




OPEN

Correlates of the country differences in the infection and mortality rates during the first wave of the COVID-19 pandemic: evidence from Bayesian model averaging

Viktor Stojkoski^{1,2,5}, Zoran Utkovski^{2,3}, Petar Jolakoski², Dragan Tevdovski¹ & Ljupcho Kocarev^{2,4}

The COVID-19 pandemic resulted in great discrepancies in both infection and mortality rates between countries. Besides the biological and epidemiological factors, a multitude of social and economic criteria also influenced the extent to which these discrepancies appeared. Consequently, there is an active debate regarding the critical socio-economic and health factors that correlate with the infection and mortality rates outcome of the pandemic. Here, we leverage Bayesian model averaging techniques and country level data to investigate whether 28 variables, which describe a diverse set of health and socio-economic characteristics, correlate with the final number of infections and deaths during the first wave of the coronavirus pandemic. We show that only a few variables are able to robustly correlate with these outcomes. To understand the relationship between the potential correlates in explaining the infection and death rates, we create a Jointness Space. Using this space, we conclude that the extent to which each variable is able to provide a credible explanation for the COVID-19 infections/mortality outcome varies between countries because of their heterogeneous features.

In order to reduce the potential enormous impact of the coronavirus disease spread (COVID-19), most governments implemented social distancing restrictions such as closure of schools, airports, borders, restaurants and shopping malls. In the most severe cases there were even lockdowns—all citizens were prohibited from leaving their homes. This subsequently led to a major economic downturn: stock markets plummeted, international trade slowed down, businesses went bankrupt and people were left unemployed. While in some countries the implemented restrictions had a significant impact on reducing the expected shock from the coronavirus, the extent of the disease spread in the population greatly varied from one economy to another.

A multitude of health, social and economic factors have been attributed as potential correlates for the observed variety in the coronavirus outcome in terms of the number of infections and/or deaths during this first wave of the pandemic. Indeed, there are numerous studies which discover various factors that affect the *within* country distribution of infections and deaths (see for example, Refs.^{1–5}). The same debate has been extended to evaluate the *between* country discrepancies. In particular, some experts say that the hardest hit countries also had an aging population^{6,7}, or an underdeveloped healthcare system^{8,9}. Others emphasize the role of the natural environment^{10,11}. Having in mind the ongoing discussion, a comprehensive empirical study of the critical health, social and economic correlates with the country level outcome of the number of infections and deaths during first

¹Faculty of Economics, Ss. Cyril and Methodius University in Skopje, Skopje, North Macedonia. ²Macedonian Academy of Sciences and Arts, Skopje, North Macedonia. ³Fraunhofer Heinrich Hertz Institute, Berlin, Germany. ⁴Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Skopje, North Macedonia. ⁵Center for Collective Learning, Artificial and Natural Intelligence Institute, Université Fédérale Toulouse Midi-Pyrénées, Toulouse, France. ✉email: vstojkoski@eccf.ukim.edu.mk

wave of the pandemic can not only aid in inferring whether there are any general rules in their potential impact, but also would offer guidance for future policies that aim at preventing the emergence of future epidemic crises.

To this end, here we perform a detailed statistical analysis on a large set of potential health and socio-economic variables and explore their potential to explain variation in the observed coronavirus total infections/deaths between countries in the *first wave* of the virus' spread. We focus on COVID-19 data that are generated only in the first wave of the pandemic, and thus do not account for various waves (we formally define the first wave in the next section). While this may be seen as a limitation of our analysis, we assert that for each subsequent wave, there was more knowledge about the spread of the virus and vaccines were available. This significantly impacted the way in which the population reacted to the potential susceptibility of the virus. Thus, each wave likely exhibited its own health, social, and economic characteristics, and therefore should be studied separately.

To construct the set of potential correlates we conduct a thorough review of the literature that describes the social and economic factors which contribute to the spread of an epidemic. We identify a total of 28 potential variables that describe a diverse ensemble of factors, including: healthcare infrastructure, societal characteristics, economic performance, demographic structure etc. To investigate the performance of each variable in explaining the coronavirus infections/deaths outcome, we collect a sample of 105 countries, the largest set of countries for which all data were available, and utilize the technique of Bayesian model averaging (BMA). BMA allows us to isolate the most important correlates by calculating the posterior probability that they truly regulate the process. At the same time, BMA provides estimates for the relative impact of the correlates and accounts for the uncertainty in their selection^{12–14}. In this respect, our analysis adds value to a growing body of literature which applies Bayesian methods for investigating the critical factors that drive a certain process and, in this particular case, the outcome of the COVID-19 pandemic¹⁵.

Based on the studied data, we observe patterns suggesting that, during the first wave of the pandemic, there were only a few variables that acted as strong and robust correlates with the final number of registered coronavirus infections and deaths in a country. These variables are related to the effect of density in social interactions and the prevalence of overweight individuals within the population. A simple correlation analysis indicates that the heterogeneity between the countries in terms of their health, social and economic nature might be the driver of this conclusion. Thus, the initial BMA results cannot capture (potentially) significant interactions between the correlates that are relevant to a particular country. To deal with this issue, we develop the coronavirus correlates Jointness Space. The Jointness Space models the interrelation between the potential correlates in explaining the coronavirus infections/deaths outcome, and can represent a statistical foundation for understanding the relationships between variables when developing policy recommendations for preventing future epidemic crises. Using this space, we find that the routes for reducing the potential negative impact of COVID-19 should focus on decreasing the prevalence of excess weight and a small number of other variables that are relevant to those studied herein. This will reduce both the registered infections and the observed deaths due to the COVID-19 disease. In the absence of models that adequately cover all relevant aspects of infections and deaths, this study provides information about the socio-economic correlates of the coronavirus pandemic.

Preliminaries

Measuring COVID-19 infections and death rates. In a formal setting, the final number of registered COVID-19 infections per million population (p.m.p.) and the number of total COVID-19 deaths p.m.p. during the first wave of the pandemic are a result of a disease spreading process^{16,17}. The extent to which a disease spreads within a population is uniquely determined by its reproduction number. This number describes the expected number of cases directly generated by one case in a population in which all individuals are susceptible to infection^{18,19}. Obviously, its magnitude depends on various natural characteristics of the disease, such as its infectivity or the duration of infectiousness, and the social distancing measures imposed by the government. Also, it depends on an abundance of health and socio-economic factors that govern the behavioral interactions within a population^{20,21}.

In general, we never observe the reproduction number, but rather the disease outcome, i.e., the number of infections/deaths. Thus, it is mathematically complex and computationally expensive to try and infer the reproduction number. To circumvent this problem, we utilize its known characteristics and derive a much simpler statistical model for the COVID-19 outcome. Here we choose a specific formulation in which the disease outcome is modeled as a dependent variable, through a linear regression framework, as either the log of accumulated number of registered COVID-19 infections p.m.p. or the log of the accumulated number of COVID-19 deaths p.m.p. of the country at the end of the first wave of the pandemic. We focus on registered quantities normalized on per capita basis for the dependent variable instead of raw values to eliminate the bias in the outcomes arising from the different population sizes in the studied countries. The accumulation of the registered infections and deaths spans from the day of observation of the first infection in the country, up until the last day of the first wave of the pandemic in that country. The last day is, in general, different for each country and is inferred on the basis of the level of daily government response. The estimation procedure used to infer the last day of the first wave will be discussed in more detail in the next section.

The log transformation of the COVID-19 infections/deaths p.m.p. reduces the skewness of the original data and makes the dependent variable real-valued and continuous. For such a dependent variable, the linear regression framework is the simplest tool that quantifies the marginal effect of a set of potential independent variables (correlates). Its advantage lies in the efficient and unbiased analytical inference of the strength of the linear relationship. As such it has been widely used in modeling the outcome of epidemiological phenomena (see for example Refs.^{22–24}).

A central question which arises in the model specification is the selection of the independent variables. While a literature review can offer a comprehensive overview of all potential correlates, in reality we are never

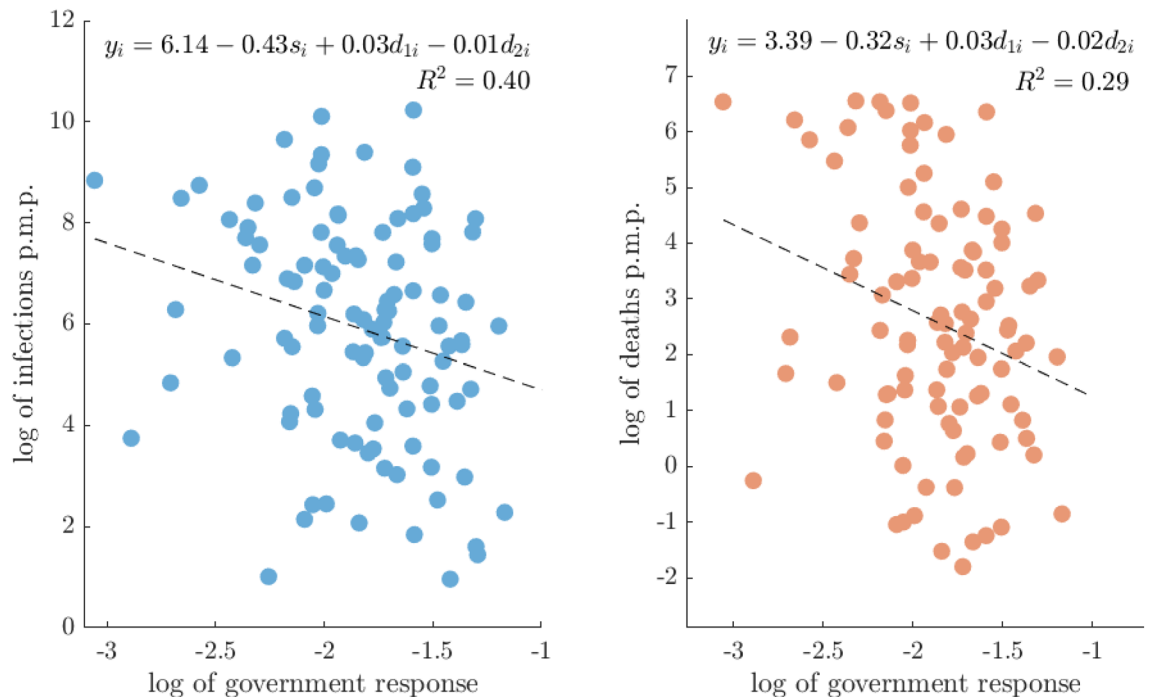


Figure 1. Explained variation in COVID-19 cases due to government response.

certain in their credibility. To reduce our uncertainty, we resort to BMA. BMA leverages Bayesian statistics to account for model uncertainty by estimating each possible specification, and thus evaluating the posterior distribution of each parameter value and probability that a particular model is the correct one²⁵. This has allowed the BMA technique to be used in various domains, ranging from studying correlates of economic growth²⁶, up to determinants of innovation processes²⁷. Recently, it was even applied for estimating the output losses during the Covid-19 pandemic²⁸.

Baseline model. The BMA method relies on the estimation of a baseline model that is used for evaluating the performance of all other models. In our case, this is the model which encompasses only variables for the state of the epidemic dynamics within the country and effect of government policies regarding social distancing, contact tracing and testing procedures.

We use two variables to quantify the possibility that countries are in a different state of the disease spreading process. The first variable simply measures the duration of epidemics (d_{1i}) in a country and is defined as the number of days since the first registered infection. In addition, we evaluate the time which the country had to prepare for the first wave of coronavirus (d_{2i}). This is given as the number of days between the first registered infection worldwide and the first infection in the country.

In order to assess the effect of government policies regarding social distancing, testing, and tracking of infections, we construct an aggregated government response index (s_i). The index quantifies the average daily variation in government responses to the epidemic dynamics. As a measure for the daily variation, we take the Oxford COVID-19 government response index²⁹. The Oxford COVID-19 government response index is a composite measure that combines the daily effect of policies on social distancing, testing and contact tracing in an economy. For each country, we construct a weighted average of the index from all available data since their first registered coronavirus infection, up until the end date, i.e., the date when the government response index is at its maximum value. This threshold is chosen as a means to capture the moment when a country gains the ability to control and stabilize the propagation of the disease. To emphasize the effect of policy responses implemented on earlier dates, we construct a weighted average by putting a larger weight on those dates. This is because earlier responses are supposed to have a bigger impact on the prevention of the spread of the virus. The procedure implemented to derive the average government response index is described in Section S1 of the Supplementary Information (SI).

Figure 1 visualizes the results from the baseline model. We observe that the countries which had more detailed response policies also had less COVID-19 infections and mortality rates, as expected. In addition, the countries with longer duration of the crisis registered more infections and deaths p.m.p., whereas the countries which had more time to prepare for the crisis also had less infections and deaths.

It is apparent that the baseline model already has a large coefficient of determination (R^2) and can significantly explain a certain amount of the cross country variations in registered COVID infections/deaths p.m.p.. However, there is still a large amount of variation that we conjecture can be attributed to various health, social, and economic correlates present within a society that are unrelated to the effects of the epidemic dynamics and government policy variables.

Variable	Measure	Source	Refs.
Healthcare infrastructure			
Medical resources	Medical resources index	WDI	30–40
Health coverage	UHC service coverage index	WDI	30–40
National health statistics			
Life expectancy	Life expectancy at birth	WDI	41–44
Mortality	Non-natural causes mortality index	WDI	45–47,49,50
Comorbidities	Comorbidities index	Our world in data	45–47,49,50
Immunization	Immunization index	WDI	30
Overweight prevalence	% of adults with BMI > 25 kg/m ²	ESG	85–87
Asthma prevalence	% of population with Asthma	Our world in data	48
Economic performance			
Economic development	GDP p.c., PPP \$	WDI	51–54,57–59
Labor market	Employment to population ratio	WDI	30,51,55,56
Government spending	Gov. health spending p.c., PPP \$	WDI	38,51–54
Income inequality	GINI index	WDI	62–66
Societal characteristics			
Social connectedness	Social connectedness index (PageRank)	DFG	88,89
Digitalization	Digitalization index	WDI	30,67–71
Education	Human capital index	WDI	41,67–71
Household size	Avg. no. of persons in a household	UN	21,72–75
Demographic structure			
Elderly population	Population age 65+ (% of total)	WDI	76–79
Young population	Population ages 0–14 (% of total)	WDI	76–79
Gender	50%+ male population (% of total)	WDI	76–79
Population size	Population, total	WM	80,81
Rural population	Rural population (% of total)	WDI	80,81
Migration	Int. migrant stock (% of population)	WDI	80,81
Population density	People per sq. km	WDI	80,81
Natural environment			
Sustainable development	Ecological footprint (gha/person)	GFN	10
Air pollution	Yearly avg P.M. 2.5 exposure	SGA	11,82,83
Weather (latitude)	Geographic coordinate: latitude	Google	84
Air transport	Yearly passengers carried	WDI	38
International Tourism	Number of tourist arrivals	WDI	38

Table 1. List of potential correlates of the COVID-19 first wave infections and mortality rates.

Health, social and economic correlates. To derive the set of potential health, social and economic correlates of the COVID-19 infection and mortality rates during the first wave of the pandemic we conduct a comprehensive literature review. From the literature review we recognize a total of 28 potential correlates, listed in Table 1. For a detailed description of the potential effect of the correlates we refer to the references given in the same table, and the references therein. We hereby point out that the data for each potential correlate corresponds to the last observed value (the value in 2019). This prevents the possible problem of endogenous independent variables in the specification of the regression.

In what follows, we only describe, briefly, the basic characteristics of the set of potential correlates.

Healthcare infrastructure. The healthcare infrastructure essentially determines both the quantity and quality with which health care services are delivered in a time of an epidemic. As measures, we include 2 variables which capture the quantity of hospital beds, nurses and medical practitioners, as well as the quality of the coverage of essential health services. On the one hand, studies report that well-structured healthcare resources positively affect a country's capacity to deal with epidemic emergencies^{30–37}. On the other hand, the healthcare infrastructure also greatly impacts the country's ability to perform testing and reporting when identifying the infected people. In this regard, economies with better structure are able to easily perform mass testing and more detailed reporting^{38–40}.

National health statistics. The physical and mental state of a person plays an important role in the degree to which the individual is susceptible to disease. In countries where a significant proportion of the population suffer from diseases highly associated with the spread of an infectious disease as well as its fatal outcomes, we would expect more severe consequences of the emergent epidemics^{41–44}. Specifically, metabolic disorders such

as diabetes may intensify epidemic complications^{45,46}, whereas it has been observed that the susceptibility to various diseases account for the majority of deaths in complex emergencies⁴⁷. In addition, there is empirical evidence that adequate hygiene greatly reduces the rate of mortality, whereas overweight or asthma prevalence in the population may increase the fatality of epidemic diseases^{48–50}. To quantify the national health characteristics, we include 6 variables that assess the general health level in the studied countries.

Economic performance. We evaluate the economic performance of a country through 4 variables. This performance often mirrors the country's ability to intervene in a case of a public health crisis^{51–56}. Variables such as GDP per capita have been used in modeling health outcomes, mortality trends, cause-specific mortality estimation and health system performance and finances^{57–59}. For poor countries, economic performance appears to improve health by providing the means to meet essential needs such as food, clean water and shelter, as well access to basic health care services. However, after a country reaches a certain threshold of development, few health benefits arise from further economic growth. It has been suggested that this is the reason why, contrary to expectations, the economic downturns during the 20th century were associated with declines in mortality rates^{60,61}. Observations indicate that what drives the health in industrialized countries is not absolute wealth or growth but how the nation's resources are shared across the population⁶². More-egalitarian income distributions are associated with better health of the population^{63–66}.

Societal characteristics. The characteristics of a society often reveal the way in which people interact, and thus spread the disease. In this aspect, properties such as education and the degree of digitalization within a society reflect the level of a person's reaction and promotion of self-induced measures for reducing the spread of the disease^{67–71}. Also, the way individuals encounter (mix) each other through their personal networks or chance encounters may influence the spread of infectious diseases^{21,72–75}. To measure the societal characteristics, we identify 4 variables.

Demographic structure. Similarly, to the national health statistics, the demographic structure may impact the average susceptibility of the population to a disease. Certain demographic groups may simply have weaker defensive health mechanisms to cope with the stress induced by the disease^{76–79}. In addition, the location of living may greatly affect the way in which the disease is spread^{80,81}. To account for these phenomena, we collect 7 variables.

Natural environment. Numerous studies discuss possible correlation between air pollution and COVID-19 infections and mortality rates^{11,82,83}. In addition, some authors note that countries where natural sustainability is deteriorated, are also more vulnerable to epidemic outbreak¹⁰. On the other hand, healthy natural environments may attract more tourists, which could drive the disease spread³⁸. Finally, weather patterns can also impact the infectiousness of the disease, especially exposure when there are very cold days in winter and very hot days in summer⁸⁴. We gather the data for 5 variables which capture the essence of this characteristic.

Results

BMA estimation. We use this set of variables and estimate two distinct BMA models. In the first model the dependent variable is the log of COVID-19 infections p.m.p., whereas in the second model we investigate the critical correlates of the log of the mortality rate due to the coronavirus. For the estimation procedure we use data on 105 countries. This is the maximal set of countries for which the data on all 28 potential correlates could be attained. The summary statistics and the data gathering and preprocessing procedures are described in SI Section S2. The mathematical background of BMA together with our inference setup is given in SI Section S2.

Figure 2 displays the respective results. In both situations, the variables are ordered according to their posterior inclusion probabilities (PIP), given in the second column. PIP quantifies the posterior probability that a given correlate belongs to the linear regression model that best describes the COVID-19 infections/mortality rates. Besides this statistic, we also provide the posterior mean (Post mean) and the posterior standard deviation (Post Std). Post mean is an estimate of the average magnitude of the effect of a correlate, whereas the Post Std evaluates the deviation from this value.

There are multiple ways which can be used to classify the correlates into groups depending on their probability to be included in the model. A standard approach is to divide the correlates on the basis of the difference between their posterior and prior inclusion probabilities¹⁴. In the inference procedure (described in SI Section S3) we initially assumed that the linear regression model which best describes the COVID-19 first wave infections and mortality rates is a result of the baseline specification and 3 additional variables. Our prior belief stems from the general observation which suggests that economies are heterogeneous, and a small number of complementing factors may contribute to the extent of the coronavirus spread, while the other potential correlates may simply behave as substitutes in terms of the socio-economic interpretation within a country. Altogether, this implies that the prior inclusion probability of each potential correlate is around 0.1. We use this attribute, together with the posterior inclusion probability of each correlate, to divide the correlates into four disjoint groups:

Correlates with strong evidence: ($PIP > 0.5$). The first group describes the correlates which have, by far, larger posterior inclusion probabilities than prior probabilities, and thus there is strong evidence that they should be included in the true model. We find two such variables related to explaining the coronavirus infections: the overweight prevalence in the country and the population density. Both variables are positively related with the number of registered COVID-19 infections p.m.p.. When investigating the critical correlates of the coronavirus

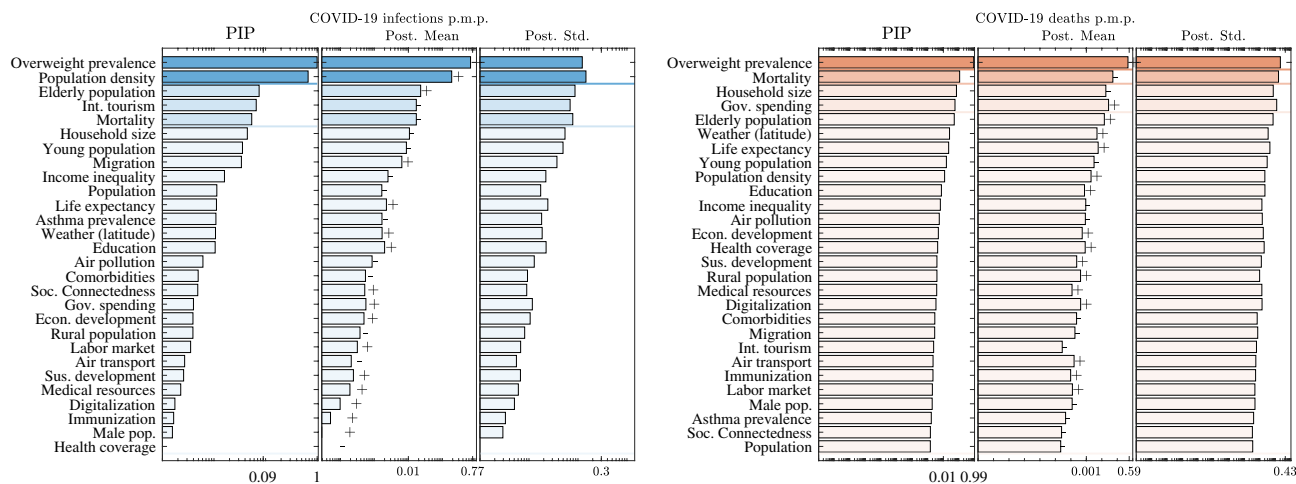


Figure 2. BMA results. Bars for the posterior inclusion probability (PIP), posterior mean (Post. Mean) and the posterior standard deviation (Post. Std.) of each potential correlate. The variables are ordered according to their PIP. The Post. Mean is in absolute value. The signs next to the bar of each variable indicate the direction of its impact. The horizontal lines divide the variables into groups according to their PIP. The horizontal axis is on a logarithmic scale. The setup used to estimate the results is described in SI Section S3.

deaths, it appears that the overweight prevalence is the only variable for which there is strong evidence to explain the outcome and has a positive impact.

Correlates with medium evidence: ($0.5 \geq \text{PIP} > 0.1$). There are no variables for which there is medium evidence to be a correlate of the COVID-19 number of infections in the first wave, whereas mortality from non-natural causes is the only variable for which there is medium evidence to be a correlate of the COVID-19 death rate, with a negative effect.

Correlates with weak evidence: ($0.1 \geq \text{PIP} > 0.05$). These are correlates which have lower posterior inclusion probability than their prior one, but still may account for some of the variations in the COVID-19 infections/deaths. For the infections per million population there are three such correlates, the fraction of elderly population, the number of international tourist arrivals and the mortality from non-natural causes. The elderly population has a positive Post Mean, whereas the other two variables have negative Post Mean. When studying the COVID-19 death rate, we find two correlates with weak evidence. They are the household size and the government health expenditure. The household size has a positive marginal effect (Post Mean), whereas the government health expenditure shows a negative effect.

Correlates with negligible evidence: ($\text{PIP} \leq 0.05$). All other variables have negligible evidence to be a true correlate of the coronavirus outcome. In total, we find negligible evidence for explaining the coronavirus infections in 23 variables and for explaining the coronavirus deaths in 24 variables.

The division of the variables into groups allows us to assess the robustness of each potential correlate—those belonging to a group described with a larger PIP also offer more credible explanation for the coronavirus infections and death rates. Nonetheless, we point out that although the comparison between posterior inclusion probabilities and prior inclusion probabilities is a common approach, its interpretation must be taken with care due to two reasons. First, there are other methods that can be used to divide the correlates into groups which may lead to different interpretation for the credibility of the correlates to explain the coronavirus cases/deaths⁹⁰. Second, the inhomogeneous nature of the specific features of the countries can drive our results. The presence of this phenomenon in our data be inferred by conducting a simple correlation analysis between the potential correlates. If the variables are highly correlated between each other then there is a problem of multicollinearity. Multicollinearity can lead to wider credible intervals that eventually produce less statistically reliable posterior inclusion probabilities in terms of the effect of independent variables in a model. As said in Ref.²⁶, even if the posterior inclusion probability is lower than the prior inclusion probability for a given variable, it might be that this particular variable is important to decision makers under certain circumstances.

In SI Section S4 we conduct several checks to confirm the robustness of our results. In the first robustness check we investigate the impact of outliers. There were several countries which were either extremely affected by the coronavirus or displayed great immunity to the epidemic crisis. To check the robustness of our results against the presence of such data we implement the following strategy. First, we remove a country from the sample. Then, we re-perform the BMA procedure with the resulting countries. We repeat this procedure for every country and recover the median results for each potential correlate. The results indicate that the findings presented here are valid even in the presence of outliers. In the same section, we display the economies which contributed most and least to the credibility of a particular variable. These are the countries which, when excluded, lead to the minimum, respectively maximum, posterior inclusion probability of the given variable. The investigation

suggests that there are multiple countries which are significant contributors to the PIP value of each correlate, thus further indicating that there is heterogeneity in the health social and economic features of the countries. In the second check, we change the end date of the pandemic to be equal to the first date after the day at which the daily government response index is at its maximum and that is at least 20% lower than the daily maximum. This effectively prolongs the duration of the first wave. Nonetheless, it still does not impact the findings. In the third check, we change the dependent variable to be the raw number of infections and deaths at the end of the first wave. In other words, now the dependent variable describes counts and the linear regression framework is not a suitable model. Instead, for the estimation of the marginal impact we use a quasi-Poisson model, which is the most often used procedure when the dependent variable is given as a count that has a large variance⁹¹. Even in this case, the results do not change. In the final robustness check, we add a spatial weighting matrix in the baseline model in order to account for the potential spatial autocorrelation in the spread of COVID-19. Multiple studies have indicated that this effect might exist (see for example⁹²). Again our findings do not significantly change.

Definitely, even if useful for presentation purposes, the mechanical application of a threshold, or a simple comparison between the prior and the posterior, should often be avoided in practice. Each BMA analysis should be coupled with an investigation for the interrelationships between the variables in explaining the dependent variable. We perform this analysis in the subsequent section.

“Jointness Space” of the COVID-19 infections/deaths correlates. The next step in deriving the linear regression model that describes best the coronavirus infections/mortality rates is to find its dimension, i.e., the number of explanatory variables included in the model. As a measure for this quantity, BMA provides the posterior size, formally defined as the posterior belief for the dimension of the model. We find that, for the coronavirus infections p.m.p. the posterior model size is 2.21 whereas for the coronavirus deaths p.m.p. it is 1.34.

After discovering the model size, we need to specify the explanatory variables. This raises the issue of how to construct the appropriate model. One possible solution is to use the correlates with the highest PIP value and regress them on the dependent variable. However, this neglects the interdependence of inclusion and exclusion of correlates in a same model. A standard approach for resolving this issue is to conduct a statistical *jointness* test. The concept of jointness has been introduced within the BMA framework with the aim to capture dependence between explanatory variables in the posterior distribution over the model space⁹³. By emphasizing dependence and conditioning on a set of one or more other variables, jointness moves away from marginal measures of variable importance and investigates the sensitivity of posterior distributions of parameters of interest to dependence across regressors. For example, if two variables are complementary in their posterior distribution over the model space, models that either include or exclude both variables together receive relatively more weight than models where only one variable is present. In our context, jointness tests will allow us to infer whether two variables are complements, i.e., tend to be included together in models with high posterior probability, or substitutes, i.e., models with high posterior probability tend to exclude the joint inclusion of both variables.

To better understand the properties of the COVID-19 infection and mortality rates during the first wave, we perform the jointness test developed by Hofmarcher et al.⁹⁴. Using this test we can estimate a metric between each pair of correlates and quantify their relationship in a range between -1 and 1 . In the two extremes, -1 indicates that the two correlates behave as perfect substitutes in the true model, whereas 1 indicates that they are included in the true model together. The resulting jointness metric between pairs of correlates can be used to construct a network (graph), which we refer to as the *Jointness Space* of the COVID-19 correlates. In this network, the nodes are the potential health, social and economic correlates, whereas the jointness values represent the edge weights. In other words, two arbitrary correlates are linked with each other by the posterior belief that both of them belong to the same linear regression model governing the coronavirus infections/mortality rate.

In theory, many possible factors may cause complementarity between the variables, such as national culture⁹⁵, the type of healthcare system⁹⁶ or political priorities⁹⁷. All of these are a priori notions of what dimension drives the relatedness between the potential correlates and assume that there is little flexibility in choosing the correct model. Instead, the Jointness Space follows an agnostic approach and uses a data-driven measure, based on the idea that, if two correlates are related because they offer contrasting information regarding the coronavirus outcome they will tend to be included in the true model in tandem, whereas variables that give similar information are less likely to be included together. Hence, the developed network offers a statistical view for the importance of the social, health and economic correlates when developing policies aimed at reducing the impact of epidemic crises.

The networks depicted in Fig. 3 visualize the Jointness Space of the correlates included in our BMA framework. To emphasize the complementary relationships, we connect only correlates with positive jointness. The full description for the procedure implemented for constructing the Jointness Space is given in SI Section S5. In the networks, the correlates which can be included in multiple models take a more central position whereas the periphery is constituted of correlates whose credibility in explaining the coronavirus outcome mostly substitutes the effect of other variables.

Interestingly, we observe that the topological form of the Jointness Space is not significantly determined by how we specify the dependent variable. In both situations, there is one large connected component with correlates where the central role is played by the overweight prevalence. Thus, the obtained maps suggest the first step in the construction of the linear-regression model for the COVID-19 infections/death rate in the first wave is by first focusing on the fraction of overweight persons in the country. Moreover, almost all other variables belong to the same component. Only in the case when the dependent variable is modelled through the COVID-19 deaths, Life expectancy and Health coverage are excluded from the component. Hence, the variables included in our analysis are complements in explaining the COVID-19 infections/death rates. Based on this finding, we once again assert that the next variables that will be included in the model, should be specific for the economy that is

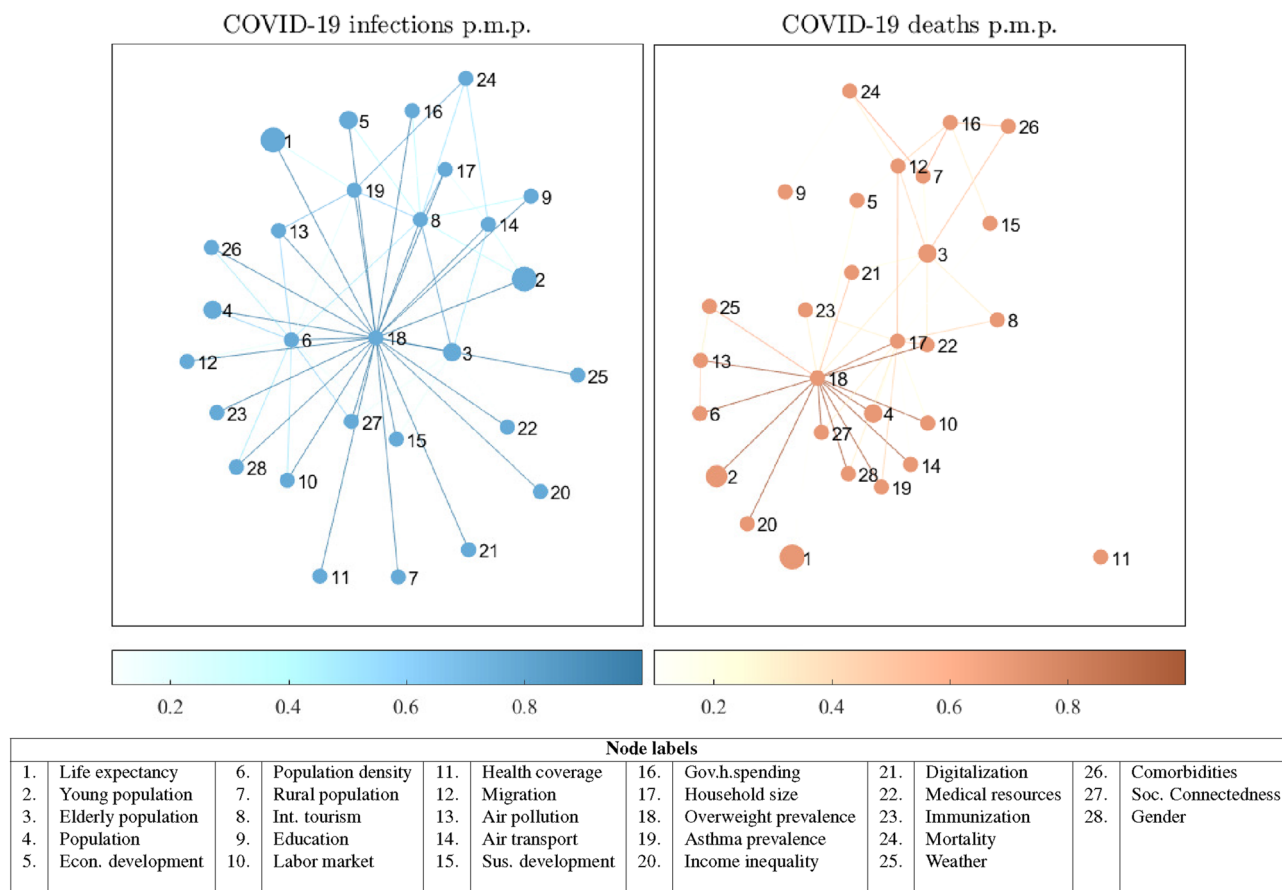


Figure 3. Jointness Space of the COVID-19 correlates. The color of the edge between a pair of correlates is proportional to their Jointness metric. To visualize the network, we use the Force-Layout drawing algorithm.

the subject of the study. Nonetheless, improving the features of the correlates that are located more centrally might yield a synergistic effect, thus significantly reducing the risk of a more negative COVID-19 infections/death rate.

Conclusion and discussion

In this work, we utilized Bayesian model averaging techniques to provide a comprehensive analysis for the health, social and economic correlates of that contributed to between country differences in the final number of infections and deaths during the first wave of the COVID-19 pandemic. Our findings suggest that government response policies, such as testing procedures, tracking of individuals and social distancing measures, and the state of the dynamics of the disease spread can significantly explain the variety in the coronavirus outcome between the countries. Aside from these variables, only a handful of additional variables are able to robustly explain the extent of the COVID-19 infection/deaths and thus provide general rules for the virus spread.

The sole variable strongly related to the coronavirus deaths is the overweight prevalence. Countries with a larger fraction of overweight population also show greater susceptibility to fatal virus outcomes. Interestingly, besides the overweight prevalence, the population density is also a strong correlate of the registered coronavirus infections per million population. More densely populated countries display higher infection rates. Plentiful explanations can provide a possible interpretation for these results. For instance, it is known that the degree of disease spread scales proportionally with population density⁹⁸. This is because, everything else considered, in denser populations typically there is more social mixing²¹. In a similar fashion, various explanations can be found for the observed effect of overweight prevalence. In particular, the prevalence of overweight people is closely related to unhealthy habits of living and, hence, larger susceptibility to both disease infections and fatal outcomes.

The robustness checks and the performed jointness analysis suggested that the insignificance of the other variables might not be the reason for their low PIP values. Instead, the variables which we studied have complementary effects in explaining the COVID-19 infections and death rates of the first wave of the pandemic. This led us to suspect that the results are driven by the heterogeneous health, social, and economic features of the countries. To this end, an interesting topic for future research would be to explore how the effect of the correlates evolved during the different waves of the pandemic. In the absence of a unifying framework covering the relevant aspects of the interrelation between the potential correlates during the various waves, the jointness analysis performed here (and the resulting Jointness Space) can provide the starting point for the development of a more comprehensive understanding of the factors determining the infection and mortality rates of the pandemic. Moreover,

with an improved understanding of the dynamics of the coronavirus pandemic, the insights obtained from this analysis can influence the development of appropriate policy recommendations.

Methods

The methods and data used in this analysis are described in detail in the Supplementary Information document. The data used in the analysis are available at <https://github.com/pero-jolak/coronavirus-socio-economic-determinants>. All experiments were performed in accordance with relevant guidelines and regulations.

Data availability

The data used in the analysis are available at <https://github.com/pero-jolak/coronavirus-socio-economic-determinants>.

Received: 24 November 2020; Accepted: 7 April 2022

Published online: 02 May 2022

References

- Singu, S., Acharya, A., Challagundla, K. & Byrareddy, S. N. Impact of social determinants of health on the emerging covid-19 pandemic in the United States. *Front. Public Health* **8**, 406 (2020).
- Galanis, G. & Hanieh, A. Incorporating social determinants of health into modelling of covid-19 and other infectious diseases: A baseline socio-economic compartmental model. *Soc. Sci. Med.* **274**, 113794 (2021).
- Rollston, R. & Galea, S. Covid-19 and the social determinants of health (2020).
- Kleitman, S. *et al.* To comply or not comply? A latent profile analysis of behaviours and attitudes during the covid-19 pandemic. *PLoS ONE* **16**, e0255268 (2021).
- Clouston, S. A., Natale, G. & Link, B. G. Socioeconomic inequalities in the spread of coronavirus-19 in the United States: A examination of the emergence of social inequalities. *Soc. Sci. Med.* **268**, 113554 (2021).
- Gardner, W., States, D. & Bagley, N. The coronavirus and the risks to the elderly in long-term care. *J. Aging Soc. Policy* **32**, 1–6 (2020).
- Lima, C. K. T. *et al.* The emotional impact of coronavirus 2019-nCoV (new coronavirus disease). *Psychiatry Res.* **287**, 112915 (2020).
- Tanne, J. H. *et al.* Covid-19: How doctors and healthcare systems are tackling coronavirus worldwide. *Bmj* **368**, m1090 (2020).
- Mikhael, E. M. & Al-Jumaili, A. A. Can developing countries alone face corona virus? An Iraqi situation. *Public Health Pract.* **1**, 100004 (2020).
- Di Marco, M. *et al.* Opinion: Sustainable development must account for pandemic risk. *Proc. Natl. Acad. Sci.* **117**, 3888–3892 (2020).
- Wu, X., Nethery, R. C., Sabath, B. M., Braun, D. & Dominici, F. Exposure to air pollution and covid-19 mortality in the United States. *medRxiv* (2020).
- Raftery, A. E., Madigan, D. & Hoeting, J. A. Bayesian model averaging for linear regression models. *J. Am. Stat. Assoc.* **92**, 179–191 (1997).
- Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. Bayesian model averaging: A tutorial. *Stat. Sci.* **14**, 382–401 (1999).
- Sala-i Martin, X., Doppelhofer, G. & Miller, R. I. Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *Am. Econ. Rev.* **94**, 813–835 (2004).
- Fragoso, T. M., Bertoli, W. & Louzada, F. Bayesian model averaging: A systematic review and conceptual classification. *Int. Stat. Rev.* **86**, 1–28 (2018).
- Wu, J. T., Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. *Lancet* **395**, 689–697 (2020).
- Kucharski, A. J. *et al.* Early dynamics of transmission and control of covid-19: A mathematical modelling study. *Lancet Infect. Dis.* **20**, 553–558 (2020).
- Bailey, N. T. *et al.* *The Mathematical Theory of Infectious Diseases and Its Applications* (Charles Griffin & Company Ltd, 1975).
- Van den Driessche, P. & Watmough, J. Further notes on the basic reproduction number. In *Mathematical Epidemiology* (eds Brauer, F. *et al.*) 159–178 (Springer, 2008).
- Keeling, M. J. & Rohani, P. *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, 2011).
- Klepac, P. *et al.* Contacts in context: Large-scale setting-specific social mixing matrices from the BBC pandemic project. *medRxiv* (2020).
- Wang, Y. & Beydoun, M. A. The obesity epidemic in the United States—gender, age, socioeconomic, racial/ethnic, and geographic characteristics: A systematic review and meta-regression analysis. *Epidemiol. Rev.* **29**, 6–28 (2007).
- Fogli, A. & Veldkamp, L. Germs, social networks and growth. Tech. Rep., National Bureau of Economic Research (2012).
- Carr, A. S., Cardwell, C. R., McCarron, P. O. & McConville, J. A systematic review of population based epidemiological studies in Myasthenia gravis. *BMC Neurol.* **10**, 46 (2010).
- Moral-Benito, E. Model averaging in economics: An overview. *J. Econ. Surv.* **29**, 46–75 (2015).
- Moral-Benito, E. Determinants of economic growth: A Bayesian panel data approach. *Rev. Econ. Stat.* **94**, 566–579 (2012).
- Santa, M., Stojkoski, V., Josimovski, M., Trpevski, I. & Kocarev, L. Robust determinants of companies' capacity to innovate: A Bayesian model averaging approach. *Technol. Anal. Strateg. Manag.* **31**, 1283–1296 (2019).
- Glocker, C. & Piribauer, P. The determinants of output losses during the covid-19 pandemic. *Econ. Lett.* **204**, 109923 (2021).
- Hale, T., Petherick, A., Phillips, T. & Webster, S. Variation in government responses to covid-19. *Blavatnik School of Government Working Paper* 31 (2020).
- Zanakis, S. H., Alvarez, C. & Li, V. Socio-economic determinants of HIV/AIDS pandemic and nations efficiencies. *Eur. J. Oper. Res.* **176**, 1811–1838 (2007).
- Itzwerth, R. L., MacIntyre, C. R., Shah, S. & Plant, A. J. Pandemic influenza and critical infrastructure dependencies: Possible impact on hospitals. *Med. J. Aust.* **185**, S70–S72 (2006).
- Whitley, R. J. & Monto, A. S. Seasonal and pandemic influenza preparedness: A global threat. *J. Infect. Dis.* **194**, S65–S69 (2006).
- Breiman, R. F., Nasidi, A., Katz, M. A., Njenga, M. K. & Vertefeuille, J. Preparedness for highly pathogenic avian influenza pandemic in Africa. *Emerg. Infect. Dis.* **13**, 1453 (2007).
- Adini, B., Goldberg, A., Cohen, R. & Bar-Dayyan, Y. Relationship between equipment and infrastructure for pandemic influenza and performance in an avian flu drill. *Emerg. Med. J.* **26**, 786–790 (2009).
- Garrett, A. L., Park, Y. S. & Redlener, I. Mitigating absenteeism in hospital workers during a pandemic. *Disaster Med. Public Health Prep.* **3**, S141–S147 (2009).

36. Oshitani, H., Kamigaki, T. & Suzuki, A. Major issues and challenges of influenza pandemic preparedness in developing countries. *Emerg. Infect. Dis.* **14**, 875 (2008).
37. Gizelis, T.-I., Karim, S., Østby, G. & Urdal, H. Maternal health care in the time of Ebola: A mixed-method exploration of the impact of the epidemic on delivery services in Monrovia. *World Dev.* **98**, 169–178 (2017).
38. Hosseini, P., Sokolow, S. H., Vandegrift, K. J., Kilpatrick, A. M. & Daszak, P. Predictive power of air travel and socio-economic data for early pandemic spread. *PLoS ONE* **5**, e12763 (2010).
39. Quinn, S. C. & Kumar, S. Health inequalities and infectious disease epidemics: A challenge for global health security. *Biosecur. Bioterror. Biodefense Strategy Pract. Sci.* **12**, 263–273 (2014).
40. Hogan, D. R., Stevens, G. A., Hosseini, A. R. & Boerma, T. Monitoring universal health coverage within the sustainable development goals: Development and baseline data for an index of essential health services. *Lancet Global Health* **6**, e152–e168 (2018).
41. Marmot, M. Social determinants of health inequalities. *Lancet* **365**, 1099–1104 (2005).
42. Chen, S.-C. & Liao, C.-M. Modelling control measures to reduce the impact of pandemic influenza among schoolchildren. *Epidemiol. Infect.* **136**, 1035–1045 (2008).
43. Kelly, E. The scourge of Asian flu in utero exposure to pandemic influenza and the development of a cohort of British children. *J. Hum. Resour.* **46**, 669–694 (2011).
44. Nguyen-Van-Tam, J. S. & Hampson, A. W. The epidemiology and clinical impact of pandemic influenza. *Vaccine* **21**, 1762–1768 (2003).
45. van Susan, D., Beulens, J. W., van der Yvonne, T. S., Grobbee, D. E. & Nealb, B. The global burden of diabetes and its complications: An emerging pandemic. *Eur. J. Cardiovasc. Prev. Rehabil.* **17**, s3–s8 (2010).
46. Allard, R., Leclerc, P., Tremblay, C. & Tannenbaum, T.-N. Diabetes and the severity of pandemic influenza A (H1N1) infection. *Diabetes Care* **33**, 1491–1493 (2010).
47. Connolly, M. A. *et al.* Communicable diseases in complex emergencies: Impact and challenges. *Lancet* **364**, 1974–1983 (2004).
48. Abrams, E. M., W't Jong, G. & Yang, C. L. Asthma and covid-19. *CMAJ* **192**, E551–E551 (2020).
49. Bassim, C. W., Gibson, G., Ward, T., Paphides, B. M. & DeNucci, D. J. Modification of the risk of mortality from pneumonia with oral hygiene care. *J. Am. Geriatr. Soc.* **56**, 1601–1607 (2008).
50. Müller, F. Oral hygiene reduces the mortality from aspiration pneumonia in frail elders. *J. Dent. Res.* **94**, 14S–16S (2015).
51. Strauss, J. & Thomas, D. Health, nutrition, and economic development. *J. Econ. Lit.* **36**, 766–817 (1998).
52. i Casasnovas, G. L. *et al.* *Health and Economic Growth: Findings and Policy Implications* (MIT Press, 2005).
53. Sachs, J. *Macroeconomics and Health: Investing in Health for Economic Development* (World Health Organization, 2001).
54. Ashraf, Q. H., Lester, A. & Weil, D. N. When does improving health raise GDP?. *NBER Macroecon. Annu.* **23**, 157–204 (2008).
55. Wobst, P. & Arndt, C. HIV/AIDS and labor force upgrading in Tanzania. *World Dev.* **32**, 1831–1847 (2004).
56. Markowitz, S., Nesson, E. & Robinson, J. The effects of employment on influenza rates. Tech. Rep., National Bureau of Economic Research (2010).
57. Preston, S. H. The changing relation between mortality and level of economic development. *Popul. Stud.* **29**, 231–248 (1975).
58. James, S. L., Gubbins, P., Murray, C. J. & Gakidou, E. Developing a comprehensive time series of GDP per capita for 210 countries from 1950 to 2015. *Popul. Health Metr.* **10**, 12 (2012).
59. Nagano, H., de Oliveira, J. A. P., Barros, A. K. & Junior, A. D. S. C. The 'heart kuznets curve'? Understanding the relations between economic development and cardiac conditions. *World Dev.* **132**, 104953 (2020).
60. Bezruchka, S. The effect of economic recession on population health. *Cmaj* **181**, 281–285 (2009).
61. Granados, J. A. T. & Ionides, E. L. The reversal of the relation between economic growth and health progress: Sweden in the 19th and 20th centuries. *J. Health Econ.* **27**, 544–563 (2008).
62. Wilkinson, R. & Pickett, K. The spirit level. *Why equality is better for everyone* (2010).
63. Ezzati, M., Friedman, A. B., Kulkarni, S. C. & Murray, C. J. The reversal of fortunes: Trends in county mortality and cross-county mortality disparities in the United States. *PLoS Med.* **5**, e66 (2008).
64. Siddiqi, A. & Hertzman, C. Towards an epidemiological understanding of the effects of long-term institutional changes on population health: A case study of Canada versus the USA. *Soc. Sci. Med.* **64**, 589–603 (2007).
65. Kawachi, I. & Kennedy, B. P. Income inequality and health: Pathways and mechanisms. *Health Serv. Res.* **34**, 215 (1999).
66. Krisberg, K. Income inequality: When wealth determines health: Earnings influential as lifelong social determinant of health (2016).
67. Putnam, R. Social capital: Measurement and consequences. *Can. J. Policy Res.* **2**, 41–51 (2001).
68. Folland, S. Does, "community social capital" contribute to population health?. *Soc. Sci. Med.* **64**, 2342–2354 (2007).
69. Lee, C.-J. & Kim, D. A comparative analysis of the validity of US state-and county-level social capital measures and their associations with population health. *Soc. Indic. Res.* **111**, 307–326 (2013).
70. Baker, D. P., Leon, J., Smith Greenaway, E. G., Collins, J. & Movit, M. The education effect on population health: A reassessment. *Popul. Dev. Rev.* **37**, 307–332 (2011).
71. Mackenbach, J. P. *et al.* Socioeconomic inequalities in health in 22 European countries. *N. Engl. J. Med.* **358**, 2468–2481 (2008).
72. Hens, N. *et al.* Estimating the impact of school closure on social mixing behaviour and the transmission of close contact infections in eight European countries. *BMC Infect. Dis.* **9**, 187 (2009).
73. Mossong, J. *et al.* Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5**, e74 (2008).
74. Melegaro, A., Jit, M., Gay, N., Zagheni, E. & Edmunds, W. J. What types of contacts are important for the spread of infections? Using contact survey data to explore European mixing patterns. *Epidemics* **3**, 143–151 (2011).
75. Prem, K., Cook, A. R. & Jit, M. Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLoS Comput. Biol.* **13**, e1005697 (2017).
76. Wallinga, J., Teunis, P. & Kretzschmar, M. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *Am. J. Epidemiol.* **164**, 936–944 (2006).
77. Erkoreka, A. The Spanish influenza pandemic in occidental Europe (1918–1920) and victim age. *Influenza Other Respir. Viruses* **4**, 81–89 (2010).
78. Armstrong, G. L., Conn, L. A. & Pinner, R. W. Trends in infectious disease mortality in the united states during the 20th century. *Jama* **281**, 61–66 (1999).
79. Ainsworth, M. & Dayton, J. The impact of the aids epidemic on the health of older persons in northwestern Tanzania. *World Dev.* **31**, 131–148 (2003).
80. Mastrandrea, R., Fournet, J. & Barrat, A. Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS ONE* **10**, e0136497 (2015).
81. Kucharski, A. J. *et al.* The contribution of social behaviour to the transmission of influenza a in a human population. *PLoS Pathog.* **10**, e1004206 (2014).
82. Braga, A., Zanobetti, A. & Schwartz, J. Do respiratory epidemics confound the association between air pollution and daily deaths?. *Eur. Respir. J.* **16**, 723–728 (2000).
83. Clay, K., Lewis, J. & Severini, E. Pollution, infectious disease, and mortality: Evidence from the 1918 Spanish influenza pandemic. *J. Econ. Hist.* **78**, 1179–1209 (2018).
84. Simiao, C. *et al.* Climate and the spread of covid-19. *Sci. Rep.* **11**, 1–6 (2021).

85. Lighter, J. *et al.* Obesity in patients younger than 60 years is a risk factor for covid-19 hospital admission. *Clin. Infect. Dis.* **71**, 896–897 (2020).
86. Sattar, N., McInnes, I. B. & McMurray, J. J. Obesity a risk factor for severe covid-19 infection: Multiple potential mechanisms. *Circulation* **142**, 4–6 (2020).
87. Stefan, N., Birkenfeld, A. L., Schulze, M. B. & Ludwig, D. S. Obesity and impaired metabolic health in patients with covid-19. *Nat. Rev. Endocrinol.* **16**, 1–2 (2020).
88. Bailey, M., Cao, R., Kuchler, T., Stroebel, J. & Wong, A. Social connectedness: Measurement, determinants, and effects. *J. Econ. Perspect.* **32**, 259–80 (2018).
89. Kuchler, T., Russel, D. & Stroebel, J. The geographic spread of covid-19 correlates with structure of social networks as measured by Facebook. Tech. Rep., National Bureau of Economic Research (2020).
90. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
91. Ver Hoef, J. M. & Boveng, P. L. Quasi-poisson versus negative binomial regression: How should we model overdispersed count data? *Ecology* **88**, 2766–2772 (2007).
92. Krisztin, T., Piribauer, P. & Wögerer, M. The spatial econometrics of the coronavirus pandemic. *Lett. Spatial Resour. Sci.* **13**, 209–218 (2020).
93. Doppelhofer, G. & Weeks, M. Jointness of growth determinants. *J. Appl. Econom.* **24**, 209–244 (2009).
94. Hofmarcher, P., Cuaresma, J. C., Grun, B., Humer, S. & Moser, M. Bivariate jointness measures in Bayesian model averaging: Solving the conundrum. *J. Macroecon.* **57**, 150–165 (2018).
95. Meeuwesen, L., van den Brink-Muinen, A. & Hofstede, G. Can dimensions of national culture predict cross-national differences in medical communication?. *Patient Educ. Couns.* **75**, 58–66 (2009).
96. Mossialos, E., Wenzl, M., Osborn, R. & Sarnak, D. *2015 International Profiles of Health Care Systems* (Canadian Agency for Drugs and Technologies in Health, 2016).
97. Boas, T. C. & Hidalgo, F. D. Electoral incentives to combat mosquito-borne illnesses: Experimental evidence from Brazil. *World Dev.* **113**, 89–99 (2019).
98. Draief, M., Ganesh, A. & Massoulié, L. Thresholds for virus spread on networks. In *Proceedings of the 1st International Conference on Performance Evaluation Methodologies and Tools*, 51–es (2006).

Author contributions

V.S.: conceptualization, methodology, software, data curation, validation, formal analysis, investigation, writing-original draft. Z.U.: conceptualization, methodology, validation, investigation, writing-original draft, writing-review and editing. P.J.: formal analysis, investigation, methodology, software, data curation and validation. D.T.: conceptualization, methodology, validation, investigation, writing-original draft, writing-review and editing. L.K.: conceptualization, supervision, writing-review and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-10894-6>.

Correspondence and requests for materials should be addressed to V.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022