Social, economic, and environmental factors influencing the basic reproduction number of COVID-19 across countries

Jude D. Kong^{1,2,+}, Edward W. Tekwa^{3,4,5,+}, Sarah A. Gignoux-Wolfsohn⁶

¹Centre for Diseases Modeling (CDM), York University, Toronto, ON M3J 1P3, Canada

²Department of Mathematics and Statistics, York University, Toronto, ON M3J 1P3, Canada

³Department of Ecology, Evolution, and Natural Resources, Rutgers University, New Brunswick, NJ 08901, USA

⁴Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544-1003, USA

⁵Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON M5S 3B2, Canada

⁶Smithsonian Environmental Research Center, Edgewater, MD 21037, USA

⁺Co-first authors

Keywords: COVID-19, basic reproduction number, social media, socio-demographic, environment

Abstract

Objective: To assess whether the basic reproduction number (R_0) of COVID-19 is different across countries and what national-level demographic, social, and environmental factors characterize initial vulnerability to the virus.

Methods: We fit logistic growth curves to reported daily case numbers, up to the first epidemic peak. This fitting estimates R_0 . We then use a generalized additive model to discern the effects, and include 5 random effect covariates to account for potential differences in testing and reporting that can bias the estimated R_0 .

Findings: We found that the mean R_0 is 1.70 (S.D. 0.57), with a range between 1.10 (Ghana) and 3.52 (South Korea). We identified four factors- population between 20-34 years old (youth), population residing in urban agglomerates over 1 million (city), social media use to organize offline action (social media), and GINI income inequality- as having strong relationships with R_0 . An intermediate level of youth and GINI inequality are associated with high R_0 , while high city population and high social media use are associated with high R_0 . Environmental and climate factors were not found to have strong relationships with R_0 .

Conclusion: Studies that aim to measure the effectiveness of interventions should account for the intrinsic differences between populations.

Introduction

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has passed the first peak in the majority of countries in the world. Scientists, health officials and citizens have tried to anticipate and explain why the epidemic initially (i.e., before novel interventions) unfolded differently among countries, but only now has the relevant data reached sufficient global reach and temporal length to begin statistical analyses. Existing studies that examine some of the factors that may contribute to differences among countries together are generally applied to metrics such as mortality, daily and cumulative case numbers, or effective reproduction number (1–4). These metrics are time varying and sensitive to reporting and testing differences, and are therefore not easily comparable across countries. For instance, decreasing testing would allow the reported cases to drop, making raw case reporting incomparable across countries.

A key metric, R_0 , has the practical advantage of being reliably estimable (5) and comparable across countries even if testing and reporting rates are different, so long as these rates are either constant or change in roughly the same way over time. R_0 is the basic reproduction number that indicates how many secondary infections are caused by an infected individual at the beginning of an epidemic (6). Without interventions, the portion of the population that is expected to be infected or immunized before the epidemic ends would be $1-1/R_0$. For example, an R_0 of 3 implies that $2/3$ of the population would have to be infected or immunized by the end of the epidemic. R_0 for COVID-19 has variably been estimated between 1.4 (7) and 8.9 (8), with a likely value of 2.5 (9). Many studies either implicitly assume or are understood to imply that R_0 is intrinsic to the infectious disease (9), but it is increasingly acknowledged that many non-interventive factors could affect heterogeneity in R_0 among local populations or countries (10). Interventive responses that occur during the initial exponential phase of COVID-19 can be understood as proximate causes of differences in R_0 across populations, but ultimately they are likely pre-adaptations anchored on existing social, demographic, and environmental factors. Later interventions generally affect R_e , the effective reproduction number at any given time during the epidemic (4).

Our goal is to use a diverse and comprehensive set of demographic, social, and environmental-climatic factors to begin explaining differences in the initial dynamics of COVID-19 across countries. The predictors are non-contemporary with COVID-19, meaning they were measured before the current epidemic began. The dependent variable is the basic reproduction number R_0 , which is derived from the maximum growth rate of COVID-19 (number of additional hosts infected per infected individual per day) within a country. R_0 can be estimated from the beginning of epidemic curves (5). The results in this study cannot be used to infer the eventual epidemic sizes among countries, which are still unfolding and can be very different from the initial dynamics due to novel interventions. We exclude proximal explanations of R_0 , such as enacted policies during the initial rise of COVID-19, because such explanations would contain statistical endogeneity - the initial epidemic growth may have partly caused the responses, therefore the responses cannot be simply used as predictors. Instead, our study focuses on how pre-existing country

characteristics can explain the initial growth phases of COVID-19, although still without implying causation. We did not attempt to include all possibly relevant covariates because of high correlations even among a limited set, and because the limited number of countries dictate that a small subset should be preselected in order to retain sufficiently positive degrees of freedom for statistical analyses. Observed correlation between the covariates tested here and R_0 may be caused by any number of other covariates that correlate with the identified covariates. Observed relationships should therefore be used for hypothesis generation and further investigation.

Covariates chosen belonged to seven categories: demographics, disease, economics, environment, habitat, health, and social. All of these categories have been suggested previously as possible factors for COVID-19 transmission. The most common factors previously studied were temperature (11–24), pollution (13,25–31), precipitation/humidity (18,32,33), population density (34,35), age structure (1,36,37), and population size (1,11,31). For these and additional covariates either previously studied or only mentioned in the media, we rely on statistics measured at a national level. A review of previously found effects on initial COVID-19 epidemic rates related to R_0 are documented in Table 1. We examined these categories simultaneously in order to better understand which group may have a larger influence on R_0 and should therefore be investigated further at both the national and other scales. This analysis is not meant to be exhaustive or definitive, but rather to help reveal baseline epidemiological differences across countries, shape the direction of future research on COVID-19, and understand infectious disease transmission in general.

Methods

All data and code are available on a Github repository (38).

Estimating the basic reproduction number of COVID-19 among countries

The basic reproduction number R_0 (the dependent variable) is given by the formula (39):

$$R_0 = e^{rT} \quad [1]$$

where T is the serial interval of COVID-19 (time delay between the symptom onset of a primary case and their secondary case) and r the initial growth rate of COVID-19. T has been estimated to be between 4–8 days; here we use 5.8 days (40–42). To estimate r , we fit the rate of change in cumulative cases of a logistic growth model, with parameters r (intrinsic growth rate) and K (theoretical epidemic size without intervention), to observe time series in daily confirmed cases (5,43). The logistic growth model is superior to fitting an exponential curve to early case numbers given that case numbers do plateau in reality. In addition, the logistic growth model performs as well or better than more complicated models when confronted with data (5,44). Mechanistic models with multiple compartments (45) and with time-dependent rates (46,47), may be more realistic for COVID-19 outbreaks that in some places exhibit multiple peaks, but such models contain more parameters, require much more data, and are statistically harder to infer reliably. Such complexity is also likely not necessary to describe the initial outbreaks, which appear qualitative logistic (Figure 1).

In the logistic growth model, the cumulative case number I is given by:

$$I(t) = \frac{K}{1 + ae^{-bt}} \quad [2]$$

② $\frac{K}{1+a}$ is the initial number of infected persons i.e. $\frac{K}{1+a} = I(0)$ $a = \frac{K-I(0)}{I(0)}$

② K is the total number of people infected at the end of the outbreak

② b is the intrinsic or the maximum infection growth rate per infected host (growth rate for short)

② $\left(\frac{\ln(a)}{b}, \frac{K}{2}\right)$ the point of maximum spread of SARS-CoV-2

② b/K is the effective between-virus competition rate, where competition is for susceptible hosts

We truncate all COVID-19 reported daily case time series (48,49) to the day with the highest daily count, because some countries have lingered near peak daily count for much longer than a logistic growth model would predict, which would pull the model peak to later than the actual date of peak incidence and thereby underestimates r . We manually checked each time series and ensured that the highest daily count only occurred during a first peak. We included all countries that were at least 6 days

into a period with at least 30 daily cases as of July 29, 2020, after truncating at the peak. We eliminated countries whose logistic growth model R^2 was less than 0.9. Countries were assigned to the regions of Asia-Australia, Africa, Eurasia, Europe, Middle East, North America, and South America. Eurasia included countries that simultaneously belong to both Asia and Europe, plus Ukraine and Uzbekistan due to geopolitical proximity. The Dominican Republic was assigned to North America, while Panama was assigned to South America, as these were the only Central American countries in the final list.

Some countries do not report daily, have variable reporting delays, and may have changed reporting methods resulting in dramatic spikes in cases for particular dates. To circumvent this inaccuracy in date, we used the 7-day rolling average (right aligned) for daily cases (48,49). While this rolling average causes data from nearby dates to be autocorrelated, it should only underestimate the p -value of the fit but not bias the parameter estimates.

Covariates

Next, we compiled data on predictors for each of the countries studied from seven categories (demographics, disease, economics, environmental, habitat, health, and social) from publicly available databases (Table 1). We chose covariates that are diverse, specific, and do not obviously covary; for example, gross domestic product per capita was not used because it covaries with many other more precise covariates. In addition, we chose covariates that are comparable across countries; for example, we chose nurses per capita over doctors per capita because in many countries, nurses are the primary caregivers. For each predictor, we used the most recent available data, which ranged from 2000-2019. When appropriate, data reported in absolute numbers were divided by total population to obtain per capita figures. Data with highly skewed distributions were log-transformed and all distributions were centred and standardized before regression. Four additional covariates were examined but were eliminated through sequential variance inflation factor (VIF) analysis based on the mixed effect generalized additive model described in Section 2.3 (adapted from the 'rms::vif' package in RStudio1.2.5033). The goal is to reduce the collinearity of the final covariate set, so that we can make better statistical attributions to how each covariate affects R_0 . In the analysis, we eliminated the covariate with the highest VIF and iterated the elimination procedure until a representative and epidemiologically reasonable set was left (the set in Table 1). The eliminated covariates were: 1. population greater than 65 years old (50), 2. life expectancy at birth (50), 3. hospital beds per capita (50), and 4. mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (51).

Statistical analysis

After compiling the variables, we fitted the generalized additive model (GAM) using the 'mgcv' package in RStudio1.2.5033, to analyze the effects of the covariates listed in Table 1, on the R_0 value across the globe. The covariates are standardized for effect comparisons. The main advantage of GAMs over traditional regression methods are their capability to model non-linear relationships (a common feature of many datasets) between a response variable and multiple covariates using non-parametric smoothers. The general formula of a GAM is:

$$g(\mu_i) = \beta + \sum_{j=1}^n f_j(X_i) + \varepsilon_i \quad [3]$$

Where $g(\mu_i)$ is a monotonous link function relating the independent variable to the given covariates, β is any strictly parametric component in the model, such as intercept, $f_j(X_i)$ is the variable explained by the nonparametric smoothing function, and ε_i is identically and independently distributed as a normal random variable.

Two sets of analyses are performed: 1. fixed effect model on R_0 ; 2. mixed effect model on R_0 , with region, the total number of days to the first 30 cases (measured from when China had her first 30th cases), gross domestic product per capita (GDP), average under-reported percentage (52), and total number of available data points as random effects. These random effects are meant to capture differences in reporting and testing standard. GDP (50) is additionally expected to correlated with many other covariates, so using it as a random effect allows us to better understand the effects of other more precise and less aggregative metrics.

There are concerns that different COVID-19 detection capabilities among nations may affect the estimated growth rates of the disease and the regression results. Some estimates of detection differences among countries have been made (53). However, we observe that if under-reporting is constant in time within countries, then the estimated r and therefore R_0 would not be affected - only K would be artificially depressed. On the other hand, if under-reporting is non-constant in time, then r would be affected (54).

For example, a country that responds strongly after the arrival of COVID-19 may ramp up testing capability, which would decrease under-reporting over time. This would cause r fitted to the reported case data to be an overestimate. On the other hand, if a country's detection capability erodes over time due to a shortage of test kits or a decision to stop testing non-severe cases, then r would be underestimated. There are ongoing efforts to correct for these temporal biases based on delayed mortality rates (52,55), but the results are currently not credible for smaller countries with poor reporting. At this point we must rely on the reported case numbers, and use random effects to partially account for possible biases.

We use the *anova()* function in R to compare the candidate models and see which one provides the best parsimonious fit of the data. Because these models differ in the use of the random variables, ANOVA will test whether or not including random effects leads to a significant improvement over using just the given covariates without any random variables. For goodness of fits test, we use a chi-squared test.

Results

Basic reproduction number of COVID-19 among countries

Figure 1 and Figures S1-S4 show growth curves fit to observe time series in daily confirmed cases across countries. Figure 2 and Table S1 summarize estimated R_0 across countries. For the countries considered, the basic reproduction number R_0 has maximum values in South Korea, Australia and Luxemburg, with 3.52, 3.35 and 3.00 and minimum values in the Dominican Republic, Ghana, Indonesia with 1.10, 1.10, 1.11. Overall, the mean R_0 is 1.70 with a standard deviation of 0.57. Belgium (1.71), Iceland (1.72), and Japan (1.79) are the closest to this mean R_0 .

Mixed effects GAM model

The explained deviance is 75.3%, this indicates that the model has a high explanatory power and predictability. The four fixed effect covariates with p -values below 0.1 are youth, city, social media, and GINI inequality. An intermediate value of youth (population between 20-34 years old) and GINI inequality are correlated with high R_0 . On the other hand, an intermediate level of city population (population in urban conglomerates over 1 million) is correlated with low R_0 . Finally, social media use to organize offline action is positively correlated with high R_0 (Figure 3, Table S2). Figure 4 shows how eight exemplary countries covering a wide range of R_0 are characterized by different demographic and social profiles. The profiles show that the countries' ranking in covariate values mostly conform to the statistical trends suggested by GAM. For example, Ghana, with the nearly lowest R_0 , has a low portion of population in large urban agglomerates, relatively low social media usage, a large youth population, and a high GINI inequality index, which conform with the profile for low R_0 . South Korea and the United States, which have high R_0 values, have a high portion of population in large urban agglomerates, a high social media usage, and an intermediate youth population, which conform with the profile for high R_0 . However, South Korea also has a relatively low GINI, while the United States has a relatively high GINI, whereas an average GINI is overall associated with the highest R_0 . This illustrates that countries with high R_0 tend to fit the statistical high R_0 profile in most but not all dimensions. Other countries examined, with lower R_0 , had profiles that diverge further from the statistical high R_0 profile (Figure 4).

Comparison with the fixed effect GAM model

For the fixed effects only model, the explained deviance is 65.5%. Figures S5 shows the effect of the covariates on R_0 , and Table S3 contains the statistical results. The ANOVA comparison of the mixed and fixed effects models shows a DF Deviance of 1.793 and $p=0.006$. This means that adding random effects to the model lead to a significantly improved fit over the fixed effects model. Some covariates have p -values below 0.1 in the fixed effects model but not in the mixed effects model (temperature, internet filtering). Conversely, some covariates have lower p -values in the mixed effects than in the fixed effects model (city, social media). These differences illustrate that random effects are important in controlling for potential biases in the raw daily COVID-19 reporting data.

Discussion

We found that across the globe, R_0 (1.70 ± 0.57 S.D.) was variable and on average slightly lower than previous estimates (8,9). However, previous studies focussed on data from China and other

countries with early epidemic onset, which our estimates show to have higher than average R_0 . We identified four factors (youth, city, social media, and GINI inequality) as having strong relationships with COVID-19 R_0 across countries. Environmental factors, which are the most common factors previously identified (temperature (11–24), pollution (13,25–31), precipitation/humidity (18,32,33)), did not have strong relationships with R_0 when other factors are considered simultaneously, although pollution, temperature, and humidity all have positive associations.

The positive relationship between social media usage and R_0 observed here has not been previously found for COVID-19. The trend may be proximally caused by the propagation of false information on social media, for example in downplaying the potential danger of COVID-19, the effectiveness of masks and social distancing, or propping up conspiracy theories on the disease (56). One study showed that more than 80% of online claims about COVID-19 were false at the beginning of the pandemic (57). These proximal mechanisms, at least at the initial onset of COVID-19, seemed to have overridden the potential benefits of social media as an accurate information spreader that allows people to assess the true risks (58). This result may be related to the finding that in social networks false information spreads faster than the truth (59). In the initial stage of COVID-19, there was an information void regarding the nature of the disease and effective interventions, so it appears that false information filled an important void for people in countries where social media and reality were tightly weaved. However, social media may help slow a contagion's spread once scientific information becomes available. Our result highlights the need to consider the dynamic role that social media plays in epidemics (60).

The quadratic relationships of youth and GINI inequality with R_0 indicate a more complex underlying tradeoff than is previously appreciated, which was either monotonically positive (36,61) or negative (1,36). A large youth population may confer resilience against the disease (36,37) while also increasing the transmission rate (62); conversely an old population may be more susceptible to the disease (63,64) but exhibits a reduced transmission rate. The synergistic result is that an intermediate level of youth is related to the highest R_0 . A high GINI inequality index, referring to the amount of income inequality across a population, may mean the physical segregation of population segments and thus initially halted Covid growth across the population (36), while a low GINI may indicate better social integrations and fewer people left at high risk exposures (61). Therefore, an intermediate GINI is related to the highest R_0 . The fact that intermediate values are related to the highest, not lowest, R_0 suggest that the hypothetical risk mechanisms - youth transmission and elder susceptibility, and social mixing and uneven risk exposures - work synergistically (rather than antagonistically) when both are present. That is, these risk mechanisms together lead to a more-than-additive increase in R_0 .

An intermediate level of city-dwelling population (population in urban conglomerates over 1 million) is related to the lowest R_0 . A high level of city dwelling is expected to increase R_0 because of high contact rates and conforms with the main empirical trend (34). However, it is unclear why a low level of city dwelling is also associated with a high R_0 , although the rise is relatively slight. In comparison to the quadratic effects of youth and GINI inequality, the effect of city dwelling appears close to monotonic.

Our analysis is based on coarse-grained country-level case data, without explicitly correcting R_0 estimates using temporal trends in testing, reporting, and mortality. The factors we analyzed hold across regions within a country to some extent, but it can be argued that each factor or its substitute can be measured more locally (10,65) and result in better statistical power. R_0 can also be estimated using less phenomenological, more mechanistic models such as multiple-compartment (eg. susceptible-exposed-infectious-recovered-susceptible) (45), social network (66), or time-varying (46,47) models. However, these approaches are more data intensive and not current available in many countries. Our country-level analysis of R_0 serves as a coarse grain baseline for future analyses pending data availability. An international perspective like the one we took here can help us understand COVID-19 in a broader context, even though we sacrifice the ability to infer local causality.

We emphasize that R_0 is not indicative of eventual outbreak sizes or the nature of subsequent waves. Given the same population, a higher R_0 can lead to a higher outbreak size, but this does not account for intervention measures that occur after the initial epidemic growth. For instance, a high initial epidemic growth may provide a strong signal to both citizens and governments, which then may mount a stronger response to limit the outbreak size than if the initial growth were weaker. For example, South Korea and Australia had high R_0 (3.52, 3.35) but low cumulative case numbers (490, 1073 per million on Oct 17, 2020 (48,49)). In contrast, countries such as Brazil and Peru have low R_0 (1.18, 1.24) and yet struggle to control the epidemic (cumulative case number=24,465, 26,156 per million on Oct 17, 2020

(48,49)). The dynamic coupling between R_0 response is one reason why it is harder to infer the effectiveness of intervention without taking into account how pre-existing characteristics relate to initial epidemic growth. It is reasonable to believe that early interventions are actually symptoms of pre-existing social, demographic, and environmental characteristics and are not easy to implement in some countries.

The factors influencing R_0 identified here reflect the naive or intrinsic factors that may determine a country's vulnerability to the novel Coronavirus. While both government and citizen interventions have since been implemented in different ways, the current study can inform both the ongoing effort to control the pandemic and efforts to anticipate and control future coronavirus epidemics. The R_0 values calculated here serve as baseline expectations for how fast COVID-19 would spread if interventions were to be prematurely lifted, given that the percent of population susceptible to COVID-19 is still relatively low. The baseline expectation for R_e without novel interventions should be the predicted R_0 of a location given updated covariates, particularly on environmental and air transport factors that have drastically changed since the initial stage. Future studies that aim to measure the effectiveness of interventions across locations should account for intrinsic factors identified here. Otherwise, interventions in countries with intrinsically low R_0 may be mistaken as more effective than they are, while effective interventions in countries with intrinsically high R_0 may be regrettably ignored.

Acknowledgements: We are grateful for feedback from Aria Ahmad and Korryn Bodner.

Funding: This work was supported by Canada's International Development Research Centre (IDRC) (Grant No. 109559-001).

Author Contributions: J.D.K., E.W.T., and S.G.W. designed research; all authors conducted literature search and data collection; J.D.K. and E.W.T. analyzed data; and all authors wrote the paper.

Disclosure statement: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of their respective institutions. The authors declare no conflict of interest.

Ethics and consent: All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

Competing Interest Statement: The authors declare no conflict of interest.

References

1. Chaudhry R, Dranitsaris G, Mubashir T, Bartoszko J, Riazi S. A country level analysis measuring the impact of government actions, country preparedness and socioeconomic factors on COVID-19 mortality and related health outcomes. *EClinicalMedicine*. 2020 Aug;25:100464.
2. Martins LD, da Silva I, Batista WV, Andrade M de F, Freitas ED de, Martins JA. How socio-economic and atmospheric variables impact COVID-19 and influenza outbreaks in tropical and subtropical regions of Brazil. *Environmental Research*. 2020 Dec;191:110184.
3. Scarpone C, Brinkmann ST, Große T, Sonnenwald D, Fuchs M, Walker BB. A multimethod approach for county-scale geospatial analysis of emerging infectious diseases: a cross-sectional case study of COVID-19 incidence in Germany. *Int J Health Geogr*. 2020 Dec;19(1):32.
4. Li Y, Campbell H, Kulkarni D, Harpur A, Nundy M, Wang X, et al. The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2: a modelling study across 131 countries. *The Lancet Infectious Diseases*. 2020 Oct;S1473309920307854.
5. Ma J, Dushoff J, Bolker BM, Earn DJD. Estimating Initial Epidemic Growth Rates. *Bull Math Biol*. 2014 Jan;76(1):245–60.
6. Ridenhour B, Kowalik JM, Shay DK. Unraveling R0: Considerations for Public Health Applications. *American Journal of Public Health*. 2014;104(2):e32–41.
7. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N Engl J Med*. 2020 Mar 26;382(13):1199–207.
8. Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R. High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2. *Emerg Infect Dis*. 2020 Jul;26(7):1470–7.
9. Petersen E, Koopmans M, Go U, Hamer DH, Petrosillo N, Castelli F, et al. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *The Lancet Infectious Diseases*. 2020 Sep;20(9):e238–44.
10. Shaw J. COVID-19 May Be Much More Contagious Than We Thought. *Harvard Magazine* [Internet]. 2020 May 13; Available from: <https://www.harvardmagazine.com/2020/05/r-nought>
11. Nakada LYK, Urban RC. COVID-19 pandemic: environmental and social factors influencing the spread of SARS-CoV-2 in São Paulo, Brazil. *Environ Sci Pollut Res* [Internet]. 2020 Sep 28 [cited 2020 Nov 2]; Available from: <http://link.springer.com/10.1007/s11356-020-10930-w>

12. Haque SE, Rahman M. Association between temperature, humidity, and COVID-19 outbreaks in Bangladesh. *Environmental Science & Policy*. 2020 Dec;114:253–5.
13. Lolli S, Chen Y-C, Wang S-H, Vivone G. Impact of meteorological conditions and air pollution on COVID-19 pandemic transmission in Italy. *Sci Rep*. 2020 Dec;10(1):16213.
14. Demongeot J, Flet-Berliac Y, Seligmann H. Temperature Decreases Spread Parameters of the New Covid-19 Case Dynamics. *Biology*. 2020 May 3;9(5):94.
15. Sajadi MM, Habibzadeh P, Vintzileos A, Shokouhi S, Miralles-Wilhelm F, Amoroso A. Temperature, Humidity, and Latitude Analysis to Estimate Potential Spread and Seasonality of Coronavirus Disease 2019 (COVID-19). *JAMA Netw Open*. 2020 Jun 11;3(6):e2011834.
16. Li H, Xu X-L, Dai D-W, Huang Z-Y, Ma Z, Guan Y-J. Air pollution and temperature are associated with increased COVID-19 incidence: A time series study. *International Journal of Infectious Diseases*. 2020 Aug;97:278–82.
17. Islam N, Bukhari Q, Jameel Y, Shabnam S, Erzurumluoglu AM, Siddique MA, et al. COVID-19 and climatic factors: A global analysis. *Environmental Research*. 2020 Oct;110355.
18. Morris DH, Yinda KC, Gamble A, Rossine FW, Huang Q, Bushmaker T, et al. The effect of temperature and humidity on the stability of SARS-CoV-2 and other enveloped viruses [Internet]. *Microbiology*; 2020 Oct [cited 2020 Nov 8]. Available from: <http://biorxiv.org/lookup/doi/10.1101/2020.10.16.341883>
19. Azuma K, Kagi N, Kim H, Hayashi M. Impact of climate and ambient air pollution on the epidemic growth during COVID-19 outbreak in Japan. *Environmental Research*. 2020 Nov;190:110042.
20. Raza A, Khan MTI, Ali Q, Hussain T, Narjis S. Association between meteorological indicators and COVID-19 pandemic in Pakistan. *Environ Sci Pollut Res* [Internet]. 2020 Oct 14 [cited 2020 Nov 8]; Available from: <http://link.springer.com/10.1007/s11356-020-11203-2>
21. Singh O, Bhardwaj P, Kumar D. Association between climatic variables and COVID-19 pandemic in National Capital Territory of Delhi, India. *Environ Dev Sustain* [Internet]. 2020 Oct 7 [cited 2020 Nov 8]; Available from: <http://link.springer.com/10.1007/s10668-020-01003-6>
22. Adhikari A, Yin J. Short-Term Effects of Ambient Ozone, PM2.5, and Meteorological Factors on COVID-19 Confirmed Cases and Deaths in Queens, New York. *IJERPH*. 2020 Jun 5;17(11):4047.
23. Yao Y, Pan J, Liu Z, Meng X, Wang W, Kan H, et al. No association of COVID-19 transmission with temperature or UV radiation in Chinese cities. *Eur Respir J*. 2020 May;55(5):2000517.

24. Ran J, Zhao S, Han L, Liao G, Wang K, Wang MH, et al. A re-analysis in exploring the association between temperature and COVID-19 transmissibility: an ecological study with 154 Chinese cities. *Eur Respir J*. 2020 Aug;56(2):2001253.
25. Comunian S, Dongo D, Milani C, Palestini P. Air Pollution and COVID-19: The Role of Particulate Matter in the Spread and Increase of COVID-19's Morbidity and Mortality. *IJERPH*. 2020 Jun 22;17(12):4487.
26. Fattorini D, Regoli F. Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy. *Environmental Pollution*. 2020 Sep;264:114732.
27. Zhang Z, Xue T, Jin X. Effects of meteorological conditions and air pollution on COVID-19 transmission: Evidence from 219 Chinese cities. *Science of The Total Environment*. 2020 Nov;741:140244.
28. Wang B, Chen H, Chan YL, Oliver BG. Is there an association between the level of ambient air pollution and COVID-19? *American Journal of Physiology-Lung Cellular and Molecular Physiology*. 2020 Sep 1;319(3):L416–21.
29. Jiang Y, Wu X-J, Guan Y-J. Effect of ambient air pollutants and meteorological variables on COVID-19 incidence. *Infect Control Hosp Epidemiol*. 2020 Sep;41(9):1011–5.
30. Frontera A, Martin C, Vlachos K, Sgubin G. Regional air pollution persistence links to COVID-19 infection zoning. *Journal of Infection*. 2020 Aug;81(2):318–56.
31. Copiello S, Grillenzoni C. The spread of 2019-nCoV in China was primarily driven by population density. Comment on “Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China” by Zhu et al. *Science of The Total Environment*. 2020 Nov;744:141028.
32. Wei J-T, Liu Y-X, Zhu Y-C, Qian J, Ye R-Z, Li C-Y, et al. Impacts of transportation and meteorological factors on the transmission of COVID-19. *International Journal of Hygiene and Environmental Health*. 2020 Sep;230:113610.
33. Sobral MFF, Duarte GB, da Penha Sobral AIG, Marinho MLM, de Souza Melo A. Association between climate variables and global transmission of SARS-CoV-2. *Science of The Total Environment*. 2020 Aug;729:138997.
34. Rader B, Scarpino SV, Nande A, Hill AL, Adlam B, Reiner RC, et al. Crowding and the shape of COVID-19 epidemics. *Nat Med* [Internet]. 2020 Oct 5 [cited 2020 Nov 6]; Available from: <http://www.nature.com/articles/s41591-020-1104-0>
35. Gargiulo C, Gaglione F, Guida C, Papa R, Zucaro F, Carpentieri G. The role of the urban settlement system in the spread of Covid-19 pandemic. The Italian case. *TeMA - Journal of Land Use. 2020 Jun 19;Mobility and Environment*:189-212 Pages.
36. Gangemi S, Billeci L, Tonacci A. Rich at risk: socio-economic drivers of COVID-19 pandemic spread. *Clin Mol Allergy*. 2020 Dec;18(1):12.

37. Diop BZ, Ngom M, Pougué Biyong C, Pougué Biyong JN. The relatively young and rural population may limit the spread and severity of COVID-19 in Africa: a modelling study. *BMJ Glob Health*. 2020 May;5(5):e002699.
38. Kong JD, Tekwa EW, Gignoux-Wolfsohn SA. Data and Code for: Social, economic, and environmental factors influencing the basic reproduction number of COVID-19 across countries [Internet]. 2020. Available from: <https://github.com/Jdkong/COVID-19>
39. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc R Soc B*. 2007 Feb 22;274(1609):599–604.
40. Du Z, Xu X, Wu Y, Wang L, Cowling BJ, Meyers LA. Serial Interval of COVID-19 among Publicly Reported Confirmed Cases. *Emerg Infect Dis* [Internet]. 2020 Jun [cited 2020 Apr 25];26(6). Available from: http://wwwnc.cdc.gov/eid/article/26/6/20-0357_article.htm
41. Park M, Cook AR, Lim JT, Sun Y, Dickens BL. A Systematic Review of COVID-19 Epidemiology Based on Current Evidence. *JCM*. 2020 Mar 31;9(4):967.
42. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* [Internet]. 2020 Apr 15 [cited 2020 May 7]; Available from: <http://www.nature.com/articles/s41591-020-0869-5>
43. Wu K, Darcet D, Wang Q, Sornette D. Generalized logistic growth modeling of the COVID-19 outbreak: comparing the dynamics in the 29 provinces in China and in the rest of the world. *Nonlinear Dyn*. 2020 Aug;101(3):1561–81.
44. Duhon J, Bragazzi N, Kong JD. The impact of non-pharmaceutical interventions, demographic, social, and climatic factors on the initial growth rate of COVID-19: A cross-country study. *Science of The Total Environment*. 2021 Mar;760:144325.
45. Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science*. 2020 May 22;368(6493):860–8.
46. Tuite AR, Fisman DN, Greer AL. Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. *CMAJ*. 2020 May 11;192(19):E497–505.
47. Matrajt L, Leung T. Evaluating the Effectiveness of Social Distancing Interventions to Delay or Flatten the Epidemic Curve of Coronavirus Disease. *Emerg Infect Dis*. 2020 Aug;26(8):1740–8.
48. European Centre for Disease Prevention and Control. Download today's data on the geographic distribution of COVID-19 cases worldwide [Internet]. 2020. Available from: <https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide>

49. Roser M, Ritchie H, Ortiz-Ospina E, Hasell J. Coronavirus Disease (COVID-19) [Internet]. OurWorldInData.org. 2020. Available from: <https://ourworldindata.org/coronavirus>
50. The World Bank. World Development Indicators [Internet]. 2017. Available from: <https://databank.worldbank.org/source/world-development-indicators>
51. Geneva, World Health Organization. Global Health Estimates 2015: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2015. [Internet]. 2016. Available from: https://www.who.int/healthinfo/global_burden_disease/estimates_regional_2000_2015/en/
52. Russell TW, Golding N, Hellewell J, Abbott S, Wright L, Pearson CAB, et al. Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections [Internet]. *Epidemiology*; 2020 Jul [cited 2020 Nov 9]. Available from: <http://medrxiv.org/lookup/doi/10.1101/2020.07.07.20148460>
53. Niehus R, De Salazar PM, Taylor A, Lipsitch M. Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan, China that depend on reported cases in international travelers [Internet]. *Infectious Diseases (except HIV/AIDS)*; 2020 Feb [cited 2020 Mar 30]. Available from: <http://medrxiv.org/lookup/doi/10.1101/2020.02.13.20022707>
54. Zhao S, Lin Q, Ran J, Musa SS, Yang G, Wang W, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases*. 2020 Mar;92:214–7.
55. Russell TW, Hellewell J, Jarvis CI, van Zandvoort K, Abbott S, Ratnayake R, et al. Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. *Eurosurveillance* [Internet]. 2020 Mar 26 [cited 2020 Apr 18];25(12). Available from: <https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.12.2000256>
56. Bavel JJV, Baicker K, Boggio PS, Capraro V, Cichocka A, Cikara M, et al. Using social and behavioural science to support COVID-19 pandemic response. *Nat Hum Behav*. 2020 May;4(5):460–71.
57. Islam MS, Sarkar T, Khan SH, Mostofa Kamal A-H, Hasan SMM, Kabir A, et al. COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. *The American Journal of Tropical Medicine and Hygiene*. 2020 Oct 7;103(4):1621–9.
58. Chan MS, Winneg K, Hawkins L, Farhadloo M, Jamieson KH, Albarracín D. Legacy and social media respectively influence risk perceptions and protective behaviors during emerging health threats: A multi-wave analysis of communications on Zika virus cases. *Social Science & Medicine*. 2018 Sep;212:50–9.
59. Vosoughi S, Roy D, Aral S. The spread of true and false news online. *Science*. 2018 Mar 9;359(6380):1146–51.

60. Sooknanan J, Comissiong DMG. Trending on Social Media: Integrating Social Media into Infectious Disease Dynamics. *Bull Math Biol.* 2020 Jul;82(7):86.
61. Oronce CIA, Scannell CA, Kawachi I, Tsugawa Y. Association Between State-Level Income Inequality and COVID-19 Cases and Mortality in the USA. *J GEN INTERN MED.* 2020 Sep;35(9):2791–3.
62. Boehmer TK, DeVies J, Caruso E, van Santen KL, Tang S, Black CL, et al. Changing Age Distribution of the COVID-19 Pandemic — United States, May–August 2020. *MMWR Morb Mortal Wkly Rep.* 2020 Oct 2;69(39):1404–9.
63. Dowd JB, Andriano L, Brazel DM, Rotondi V, Block P, Ding X, et al. Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proc Natl Acad Sci USA.* 2020 May 5;117(18):9696–8.
64. CMMID COVID-19 working group, Davies NG, Klepac P, Liu Y, Prem K, Jit M, et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat Med.* 2020 Aug;26(8):1205–11.
65. Cowling BJ, Ali ST, Ng TWY, Tsang TK, Li JCM, Fong MW, et al. Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: an observational study. *The Lancet Public Health.* 2020 May;5(5):e279–88.
66. Hilton J, Keeling MJ. Estimation of country-level basic reproductive ratios for novel Coronavirus (SARS-CoV-2/COVID-19) using synthetic contact matrices. Flegg JA, editor. *PLoS Comput Biol.* 2020 Jul 2;16(7):e1008031.
67. Kreutz R, Algharably EAE-H, Azizi M, Dobrowolski P, Guzik T, Januszewicz A, et al. Hypertension, the renin–angiotensin system, and the risk of lower respiratory tract infections and lung injury: implications for COVID-19. *Cardiovascular Research.* 2020 Aug 1;116(10):1688–99.
68. Ghosh D, Bernstein JA, Mersha TB. COVID-19 pandemic: The African paradox. *Journal of Global Health.* 2020 Dec;10(2):020348.
69. Central Intelligence Agency. GINI Index [Internet]. The World Factbook. 2017. Available from: <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2172rank.html>
70. The World Bank. World Development Indicators [Internet]. 2019. Available from: <https://databank.worldbank.org/source/world-development-indicators>
71. The World Bank. Climate Data API [Internet]. 2011. Available from: <https://datahelpdesk.worldbank.org/knowledgebase/articles/902061-climate-data-api>
72. Nuclear Threat Initiative, Johns Hopkins Center for Health Security, The Economist Intelligence Unit. GHS Index [Internet]. 2019. Available from: <https://www.ghsindex.org/>

- 525 73. V-Dem Institute. V-dem (version 10) [Internet]. 2020. Available from: [https://www.v-](https://www.v-dem.net/)
526 dem.net/
- 527 74. Pequeno P, Mendel B, Rosa C, Bosholn M, Souza JL, Baccaro F, et al. Air transportation,
528 population density and temperature predict the spread of COVID-19 in Brazil. PeerJ. 2020
529 Jun 3;8:e9322.
- 530

Figures and Tables

Figure 1. The COVID-19 daily cases. Dots represent daily cases averaged over a 7-day window, and curves are fitted based on the logistic growth model. Example countries are arranged from top left to bottom right in order of increasing basic reproduction number (R_0).

Figure 2. Estimated basic reproduction numbers (R_0) for countries across the globe. Gray countries are not included in our analysis.

Figure 3. Mixed GAM derived partial effects (smoother plot) of the covariates, on R_0 . Circles are partial residuals, and red shades are 95% confidence intervals.

Figure 4. Country profiles. The four characteristics (youth, city, social media, and GINI inequality) with the lowest p -values in the mixed effect GAM are plotted (centred and standardized) for 8 countries representing, from top left to bottom right in the legend, increasing R_0 . Red dashed lines represent alternative high R_0 profiles based on the mixed effects GAM model.

Table 1. Covariates and previous findings. Data sources are cited under the covariate column. Previous effects on epidemic rates are not necessarily on basic reproduction number R_0 , but rather on cumulative case load, daily cases at certain stages, or effective reproduction number. Effects on epidemic rates are recorded as positive (+), negative (-), insignificant (0), or non-monotonic (u-shape or n-shape). Effects accompanied by (?) are theoretical.

Category	Covariate	Previously found effects
demographics	Youth: Population ages 20-35 (% of total population) (50)	(+) (36,37) (-) (1)
demographics	Total Pop: Population total (50)	(+) (11,31) (0) (1)
disease	Mort Resp: Mortality rate from lower respiratory infections (per 100,000) (51)	(0) (67)
disease	Mort Infect: Mortality rate from infectious and parasitic diseases (per 100,000) (51)	(-) (68)
economic	GINI: GINI index (income inequality, 100=high) (69)	(+) (61) (-) (36)
economic	Business: Ease of doing business index 2019 (1=most business-friendly regulations) (70)	
environmental	Temperature: °C January-March (71)	(-) (11–18) (+) (19–22) (0) (23) (n-shape) (24)
environmental	Precipitation: mm January-March (71)	(+) (32,33) (u-shape) (18)
environmental	Pollution: PM2.5 air pollution, mean annual exposure (micrograms per cubic meter) (50)	(+) (13,25–30) (-) (31)
habitat	City: Population in urban agglomerations of more than 1 million (% of total population) (50)	(+) (34,35)
habitat	Urbanization: Urban population (% of total population) (50)	(+) (34,35)
health	GHS: Global Health Security detection index (72)	(0) (1)
health	Nurses: Nurses and midwives (per 1,000 people) (50)	(-) (1)
social	Social Media: Average People's Use Of Social Media To Organize Offline Action (4=high) (73)	(-?) (58) (+?) (56)
social	Internet Filtering: Government Internet filtering in practice (4=low) (73)	
social	Air Transport: passengers carried per capita (50)	(+) (36) (0) (74)