COVID-19 transmission risk factors

Alessio Notari[1*] and Giorgio Torrieri[2†]
[1] *Departament de Física Quàntica i Astrofisíca & Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain and*
[2] *Instituto de Fisica Gleb Wataghin (IFGW), Unicamp, 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 α 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), UV index (0.004, smaller sample, 73 countries), low Vitamin D serum levels ($0.002 - 0.006$, smaller sample, ~ 50 countries). There is highly significant correlation also with blood type: positive correlation with types RH- ($2{\cdot}10^{-5}$) and A+ ($2{\cdot}10^{-3}$), 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. Other variables are found to have a counterintuitive *negative* correlation, which may be explained due their strong negative correlation with life expectancy: *slower* spread of COVID-19 is correlated with high death-rate due to pollution, prevalence of anemia and hepatitis B, high blood pressure in females. 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 discuss correlation with the above variables.

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 in order to find causal factors and it can help policy makers in their decisions.

Some factors are intuitive and have been found in other studies, such as temperature [1, 3–9] (see also [10] for a different conclusion) and air travel [2, 10]; 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.

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.

*Electronic address: notari@fqa.ub.edu
†Electronic address: torrieri@ifi.unicamp.br

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 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 [12] on April 15th. 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 $N_i = 30$ [54];
- 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. In the fit we used Poissonian errors, given by $\sqrt{N}$, on the daily counting of cases.

Note that the statistical errors on the exponents α, considering Poissonian errors on the daily counting of cases, 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 [13], 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);
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);
14. Vitamin D serum levels (negative correlation, annual values p-value 0.006, seasonal values 0.002; smaller sample, ~ 50 countries).

3. *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 0.023);
3. Vaccination coverage for Tuberculosis (BCG) (negative correlation, p-value 0.028).

4. *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 $2.4 \cdot 10^{-4}$);
5. PM2.5 air pollution (negative correlation, p-value 0.029).

A. Bias due to GDP: lack of testing?

We also find a correlation with GDP per capita, which we 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.

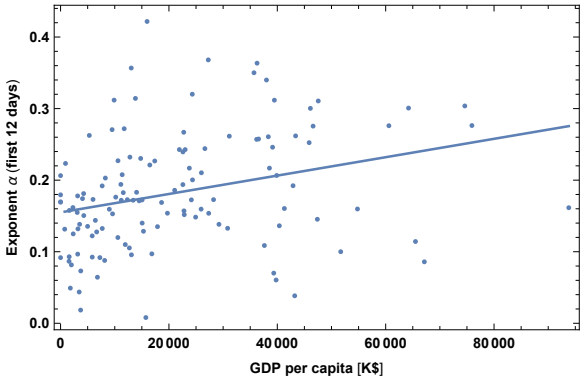

Figure 1: Exponent α for each country vs. GDP per capita. We show the data points and the best-fit for the linear interpolation.

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

R^2	0.087
N	121

Table I: 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.

We performed 2-variables fits, including GDP and each of the above significant variables, in order 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:

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).

6. Starting day of the epidemic (negative correlation, p-value 0.0037);

7. Lung cancer death rates (positive correlation, p-value 0.0039);

8. Life expectancy (positive correlation, p-value 0.0048);

2. Significant variables, moderate evidence

We find moderate evidence for:

1. UV index (negative correlation, p-value 0.01; smaller sample, 73 countries);

2. Type-I diabetes in children, 0-19 years-old (negative correlation, p-value 0.01);

3. Vaccination coverage for Tuberculosis (BCG) (negative correlation, p-value 0.023);

4. Obesity in males (positive correlation, p-value 0.02);

5. CO_2 emissions (positive correlation, p-value 0.02);

6. Alcohol consumption (positive correlation, p-value 0.03);

7. Daily smoking prevalence (positive correlation, p-value 0.03);

8. Share of population in urban areas (positive correlation, p-value 0.04);

3. 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);

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). 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

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 fig. 2 and Table II.

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.

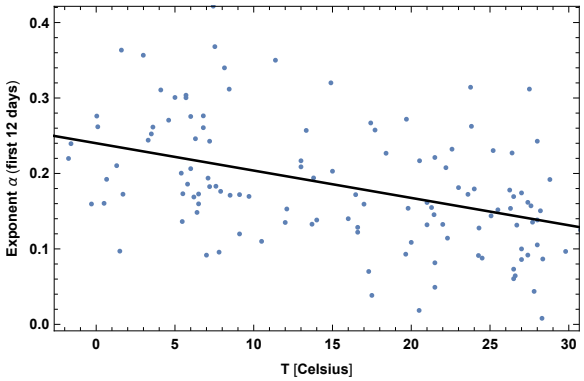

Figure 2: Exponent α for each country vs. average temperature T, for the relevant period of time, as defined in [1]. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.239	0.0124	19.2	4.64×10^{-39}	
T	-0.00359	0.000676	-5.32	4.73×10^{-7}	7.1×10^{-5}

R^2	0.186
N	126

	Estimate	Standard Error	t-Statistic	p-value
1	0.223	0.0188	11.9	7.45×10^{-22}
GDP	5.62×10^{-7}	3.93×10^{-7}	1.43	0.155
T	-0.0033	0.000761	-4.33	0.0000311

R^2	0.212
N	121
Cross-correlation	0.425

Table II: 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 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 α 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.

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 fig. 3 and Table III. This is an interesting finding, since it suggests that 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 children, is more immune to COVID-19, or less contagious.

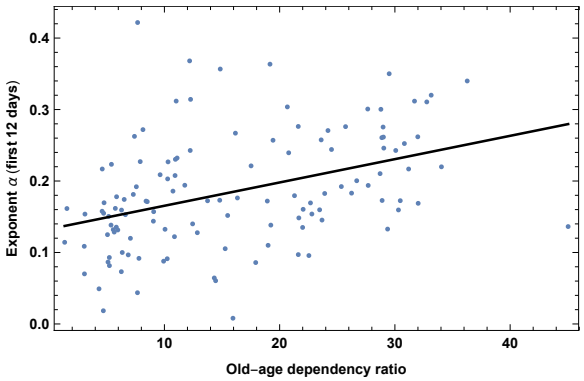

Figure 3: Exponent α for each country vs. old-age dependency ratio, as defined in the text. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.132	0.0126	10.5	9.03×10^{-19}	
OLD	0.00326	0.000669	4.87	3.37×10^{-6}	0.0025

R^2	0.164
N	123

	Estimate	Standard Error	t-Statistic	p-value
1	0.126	0.0132	9.56	2.42×10^{-16}
GDP	7.18×10^{-7}	4.04×10^{-7}	1.78	0.0778
OLD	0.0027	0.00076	3.55	0.000557

R^2	0.188
N	120
Cross-correlation	-0.4561

Table III: 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.

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 [13] 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 fig. 4 and Table IV.

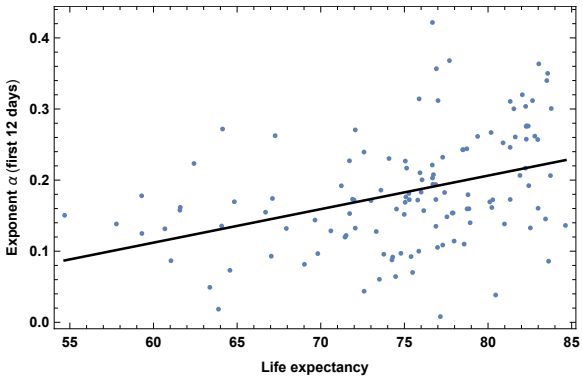

Figure 4: Exponent α for each country vs. life expectancy. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	-0.147	0.0716	-2.05	0.0424	
LIFE	0.00446	0.00096	4.65	8.56×10^{-6}	0.0041

R^2	0.151
N	125

	Estimate	Standard Error	t-Statistic	p-value
1	-0.11	0.0929	-1.19	0.237
LIFE	0.00386	0.00134	2.87	0.00485
GDP	3.99×10^{-7}	4.98×10^{-7}	0.801	0.424

R^2	0.160
N	120
Cross-correlation	-0.679

Table IV: 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.

4. International tourism: number of arrivals

The dataset is for year 2016. Results are shown in fig. 5 and Table V. As expected, more tourists correlate with higher speed of contagion. This is in agreement with [2, 10], 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.

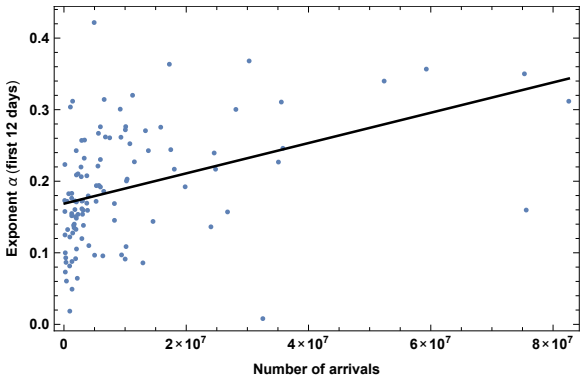

Figure 5: Exponent α for each country vs. number of tourist arrivals. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.168	0.00846	19.9	7.38×10^{-38}	
ARR	2.13×10^{-9}	4.55×10^{-10}	4.69	8.12×10^{-6}	0.0002

R^2	0.169
N	110

	Estimate	Standard Error	t-Statistic	p-value
1	0.146	0.0116	12.6	1.48×10^{-22}
GDP	1.11×10^{-6}	4.02×10^{-7}	2.76	0.00691
ARR	1.77×10^{-9}	4.68×10^{-10}	3.78	0.000265

R^2	0.226
N	107
Cross-correlation	-0.288

Table V: 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 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 α 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.

5. *Starting date of the epidemic*

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

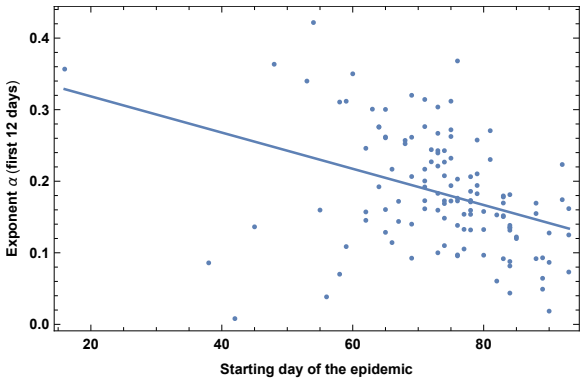

Figure 6: Exponent α for each country vs. starting date of the analysis of the epidemic, DATE, defined as the day when the positive cases reached $N = 30$. Days are counted from Dec 31st 2019. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.369	0.0421	8.76	1.21×10^{-14}	
DATE	-0.00253	0.000565	-4.48	0.0000168	0.011

R^2	0.139
N	126

	Estimate	Standard Error	t-Statistic	p-value
1	0.32	0.0568	5.64	1.18×10^{-7}
GDP	5.85×10^{-7}	4.38×10^{-7}	1.33	0.185
DATE	-0.00204	0.000689	-2.96	0.00367

R^2	0.151
N	121
Cross-correlation	0.539

Table VI: 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 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 α 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.

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 [28].

1. No or little physical contact, bowing. In this group we have: Bangladesh, Cambodia, Japan, Korea South, Sri Lanka, 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, 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, Kingdom United, States United.

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, Uzbekistan.

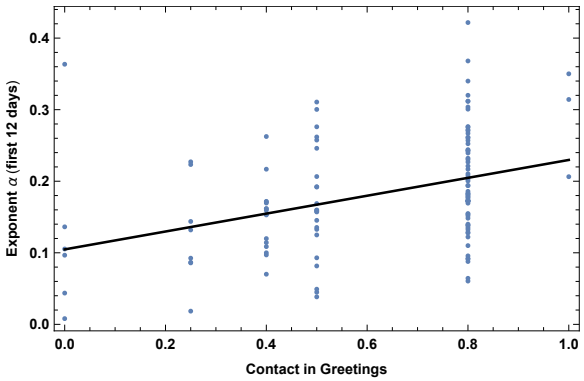

Figure 7: Exponent α for each country vs. level of contact in greeting habits, GRE, as defined in the text. We show the data points and the best-fit for the linear interpolation.

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, Venezuela.

6. Handshaking and kissing. In this group we have: Andorra, Brazil, 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 Fig. 7 and Table VII.

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 $R^2 \approx 0.23$.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.111	0.019	5.8	5.47×10^{-8}	
GRE	0.12	0.0287	4.2	0.000051	0.0029

R^2	0.129
N	121

	Estimate	Standard Error	t-Statistic	p-value
1	0.079	0.0204	3.87	0.000184
GDP	1.26×10^{-6}	3.65×10^{-7}	3.45	0.000794
GRE	0.128	0.0282	4.53	0.0000149

R^2	0.22
N	116
Cross-correlation	0.00560

Table VII: 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.

7. *Lung cancer death rates*

This dataset refers to year 2002. Results are shown in Fig. 8 and Table VIII. 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 [14].

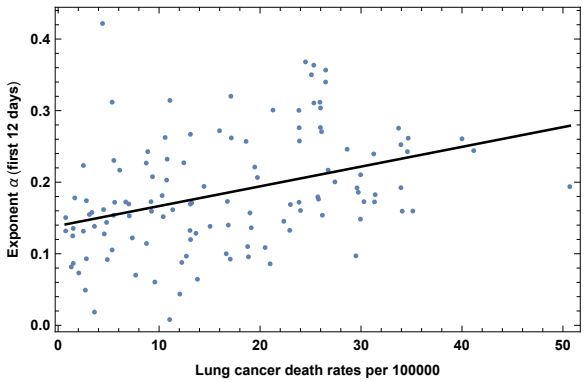

Figure 8: Exponent α for each country vs. lung cancer death rates. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.143	0.0121	11.8	9.96×10^{-22}	
LUNG	0.00159	0.000381	4.17	0.0000572	0.024

R^2	0.127
N	121

	Estimate	Standard Error	t-Statistic	p-value
1	0.133	0.013	10.2	7.48×10^{-18}
GDP	8.93×10^{-7}	4.02×10^{-7}	2.22	0.0282
LUNG	0.00122	0.000416	2.94	0.00396

R^2	0.164
N	118
Cross-correlation	-0.403

Table VIII: 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.

8. *Obesity in males*

This refers to the prevalence of obesity in adult *males*, measured in 2014. Results are shown in fig. 9 and Table IX. 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 when excluding countries below 5K$ 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 [15] for increased risk of severe COVID-19 symptoms for obese patients.

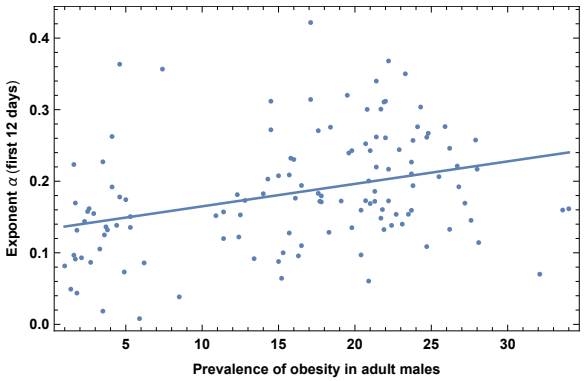

Figure 9: Exponent α for each country vs. prevalence of obesity in adult males. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.134	0.0144	9.29	6.9×10^{-16}	
OBE	0.00314	0.000788	3.98	0.000115	0.12

R^2	0.114
N	125

	Estimate	Standard Error	t-Statistic	p-value
1	0.132	0.0148	8.89	8.35×10^{-15}
GDP	5.6×10^{-7}	4.84×10^{-7}	1.16	0.25
OBE	0.00249	0.00106	2.35	0.0202

R^2	0.128
N	121
Cross-correlation	-0.634

Table IX: 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 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 α 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.

9. Urbanization

This is the share of population living in urban areas, collected in year 2017. Results are shown in Fig. 10 and Table X. This is an expected correlation, in agreement with [19, 20].

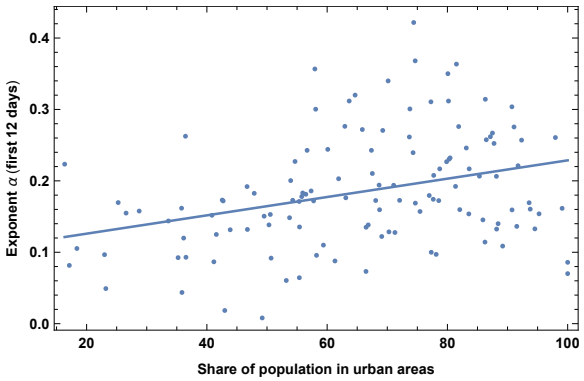

Figure 10: Exponent α for each country vs. share of population in urban areas. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.1	0.0228	4.4	0.0000231	
URB	0.00128	0.000331	3.88	0.000173	0.057

R^2	0.109
N	124

	Estimate	Standard Error	t-Statistic	p-value
1	0.108	0.0247	4.37	0.0000267
GDP	7.57×10^{-7}	4.74×10^{-7}	1.6	0.113
URB	0.000916	0.000439	2.09	0.0391

R^2	0.133
N	120
Cross-correlation	-0.6216

Table X: 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.

10. Alcohol consumption

This dataset refers to year 2016. Results are shown in Fig. 11 and Table XI. 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.

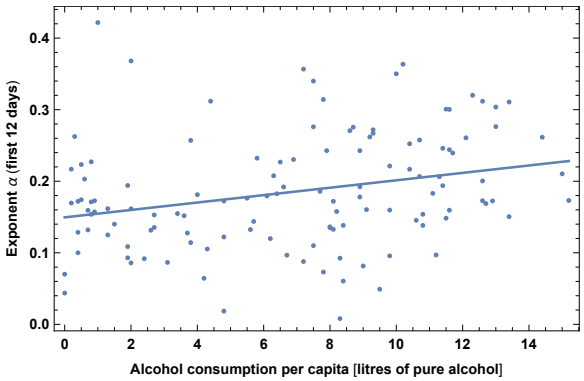

Figure 11: Exponent α for each country vs. alcohol consumption. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.15	0.0131	11.4	7.8×10^{-21}	
ALCO	0.00518	0.00164	3.17	0.00195	0.012

R^2	0.076
N	126

	Estimate	Standard Error	t-Statistic	p-value
1	0.134	0.0141	9.48	3.75×10^{-16}
GDP	1.12×10^{-6}	3.86×10^{-7}	2.91	0.00437
ALCO	0.00382	0.0017	2.25	0.0264

R^2	0.138
N	120
Cross-correlation	-0.286

Table XI: 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.

11. Smoking

This dataset refers to year 2012. Results are shown in Fig. 12 and Table XII. 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 5K$ 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 negative effects on conditions of lungs that facilitates contagion or that it contributes to increased transmission of virus from hand to mouth [16]. Interestingly, note that our finding is in contrast with claims of a possible protective effect of nicotine and smoking against COVID-19 [17, 18].

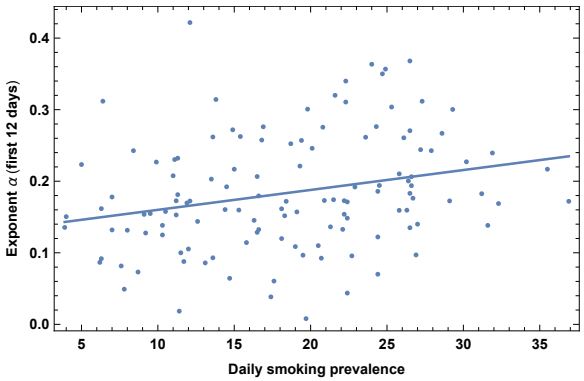

Figure 12: Exponent α for each country vs. daily smoking prevalence. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.132	0.0187	7.07	1.07×10^{-10}	
SMOK	0.00278	0.000937	2.97	0.00361	0.19

R^2	0.067
N	124

	Estimate	Standard Error	t-Statistic	p-value
1	0.121	0.019	6.38	3.68×10^{-9}
GDP	1.06×10^{-6}	3.89×10^{-7}	2.73	0.00729
SMOK	0.00209	0.000966	2.16	0.0328

R^2	0.122
N	121
Cross-correlation	-0.2646

Table XII: 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.

12. 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 [21], 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 Fig. 13 and Table XIII. 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 both.

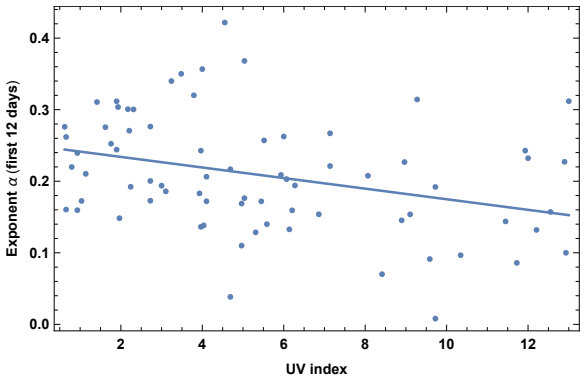

Figure 13: Exponent α for each country vs. UV index for the relevant month of the epidemic. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.249	0.0163	15.3	3.7×10^{-24}	
UV	-0.0074	0.00249	-2.97	0.00408	0.012

R^2	0.110
N	73

	Estimate	Standard Error	t-Statistic	p-value
1	0.242	0.0269	9.01	2.88×10^{-13}
GDP	1.92×10^{-7}	5.8×10^{-7}	0.33	0.742
UV	-0.00707	0.00274	-2.58	0.0119

R^2	0.112
N	73
Cross-correlation	0.383

Table XIII: 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.

13. Vitamin D serum concentration

Another relevant variable is the amount of serum Vitamin D. We have 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 Fig. 14 and Table XXXII for the annual levels and in Fig. 14 and Table XXXII for the seasonal levels.

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

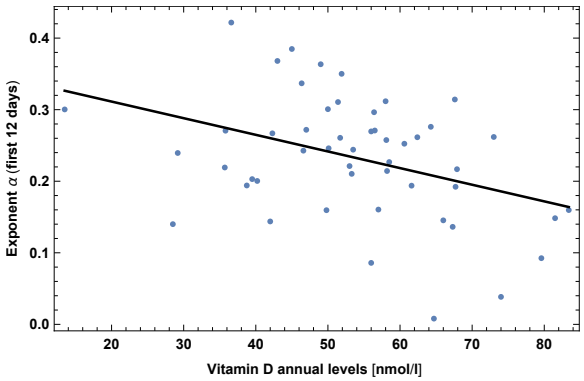

Figure 14: Exponent α for each country vs. annual levels of vitamin D, for the relevant period of time, as defined in the text, for the base set of 42 countries. We show the data points and the best-fit for the linear interpolation.

COVID-19 [45–47].

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.356	0.0443	8.02	2.02×10^{-10}	
D	-0.00231	0.000801	-2.88	0.00586	0.0059

R^2	0.147
N	50

	Estimate	Standard Error	t-Statistic	p-value
1	0.342	0.0457	7.48	1.52×10^{-9}
GDP	8.29×10^{-7}	6.99×10^{-7}	1.19	0.242
D	-0.00255	0.000822	-3.1	0.00328

R^2	0.172
N	50
Cross-correlation	-0.243

Table XIV: 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 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 α 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.

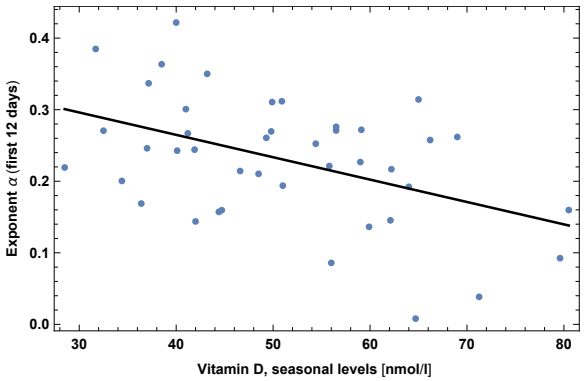

Figure 15: Exponent α for each country vs. seasonal levels of vitamin D, for the relevant period of time, as defined in the text, for the base set of 42 countries. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.385	0.0499	7.72	1.91×10^{-9}	
D_s	-0.00305	0.000949	-3.22	0.00256	0.0024

R^2	0.206
N	42

	Estimate	Standard Error	t-Statistic	p-value
1	0.375	0.0533	7.03	1.94×10^{-8}
GDP	4.66×10^{-7}	8.02×10^{-7}	0.581	0.565
D_s	-0.00314	0.00097	-3.24	0.00243

R^2	0.212
N	42
Cross-correlation	-0.162

Table XV: 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 mean annual 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 α 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.

14. 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 Fig. 16 and Table XVI.

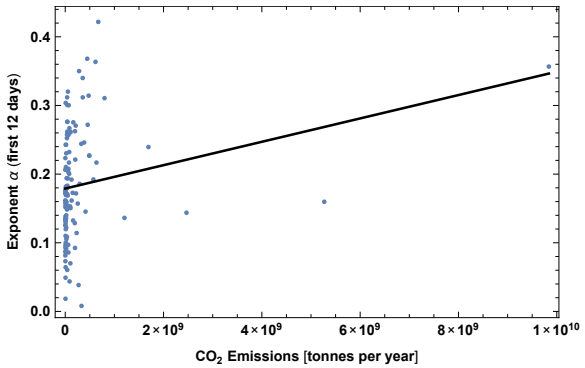

Figure 16: Exponent α for each country vs. CO_2 emissions. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.18	0.0073	24.	9.6×10^{-49}	
CO_2	1.7×10^{-11}	6.9×10^{-12}	2.5	0.015	0.048

R^2	0.048
N	110

	Estimate	Standard Error	t-Statistic	p-value
1	0.152	0.011	13.8	1.82×10^{-26}
GDP	1.23×10^{-6}	3.75×10^{-7}	3.29	0.00133
CO2	1.57×10^{-11}	6.74×10^{-12}	2.33	0.0214

R^2	0.126
N	109
Cross-correlation	-0.0321

Table XVI: 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 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 α 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.

15. Prevalence of type-1 Diabetes

This is the prevalence of type-1 Diabetes in children, 0-19 years-old, taken from [23]. Results are shown in Fig. 17 and Table XXIII. Note however that significance becomes very small when restricting to countries with GDP per capita larger than 5K$. Note also that in the case of diabetes of *any* kind we do *not* find a correlation with COVID-19. Such correlation, even if not highly significant, could be non-trivial and could constitute useful information for clinical and genetic research. See also [49].

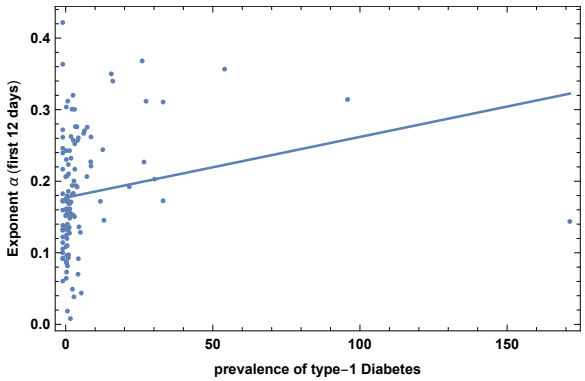

Figure 17: Exponent α for each country vs. prevalence of type-1 Diabetes. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.177	0.00793	22.4	2.35×10^{-41}	
DIAB	0.00087	0.000367	2.37	0.0198	0.072

R^2	0.052
N	105

	Estimate	Standard Error	t-Statistic	p-value
1	0.142	0.0116	12.3	1.52×10^{-21}
GDP	1.58×10^{-6}	3.88×10^{-7}	4.08	0.000092
DIAB	0.000932	0.000349	2.67	0.00889

R^2	0.189
N	101
Cross-correlation	0.0415

Table XVII: 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.

16. Tuberculosis (BCG) vaccination coverage

This dataset is the vaccination coverage for tuberculosis for year 2015 [55]. Results are shown in Fig. 18 and Table XVIII. 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 [29–31]) and even for vaccine development [32–34].

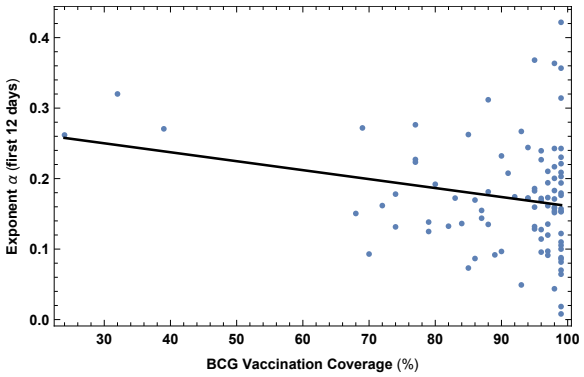

Figure 18: Exponent α for each country vs. BCG vaccination coverage. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.289	0.053	5.46	3.99×10^{-7}	
BCG	-0.00129	0.000579	-2.22	0.0286	0.011

R^2	0.051
N	94

	Estimate	Standard Error	t-Statistic	p-value
1	0.28	0.0534	5.24	1.07×10^{-6}
GDP	8.53×10^{-7}	4.79×10^{-7}	1.78	0.0781
BCG	-0.00134	0.00058	-2.3	0.0235

R^2	0.084
N	92
Cross-correlation	-0.0367

Table XVIII: 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 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 α 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.

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 Fig. 19 and Table XIX. Contrary to naive expectations and to claims in the opposite direction [48], we find that countries with larger death rate from air pollution actually have *slower* COVID-19 contagion.

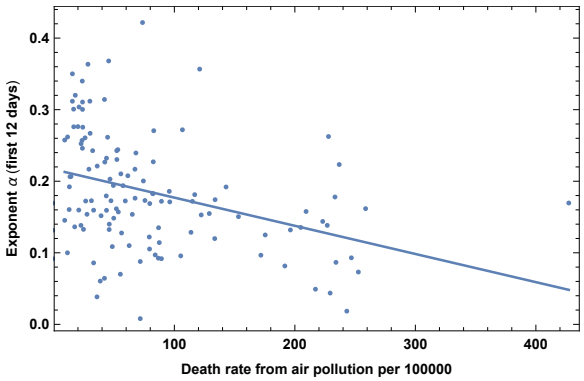

Figure 19: Exponent α for each country vs. death rate from air pollution per 100000. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.216	0.0102	21.3	8.97×10^{-43}	
POLL	-0.000394	0.0000917	-4.29	0.0000357	0.040

R^2	0.132
N	123

	Estimate	Standard Error	t-Statistic	p-value
1	0.211	0.0204	10.3	3.75×10^{-18}
GDP	3.13×10^{-7}	4.77×10^{-7}	0.655	0.514
POLL	-0.000418	0.000131	-3.19	0.00185

R^2	0.158
N	120
Cross-correlation	0.630

Table XIX: 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 death rate from air pollution per 100000 (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 α 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.

2. High blood pressure in females

This dataset is for year 2015. Results are shown in Fig. 20 and Table XX. 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*.

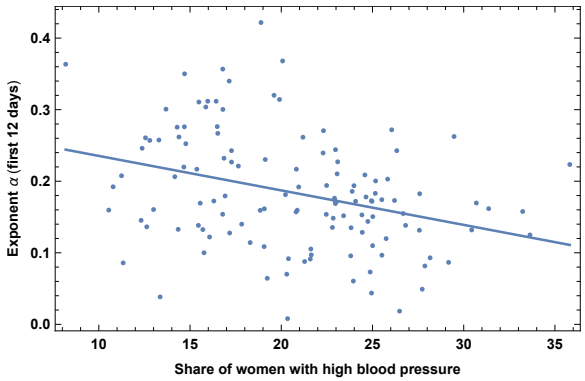

Figure 20: Exponent α for each country vs. share of women with high blood pressure. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.284	0.0265	10.7	2.76×10^{-19}	
PRE	-0.00482	0.00124	-3.89	0.000164	0.028

R^2	0.109
N	125

	Estimate	Standard Error	t-Statistic	p-value
1	0.248	0.0432	5.75	7.2×10^{-8}
GDP	5.85×10^{-7}	4.89×10^{-7}	1.2	0.234
PRE	-0.00372	0.00168	-2.22	0.0284

R^2	0.123
N	121
Cross-correlation	0.642

Table XX: 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 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 α 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.

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 Fig. 21 and Table XXI. Countries with higher incidence of hepatitis B have *slower* contagion of COVID-19.

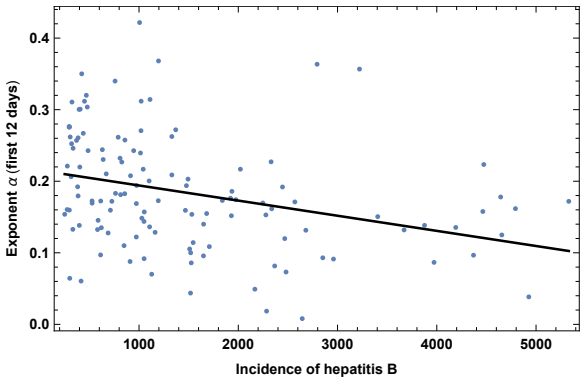

Figure 21: Exponent α for each country vs. incidence of Hepatitis B, for the relevant period of time, as defined in the text, for the base set of 42 countries. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.215	0.0107	20.1	1.16×10^{-40}	
HEP	-0.000021	5.56×10^{-6}	-3.78	0.000243	0.016

R^2	0.104
N	125

	Estimate	Standard Error	t-Statistic	p-value
1	0.189	0.017	11.1	5.02×10^{-20}
GDP	8.39×10^{-7}	4.1×10^{-7}	2.05	0.0429
HEP	-0.0000161	6.21×10^{-6}	-2.58	0.011

R^2	0.136
N	121
Cross-correlation	0.420

Table XXI: 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 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 α 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.

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 Fig. 22 and Table XXII. The 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.

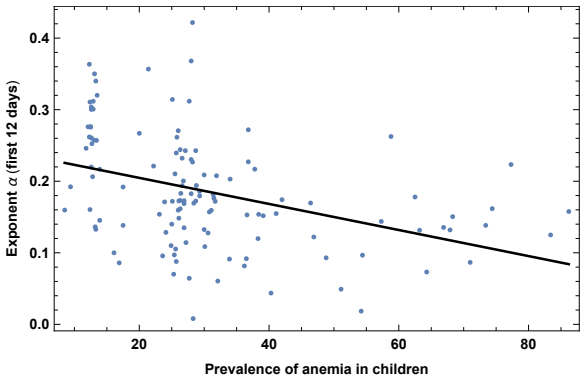

Figure 22: Exponent α for each country vs. prevalence of anemia in children. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.24	0.0136	17.7	1.58×10^{-35}	
ANE	-0.00181	0.000386	-4.7	6.95×10^{-6}	0.0014

R^2	0.153
N	124

	Estimate	Standard Error	t-Statistic	p-value
1	0.22	0.0249	8.83	1.23×10^{-14}
GDP	4.97×10^{-7}	4.74×10^{-7}	1.05	0.297
ANE	-0.00148	0.000514	-2.89	0.0046

R^2	0.161
N	120
Cross-correlation	0.638

Table XXII: 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.

B. Blood types

Blood types are not equally distributed in the world and thus we have correlated them with α. Data were taken from [50]. 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 [26], (ii) we find higher risk for group 0- and no correlation for group 0+ (while [26] finds lower risk higher risk for groups 0), (iii) we have a strong significance for lower risk for RH+ 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.

1. *Type A+*

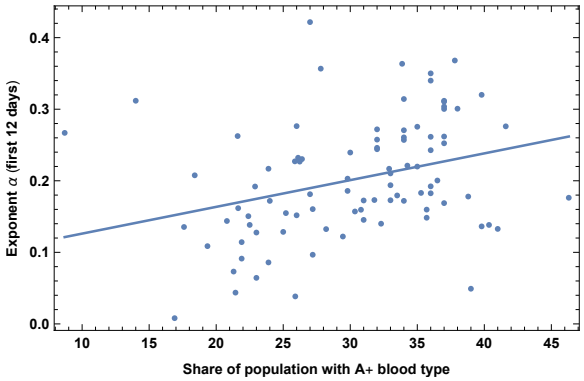

Figure 23: Exponent α for each country vs. percentage of population with blood type A+. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.0892	0.0367	2.43	0.0172	
A+	0.00372	0.00118	3.15	0.00225	0.011

R^2	0.100
N	91

	Estimate	Standard Error	t-Statistic	p-value
1	0.0907	0.0363	2.5	0.0144
GDP	9.42×10^{-7}	4.8×10^{-7}	1.96	0.0528
A+	0.00292	0.00124	2.34	0.0214

R^2	0.139
N	90
Cross-correlation	-0.340

Table XXIII: 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.

2. *Type B+*

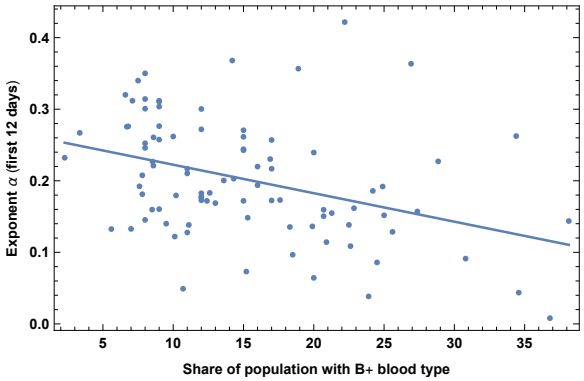

Figure 24: Exponent α for each country vs. percentage of population with blood type B+. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.262	0.0176	14.9	7.3×10^{-26}	
B+	-0.00398	0.00104	-3.84	0.000233	0.0014

R^2	0.142
N	91

	Estimate	Standard Error	t-Statistic	p-value
1	0.232	0.0237	9.78	1.15×10^{-15}
GDP	9.15×10^{-7}	4.6×10^{-7}	1.99	0.0495
B+	-0.00341	0.00107	-3.18	0.00202

R^2	0.181
N	90
Cross-correlation	0.285

Table XXIV: 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 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 α 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.

3. Type 0-

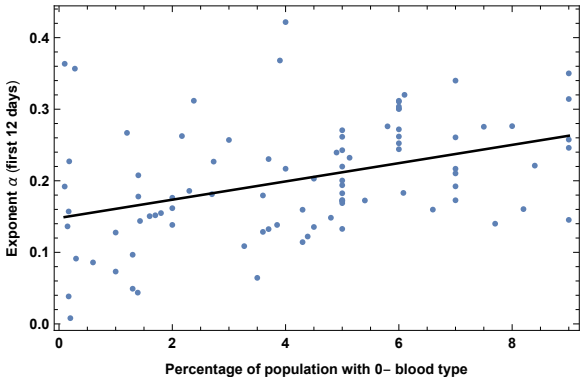

Figure 25: Exponent α for each country vs. percentage of population with blood type 0-. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.148	0.0155	9.52	3.1×10^{-15}	
0-	0.0127	0.00316	4.04	0.000115	0.0058

R^2	0.155
N	91

	Estimate	Standard Error	t-Statistic	p-value
1	0.14	0.0166	8.45	5.86×10^{-13}
GDP	6.84×10^{-7}	4.9×10^{-7}	1.4	0.166
0-	0.0106	0.0035	3.04	0.00315

R^2	0.173
N	90
Cross-correlation	-0.430

Table XXV: 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.

4. *Type A-*

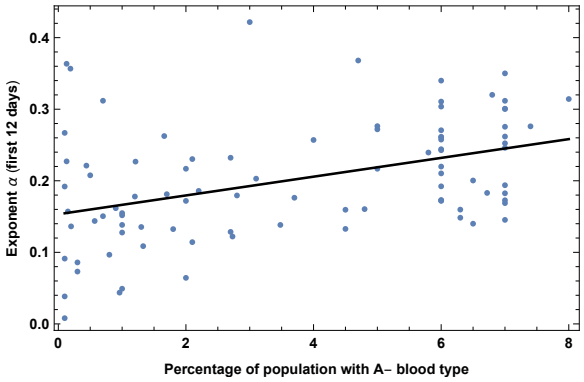

Figure 26: Exponent α for each country vs. percentage of population with blood type A-. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.153	0.0136	11.3	7.3×10^{-19}	
A-	0.0131	0.00298	4.38	0.0000327	0.0027

R^2	0.177
N	91

	Estimate	Standard Error	t-Statistic	p-value
1	0.146	0.0151	9.6	2.59×10^{-15}
GDP	$6. \times 10^{-7}$	4.88×10^{-7}	1.23	0.222
A-	0.0112	0.00334	3.37	0.00113

R^2	0.191
N	90
Cross-correlation	-0.441

Table XXVI: 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.

5. *Type B-*

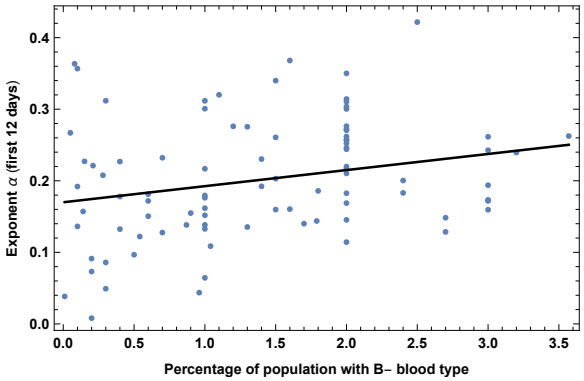

Figure 27: Exponent α for each country vs. percentage of population with blood type B-. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.17	0.0154	11.	2.35×10^{-18}	
B-	0.0224	0.00908	2.47	0.0156	0.14

R^2	0.064
N	91

	Estimate	Standard Error	t-Statistic	p-value
1	0.147	0.0176	8.36	9.28×10^{-13}
GDP	1.16×10^{-6}	4.62×10^{-7}	2.51	0.0139
B-	0.0183	0.00901	2.03	0.0451

R^2	0.127
N	90
Cross-correlation	-0.175

Table XXVII: 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 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 α 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.

6. *Type AB-*

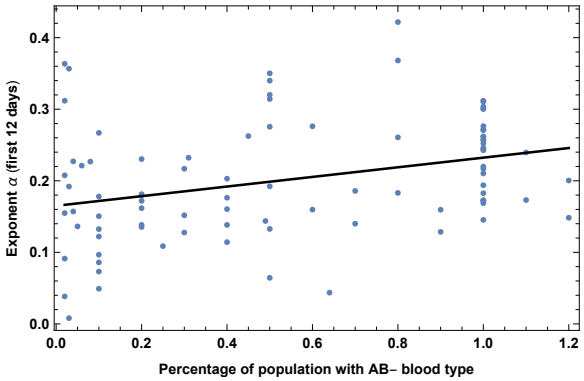

Figure 28: Exponent α for each country vs. percentage of population with blood type AB-. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.165	0.014	11.8	8.29×10^{-20}	
AB-	0.0668	0.0208	3.22	0.00181	0.027

R^2	0.104
N	91

	Estimate	Standard Error	t-Statistic	p-value
1	0.151	0.0159	9.48	4.73×10^{-15}
GDP	9.18×10^{-7}	4.84×10^{-7}	1.9	0.0612
AB-	0.0517	0.0222	2.33	0.0221

R^2	0.139
N	90
Cross-correlation	-0.360

Table XXVIII: 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.

7. *RH-positive*

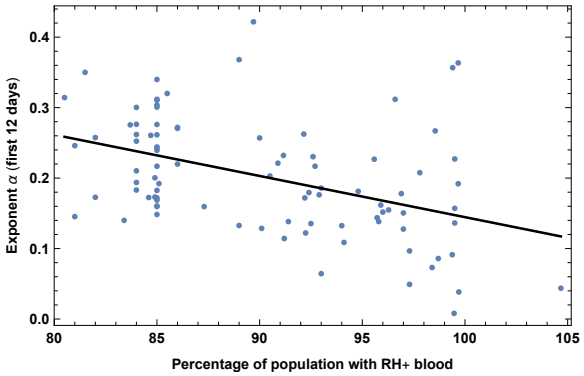

Figure 29: Exponent α for each country vs. percentage of population with RH-positive blood. We show the data points and the best-fit for the linear interpolation.

	Estimate	Standard Error	t-Statistic	p-value	p-value, GDP>5K$
1	0.727	0.116	6.28	1.22×10^{-8}	
RH+	-0.00583	0.00128	-4.55	0.0000171	0.0035

R^2	0.188
N	91

	Estimate	Standard Error	t-Statistic	p-value
1	0.646	0.135	4.79	6.8×10^{-6}
GDP	5.74×10^{-7}	4.84×10^{-7}	1.19	0.238
RH+	-0.00508	0.00143	-3.55	0.000625

R^2	0.201
N	90
Cross-correlation	0.437

Table XXIX: 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.

V. CROSS-CORRELATIONS

In this section we first perform linear fits of α with each possible pair of variables (excluding the ones which were not significant when combined with GDP per capita and considering only RH+ and B+ for blood types). We show the correlation coefficients between the two variables, for each pair, in Fig. 30. We also show the p-value of the t-statistic of each pair of variables and the total R^2 of such fits in Fig. 31.

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

	T	OLD	LIFE	ARR	DATE	GRE	LUNG	OBE	URB	UV	GDP	ALCO	SMOK	ANE	POLL	HEP	PRE	D	D_s	CO_2	DIAB	BCG	RH+	$B+$	α	$p-$value
T	1	0.7	0.53	0.27	−0.38	0.25	0.72	0.5	0.33	−0.93	0.43	0.54	0.61	−0.61	−0.42	−0.45	−0.32	−0.071	−0.29	0.18	5×10^{-3}	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	0.69	−0.46	−0.71	−0.57	0.66	0.59	0.6	0.49	−0.25	0.018	−0.054	0.011	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	0.41	−0.68	−0.32	−0.48	0.9	0.84	0.74	0.78	−0.29	−0.16	−0.076	−0.035	−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	0.13	−0.29	−0.18	−0.23	0.34	0.25	0.16	0.34	−0.051	4×10^{-3}	−0.54	−0.29	−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	−0.031	0.54	0.2	0.31	−0.57	−0.46	−0.19	−0.57	0.14	-6×10^{-3}	0.54	0.26	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	0.38	6×10^{-3}	−0.34	−0.2	0.27	0.36	0.47	0.13	0.2	0.19	0.11	0.026	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	0.63	−0.4	−0.6	−0.76	0.6	0.5	0.48	0.32	−0.08	0.15	−0.11	0.036	−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	0.5	−0.63	−0.34	−0.44	0.69	0.73	0.65	0.56	-7×10^{-4}	0.032	-3×10^{-4}	0.067	−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	0.16	−0.62	−0.17	−0.29	0.63	0.71	0.59	0.77	−0.11	−0.13	−0.031	0.031	−0.071	0.36	0.5	0.33	2×10^{-4}
UV	−0.93	0.69	0.41	0.13	−0.031	0.38	0.63	0.5	0.16	1	0.38	0.6	0.47	−0.49	−0.33	−0.36	−0.11	−0.077	−0.25	0.022	−0.18	−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	0.38	1	−0.29	−0.26	0.64	0.63	0.42	0.64	−0.24	−0.16	−0.061	0.042	−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	0.6	−0.29	1	−0.41	0.4	0.36	0.39	0.3	−0.28	−0.064	−0.05	-2×10^{-3}	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	0.47	−0.26	−0.41	1	0.54	0.43	0.4	0.22	0.31	0.38	−0.073	0.042	−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	−0.49	0.64	0.4	0.54	1	−0.82	−0.77	−0.78	0.29	0.23	0.12	-7×10^{-3}	0.12	−0.51	−0.43	−0.39	6×10^{-6}
POLL	−0.42	0.59	0.84	0.25	−0.46	0.36	0.5	0.73	0.71	−0.33	0.63	0.36	0.43	−0.82	1	−0.66	−0.75	0.28	0.28	4×10^{-3}	−0.04	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	−0.36	0.42	0.39	0.4	−0.77	−0.66	1	−0.57	−0.077	−0.1	−0.042	0.07	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	−0.11	0.64	0.3	0.22	−0.78	−0.75	−0.57	1	0.32	0.32	0.17	0.029	0.065	−0.15	−0.41	−0.33	2×10^{-4}
D	−0.071	−0.25	−0.29	−0.051	0.14	0.2	−0.08	-7×10^{-4}	−0.11	−0.077	−0.24	−0.28	0.31	0.29	0.28	−0.077	0.32	1	−0.91	−0.012	0.083	0.14	−0.047	3×10^{-3}	−0.38	6×10^{-3}
D_s	−0.29	0.018	−0.16	4×10^{-3}	-6×10^{-3}	0.19	0.15	0.032	−0.13	−0.25	−0.16	−0.064	0.38	0.23	0.28	−0.1	0.32	−0.91	1	0.062	0.074	0.15	−0.035	-5×10^{-3}	−0.46	2×10^{-3}
CO_2	0.18	−0.054	−0.076	−0.54	0.54	0.11	−0.11	-3×10^{-4}	−0.031	0.022	−0.061	−0.05	−0.073	0.12	4×10^{-3}	−0.042	0.17	−0.012	0.062	1	−0.45	−0.062	−0.12	−0.097	0.22	0.016
DIAB	5×10^{-3}	0.011	−0.035	−0.29	0.26	0.026	0.036	0.067	0.031	−0.18	0.042	-2×10^{-3}	0.042	-7×10^{-3}	−0.04	0.07	0.029	0.083	0.074	−0.45	1	−0.029	0.021	−0.18	0.23	0.02
BCG	2×10^{-3}	0.15	−0.2	−0.025	0.11	0.035	−0.11	−0.14	−0.071	−0.23	−0.037	0.13	−0.13	0.12	0.13	0.075	0.065	0.14	0.15	−0.062	−0.029	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	−0.67	0.44	0.55	0.38	−0.51	−0.5	−0.55	−0.15	−0.047	−0.035	−0.12	0.021	−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	−0.34	0.29	0.38	0.056	−0.43	−0.49	−0.45	−0.41	3×10^{-3}	-5×10^{-3}	−0.097	−0.18	−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	−0.33	0.29	0.27	0.26	−0.39	−0.37	−0.32	−0.33	−0.38	−0.46	0.22	0.23	−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}	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}	2×10^{-3}	0.016	0.02	0.031	2×10^{-5}	2×10^{-4}	0	

Figure 30: Correlation coefficients between each pair of variables. 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, 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+), percentage with blood type B+ (B+).

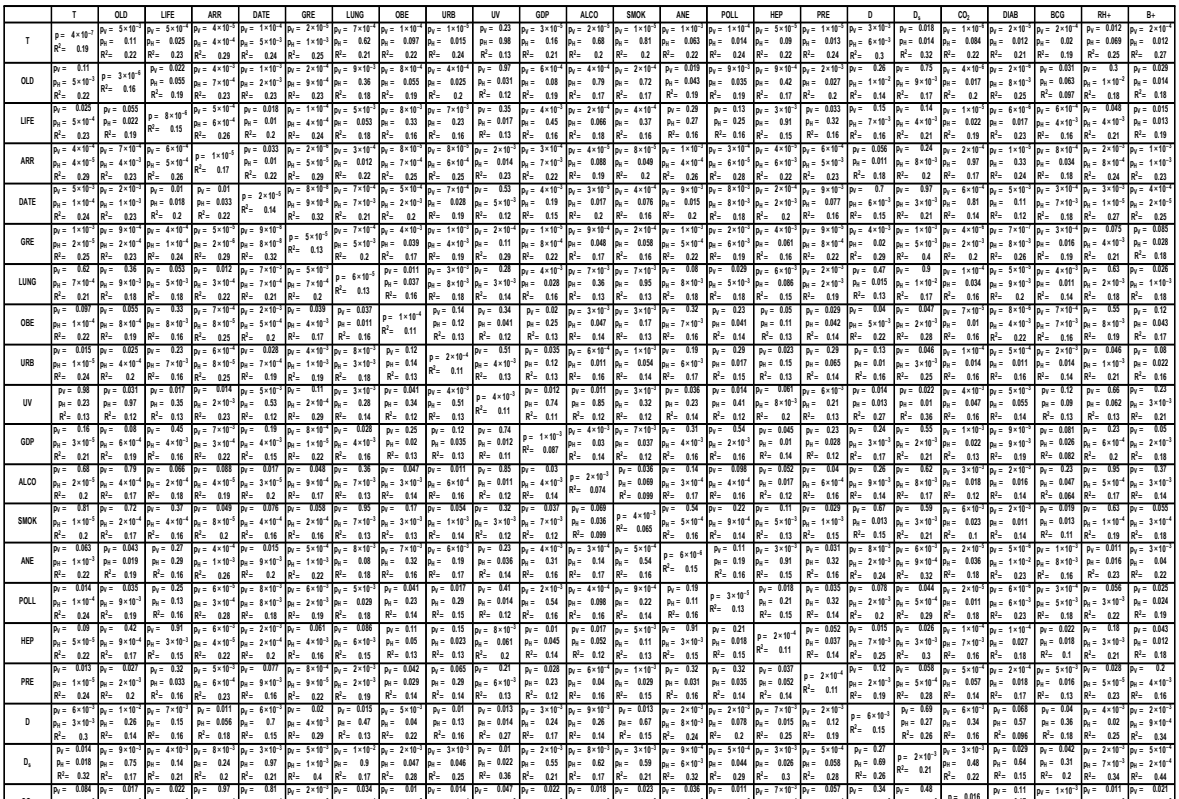

Figure 31: Significance and R^2 for two-variable linear fits of α, with each variable chosen in pairs in each entry. Variables names are the same as in legend of Fig. 30. Each cell in the table gives R^2 and the p-value of the each of the two variables, using t-statistic, labeled as p_H for horizontal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

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 percentage of RH- blood types, earlier starting date of the epidemic, high physical contact in greeting habits, 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, 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 a promising interpretation is indeed that BCG-vaccinated people could be more protected against COVID-19 [29–31].

Lung cancer and alcohol consumption remain rather significant, close to a p-value of 0.05, even after taking into account for 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 (and with old age dependency ratio), however it remains moderately significant when combined with it.

Finally vitamin D levels, which are measured on a smaller sample, also have little correlations with the main factors and might indeed constitute an independent factor. This factor is also quite interesting, since it may open avenues for research on protective factors. It is quite possible that high levels of Vitamin D have an impact on the immune response to COVID-19 [45–47].

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 a 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 3 parameters.

	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
ARR	1.72×10^{-9}	4.24×10^{-10}	4.05	0.0000976
GRE	0.0972	0.0276	3.52	0.000651

R^2	0.36
N	107

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

Table XXX: 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.

2. *Temperature+Arrivals+Greetings+Starting date*

Here we show an example of a fit with 4 parameters. As we combine more than 3 parameters, typically at least one of them becomes less significant.

	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
ARR	7.69×10^{-10}	4.99×10^{-10}	1.54	0.126
GRE	0.13	0.0283	4.59	0.0000127
DATE	-0.00237	0.000728	-3.26	0.00154

R^2	0.42
N	107

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

Table XXXI: 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.

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, CO_2. These variables are defined for a sample of $N = 102$ 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. Moreover clearly many of these variables are redundant. Indeed by diagonalizing the covariance matrix one finds that only about 7 independent linear combinations of the factors are significant.

Rather than reporting such factors, which is not particularly illuminating, we checked that for instance eliminating the following factors reduces the R^2 very little, to about $R^2 \approx 0.47$: ALCO, SMOK, ANE, POLL, CO_2, OBE. Moreover, reducing just to 6 variables, e.g. T, LIFE, ARR, DATE, GRE, LUNG, brings R^2 down to about 0.44. Such a choice however is not unique, due to redundancy of our variables.

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 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), low vitamin D serum levels (annual values p-value 0.006, seasonal values 0.002; smaller sample, ~ 50 countries).

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 $2 \cdot 10^{-4}$), 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. Pre-existing 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 co-morbidities and old age affect immune response and thus may directly increase the growth rate of the contagion. Similarly it is likely that individuals with co-morbidities 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 [53], 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, pre-existing 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 ("Sus-

pended 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 [48]). 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.

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 to increase the intake of vitamin D, 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), type 1 diabetes. This definitely deserves further study, also of correlational type using data from patients.

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

Acknowledgments

GT acknowledges support from FAPESP proc. 2017/06508-7, partecipation 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.

Appendix A: Vitamin D

We collected most data on vitamin D from [35–39] 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 XXXII.

Country	Vit D (annual)	Vit D (seasonal)
Argentina	53	55.8
Australia	66	62.1
Austria	13.5	N/A
Belgium	51.7	49.3
Brazil	67.6	65.
Canada	67.7	64.
Chile	42.3	41.2
China	45	31.7
Croatia	46.6	40.1
Czech Republic	62.4	N/A
Denmark	60.6	54.4
Estonia	49.8	44.7
Finland	58.2	46.6
France	58	50.9
Germany	51.4	49.9
Greece	67.9	62.2
Hungary	61.6	51.
Iceland	57	N/A
India	42	42.
Iran	36.6	40.
Ireland	56.4	N/A
Israel	56.5	56.5
Italy	46.4	37.2
Japan	67.3	59.9
Lebanon	28.5	N/A

Country	Vit D (annual)	Vit D (seasonal)
Lithuania	53.3	48.5
Mexico	58.5	59.
Morocco	39.5	N/A
Netherlands	56	49.8
New Zealand	58.1	66.2
Norway	64.27	56.5
Poland	53.5	41.9
Romania	40.2	34.4
Russia	29.2	N/A
Arabia Saudi	35.7	28.5
Singapore	56	56.
Slovakia	81.5	N/A
South Africa	47	59.1
South Korea	49	38.5
Spain	51.9	43.2
Sweden	73	69.
Switzerland	50	41.
Taiwan	74	71.2
Thailand	64.7	64.7
Tunisia	38.8	N/A
Turkey	43	N/A
Ukraine	35.8	32.5
UK	50.1	37.
USA	83.4	80.5
Vietnam	79.6	79.6

Table XXXII: Vitamin D serum levels (in nmol/l) obtained with a weighted average from refs. [35–39] 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.

[1] Alessio Notari; medRxiv 2020.03.26.20044529; arXiv:2003.12417v4 [q-bio.PE]; doi: https://doi.org/10.1101/2020.03.26.20044529.
[2] T. Fiolet, "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, doi = 10.5281/zenodo.3784948, url = https://doi.org/10.5281/zenodo.3784948;
[3] Demongeot, J.; Flet-Berliac, Y.; Seligmann, H. Temperature Decreases Spread Parameters of the New COVID-19 Case Dynamics. Biology 2020, 9, 94.
[4] Wang, Mao; Jiang, Aili; Gong, Lijuan; Luo, Lina; Guo, Wenbin; Li, Chuyi; Li, Chaoyong; Yang, Bixing; Zeng, Jietong; Chen, Youping; Zheng, Ke; Li, Hongyan. (2020). Temperature significant change COVID-19 Transmission in 429 cities. https://www.medrxiv.org/content/10.1101/2020.02.22.20025791v1. doi: 10.1101/2020.02.22.20025791.
[5] Luo, Wei and Majumder, Maimuna S and Liu, Dianbo and Poirier, Canelle and Mandl, Kenneth D and Lipsitch, Marc and Santillana, Mauricio, "The role of absolute humidity on transmission rates of the COVID-19 outbreak". https://www.medrxiv.org/content/early/2020/02/17/2020.02.12.20022467; doi = 10.1101/2020.02.12.20022467.
[6] Araujo, Miguel B. and Naimi, Babak, "Spread of SARS-CoV-2 Coronavirus likely to be constrained by climate". https://www.medrxiv.org/content/early/2020/03/16/2020.03.12.20034728; doi =

10.1101/2020.03.12.20034728.

[7] Bukhari, Qasim and Jameel, Yusuf, "Will Coronavirus Pandemic Diminish by Summer?" ; Available at SSRN: https://ssrn.com/abstract=3556998; doi = http://dx.doi.org/10.2139/ssrn.3556998.

[8] Jingyuan Wang, Ke Tang, Kai Feng, Weifeng Lv, "High Temperature and High Humidity Reduce the Transmission of COVID-19". 2020. arXiv:2003.05003 [q-bio.PE], doi = http://dx.doi.org/10.2139/ssrn.3556998.

[9] Sajadi, Mohammad M. and Habibzadeh, Parham and Vintzileos, Augustin and Shokouhi, Shervin and Miralles-Wilhelm, Fernando and Amoroso, Anthony, "Temperature, Humidity and Latitude Analysis to Predict Potential Spread and Seasonality for COVID-19"; Available at SSRN: https://ssrn.com/abstract=3550308 or http://dx.doi.org/10.2139/ssrn.3550308.

[10] Marco Tulio Pacheco Coelho, Joao Fabricio Mota Rodrigues, Anderson Matos Medina, Paulo Scalco, Levi Carina Terribile, Bruno Vilela, Jose Alexandre Felizola Diniz-Filho, Ricardo Dobrovolski; medRxiv 2020.04.02.20050773; doi: https://doi.org/10.1101/2020.04.02.20050773.

[11] 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.

[12] `https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases`

[13] Taken from `https://ourworldindata.org/charts`

[14] Clinical characteristics of COVID-19-infected cancer patients: a retrospective case study in three hospitals within Wuhan, China Zhang, L. et al. Annals of Oncology, Volume 0, Issue 0. DOI: https://doi.org/10.1016/j.annonc.2020.03.296.

[15] Kassir, R. Risk of COVID?19 for patients with obesity. Obesity Reviews. 2020; 21:e13034. https://doi.org/10.1111/obr.13034.

[16] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-on-smoking-and-COVID-19

[17] "A nicotinic hypothesis for COVID-19 with preventive and therapeutic implications" Jean-Pierre Changeux, Zahir Amoura, Felix Rey, Makoto Miyara. Qeios ID: FXGQSB.2. https://doi.org/10.32388/FXGQSB.2

[18] "Low incidence of daily active tobacco smoking in patients with symptomatic COVID-19"; Makoto Miyara, Florence Tubach, Valerie Pourcher, Capucine Morelot-Panzini, Julie Pernet, Julien Haroche, Said Lebbah, Elise Morawiec, Guy Gorochov, Eric Caumes, Pierre Hausfater, Alain COMBES, Thomas Similowski, Zahir Amoura. Qeios ID: WPP19W.3. https://doi.org/10.32388/WPP19W.3

[19] "A Novel Methodology for Epidemic Risk Assessment: the case of COVID-19 outbreak in Italy", A. Pluchino and A. E. Biondo and N. Giuffrida and G. Inturri and V. Latora and R. Le Moli and A. Rapisarda and G. Russo and C. Zappala', eprint 2004.02739, arXiv, physics.soc-ph.

[20] Stier, Andrew and Berman, Marc and Bettencourt, Luis, COVID-19 Attack Rate Increases with City Size (March 30, 2020). Mansueto Institute for Urban Innovation Research Paper. Available at SSRN: https://ssrn.com/abstract=3564464.

[21] 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`

[22] V.Fioletov and B. Kerr, Canadian journal of public health. Revue canadienne de sante publique **101 (4)** I5-9 (July 2010)

[23] `https://diabetesatlas.org/data/en/indicators/12/`

[24] Carlo Mengoli, Carlo Bonfanti, Chiara Rossi and Massimo Franchini Blood Transfus. **13(2)**: 313-317 (apr. 2015) doi: 10.2450/2014.0159-14 PMCID: PMC4385082 PMID: 25369594

[25] Amos Grunbaum, BabyMed `https://www.babymed.com/pregnancy/blood-type-and-rh-rhesus-status-countries`

[26] Jiao Zhao et al, medRxiv 2020.03.11.20031096; doi: https://doi.org/10.1101/2020.03.11.20031096

[27] Swetaprovo Chaudhuri et al., arXiv:2004.10929

[28] `https://guide.culturecrossing.net/basics_business_student_details.php` and `https://guide.culturecrossing.net/`

[29] Anita Shet et al., `https://doi.org/10.1101/2020.04.01.20049478`

[30] Paul Hegarty et al, `https://doi.org/10.1101/2020.04.07.20053272doi`

[31] "Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study". Aaron Miller, Mac Josh Reandelar, Kimberly Fasciglione, Violeta Roumenova, Yan Li, Gonzalo H Otazu medRxiv 2020.03.24.20042937; doi: https://doi.org/10.1101/2020.03.24.20042937.

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

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

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

[35] 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". Archives of Osteoporosis 2012. https://www.iofbonehealth.org/facts-and-statistics/vitamin-d-studies-map.

[36] Spiro, A. and Buttriss, J.L. (2014), Vitamin D status and intake in Europe. Nutrition Bulletin, 39: 322-350. doi:10.1111/nbu.12108

[37] Lips, P., Cashman, K., Lamberg-Allardt, C., Bischoff-Ferrari, H., Obermayer-Pietsch, B., Bianchi, M., Stepan, J., El-Hajj Fuleihan, G., Bouillon, R. (2019). 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, European Journal of Endocrinology, 180(4), P23-P54. Retrieved May 7, 2020, from https://eje.bioscientifica.com/view/journals/eje/180/4/EJE-18-0736.xml

[38] Pludowski P, Grant WB, Bhattoa HP, et al. Vitamin d status in central europe. Int J Endocrinol. 2014;2014:589587. doi:10.1155/2014/589587.

[39] Kuchuk, N.O., van Schoor, N.M., Pluijm, S.M., Chines, A. and Lips, P. (2009), Vitamin D Status, Parathyroid Function, Bone Turnover, and BMD in Postmenopausal Women With Osteoporosis: Global Perspective. J Bone Miner Res, 24: 693-701. doi:10.1359/jbmr.081209

[40] Matthias Wacker and Michael Holick, Dermatoendocrinol. **5(1)** 51-108 (2013)

[41] MH Edwards et.al., JARLIFE (The Journal of Aging Research and Lifestyle), `http://www.jarlife.net/703-the-global-epidemiology-of-vitamin-d-status.html`

[42] N.Kuchuk et al, Journal of Bone and Mineral Research **24** 4 693-701 (2009)

[43] Richard Semba et al, Eur J Clin Nutr. **64(2)** 2010 203-209.

[44] Denys Wahl, Archives of Osteoporosis**7(12)**:155-72 2012

[45] G.Isaia,E.Medico, 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`

[46] William B. Grant et al Nutrients **12(4)**, 988; https://doi.org/10.3390/nu12040988 (2020)

[47] Petre Cristian Ilie, Simina Stefanescu, Lee Smith 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 [+https://doi.org/10.21203/rs.3.rs-21211/v1+].

[48] Exposure to air pollution and COVID-19 mortality in the United States. Xiao Wu, Rachel C. Nethery, Benjamin M. Sabath, Danielle Braun, Francesca Dominici. medRxiv 2020.04.05.20054502; doi: https://doi.org/10.1101/2020.04.05.20054502.

[49] Lei Fang, George Karakiulakis, Michael Roth. https://www.thelancet.com/pdfs/journals/lanres/PIIS2213-2600(20)30116-8.pdf.

[50] https://en.wikipedia.org/wiki/Blood_type_distribution_by_country

[51] Kimball A, Hatfield KM, Arons M, James A, et al. MMWR, **69(13)**;377-381 (2020)

[52] Jou R, Huang WL, Su WJ. Tokyo-172 BCG vaccination complications, Taiwan. Emerg Infect Dis. 2009;15(9):1525?1526. doi:10.3201/eid1509.081336

[53] Andrew T. Jebb, Louis Tay, Ed Diener and Shigehiro Oishi Nature Human Behaviour 2(1):33-38 January 2018,10.1038/s41562-017-0277-0

[54] 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.

[55] Countries were taken from [13], plus Taiwan added from [52].