019	5.8	5.47×10^{-8}		R^2	0.129
GRE	0.12	0.0287	4.2	0.000051	0.0029	N	121
	Estimate	Standard Error	t-Statistic	p-value			
1	0.079	0.0204	3.87	0.000184		R^2	0.22
GDP	1.26×10^{-6}	3.65×10^{-7}	3.45	0.000794		N	116
GRE	0.128	0.0282	4.53	0.0000149		Cross-correlation	0.00560

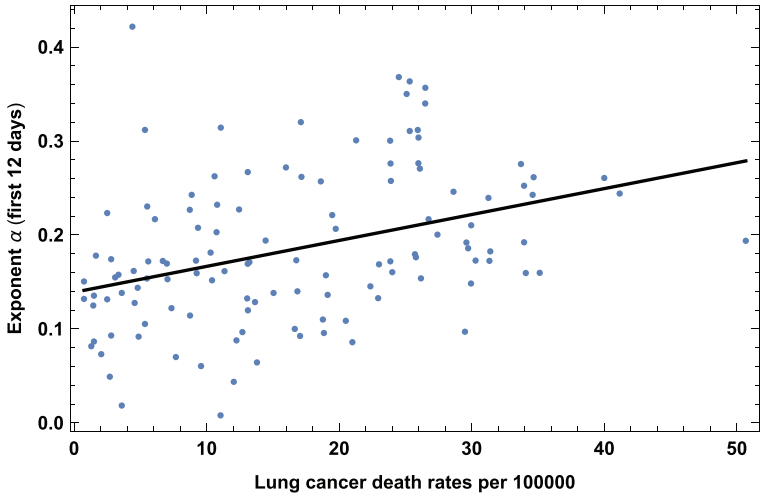

Figure 8. Exponent α for each country vs. lung cancer death rates. We show the data points and the best-fit for the linear interpolation.

Table 8. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with lung cancer death rates, LUNG. We also show the p-value, excluding countries below 5 thousand \$ GDP per capita. In the left bottom panel: same quantities for correlation of α with LUNG and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K\$		
1	0.143	0.0121	11.8	9.96×10^{-22}		R^2	0.127
LUNG	0.00159	0.000381	4.17	0.0000572	0.024	N	121
	Estimate	Standard Error	t-Statistic	p-value			
1	0.133	0.013	10.2	7.48×10^{-18}		R^2	0.164
GDP	8.93×10^{-7}	4.02×10^{-7}	2.22	0.0282		N	118
LUNG	0.00122	0.000416	2.94	0.00396		Cross-correlation	−0.403

7. Lung cancer death rates

This dataset refers to year 2002. Results are shown in Figure 8 and Table 8. Such results are interesting and could be interpreted a priori in two ways. A first interpretation is that COVID-19 contagion might correlate to lung cancer, simply due to the fact that lung cancer is more prevalent in countries with more old people. Such a simplistic interpretation is somehow contradicted by the case of generic cancer death rates, discussed in section III, which is indeed less significant than lung cancer. A better interpretation is therefore that lung cancer may be a specific risk factor for COVID-19 contagion. This is supported also by the observation of high rates of lung cancer in COVID-19 patients [18].

8. Obesity in males

This refers to the prevalence of obesity in adult *males*, measured in 2014. Results are shown in Figure 9 and Table 9. Note that this effect is mostly due to the difference between very poor countries and the rest of the world; indeed this becomes non-significant

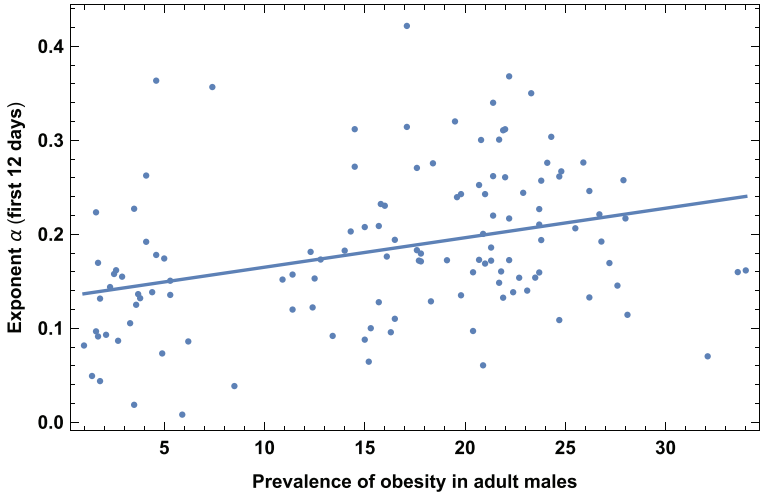

Figure 9. Exponent *a* for each country vs. prevalence of obesity in adult males. We show the data points and the best-fit for the linear interpolation.

Table 9. In the left top panel: best-estimate, standard error (σ), t-statistic, and *p*-value for the parameters of the linear interpolation, for correlation of *a* with prevalence of obesity in adult males (OBE). We also show the *p*-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of *a* with OBE and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries *N*. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	*p*-value	*p*-value, GDP>5 K$		
1	0.134	0.0144	9.29	6.9×10^{-16}		R^2	0.114
OBE	0.00314	0.000788	3.98	0.000115	0.12	*N*	125
	Estimate	Standard Error	t-Statistic	*p*-value			
1	0.132	0.0148	8.89	8.35×10^{-15}		R^2	0.128
GDP	5.6×10^{-7}	4.84×10^{-7}	1.16	0.25		*N*	121
OBE	0.00249	0.00106	2.35	0.0202		Cross-correlation	−0.634

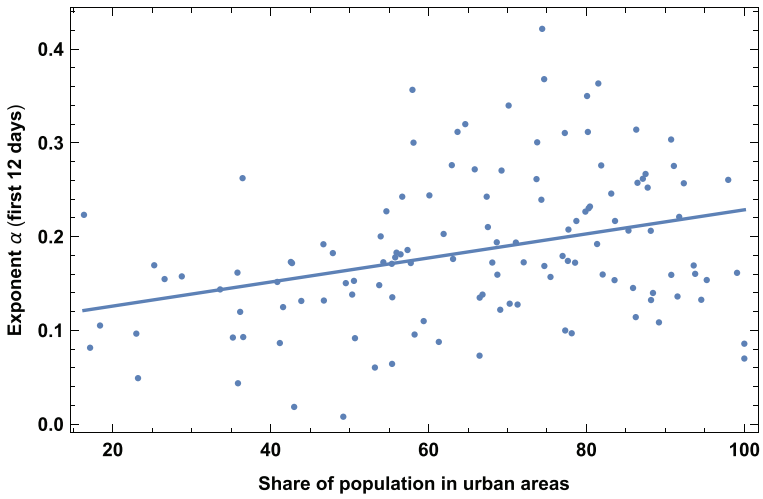

Figure 10. Exponent *a* for each country vs. share of population in urban areas. We show the data points and the best-fit for the linear interpolation.

when excluding countries below 5 K$ GDP per capita. Note also that obesity in *females* instead is *not* correlated with growth rate of COVID-19 contagion in our sample. See also [19] for increased risk of severe COVID-19 symptoms for obese patients.

9. Urbanization

This is the share of population living in urban areas, collected in year 2017. Results are shown in Figure 10 and Table 10. This is an expected correlation, in agreement with [20,21].

Table 10. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with share of population in urban areas, URB, as defined in the text. We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of α with URB and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.1	0.0228	4.4	0.0000231		R^2	0.109
URB	0.00128	0.000331	3.88	0.000173	0.057	N	124
	Estimate	Standard Error	t-Statistic	p-value			
1	0.108	0.0247	4.37	0.0000267		R^2	0.133
GDP	7.57×10^{-7}	4.74×10^{-7}	1.6	0.113		N	120
URB	0.000916	0.000439	2.09	0.0391		Cross-correlation	−0.6216

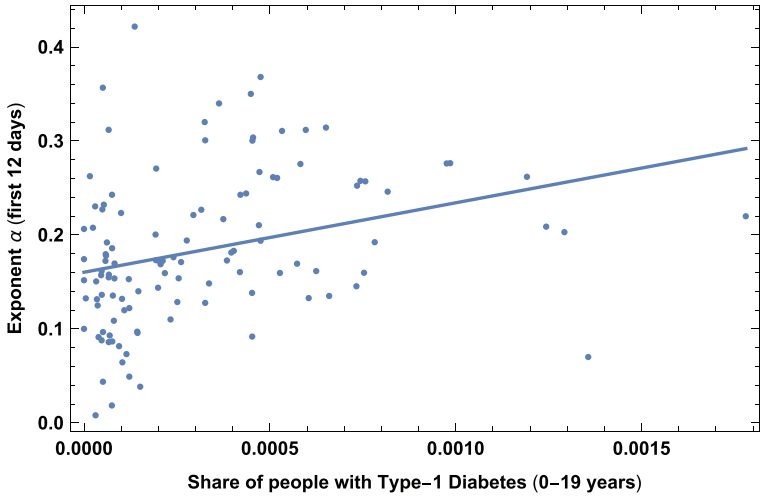

Figure 11. Exponent α for each country vs. prevalence of type-1 Diabetes (0–19 years). We show the data points and the best-fit for the linear interpolation.

Table 11. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with prevalence of type-1 Diabetes, DIA. We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of α with DIA and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.16	0.00986	16.2	2.41×10^{-31}		R^2	0.095
DIAB	73.9	21.5	3.44	0.00082	0.03	N	115
	Estimate	Standard Error	t-Statistic	p-value			
1	0.146	0.0116	12.6	5.06×10^{-23}		R^2	0.138
GDP	9.95×10^{-7}	4.49×10^{-7}	2.22	0.0288		N	111
DIAB	47.8	25.2	1.9	0.0605		Cross-correlation	−0.508

10. Prevalence of type-1 Diabetes

This is the share of people with type-1 Diabetes in 0–19 years population taken from [22]. Results are shown in Figure 11 and Table 23. Note, however, that significance decreases a lot when restricting to countries with GDP per capita larger than 5 K$ and goes even beyond 0.05 when combining when GDP in a two-variables fit. Such a correlation, even if not very robust, could be non-trivial and could constitute useful information for clinical and genetic research. See also [23].

11. Alcohol consumption

This dataset refers to year 2016. Results are shown in Figure 12 and Table 12. Note that this variable is highly correlated with old-age dependency ratio, as discussed in section V. While the correlation with alcohol consumption may be simply due to correlation with other variables, such as old-age dependency ratio, this finding deserves anyway more research, to assess whether it may be at least partially due to the deleterious effects of alcohol on the immune system.

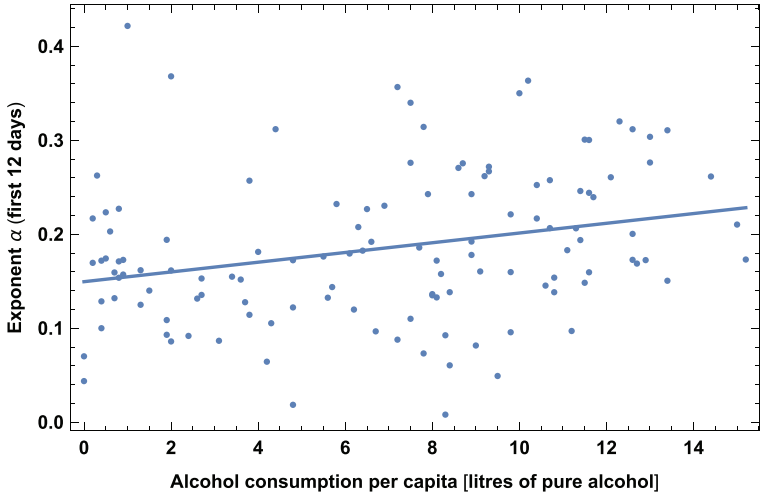

Figure 12. Exponent α for each country vs. alcohol consumption. We show the data points and the best-fit for the linear interpolation.

Table 12. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with alcohol consumption (ALCO). We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of α with ALCO and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.15	0.0131	11.4	7.8×10^{-21}		R^2	0.076
ALCO	0.00518	0.00164	3.17	0.00195	0.012	N	126
	Estimate	Standard Error	t-Statistic	p-value			
1	0.134	0.0141	9.48	3.75×10^{-16}		R^2	0.138
GDP	1.12×10^{-6}	3.86×10^{-7}	2.91	0.00437		N	120
ALCO	0.00382	0.0017	2.25	0.0264		Cross-correlation	−0.286

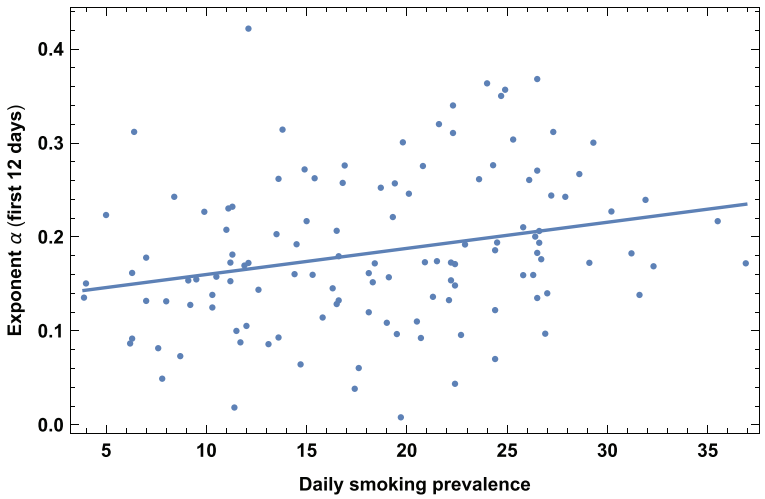

Figure 13. Exponent α for each country vs. daily smoking prevalence. We show the data points and the best-fit for the linear interpolation.

12. Smoking,

This dataset refers to year 2012. Results are shown in Figure 13 and Table 13. As expected this variable is highly correlated with lung cancer, as discussed in section V. We find that COVID-19 spreads more rapidly in countries with higher daily smoking prevalence. Note, however, that this becomes non-significant when excluding countries below 5 K$ GDP per capita.

Correlation of α with smoking thus could be simply due to correlation with other variables or to a bias due to lack of testing in very poor countries. Alternative interpretations are that smoking has

Table 13. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with daily smoking prevalence (SMOK). We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of α with SMOK and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.132	0.0187	7.07	1.07×10^{-10}		R^2	0.067
SMOK	0.00278	0.000937	2.97	0.00361	0.19	N	124
	Estimate	Standard Error	t-Statistic	p-value			
1	0.121	0.019	6.38	3.68×10^{-9}		R^2	0.122
GDP	1.06×10^{-6}	3.89×10^{-7}	2.73	0.00729		N	121
SMOK	0.00209	0.000966	2.16	0.0328		Cross-correlation	−0.2646

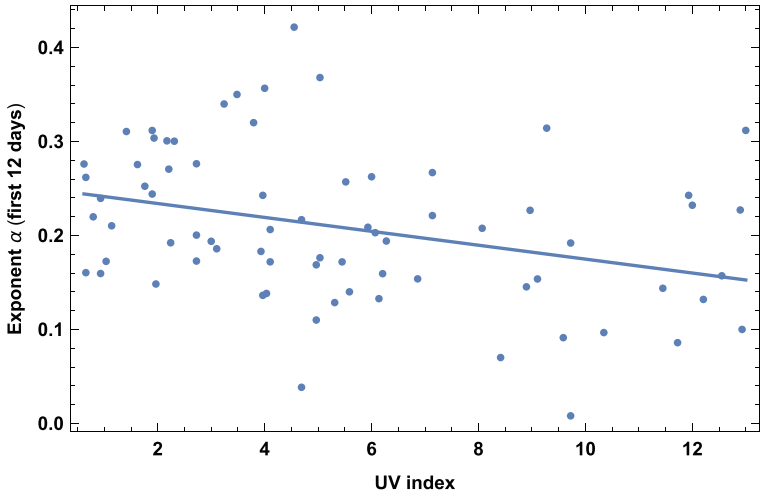

Figure 14. Exponent α for each country vs. UV index for the relevant month of the epidemic. We show the data points and the best-fit for the linear interpolation.

Table 14. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with UV index for the relevant period of time of the epidemic (UV). We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of α with UV and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.249	0.0163	15.3	3.7×10^{-24}		R^2	0.110
UV	−0.0074	0.00249	−2.97	0.00408	0.012	N	73
	Estimate	Standard Error	t-Statistic	p-value			
1	0.242	0.0269	9.01	2.88×10^{-13}		R^2	0.112
GDP	1.92×10^{-7}	5.8×10^{-7}	0.33	0.742		N	73
UV	−0.00707	0.00274	−2.58	0.0119		Cross-correlation	0.383

negative effects on conditions of lungs that facilitates contagion or that it contributes to increased transmission of virus from hand to mouth [24]. Interestingly, note that our finding is in contrast with claims of a possible protective effect of nicotine and smoking against COVID-19 [25,26].

13. UV index

This is the UV index for the relevant period of time of the epidemic. In particular, the UV index has been collected from [27], as a monthly average, and then with a linear interpolation we have used the average value during the 12 days of the epidemic growth, for each country. Results are shown in Figure 14 and Table 14. Not surprisingly in section V, we will see that such quantity is very highly correlated with T (correlation coefficient 0.93). Note also that here the sample size is smaller (73) than in the case of other variables, so it is not strange that the significance of a correlation with α here is not as high as in the case of α with Temperature. More research is required to answer more specific questions, for instance whether the virus survives less in an environment with high UV index, or whether a high UV index stimulates vitamin D production that may help the immune system, or

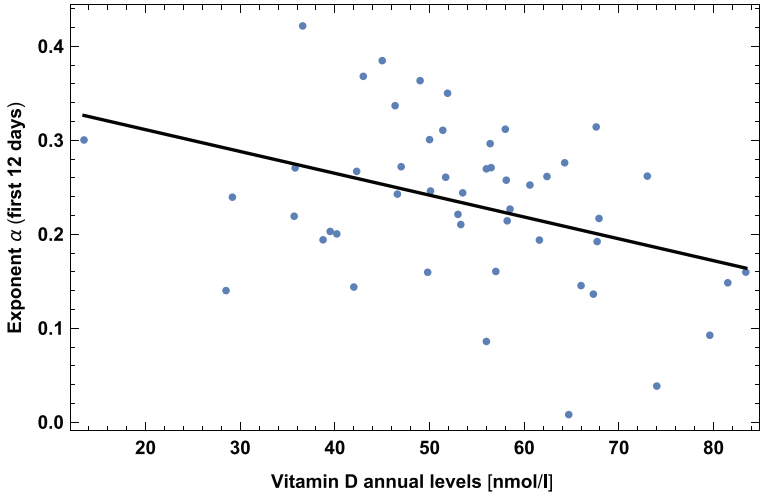

Figure 15. Exponent *a* for each country vs. annual levels of vitamin *D*, for the relevant period of time, as defined in the text, for the base set of 42 countries. We show the data points and the best-fit for the linear interpolation.

Table 15. In the left top panel: best-estimate, standard error (σ), t-statistic, and *p*-value for the parameters of the linear interpolation, for correlation of *a* with mean annual levels of vitamin D (variable name: D). We also show the *p*-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of *a* with D and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries *N*. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	*p*-value	*p*-value, GDP>5 K$		
1	0.356	0.0443	8.02	2.02×10^{-10}		R^2	0.147
D	−0.00231	0.000801	−2.88	0.00586	0.0059	*N*	50
	Estimate	Standard Error	t-Statistic	*p*-value			
1	0.342	0.0457	7.48	1.52×10^{-9}		R^2	0.172
GDP	8.29×10^{-7}	6.99×10^{-7}	1.19	0.242		*N*	50
D	−0.00255	0.000822	−3.1	0.00328		Cross-correlation	−0.243

both. Very recent results [28] consistently find that indeed SARS-CoV-2 is deactivated efficiently under simulated sunlight in few minutes, due to UVB.

14. Vitamin D serum concentration

Another relevant variable is the amount of serum Vitamin D. We collected data in the literature for the average annual level of serum Vitamin D and for the seasonal level (D_s). The seasonal level is defined as: the amount during the month of March or during winter for northern hemisphere, *or* during summer for southern hemisphere *or* the annual level for countries with little seasonal variation. The dataset for the annual D was built with the available literature, which is unfortunately quite inhomogeneous as discussed in Appendix A. For many countries, several studies with quite different values were found and in this case we have collected the mean and the standard error and a weighted average has been performed. The countries included in this dataset are 50, as specified in Appendix. The dataset for the seasonal levels is more restricted, since the relative literature is less complete, and we have included 42 countries.

Results are shown in Figure 15 and Table 34 for the annual levels and in Figure 15 and Table 34 for the seasonal levels. Note, however, that such results are preliminary and based on inhomogeneous data, and have to be confirmed on a larger sample.

Interestingly, in section V, we will see that D is *not* highly correlated with T or UV index, as one naively could expect, due to different food consumption in different countries. A slightly higher correlation, as it should be, is present between T and D_s. Note that our results are in agreement with the fact that increased vitamin D levels have been proposed to have a protective effect against COVID-19 [29–31].Figure 16

15. CO_2 emissions

This is the data for year 2017. We have also checked that this has very high correlation with SO emissions (about 0.9 correlation coefficient). We show here only the case for CO_2, but the reader should keep in mind that a very similar result applies also to SO emissions. Note also that this is expected to have a high correlation with the number of international tourist arrivals, as we will show in section V. Results are shown in Figure 17 and Table 17.

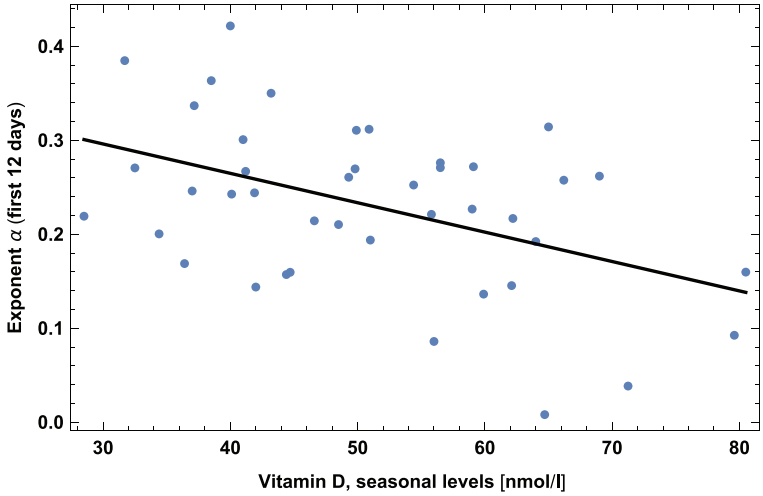

Figure 16. Exponent a for each country vs. seasonal levels of vitamin D, for the relevant period of time, as defined in the text, for the base set of 42 countries. We show the data points and the best-fit for the linear interpolation.

Table 16. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of a with seasonal levels of vitamin D (variable name: D_s). We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of a with D_s and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.385	0.0499	7.72	1.91×10^{-9}		R^2	0.206
D_s	−0.00305	0.000949	−3.22	0.00256	0.0024	N	42
	Estimate	Standard Error	t-Statistic	p-value			
1	0.375	0.0533	7.03	1.94×10^{-8}		R^2	0.212
GDP	4.66×10^{-7}	8.02×10^{-7}	0.581	0.565		N	42
D_s	−0.00314	0.00097	−3.24	0.00243		Cross-correlation	−0.162

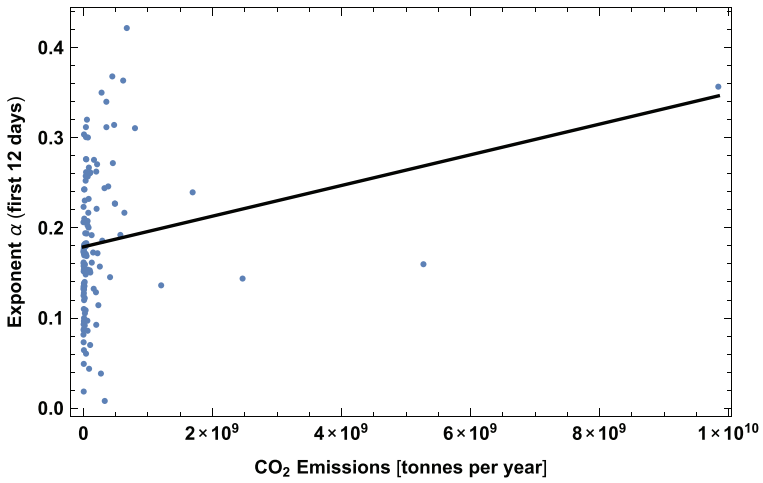

Figure 17. Exponent a for each country vs. CO_2 emissions. We show the data points and the best-fit for the linear interpolation.

Table 17. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of a with CO_2 emissions. We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of a with CO_2 and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.18	0.0073	24.	9.6×10^{-49}		R^2	0.048
CO_2	1.7×10^{-11}	6.9×10^{-12}	2.5	0.015	0.048	N	110
	Estimate	Standard Error	t-Statistic	p-value			
1	0.152	0.011	13.8	1.82×10^{-26}		R^2	0.126
GDP	1.23×10^{-6}	3.75×10^{-7}	3.29	0.00133		N	109
CO2	1.57×10^{-11}	6.74×10^{-12}	2.33	0.0214		Cross-correlation	−0.0321

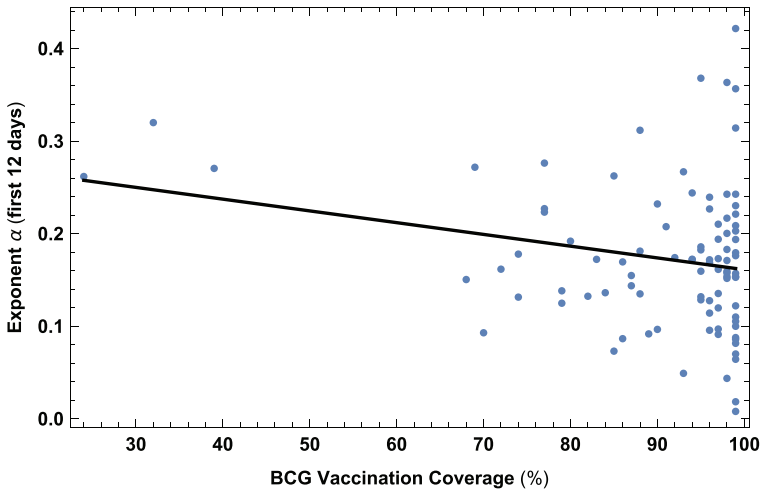

Figure 18. Exponent a for each country vs. BCG vaccination coverage. We show the data points and the best-fit for the linear interpolation.

Table 18. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of a with BCG vaccination coverage (BCG). We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of a with BCG and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.289	0.053	5.46	3.99×10^{-7}		R^2	0.051
BCG	−0.00129	0.000579	−2.22	0.0286	0.011	N	94
	Estimate	Standard Error	t-Statistic	p-value			
1	0.28	0.0534	5.24	1.07×10^{-6}		R^2	0.084
GDP	8.53×10^{-7}	4.79×10^{-7}	1.78	0.0781		N	92
BCG	−0.00134	0.00058	−2.3	0.0235		Cross-correlation	−0.0367

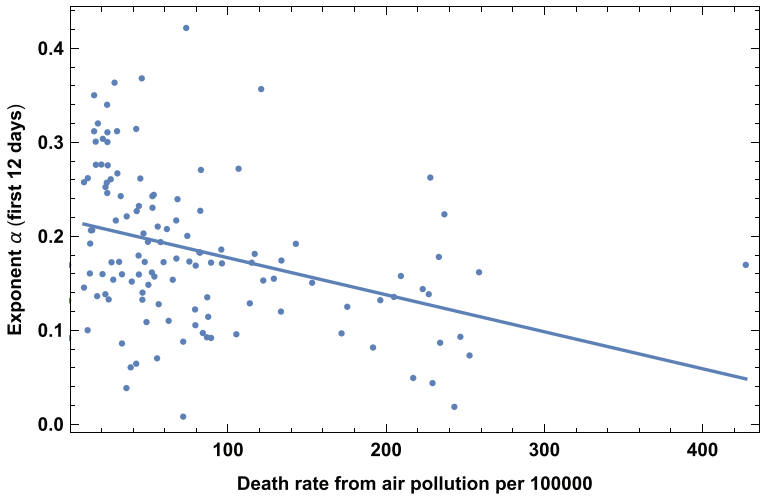

Figure 19. Exponent a for each country vs. death rate from air pollution per 100,000. We show the data points and the best-fit for the linear interpolation.

16. Tuberculosis (BCG) vaccination coverage

This dataset is the vaccination coverage for tuberculosis for year 2015.[2] Results are shown in Figure 18 and Table 18. We find a negative correlation. Note that this depends mostly on the three countries with low coverage present in the plot: by excluding them the correlation becomes non-significant. Note also that several countries with no compulsory vaccination (such as USA or Italy) do not have an estimate for BCG coverage

[2]Countries were taken from [15], plus Taiwan added from [64].

Table 19. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of a with death rate from air pollution per 100,000 (POLL). We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of a with POLL and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.216	0.0102	21.3	8.97×10^{-43}		R^2	0.132
POLL	−0.000394	0.0000917	−4.29	0.0000357	0.040	N	123
	Estimate	Standard Error	t-Statistic	p-value			
1	0.211	0.0204	10.3	3.75×10^{-18}		R^2	0.158
GDP	3.13×10^{-7}	4.77×10^{-7}	0.655	0.514		N	120
POLL	−0.000418	0.000131	−3.19	0.00185		Cross-correlation	0.630

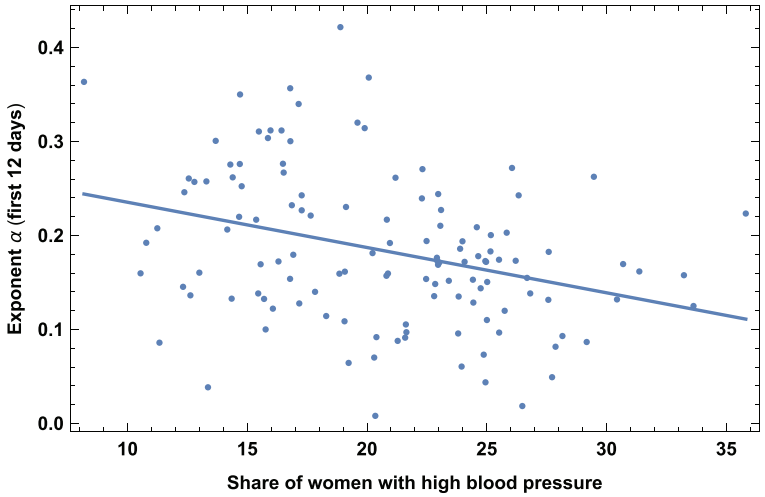

Figure 20. Exponent a for each country vs. share of women with high blood pressure. We show the data points and the best-fit for the linear interpolation.

Table 20. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of a with share of women with high blood pressure (PRE). We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of a with PRE and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.284	0.0265	10.7	2.76×10^{-19}		R^2	0.109
PRE	−0.00482	0.00124	−3.89	0.000164	0.028	N	125
	Estimate	Standard Error	t-Statistic	p-value			
1	0.248	0.0432	5.75	7.2×10^{-8}		R^2	0.123
GDP	5.85×10^{-7}	4.89×10^{-7}	1.2	0.234		N	121
PRE	−0.00372	0.00168	−2.22	0.0284		Cross-correlation	0.642

and were not included in the analysis. Including them, with very low coverage, would probably affect a lot the significance. This correlation, even if not highly significant and to be confirmed by more data, is also quite non-trivial and could be useful information for clinical and genetic research (see also [32–34]) and even for vaccine development [35–37]. Figure 23

A. 'Counterintuitive' correlations

We show here other correlations that are somehow counterintuitive, since they go in the opposite direction than from a naive expectation. We will try to interpret these results in section V.

1. Death rate from air pollution

This dataset is for year 2015. Results are shown in Figure 19 and Table 19. Contrary to naive expectations and to claims in the opposite direction [38], we find that countries with larger death rate from air pollution actually have *slower* COVID-19 contagion.

2. High blood pressure in females

This dataset is for year 2015. Results are shown in Figure 20 and Table 20. Countries with larger share of high blood pressure in females have *slower* COVID-19 contagion. Note that we do *not* find a significant correlation instead with high blood pressure in *males*.

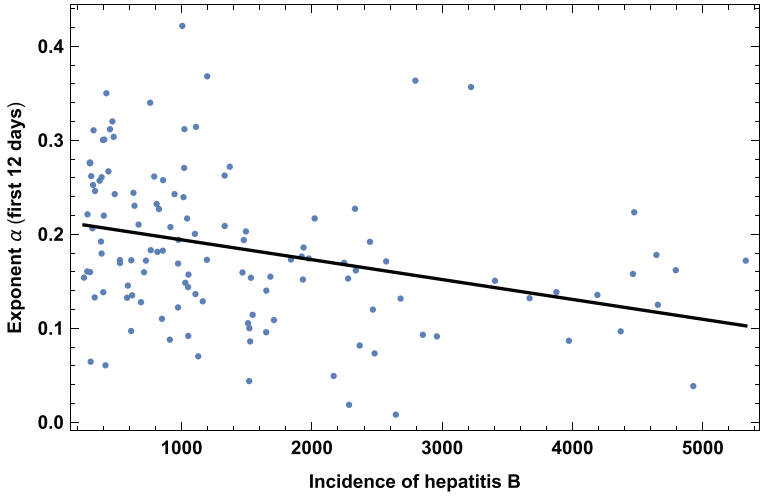

Figure 21. Exponent a for each country vs. incidence of Hepatitis B, for the relevant period of time, as defined in the text, for the base set of 42 countries. We show the data points and the best-fit for the linear interpolation.

Table 21. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of a with incidence of Hepatitis B (HEP). We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of a with HEP and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.215	0.0107	20.1	1.16×10^{-40}		R^2	0.104
HEP	−0.000021	5.56×10^{-6}	−3.78	0.000243	0.016	N	125
	Estimate	Standard Error	t-Statistic	p-value			
1	0.189	0.017	11.1	5.02×10^{-20}		R^2	0.136
GDP	8.39×10^{-7}	4.1×10^{-7}	2.05	0.0429		N	121
HEP	−0.0000161	6.21×10^{-6}	−2.58	0.011		Cross-correlation	0.420

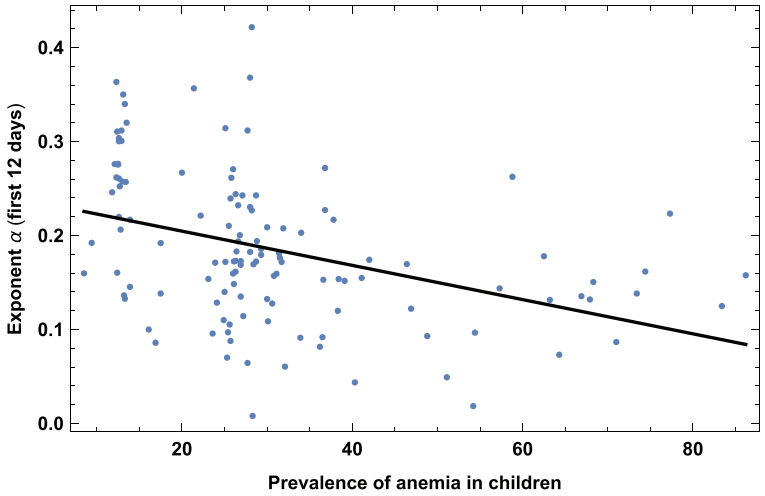

Figure 22. Exponent a for each country vs. prevalence of anemia in children. We show the data points and the best-fit for the linear interpolation.

3. Hepatitis B incidence rate

This is the incidence of hepatitis B, measured as the number of new cases of hepatitis B per 100,000 individuals in a given population, for the year 2015. Results are shown in Figure 21 and Table 21. Countries with higher incidence of hepatitis B have *slower* contagion of COVID-19.

4. Prevalence of anemia

Prevalence of anemia in children in 2016, measured as the share of children under the age of five with hemoglobin levels less than 110 grams per liter at sea level. A similar but less significant correlation is found also with anemia in adults, which we do not report here. Results are shown in Figure 22 and Table 22. The

Table 22. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with prevalence of anemia in children (ANE). We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of α with ANE and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.24	0.0136	17.7	1.58×10^{-35}		R^2	0.153
ANE	−0.00181	0.000386	−4.7	6.95×10^{-6}	0.0014	N	124
	Estimate	Standard Error	t-Statistic	p-value			
1	0.22	0.0249	8.83	1.23×10^{-14}		R^2	0.161
GDP	4.97×10^{-7}	4.74×10^{-7}	1.05	0.297		N	120
ANE	−0.00148	0.000514	−2.89	0.0046		Cross-correlation	0.638

1. *Type A+*

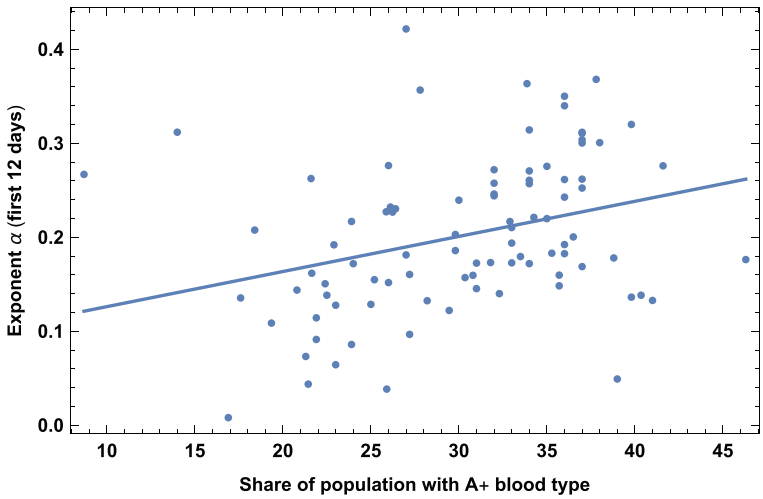

Figure 23. Exponent α for each country vs. percentage of population with blood type A+. We show the data points and the best-fit for the linear interpolation.

Table 23. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with percentage of population with blood type A+. We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of α with A+ blood and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$	R^2	0.093
1	0.0921	0.0372	2.48	0.0152		N	91
A+	0.00362	0.0012	3.03	0.0032	0.011		
	Estimate	Standard Error	t-Statistic	p-value		R^2	0.136
1	0.0935	0.0367	2.54	0.0127		N	90
GDP	9.75×10^{-7}	4.8×10^{-7}	2.03	0.0454		Cross-correlation	−0.334
A+	0.0028	0.00125	2.23	0.0282			

significance is quite high, but could be interpreted as due to a high correlation with life expectancy, as we explain in section V. A different hypothesis is that this might be related to genetic factors, which might affect the immune response to COVID-19.

B. Blood types

Blood types are not equally distributed in the world and thus we have correlated them with α. Data were taken from [39]. Very interestingly, we find significant correlations, especially for blood types B+ (slower COVID-19 contagion) and A− (faster COVID-19 contagion). In general, also all RH-negative blood types correlate with faster COVID-19 contagion. It is interesting to compare with findings in clinical data: (i) our finding that blood type A is associated with a higher risk for acquiring COVID-19 is in good agreement with [40], (ii) we find higher risk for group 0- and no correlation for group 0+ (while [40] finds lower risk for groups 0), and (iii) we have a strong significance for lower risk for RH+

2. Type B+

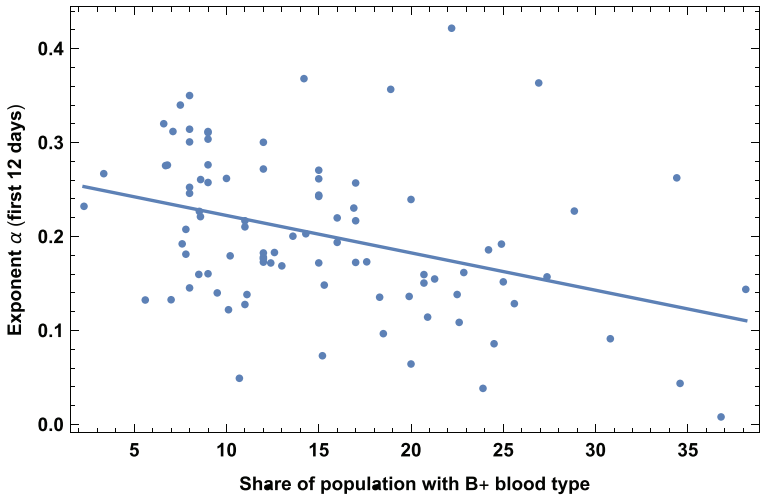

Figure 24. Exponent *a* for each country vs. percentage of population with blood type B+. We show the data points and the best-fit for the linear interpolation.

Table 24. In the left top panel: best-estimate, standard error (σ), t-statistic, and *p*-value for the parameters of the linear interpolation, for correlation of a with percentage of population with blood type B + . We also show the *p*-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of a with B+ and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries *N*. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	*p*-value	*p*-value, GDP>5 K$		
1	0.262	0.0177	14.8	8.76×10^{-26}		R^2	0.141
B+	−0.00398	0.00104	−3.82	0.000246	0.00138	*N*	91
	Estimate	Standard Error	t-Statistic	*p*-value			
1	0.231	0.0237	9.76	1.27×10^{-15}		R^2	0.181
GDP	9.18×10^{-7}	4.6×10^{-7}	2.	0.0489		*N*	90
B+	−0.00341	0.00107	−3.17	0.00211		Cross-correlation	0.284

3. Type 0-

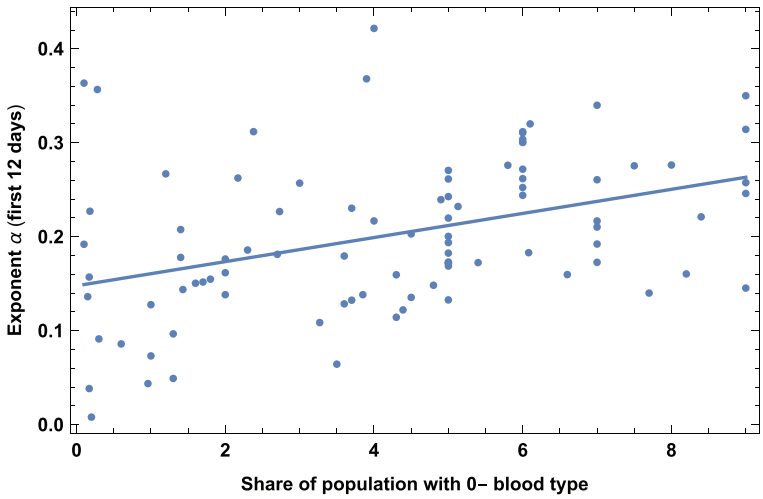

Figure 25. Exponent *a* for each country vs. percentage of population with blood type 0−. We show the data points and the best-fit for the linear interpolation.

Table 25. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with percentage of population with blood type 0 $-$. We also show the p-value, excluding countries below 5 thousand \$ GDP per capita. In the left bottom panel: same quantities for correlation of α with 0 $-$ and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K\$		
1	0.148	0.0155	9.55	2.77×10^{-15}		R^2	0.157
0-	0.0128	0.00315	4.08	0.0000991		N	91
	Estimate	Standard Error	t-Statistic	p-value			
1	0.14	0.0165	8.47	5.42×10^{-13}		R^2	0.175
GDP	6.84×10^{-7}	4.9×10^{-7}	1.39	0.167		N	90
0-	0.0107	0.00349	3.07	0.00285		Cross-correlation	−0.431

4. Type A-

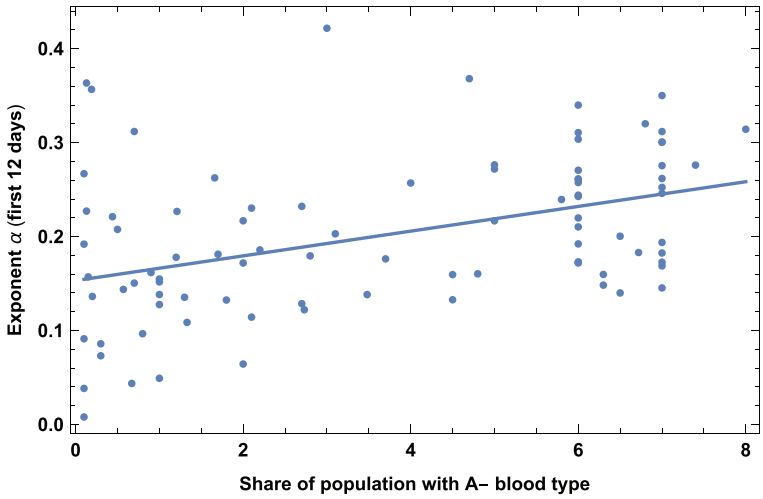

Figure 26. Exponent α for each country vs. percentage of population with blood type A−. We show the data points and the best-fit for the linear interpolation.

Table 26. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with percentage of population with blood type A−. We also show the p-value, excluding countries below 5 thousand \$ GDP per capita. In the left bottom panel: same quantities for correlation of α with A− and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K\$		
1	0.153	0.0135	11.3	6.71×10^{-19}		R^2	0.18
A−	0.0132	0.00298	4.42	0.0000278	0.0027	N	91
	Estimate	Standard Error	t-Statistic	p-value			
1	0.145	0.0151	9.62	2.46×10^{-15}		R^2	0.193
GDP	5.99×10^{-7}	4.87×10^{-7}	1.23	0.222		N	90
A−	0.0113	0.00333	3.4	0.00101		Cross-correlation	−0.442

types and in particular lower risk for group B+, which is probably a new finding, to our knowledge. These are also non-trivial findings which should stimulate further medical research on the immune response of different blood-types against COVID-19.

V. Cross-correlations

In this section, we first perform linear fits of α with each possible pair of variables (for blood types, we considered only RH+ and B+). We show the correlation coefficients between the two variables, for each pair, in

5. Type B-

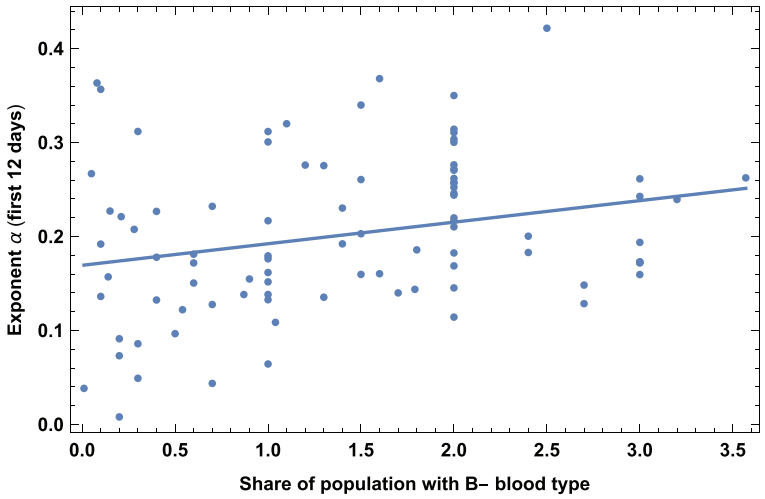

Figure 27. Exponent *a* for each country vs. percentage of population with blood type B–. We show the data points and the best-fit for the linear interpolation.

Table 27. In the left top panel: best-estimate, standard error (σ), t-statistic, and *p*-value for the parameters of the linear interpolation, for correlation of a with percentage of population with blood type B–. We also show the *p*-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of a with B– and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	*p*-value	*p*-value, GDP>5 K$		
1	0.169	0.0153	11.	2.27×10^{-18}		R^2	0.0674
B–	0.0229	0.00905	2.54	0.013	0.146	N	91
	Estimate	Standard Error	t-Statistic	*p*-value			
1	0.147	0.0176	8.36	8.98×10^{-13}		R^2	0.13
GDP	1.16×10^{-6}	4.62×10^{-7}	2.51	0.0139		N	90
B–	0.0188	0.00899	2.09	0.0392		Cross-correlation	−0.179

6. Type AB-

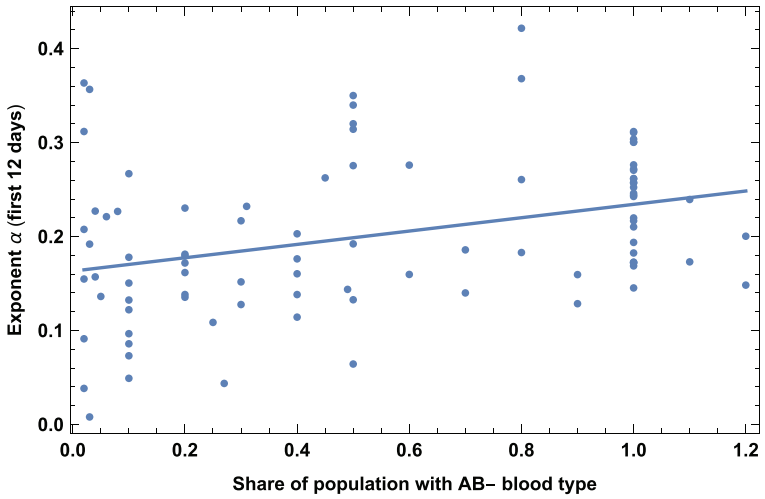

Figure 28. Exponent *a* for each country vs. percentage of population with blood type AB–. We show the data points and the best-fit for the linear interpolation.

Table 28. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with percentage of population with blood type AB−. We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of α with AB− and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.163	0.0139	11.8	7.32×10^{-20}		R^2	0.118
AB−	0.0711	0.0206	3.46	0.000844	0.0278	N	91
	Estimate	Standard Error	t-Statistic	p-value			
1	0.15	0.0157	9.53	3.75×10^{-15}		R^2	0.15
GDP	8.75×10^{-7}	4.84×10^{-7}	1.81	0.074		N	90
AB−	0.0563	0.0221	2.55	0.0126		Cross-correlation	−0.371

7. *RH-positive*

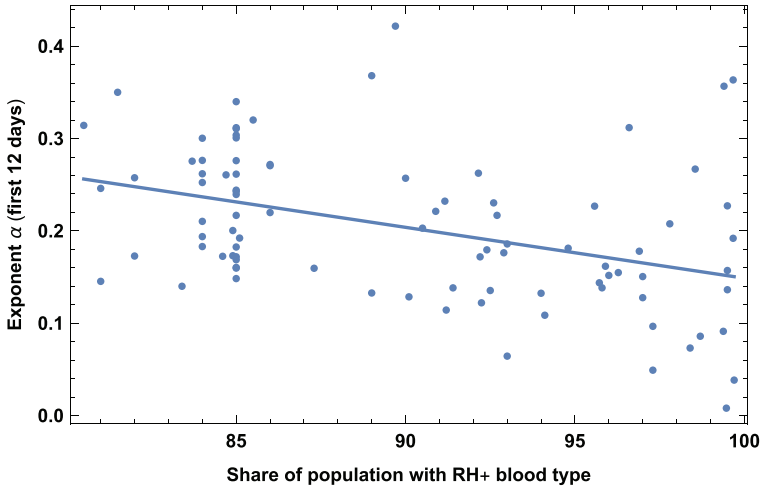

Figure 29. Exponent α for each country vs. percentage of population with RH-positive blood. We show the data points and the best-fit for the linear interpolation.

Table 29. In the left top panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation, for correlation of α with percentage of population with RH-positive blood (RH+). We also show the p-value, excluding countries below 5 thousand $ GDP per capita. In the left bottom panel: same quantities for correlation of α with RH+ and GDP per capita. In the right panels: R^2 for the best-estimate and number of countries N. We also show the correlation coefficient between the 2 variables in the two-variable fit.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5 K$		
1	0.723	0.119	6.05	3.4×10^{-8}		R^2	0.176
RH+	−0.00578	0.00132	−4.37	0.0000341	0.00355	N	91
	Estimate	Standard Error	t-Statistic	p-value			
1	0.633	0.138	4.57	0.0000158		R^2	0.191
GDP	6.27×10^{-7}	4.85×10^{-7}	1.29	0.2		N	90
RH+	−0.00495	0.00147	−3.36	0.00114		Cross-correlation	0.432

Figure 30. We also show the p-value of the t-statistic of each variable in a pair and the total R^2 of such fits in Figure 31.

We give here below possible interpretations of the redundancy among our variables, and we perform multiple variable fits in the following subsections.

A. Possible interpretations

The set of most significant variables, i.e. with smallest p-value, which correlate with faster propagation of COVID-19 are the following: low temperature, high percentage of old vs. working people and life expectancy, number of international tourist arrivals, high

	T	OLD	LIFE	ARR	DATE	GRE	LUNG	OBE	URB	DIAB	UV	GDP	ALCO	SMOK	ANE	POLL	HEP	PRE	D	D_s	CO_2	BCG	RH+	$B+$	α	$p-$ value
T	1	−0.7	−0.53	−0.27	0.38	−0.25	−0.72	−0.5	−0.33	−0.48	0.93	−0.43	−0.54	−0.61	0.61	0.42	0.45	0.32	0.072	−0.15	−0.18	-2×10^{-3}	0.66	0.34	−0.43	4×10^{-7}
OLD	−0.7	1	0.68	0.35	−0.41	0.25	0.74	0.47	0.4	0.44	−0.69	0.46	0.71	0.57	−0.66	−0.59	−0.6	−0.49	0.25	0.34	0.054	−0.15	−0.64	−0.4	0.4	3×10^{-6}
LIFE	−0.53	0.68	1	0.31	−0.58	0.19	0.56	0.69	0.65	0.51	−0.41	0.68	0.32	0.48	−0.9	−0.84	−0.74	−0.78	0.29	0.19	0.076	0.2	−0.46	−0.38	0.39	8×10^{-6}
ARR	−0.27	0.35	0.31	1	−0.61	−0.036	0.3	0.19	0.19	0.15	−0.13	0.29	0.18	0.23	−0.34	−0.25	−0.16	−0.34	0.051	0.32	0.54	0.025	−0.11	−0.045	0.41	1×10^{-5}
DATE	0.38	−0.41	−0.58	−0.61	1	0.18	−0.37	−0.26	−0.42	−0.23	0.031	−0.54	−0.2	−0.31	0.57	0.46	0.19	0.57	−0.14	−0.22	−0.54	−0.11	0.044	−0.1	−0.37	2×10^{-5}
GRE	−0.25	0.25	0.19	−0.036	0.18	1	0.28	0.45	0.29	0.17	−0.38	-6×10^{-3}	0.34	0.2	−0.27	−0.36	−0.47	−0.13	−0.2	7×10^{-3}	−0.11	−0.035	−0.49	−0.57	0.36	5×10^{-5}
LUNG	−0.72	0.74	0.56	0.3	−0.37	0.28	1	0.5	0.35	0.3	−0.63	0.4	0.6	0.76	−0.6	−0.5	−0.48	−0.32	0.08	0.32	0.11	0.11	−0.57	−0.17	0.36	6×10^{-5}
OBE	−0.5	0.47	0.69	0.19	−0.26	0.45	0.5	1	0.74	0.56	−0.5	0.63	0.34	0.44	−0.69	−0.73	−0.65	−0.56	7×10^{-4}	0.048	3×10^{-4}	0.14	−0.72	−0.62	0.34	1×10^{-4}
URB	−0.33	0.4	0.65	0.19	−0.42	0.29	0.35	0.74	1	0.43	−0.16	0.62	0.17	0.29	−0.63	−0.71	−0.59	−0.77	0.11	0.018	0.031	0.071	−0.36	−0.5	0.33	2×10^{-4}
DIAB	−0.48	0.44	0.51	0.15	−0.23	0.17	0.3	0.56	0.43	1	−0.42	0.51	0.24	0.19	−0.47	−0.43	−0.43	−0.36	0.16	−0.18	-4×10^{-4}	−0.093	−0.6	−0.42	0.31	8×10^{-4}
UV	0.93	−0.69	−0.41	−0.13	0.031	−0.38	−0.63	−0.5	−0.16	−0.42	1	−0.38	−0.6	−0.47	0.49	0.33	0.36	0.11	0.077	−0.079	−0.022	0.23	0.67	0.34	−0.33	4×10^{-3}
GDP	−0.43	0.46	0.68	0.29	−0.54	-6×10^{-3}	0.4	0.63	0.62	0.51	−0.38	1	0.29	0.26	−0.64	−0.63	−0.42	−0.64	0.24	0.13	0.061	0.037	−0.44	−0.29	0.29	1×10^{-3}
ALCO	−0.54	0.71	0.32	0.18	−0.2	0.34	0.6	0.34	0.17	0.24	−0.6	0.29	1	0.41	−0.4	−0.36	−0.39	−0.3	0.28	0.35	0.05	−0.13	−0.55	−0.38	0.27	2×10^{-3}
SMOK	−0.61	0.57	0.48	0.23	−0.31	0.2	0.76	0.44	0.29	0.19	−0.47	0.26	0.41	1	−0.54	−0.43	−0.4	−0.22	−0.31	0.42	0.073	0.13	−0.38	−0.056	0.26	4×10^{-3}
ANE	0.61	−0.66	−0.9	−0.34	0.57	−0.27	−0.6	−0.69	−0.63	−0.47	0.49	−0.64	−0.4	−0.54	1	0.82	0.77	0.78	−0.29	−0.22	−0.12	−0.12	0.51	0.43	−0.39	6×10^{-4}
POLL	0.42	−0.59	−0.84	−0.25	0.46	−0.36	−0.5	−0.73	−0.71	−0.43	0.33	−0.63	−0.36	−0.43	0.82	1	0.66	0.75	−0.28	−0.19	-4×10^{-3}	−0.13	0.5	0.49	−0.37	3×10^{-5}
HEP	0.45	−0.6	−0.74	−0.16	0.19	−0.47	−0.48	−0.65	−0.59	−0.43	0.36	−0.42	−0.39	−0.4	0.77	0.66	1	0.57	0.077	−0.18	0.042	−0.075	0.55	0.45	−0.32	2×10^{-4}
PRE	0.32	−0.49	−0.78	−0.34	0.57	−0.13	−0.32	−0.56	−0.77	−0.36	0.11	−0.64	−0.3	−0.22	0.78	0.75	0.57	1	−0.32	−0.21	−0.17	−0.065	0.15	0.41	−0.33	2×10^{-4}
D	0.072	0.25	0.29	0.051	−0.14	−0.2	0.08	7×10^{-4}	0.11	0.16	0.077	0.24	0.28	−0.31	−0.29	−0.28	0.077	−0.32	1	−0.065	0.012	−0.14	0.047	-3×10^{-3}	−0.38	6×10^{-3}
D_s	−0.15	0.34	0.19	0.32	−0.22	7×10^{-3}	0.32	0.048	0.018	−0.18	−0.079	0.13	0.35	0.42	−0.22	−0.19	−0.18	−0.21	−0.065	1	0.23	0.28	−0.09	−0.12	0.16	0.46
CO_2	−0.18	0.054	0.076	0.54	−0.54	−0.11	0.11	3×10^{-4}	0.031	-4×10^{-4}	−0.022	0.061	0.05	0.073	−0.12	-4×10^{-3}	0.042	−0.17	0.012	0.23	1	0.062	0.12	0.097	0.22	0.016
BCG	-2×10^{-3}	−0.15	0.2	0.025	−0.11	−0.035	0.11	0.14	0.071	−0.093	0.23	0.037	−0.13	0.13	−0.12	−0.13	−0.075	−0.065	−0.14	0.28	0.062	1	0.13	0.15	−0.22	0.031
RH+	0.66	−0.64	−0.46	−0.11	0.044	−0.49	−0.57	−0.72	−0.36	−0.6	0.67	−0.44	−0.55	−0.38	0.51	0.5	0.55	0.15	0.047	−0.09	0.12	0.13	1	0.55	−0.44	2×10^{-5}
$B+$	0.34	−0.4	−0.38	−0.045	−0.1	−0.57	−0.17	−0.62	−0.5	−0.42	0.34	−0.29	−0.38	−0.056	0.43	0.49	0.45	0.41	-3×10^{-3}	−0.12	0.097	0.15	0.55	1	−0.38	2×10^{-4}
α	−0.43	0.4	0.39	0.41	−0.37	0.36	0.36	0.34	0.33	0.31	−0.33	0.29	0.27	0.26	−0.39	−0.37	−0.32	−0.33	−0.38	0.16	0.22	−0.22	−0.44	−0.38	1	0
$p-$ value	4×10^{-7}	3×10^{-6}	8×10^{-6}	1×10^{-5}	2×10^{-5}	5×10^{-5}	6×10^{-5}	1×10^{-4}	2×10^{-4}	8×10^{-4}	4×10^{-3}	1×10^{-3}	2×10^{-3}	4×10^{-3}	6×10^{-6}	3×10^{-5}	2×10^{-4}	2×10^{-4}	6×10^{-3}	0.46	0.016	0.031	2×10^{-5}	2×10^{-4}	0	

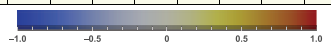

Figure 30. Correlation coefficients between each variable in a pair. Such coefficient corresponds to the off-diagonal entry of the (normalized) covariance matrix, multiplied by -1. In the last column and row we show the p-value of each variable when performing a one-variable linear fit for the growth rate α. Note also that the fits that include vitamin D variables (D and D_s) and UV index are based on smaller samples than for the other fits and were collected with rather with inhomogeneous data, and so have to be confirmed on a larger sample, as explained in the text. The variables considered here are: Temperature (T), Old age dependency ratio (OLD), Life expectancy (LIFE), Number of tourist arrivals (ARR), Starting date of the epidemic (DATE), Amount of contact in greeting habits (GRE), Lung cancer (LUNG), Obesity in males (OBE), Urbanization (URB), UV Index (UV), GDP per capita (GDP), Alcohol consumption (ALCO), Daily smoking prevalence (SMOK), Prevalence of anemia in children (ANE), Death rate due to pollution (POLL), Prevalence of hepatitis B (HEP), High blood pressure in females (PRE), average vitamin D serum levels (D), seasonal vitamin D serum levels (D_s), CO_2 emissions (CO_2), type 1 diabetes prevalence (DIAB), BCG vaccination (BCG), percentage with blood of RH+ type (RH+), and percentage with blood type B+ (B+).

percentage of RH- blood types, earlier starting date of the epidemic, high physical contact in greeting habits, and prevalence of lung cancer. Such variables are, however, correlated with each other and we analyze them together below. Most other variables are mildly/strongly correlated with the previous set of variables, except for: prevalence of Type-I diabetes, BCG vaccination, Vitamin D levels (note, however, that the latter is based on more inhomogeneous data and should be confirmed on a larger sample), which might indeed be considered as almost independent factors.

From the above table one can verify that all the 'counterintuitive' variables (death-rate due to pollution, POLL, prevalence of anemia, ANE, prevalence of hepatitis B, HEP, high blood pressure in women, PRE) have a strong negative correlation with life expectancy, LIFE. This offers a neat possible interpretation: since countries with high deaths by pollution or high prevalence of anemia or hepatitis B or high blood pressure in women have a younger population, then the virus spread is slower. Therefore, one expects that when performing a fit with any of such variable and LIFE together, one of the variables will turn out to be non-significant, i.e. redundant. Indeed, one may verify from Table 31 that this happens for all of the four above variables.

Redundancy is also present when one of the following variables is used together with life expectancy in a 2 variables fit: smoking, urbanization, obesity in males. Also, old age dependency and life expectancy are obviously quite highly correlated.

Other variables instead do *not* have such an interpretation: BCG vaccination, type-1 diabetes in children and vitamin D levels. In this case other interpretations have to be looked for. Regarding the vaccination

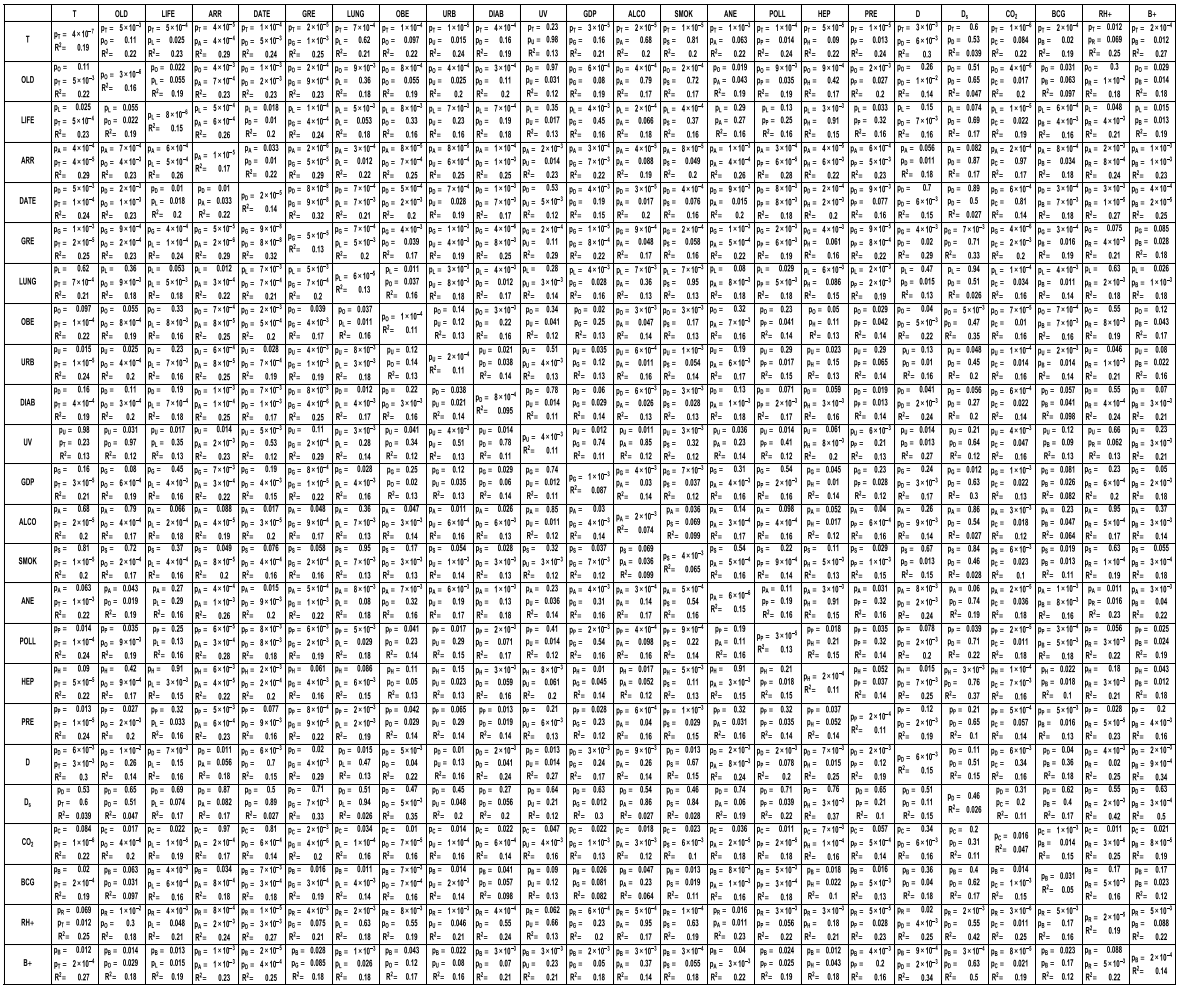

Figure 31. Significance and R^2 for linear fits of α as a function of two-variables. Variables names are the same as in Figure 30. Each cell in the table gives R^2 and the p-value of the each of the two variables, using t-statistic, labeled as p_H for horizontal and p_V for vertical.

Table 30. In the left upper panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear fit. In the right panels: R^2 for the best-estimate and number of countries N. Below we show the correlation matrix for all variables.

	Estimate	Standard Error	t-Statistic	p-value		
1	0.149	0.0247	6.02	2.67×10^{-8}		
T	−0.00247	0.000709	−3.48	0.000741	R^2	0.36
ARR	1.72×10^{-9}	4.24×10^{-10}	4.05	0.0000976	N	107
GRE	0.0972	0.0276	3.52	0.000651		

$$\begin{pmatrix} 1. & -0.65 & -0.37 & -0.83 \\ -0.65 & 1. & 0.28 & 0.24 \\ -0.37 & 0.28 & 1. & 0.1 \\ -0.83 & 0.24 & 0.1 & 1. \end{pmatrix}$$

a promising interpretation is indeed that BCG-vaccinated people could be more protected against COVID-19 [32–34].

Lung cancer and alcohol consumption remain rather significant, close to a p-value of 0.05, even after taking into account life expectancy, but they become non-significant when combining with old age dependency. In this case, it is also difficult to disentangle them from the fact that old people are more subject to COVID-19 infection.

Blood type RH+ is also quite correlated with T; however, it remains moderately significant when combined with it.

Finally vitamin D, which is measured on a smaller sample, also has little correlations with the main factors. This is quite interesting, since it may open avenues for research on protective factors and health policies. It is quite possible that high Vitamin D helps the immune response to COVID-19 [29–31]. Note, however, that this finding

Table 31. In the left upper panel: best-estimate, standard error (σ), t-statistic, and p-value for the parameters of the linear interpolation. In the right panels: R^2 for the best-estimate and number of countries N. Below we show the correlation matrix for all variables.

	Estimate	Standard Error	t-Statistic	p-value		
1	0.299	0.0519	5.77	8.66×10^{-8}		
T	−0.00169	0.000718	−2.36	0.0203	R^2	0.42
ARR	7.69×10^{-10}	4.99×10^{-10}	1.54	0.126	N	107
GRE	0.13	0.0283	4.59	0.0000127		
DATE	−0.00237	0.000728	−3.26	0.00154		

$$\begin{pmatrix} 1. & 0.016 & -0.66 & -0.04 & -0.89 \\ 0.016 & 1. & 0.025 & 0.33 & -0.33 \\ -0.66 & 0.025 & 1. & -0.13 & 0.58 \\ -0.04 & 0.33 & -0.13 & 1. & -0.35 \\ -0.89 & -0.33 & 0.58 & -0.35 & 1. \end{pmatrix}$$

is based on more inhomogeneous data and should be confirmed on a larger and more homogeneous sample.

B. Multiple variable fits

It is not too difficult to identify redundant variables, looking at very strongly correlated pairs in Table 30. It is generically harder, instead, to extract useful information when combining more than 2 or 3 variables, since we have many variables with comparable predictive power (individual R^2 are at most around 0.2) and several of them exhibit mild/strong correlations. In the following we perform examples of fits with some of the most predictive variables, trying to keep small correlation between them.

1. Temperature+arrivals+greetings

Here we show an example of a fit with three parameters.

2. Temperature+arrivals+greetings+starting date

Here we show an example of a fit with four parameters. As we combine more than three parameters, typically at least one of them becomes less significant.

3. Many variables fit

One may think of combining *all* our variables. This is, however, not a straightforward task, because we do not have data on the same number N of countries for all variables. As a compromise we may restrict to a large number of variables, but still keeping a large number of countries. For instance we may choose the following set: T, OLD, LIFE, ARR, DATE, GRE, LUNG, OBE, URB, GDP, ALCO, SMOK, ANE, POLL, HEP, PRE, and CO_2. These variables are defined for a sample of $N = 103$ countries. By combining all of them we get $R^2 = 0.48$, which tells us that only about half of the variance is described by these variables. However, clearly, many of these variables are redundant. We perform thus now a Principal Component Analysis (PCA), i.e. we look for linear combinations of such variables that diagonalize the covariance matrix.

We perform the PCA analysis in two different ways:

(1) First, we we exclude the variables with a counterintuitive behavior, see section IV A. Moreover we also exclude LIFE, since its meaning is also captured by the similar variable OLD. We performed thus a Principal Component Analysis, fitting with linear combinations of: 1, T, OLD, ARR, GRE, OBE, URB, SMOK, CO_2, DATE, LUNG, GDP, and ALCO.
(2) Second, we include also the 'counterintuitive' variables, using thus linear combinations of: 1, T, OLD, ARR, GRE, OBE, URB, SMOK, CO_2, DATE, LUNG, GDP, ALCO, ANE, HEP, POLL, and PRE.

In both cases we fit with

$$a = \sum_i \beta_i v_i \,, \tag{1}$$

where v_i are linear combinations of the above variables.

In case (1), we find a fit with total $R^2 = 0.46$, with $N = 103$. There are, however, only six significant independent orthogonal linear combinations. Such six combinations, which have $R^2 = 0.42$, are:

$$v_1 = \text{GRE}\,,$$

Table 32. Best-estimate, standard error, t-statistic, and p-value for the parameters of the Principal Components, see eqs. (1–2). In the right panel: R^2 for the best-estimate and number of countries N.

	Estimate	Standard Error	t-Statistic	p-value		
β_1	−0.14	0.034	−4.1	0.000075		
β_2	−0.00038	0.000099	−3.9	0.00021	R^2	0.42
β_3	1.5×10^{-9}	3.6×10^{-10}	4.2	0.000057	N	103
β_4	0.0011	0.00033	3.3	0.0012		
β_5	0.0011	0.00043	2.5	0.014		
β_6	−0.002	0.001	−2.	0.051		

Table 33. Best-estimate, standard error, t-statistic, and *p*-value for the parameters of the Principal Components, see eqs. (1–3). In the right panel: R^2 for the best-estimate and number of countries *N*.

	Estimate	Standard Error	t-Statistic	*p*-value		
β_1	0.000019	4.1×10^{-6}	4.6	0.000015		
β_2	−0.16	0.038	−4.3	0.000045	R^2	0.42
β_3	-1.4×10^{-9}	4.1×10^{-10}	−3.4	0.001	*N*	103
β_4	−0.003	0.0011	−2.8	0.0054		
β_5	-1.5×10^{-11}	5.5×10^{-12}	−2.7	0.0083		
β_6	−0.00042	0.00019	−2.2	0.032		

$$
\begin{aligned}
v_2 &= 0.11\,\mathrm{OBE} + 0.11\,\mathrm{OLD} + 0.18\,\mathrm{SMOK} + 0.18\,T \\
&\quad + 0.2\,\mathrm{LUNG} + 0.51\,\mathrm{URB} + 0.78\,\mathrm{DATE},
\end{aligned}
$$

$$
v_3 = \mathrm{ARR},
$$

$$
\begin{aligned}
v_4 &= 0.1\,\mathrm{ALCO} + 0.11\,\mathrm{OBE} + 0.14\,\mathrm{URB} + 0.26\,\mathrm{SMOK} \\
&\quad + 0.29\mathrm{OLD} - 0.32\,\mathrm{DATE} - 0.4\,T + 0.74\,\mathrm{LUNG},
\end{aligned}
$$

$$
v_5 = 0.12\,\mathrm{OBE} - 0.29\,\mathrm{LUNG} - 0.45\,\mathrm{DATE} + 0.82\,\mathrm{URB},
$$

$$
\begin{aligned}
v_6 &= -0.23\,\mathrm{DATE} + 0.28\,\mathrm{OLD} + 0.45\,\mathrm{SMOK} + 0.56\,T \\
&\quad - 0.59\,\mathrm{OBE},
\end{aligned}
\tag{2}
$$

where we omitted variables whose coefficients ('loadings') on v_1, v_2, v_3, v_4, v_6 are less than 0.1. The variables with loadings of at least 0.3 are: GRE, URB, DATE, ARR, T, LUNG, SMOK, and OBE. The significance of the principal components is given in Table 32.

In case (2), we find a fit with total $R^2 = 0.48$, with $N = 103$. There are again only six significant independent orthogonal linear combinations. Such 6 combinations, which have $R^2 = 0.42$, are:

$$
v_1 = -\mathrm{HEP},
$$

$$
v_2 = -\mathrm{GRE},
$$

$$
v_3 = -\mathrm{ARR},
$$

$$
\begin{aligned}
v_4 &= -0.49\,\mathrm{ANE} + 0.27\,\mathrm{LUNG} - 0.2\,\mathrm{OBE} - 0.35\,\mathrm{OLD} \\
&\quad + 0.17\,\mathrm{SMOK} + 0.69\,T,
\end{aligned}
$$

$$
v_5 = \mathrm{CO}_2,
$$

$$
\begin{aligned}
v_6 &= -0.25\,\mathrm{DATE} - 0.24\,\mathrm{LUNG} - 0.15\,\mathrm{OBE} - 0.12\,\mathrm{OLD} \\
&\quad + 0.74\,\mathrm{POLL} - 0.15\,\mathrm{SMOK} - 0.52\,\mathrm{URB},
\end{aligned}
$$

where, again, we omitted variables whose loadings are less than 0.1. The variables with loadings of at least 0.3 are now: HEP, GRE, ARR, ANE, OLD, T, CO_2, and URB. The significance of the principal components is given in Table 33.

Table 34. Vitamin D serum levels (in nmol/l) obtained with a weighted average from refs. [52–56] and references therein. The 'annual' level refers to an average over the year. The 'seasonal' level refers to the value present in the literature, which is closer to the months of January-March: either the amount during such months or during winter for northern hemisphere, *or* during summer for southern hemisphere *or* the annual level for countries with little seasonal variation.

Country	Vit D (annual)	Vit D (seasonal)	Country	Vit D (annual)	Vit D (seasonal)
Argentina	53	55.8	Lithuania	53.3	48.5
Australia	66	62.1	Mexico	58.5	59.
Austria	13.5	N/A	Morocco	39.5	N/A
Belgium	51.7	49.3	Netherlands	56	49.8
Brazil	67.6	65.	New Zealand	58.1	66.2
Canada	67.7	64.	Norway	64.27	56.5
Chile	42.3	41.2	Poland	53.5	41.9
China	45	31.7	Romania	40.2	34.4
Croatia	46.6	40.1	Russia	29.2	N/A
Czech Republic	62.4	N/A	Arabia Saudi	35.7	28.5
Denmark	60.6	54.4	Singapore	56	56.
Estonia	49.8	44.7	Slovakia	81.5	N/A
Finland	58.2	46.6	South Africa	47	59.1
France	58	50.9	South Korea	49	38.5
Germany	51.4	49.9	Spain	51.9	43.2
Greece	67.9	62.2	Sweden	73	69.
Hungary	61.6	51.	Switzerland	50	41.
Iceland	57	N/A	Taiwan	74	71.2
India	42	42.	Thailand	64.7	64.7
Iran	36.6	40.	Tunisia	38.8	N/A
Ireland	56.4	N/A	Turkey	43	N/A
Israel	56.5	56.5	Ukraine	35.8	32.5
Italy	46.4	37.2	UK	50.1	37.
Japan	67.3	59.9	USA	83.4	80.5
Lebanon	28.5	N/A	Vietnam	79.6	79.6

VI. Conclusions

We have collected data for countries that had at least 12 days of data after a starting point, which we fixed to be at the threshold of 30 confirmed cases. We considered a dataset of 126 countries, collected on April 15th. We have fit the data for each country with an exponential and extracted the exponents α, for each country. Then we have correlated such exponents with several variables, one by one.

We first performed a bivariate analysis, where we correlated the exponents with many different variables. However, many such factors are intercorrelated, and so we also performed a Principal Component Analysis (PCA), which leads to a number of few independent orthogonal combinations. From the bivariate analysis, we found a positive correlation with *high confidence* level with the following variables, with respective *p*-value: low temperature (negative correlation, *p*-value $4 \cdot 10^{-7}$), high ratio of old people vs. people in the working-age (15–64 years) (*p*-value $3 \cdot 10^{-6}$), life expectancy (*p*-value $8 \cdot 10^{-6}$), international tourism: number of arrivals (*p*-value $1 \cdot 10^{-5}$), earlier start of the epidemic (*p*-value $2 \cdot 10^{-5}$), high amount of contact in greeting habits (positive correlation, *p*-value $5 \cdot 10^{-5}$), lung cancer death rates (*p*-value $6 \cdot 10^{-5}$), obesity in males (*p*-value $1 \cdot 10^{-4}$), share of population in urban areas (*p*-value $2 \cdot 10^{-4}$), share of population with cancer (*p*-value $2.8 \cdot 10^{-4}$), alcohol consumption (*p*-value 0.0019), daily smoking prevalence (*p*-value 0.0036), low UV index (*p*-value 0.004; smaller sample, 73 countries), and low vitamin D serum levels (annual values *p*-value 0.006, seasonal values 0.002; smaller sample, ~50 countries).

We also find strong evidence for correlation with blood types: RH + blood group system (negative correlation, *p*-value $3 \cdot 10^{-5}$); A+ (positive correlation, *p*-value $3 \cdot 10^{-3}$); B + (negative correlation, *p*-value $2 \cdot 10^{-4}$); A– (positive correlation, *p*-value $3 \cdot 10^{-5}$); 0- (positive correlation, *p*-value $8 \cdot 10^{-4}$); and AB– (positive correlation, *p*-value 0.028). We find moderate evidence for correlation with: B– (positive correlation, *p*-value 0.013).

We find *moderate* evidence for positive correlation with: CO_2 (and SO) emissions (*p*-value 0.015), type-1 diabetes in children (*p*-value 0.023), vaccination coverage for Tuberculosis (BCG) (*p*-value 0.028).

Counterintuitively we also find negative correlations, in a direction opposite to a naive expectation, with: death rate from air pollution (*p*-value $3{\cdot}10^{-5}$), prevalence of anemia, adults and children, (*p*-value $1 \cdot 10^{-4}$ and $7 \cdot 10^{-6}$, respectively), share of women with high-blood-pressure (*p*-value $2 \cdot 10^{-4}$), incidence of Hepatitis B (*p*-value $2 \cdot 10^{-4}$), and PM2.5 air pollution (*p*-value 0.029).

As is clear from the figures, the data present a high amount of dispersion, for all fits that we have performed. This is of course unavoidable, given the existence of many systematic effects. One obvious factor is that the data are collected at *country* level, whereas many of the factors considered are regional. This is obvious from empirical data (see for instance the difference between the epidemic development in Lombardy vs. other regions in Italy, or New York vs. more rural regions), and also sometimes has obvious explanations (climate, health factors vary a lot region by region) as well as not so obvious ones. Because of this, we consider R^2 values as at least as important as *p*-values and correlation coefficients: an increase of the R^2 after a parameter is included means that the parameter has a systematic effect in reducing the dispersion ('more data points are explained').

Several of the above variables are correlated with each other and so they are likely to have a common interpretation and it is not easy to disentangle them. The correlation structure is quite rich and non-trivial, and we encourage interested readers to study the tables in detail, giving both R^2, *p*-values and correlation estimates. Note that some correlations are 'obvious', for example between temperature and UV radiation. Others are accidental, historical and sociological. For instance, social habits like alcohol consumption and smoking are correlated with climatic variables. In a similar vein correlation of smoking and lung cancer is very high, and this is likely to contribute to the correlation of the latter with climate. Historical reasons also correlate climate with GDP per capita.

Other variables are found to have a counterintuitive *negative* correlation, which can be explained due their strong negative correlation with life expectancy: death-rate due to pollution, prevalence of anemia, Hepatitis B and high blood pressure for women.

We also analyzed the possible existence of a bias: countries with low GDP-per capita, typically located in warm regions, might have less intense testing and we discussed the correlation with the above variables, showing that most of them remain significant, even after taking GDP into account. In this respect, note that in countries where testing is not prevalent, registration of the illness is dependent on the development of severe symptoms. Hence, while this study is about *infection rates* rather than *mortality*, in quite a few countries, we are actually measuring a proxy of mortality rather than infection rate. Hence, effects affecting mortality will be more relevant. Preexisting lung conditions, diabetes, smoking and health indicators in general as well as pollution are likely to be important in this respect, perhaps not affecting α per se but the detected amount of α. These are in turn generally correlated with GDP and temperature for historical reasons. Other interpretations, which may be complementary, are that comorbidities and old age affect immune response and thus may directly increase the growth rate of the contagion. Similarly it is likely that

individuals with comorbidities and old age, developing a more severe form of the disease, are also more contagious than younger or asymptomatic individuals, producing thus an increase in α. In this regard, we wish to point the reader's attention to the relevant differences in correlations once we apply a threshold on GDP per capita. It has long been known that human wellness (we refer to a psychological happiness study [41], but the point is more general) depends non-linearly on material resources, being strongly correlated when resources are low and reaching a plateau after a critical limit. The biases described above (weather comorbidity, testing facilities, preexisting conditions and environmental factors) seem to reflect this, changing considerably in the case our sample has a threshold w.r.t. a more general analysis without a threshold.

About pollution our findings are mixed. We find no correlation with generic air pollution ('Suspended particulate matter (SPM), in micrograms per cubic metre'). We find higher contagion to be moderately correlated only with and CO_2/SO emissions. Instead, we find a *negative* correlation with death rates due to air pollution and PM2.5 concentration (in contrast with [38], see also [42]). Note, however, that correlation with PM2.5 becomes non significant when combined with GDP per capita, while CO_2/SO becomes non significant when combined with tourist arrivals. Finally, death rates due to air pollution is also redundant when correlating with life expectancy.

Given the existence of intercorrelations among our factors, we performed a Principal Component Analysis, to disentangle the minimal number of significant non-redundant factors. We selected a sample of N = 103 countries, where: (1) we omitted all variables with a counterintuitive behavior, (2) we omitted Vitamin D, BCG, UV and Blood variables, as we did not have data for so many countries, (3) we also omitted LIFE, since its meaning is already described by the similar variable OLD. Therefore we used the following variables: 1, T, OLD, ARR, GRE, OBE, URB, SMOK, CO2, DATE, LUNG, GDP, and ALCO. As a result we found that only 6 linear orthogonal combinations are significant, and the variables with a loading of at least 0.3 on any of such 6 combinations are: GRE, URB, DATE, ARR, T, LUNG, SMOK, and OBE.

Including also the counterintuitive variables we find again only 6 significant linear orthogonal combinations, but the variables with a loading of at least 0.3 are now: HEP, GRE, ARR, ANE, OLD, T, CO_2, and URB.

Some of the above significant factors have been examined in the literature, with the status of a causal relationship explanation varying greatly. Some factors, such as GRE and ARR, quite obviously are expected to have some effect. To our knowledge we established for the first time at a global level. Some previous studies at a local level were made in [43,44].

URB is naively 'obvious' but a bit more subtle, since it does not refer to population density (which we found to be among the non-significant variables) but specifically to Urbanization. Previous studies have already stressed that urban areas are more subject to generic outbreaks [45] and also specifically to COVID-19 (see [21] for a study involving US cities). There are two main factors at work here: a 'trivial' one is that urban societies, both because of the nature of their economic activities and social factors are much more interconnected. The power-law scaling seen in comparative city studies [21] is exactly what one expects from the hypothesis that urban contagion patterns are determined by interconnectedness. However, the greater health-care density in urban areas [21] will also lead to a quicker identification and reporting of cases [20], which would also explain part of the dependence we observe. Reporting bias described in [20], together with the growth of the awareness of the pandemic (more intense use of social distances and masks in countries where the epidemic started later) over the period described in our work, is also at work for DATE (see also [46] for a study, subsequent to our work, that confirms similar conclusions).

LUNG, SMOK, and OBE are probably dominated by the bias toward symptomatic cases, and that has been extensively examined in the medical literature. To our knowledge, we established for the first time their correlation with COVID-19 spread at a global level. In particular, OBE is well known to be linked to a compromised immune system [47], and it has been linked to the number of hospitalizations due to COVID-19 [48]. Similar but more general reasons (poor lung function and link to poor hygiene habits) could account for SMOKE [49] and the link to LUNG has been studied in detail, including by detailed modeling [50]. Which brings us to T, the variable which more clearly has an 'environmental' rather than 'social' or 'medical' origin. As shown in [51], if the chemical reactions behind the viral activation obey usual kinetic equations, a strong positive correlation of temperature with virus growth is expected.

Some of the variables that we have studied cannot be arbitrarily changed but can be taken into account by public health policies, such as temperature, amount of old people and life expectancy, by implementing stronger testing and tracking policies, and possibly lockdowns, both with the arrival of the cold seasons and for the old aged population.

Other variables instead can be controlled by governments: testing and isolating international travelers and reducing number of flights in more affected regions; promoting social distancing habits as long as the virus is spreading, such as campaigns for reducing physical contact in greeting habits; campaigns against vitamin D deficiency, decrease smoking and obesity.

We also emphasize that some variables are useful to inspire and support medical research, such as correlation of contagion with: lung cancer, obesity, low vitamin D levels, blood types (higher risk for all RH- types,

A types, lower risk for B+ type), and type 1 diabetes. This definitely deserves further study, also of correlational type using data from patients.

In conclusion, our findings could thus be very useful both for policy makers and for further experimental research.

Acknowledgments

GT acknowledges support from FAPESP proc. 2017/06508-7, participation in FAPESP tematico 2017/05685-2 and CNPQ bolsa de produtividade 301432/2017-1. We would like to acknowledge Alberto Belloni, Jordi Miralda and Miguel Quartin for useful discussions and comments.

Disclosure statement

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

Funding

This work was supported by the CNPQ Bolsa de produtividade [306152/2020-7]; FAPESP Bolsa de pesquisa [2021/01700-2].

ORCID

Alessio Notari http://orcid.org/0000-0002-8854-3987

References

[1] Notari A. Temperature dependence of COVID-19 transmission. Sci Total Environ. 2021 Apr 1;763:144390. doi:10.1016/j.scitotenv.2020.144390. Epub 2020 Dec 13. PMID: 33373782; PMCID: PMC7733690.
[2] Tian H, Liu Y, Li Y, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science. 2020;368 (6491):638–642.
[3] Sun GQ, Wang SF, Li MT, et al. Transmission dynamics of COVID-19 in Wuhan, China: effects of lockdown and medical resources. Nonlinear Dyn. 2020;1–13. doi:10.1007/s11071-020-05770-9
[4] Ming-Tao L, Sun G-Q, Zhang J, et al. Analysis of COVID-19 transmission in Shanxi Province with discrete time imported cases. Math Biosci Eng. 2020;17 (4):3710–3720.
[5] Demongeot J, Flet-Berliac Y, Seligmann H. Temperature decreases spread parameters of the new COVID-19 case dynamics. Biology (Basel). 2020;9(5):94.
[6] Wang M, Jiang A, Gong L, et al. Temperature significant change COVID-19 Transmission in 429 cities; 2020. 10.1101/2020.02.22.20025791.
[7] Luo W, Majumder MS, Liu D, et al. The role of absolute humidity on transmission rates of the COVID-19 outbreak. 10.1101/2020.02.12.20022467.
[8] Araujo MB, Naimi B. Spread of SARS-CoV-2 coronavirus likely to be constrained by climate. 10.1101/2020.03.12.20034728.
[9] Bukhari Q, Jameel Y. Will coronavirus pandemic diminish by summer? http://dx.doi.org/10.2139/ssrn.3556998.
[10] Wang J, Tang K, Feng K, et al. High temperature and high humidity reduce the transmission of COVID-19; 2020. arXiv:2003.05003 [q-bio.PE] 10.2139/ssrn.3556998.
[11] Sajadi MM, Habibzadeh P, Vintzileos A, et al. Temperature, humidity and latitude analysis to predict potential spread and seasonality for COVID-19. 10.2139/ssrn.3550308.
[12] Coelho MTP, Rodrigues JFM, Anderson Matos M, et al. medRxiv 2020.04.02.20050773. 10.1101/2020.04.02.20050773.
[13] Fiolet T. Ecological Study on COVID-19: associations between the early growth rate and historical environmental and socio-economic factors in 96 countries using GAM (Generalized Additive models). Zenodo. 10.5281/zenodo.3784948 .
[14] https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases
[15] Taken from https://ourworldindata.org/charts
[16] Blyuss KB, Kyrychko YN. Effects of latency and age structure on the dynamics and containment of COVID-19. J Theor Biol. 2021;513:110587.
[17] https://guide.culturecrossing.net/basics_business_student_details.php and https://guide.culturecrossing.net/
[18] Zhang L, Zhu F, Xie L, Wang C, Wang J, Chen R, Jia P, Guan HQ, Peng L, Chen Y, Peng P, Zhang P, Chu Q, Shen Q, Wang Y, Xu SY, Zhao JP, Zhou M. Clinical characteristics of COVID-19-infected cancer patients: a retrospective case study in three hospitals within Wuhan. Ann Oncol. 2020 Jul;31(7):894–901. DOI:10.1016/j.annonc.2020.03.296. Epub 2020 Mar 26. PMID: 32224151; PMCID: PMC7270947.
[19] Kassir R. Risk of COVID-19 for patients with obesity. Obesity Rev. 2020;21(6):e13034.
[20] Pluchino A, Biondo AE, Giuffrida N, et al. A novel methodology for epidemic risk assessment of COVID-19 outbreak. Sci Rep. 2021;11(1):5304.
[21] Stier A, Berman M, Bettencourt L. COVID-19 Attack Rate Increases with City Size March 30, 2020. Mansueto Institute for Urban Innovation Research Paper. https://ssrn.com/abstract=3564464 .
[22] https://diabetesatlas.org/data/en/indicators/12/
[23] Fang L, Karakiulakis G, Roth M. https://www.thelancet.com/pdfs/journals/lanres/PIIS2213-2600(20)30116-8.pdf .
[24] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-on-smoking-and-COVID-19
[25] Changeux J-P, Amoura Z, Rey F, et al. A nicotinic hypothesis for COVID-19 with preventive and therapeutic implications. Qeios ID: FXGQSB.2. 10.32388/FXGQSB.2
[26] Miyara M, Tubach F, Pourcher V, Morelot-Panzini C, et al. Low incidence of daily active tobacco smoking in patients with symptomatic COVID-19. Qeios. DOI:10.32388/WPP19W.3. doi:10.32388/WPP19W.4
[27] UV radiation monitoring archive, http://www.temis.nl/uvradiation/UVarchive.html and WHO, https://www.who.int/uv/intersunprogramme/activities/uv_index/en/index3.html
[28] Ratnesar-Shumate S, Williams G, Green B, Krause M, Holland B, Wood S, Bohannon J, Boydston J, Freeburger D, Hooper I, Beck K, Yeager J, Altamura LA, Biryukov J, Yolitz J, Schuit M, Wahl V, Hevey M,

Dabisch P. Simulated sunlight rapidly inactivates SARS-CoV-2 on surfaces. J Infect Dis. 2020 Jul 15:222 (2):214–222. doi:10.1093/infdis/jiaa274

[29] Isaia G, Medico E. Comunicato all'accademia di Medicina dell'Universita' di Torino. https://www.unitonews.it/storage/2515/8522/3585/Ipovitaminosi_D_e_Coronavirus_25_marzo_2020.pdf

[30] Grant WB, Lahore H, McDonnell S, et al. Evidence that vitamin D supplementation could reduce risk of influenza and COVID-19 Infections and deaths. Nutrients. 2020;12(4):988.

[31] Ilie PC, Stefanescu S, Smith L et al. The role of vitamin D in the prevention of coronavirus disease 2019 infection and mortality, 08 April 2020, PREPRINT (Version 1) available at Research Square. 10.21203/rs.3.rs-21211/v1.

[32] Shet A, Ray D, Malavige N, Santosham M, Bar-Zeev N. medRxiv 2020.04.01.20049478; doi: https://doi.org/10.1101/2020.04.01.20049478

[33] Hegarty P, Dayal D, Gupta S, et al. Connecting BCG vaccination and COVID-19: additional data. medRxiv2020.04.07.20053272. doi:10.1101/2020.04.07.20053272

[34] Miller A, Reandelar MJ, Fasciglione K, et al. Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study. medRxiv 2020.03.24.20042937. 10.1101/2020.03.24.20042937.

[35] https://clinicaltrials.gov/ct2/show/NCT04327206

[36] https://clinicaltrials.gov/ct2/show/NCT04328441

[37] https://clinicaltrials.gov/ct2/show/NCT04348370

[38] Wu X, Nethery RC, Sabath BM, et al. Exposure to air pollution and COVID-19 mortality in the United States. medRxiv 2020.04.05.20054502. 10.1101/2020.04.05.20054502.

[39] Due to a mistake for Bangladesh (data with RH+ sum to more than 100%). we took data from https://en.wikipedia.org/wiki/Blood_type_distribution_by_country .

[40] Zhao J et al., medRxiv 2020.03.11.20031096 10.1101/2020.03.11.20031096

[41] Jebb AT, Tay L, Diener E, et al. Happiness, income satiation and turning points around the world. Nat Human Behav. 2018;2(1):33–38.

[42] Pansini R, Fornacca D. Higher virulence of COVID-19 in the air-polluted regions of eight severely affected countries medRxiv 2020.04.30.20086496. 10.1101/2020.04.30.20086496

[43] Zheng R, Xu Y, Wang W, et al. Spatial transmission of COVID-19 via public and private transportation in China. Travel Med Infect Dis. 2020;34:101626.

[44] Wells CR, Sah P, Moghadas SM, et al. Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. Proc Nat Acad Sci. 2020 Mar;117(13):7504–7509.

[45] Alirol E, Getaz L, Stoll B, et al. Urbanisation and infectious diseases in a globalised world. Lancet Infect Dis. 2011 Feb;11(2):131–141.

[46] Djordjevic M, Salom I, Markovic S, et al. Inferring the main drivers of SARS-CoV-2 global transmissibility by feature selection methods. 10.1029/2021GH000432

[47] Johnston R, Suratt B. Mechanisms and manifestations of obesity in lung disease. London: Academic Press; 2018.

[48] Finer N, Garnett Sp, Bruun JM. COVID-19 and obesity. Clin Obes. 2020;10(3):e12365.

[49] Ahmed N, Maqsood A, Abduljabbar T, et al. Tobacco smoking a potential risk factor in transmission of COVID-19 infection. Pak J Med Sci. 2020;36(COVID19–S4):COVID19-S104-S107.

[50] Herrmann J, Mori V, Bates JHT, et al. Modeling lung perfusion abnormalities to explain early COVID-19 hypoxemia. Nat Commun. 2020;11(1):4883.

[51] Morris DH, Yinda KC, Gamble A, et al. Mechanistic theory predicts the effects of temperature and humidity on inactivation of SARS-CoV-2 and other enveloped viruses. Preprint. bioRxiv. 2020;2020.10.16.341883; 2020 Dec 18. 10.1101/2020.10.16.341883

[52] Vitamin D map developed by D.A. Wahl et al. on behalf of International Osteoporosis Foundation (IOF). A global representation of vitamin D status in healthy populations. Arch Osteoporos. 2012. https://www.iofbonehealth.org/facts-and-statistics/vitamin-d-studies-map

[53] Spiro A, Buttriss JL. Vitamin D: an overview of vitamin D status and intake in Europe. Nutr Bull. 2014;39 (4):322–350.

[54] Lips P, Cashman K, Lamberg-Allardt C, et al. Current vitamin D status in European and Middle East countries and strategies to prevent vitamin D deficiency: a position statement of the European Calcified Tissue Society. Eur J Endocrinol. 2019;180(4):P23–P54.

[55] Pludowski P, Grant WB, Bhattoa HP, et al. Vitamin d status in central Europe. Int J Endocrinol. 2014;2014:589587.

[56] Kuchuk NO, van Schoor NM, Pluijm SM, et al. Vitamin D status, parathyroid function, bone turnover, and BMD in postmenopausal women with osteoporosis: global perspective. J Bone Miner Res. 2009;24(4):693–701.

[57] In practice we choose, as the first day, the one in which the number of cases N_i is closest to 30. In some countries, such a number N_i is repeated for several days; in such cases we choose the last of such days as the starting point. For the particular case of China, we started from January 16th, with 59 cases, since the number before that day was essentially frozen.

[58] Fioletov V, Kerr B. Canadian journal of public health. Revue canadienne de sante publique. 2010;101(4):I5–9.

[59] Grunbaum A. BabyMed https://www.babymed.com/pregnancy/blood-type-and-rh-rhesus-status-countries

[60] Wacker M, Holick M. Sunlight and Vitamin D. Dermatoendocrinol. 2013;5(1):51–108.

[61] Semba RD, Houston DK, Bandinelli S, et al. Relationship of 25-hydroxyvitamin D with all-cause and cardiovascular disease mortality in older community-dwelling adults. Eur J Clin Nutr. 2010;64(2):203–209.

[62] Wahl D, Cooper C, Ebeling PR. A global representation of vitamin D status in healthy populations. Arch Osteoporos. 2012;7(12):155–172.

[63] Dipta T, Iqbal M, Hossain A, et al. Distribution Of Phenotypic And Genotypic Abo And Rhesus Blood Groups Among Bangladeshi Population. Ibrahim Med College J. 1970;5(2):59–62.

[64] Jou R, Huang W-L, Su W-J. Tokyo-172 BCG vaccination complications, Taiwan. Emerg Infect Dis. 2009;15 (9):1525?1526.

Appendix A: Vitamin D

We collected most data on vitamin D from [52–64] and from references therein. For a first dataset of 50 countries, we have collected annual averages. For many countries, several studies with different values were found, and in this case, we have collected the mean and the standard error (when available) and a weighted average has been performed. The resulting values that we have used are listed in Table 34. The sample size here is smaller and such dataset is based on quite inhomogeneous research and thus should be confirmed by a more complete dataset.