In the face of the pandemic, are all equal? On the suitability of the Gini index to monitor time and geographic trends in incidence and death during the SARS-CoV-2 pandemic

Kirsi Manz[1,*] **and Ulrich Mansmann**[1]

[1]Institute of Medical Information Processing, Biometry, and Epidemiology, Ludwig-Maximilians University Munich, Munich, 81377, Germany
[*]manz@ibe.med.uni-muenchen.de

ABSTRACT

Health inequalities across nations reinforce social and economic differences. The European Commission has outlined its commitment to reducing health inequalities.The SARS-CoV-2 pandemic is a disruptive event with global, national, and local effects on every aspect of a society's life. Therefore, it is of interest to quantify and display health (in)equality induced by COVID-19 over time and on different geographic scales (global, continental, national, regional).
We aimed to perform a study of (in)equality regarding COVID-19 related incidences and deaths using the Gini index (GI) that was developed to explore inequality in wealth within or between nations. The GI is a relative measure which quantifies the ratio of the mean absolute difference between observations and their mean: GI=0 describes a situation where everybody owns the same, GI=1 corresponds to one person owning the entire wealth of a society. A specific focus of our analysis is the (in)equality of case-fatality rates. We use global, nation- and statewide data to evaluate the heterogeneity at different levels. We also reflect on the suitability of the Gini index to monitor equality aspects of the pandemic.
Data from the SARS-CoV-2 pandemic show a considerable amount of heterogeneity in the global distribution of incidences and deaths. Only few countries contribute largely to the total incidence and death toll. Inequality may result from heterogeneous reporting systems, heterogeneous policy making to fight the pandemic, heterogeneous health systems, and demographic structures. Inequality decreases when going from continental to national or regional levels. Time trends on national level reflect local outbreaks. There is heterogeneity in case fatality rates, too. There is also a specific numeric artefact of the GI: Even if many smaller regions have no COVID-19 related deaths at all, and one region has low (but still non-negligible) number of deaths the corresponding GI is near to 1.
We used the Gini index to reflect inequality between incidence and death rates of nations or geographic regions. The interpretation of the differences between reported GIs is not straightforward. However, combined with the information presented by maps, the GI is a useful tool.

Introduction

Increasing health inequalities both within and between countries continue to remain a challenge both globally and for the European Union. For example, a review of health inequalities for the WHO European region found that life expectancy differs significantly across the regions and even in affluent countries inequities have increased[1]. Health 2020 is the new European health policy framework. It aims to support actions across governments and societies to "significantly improve the health and well-being of populations, reduce health inequalities, strengthen public health and ensure people-centred health systems that are universal, equitable, sustainable and of high quality".

The SARS-CoV-2 pandemic is a disruptive event with global, national, and local effects on every aspect of a society's life. Therefore, it is of interest to quantify and display COVID-19 related health (in)equality over time and on different geographic scales (global, continental, national, regional) regarding COVID-19 using incidence and death rates.

In general, incidence and/or death rates are presented on maps or listings, but not accompanied by measures of inequality. Such measures help to understand the graphical representations and changes over time and additionally allow for comparisons between and within sub-regions.

In this paper we propose and discuss the Gini index[2] as a possible candidate to quantify inequality (heterogeneity) between populations of interest. The Gini index (GI) was invented to measure heterogeneity between individual wealth within a society.

It is applied to a set of non-negative observations, such as individual fortunes within a society (while not counting negative fortunes representing individual debts). The Gini index is a relative measure: the ratio of the mean absolute difference between two observations within a population and their mean. It is a number between zero and one which describe perfect equality (GI = 0) and perfect inequality (GI = 1). GI = 0 holds if each individual within a society owns the same part of the society's wealth, GI = 1 holds if only one individual owns everything.

Inequality is a different concept as *variability*. Following the example of the last paragraph, we compare two data sets: (1) a set of 100 observations containing 99 times the value of zero and one observation with the value 100, (2) a set of 100 observations containing 99 times the value of 100 and one observation with the value zero. Both data sets have the same variability, since the empirical standard deviation is 10, but the GIs differ to a substantial degree: Whereas the empirical GI for the first set is 0.99 quantifying extensive inequality, the empirical GI for the second set is 0.01 quantifying nearly perfect equality. Applying this idea to the current pandemic leads to the following considerations: According to the first example we have 100 countries where COVID-19 is present, but of these 99 countries only one country has reported deaths and 99 countries have no reported deaths. This "good case" leads to a very high GI. The second example would be rephrased as referring to 99 heavily affected countries (with a death rate of 100) and one country without COVID-19 related deaths (and thus death rate of 0). In this "bad case" the Gini index would indicate very low heterogeneity, which we call "equality in bad".

A property of the GI is that it is invariant with regard to multiplication: $\mathrm{GI}_X = \mathrm{GI}_Y$ for $X = a \cdot Y$ where X, Y are non-negative random variables and a is a positive number. Another property of the Gini index is that if a constant amount is added (subtracted) to (from) each value the Gini index decreases (increases). This leads to a question what happens if many regions are able to reduce their COVID-19 related deaths to zero by a shift of values towards zero. This may result in many zero observations and only a few regions being responsible for the death load of the pandemic. Following the definition of GI, this reflects increasing inequality between death rates and at the same time a reduced total amount of pandemic related deaths. More technical details on the Gini index are presented in the Methods section.

The multiplicative invariance of the Gini index has a straightforward application to the case fatality rate (CFR) of COVID-19: If the CFR is the same for all infected (over the range of populations considered) the Gini indices for incidence and death rates are the same: $\mathrm{GI}_{\mathrm{Incidence}} = \mathrm{GI}_{\mathrm{Death}}$. Heterogeneity of demographic features such as age or preexisting comorbidities between regions results in region-specific case fatality rates. This will be reflected in the corresponding indices. Therefore, our comparison of GIs uses standardized rates. Given age and sex standardized rates, what is the meaning of an observation $\mathrm{GI}_{\mathrm{Incidence}} < \mathrm{GI}_{\mathrm{Death}}$?

The aim of this paper is to gain insight into the global and regional heterogeneity in the distribution of COVID-19 related incidences and deaths by evaluating the Gini index over time and within more or less extensive geographical regions. Further, we asked ourselves, which relevant information is hidden in the Gini index and what can be learned from the comparisons. In the following paper the term *heterogeneity* is used in the context of *inequality* and not, as often used, for *variability*.

Results

The global incidence and death data

Up to July 28, 2020, globally around 16.68 million (16,681,275) confirmed SARS-CoV-2 cases have been reported. The transmission of the virus has been documented in total of 186 countries and the reported number of cases per country varies between 10 (Western Sahara) and 4.35 million cases (United States). Deaths due to the virus have been documented in 170 countries, whereas the number of deaths per country varies between 0 and 149,256. The US has worldwide the highest total number of confirmed incidences and deaths (4,351,997 and 149,256), followed by Brazil (2,483,191 and 88,539) and India (2,294,438 and 45,597).

Gini index for global and international data

The Gini indices between nations for reported deaths and incidences over the time course of the pandemic were studied in two ways: By a *cumulative approach* referring to the development of the numbers within a nation aggregated from the beginning up to time t and by a *periodic approach* presenting the numbers within a nation aggregated over a short time period of past seven days.

Figure 1 provides the global view over 186 countries (panels a and b) and the situation in the People's Republic of China (PRC, Mainland China, panels c and d). The left panels show the cumulative and the right ones the periodic approach. The cumulative data show higher GIs for the death rates compared to the incidence rates: $\mathrm{GI}_{\mathrm{Incidence}} < \mathrm{GI}_{\mathrm{Death}}$. At the beginning of the epidemic, there are only a few cases in a specific region and most other regions have no infected or deaths. Therefore, at the beginning of the pandemic dynamics it is natural that GI ~ 1. As the pandemic spreads into more regions, the Gini index starts to decrease from its starting value. Still, globally both indices remain at a high level (Figure 1a).

The observations from mainland China reflect inequality between the provinces of PRC and demonstrate a specific property of the GI: Incidences were mainly observed in a few number of provinces, reflecting high inequality. The difference between

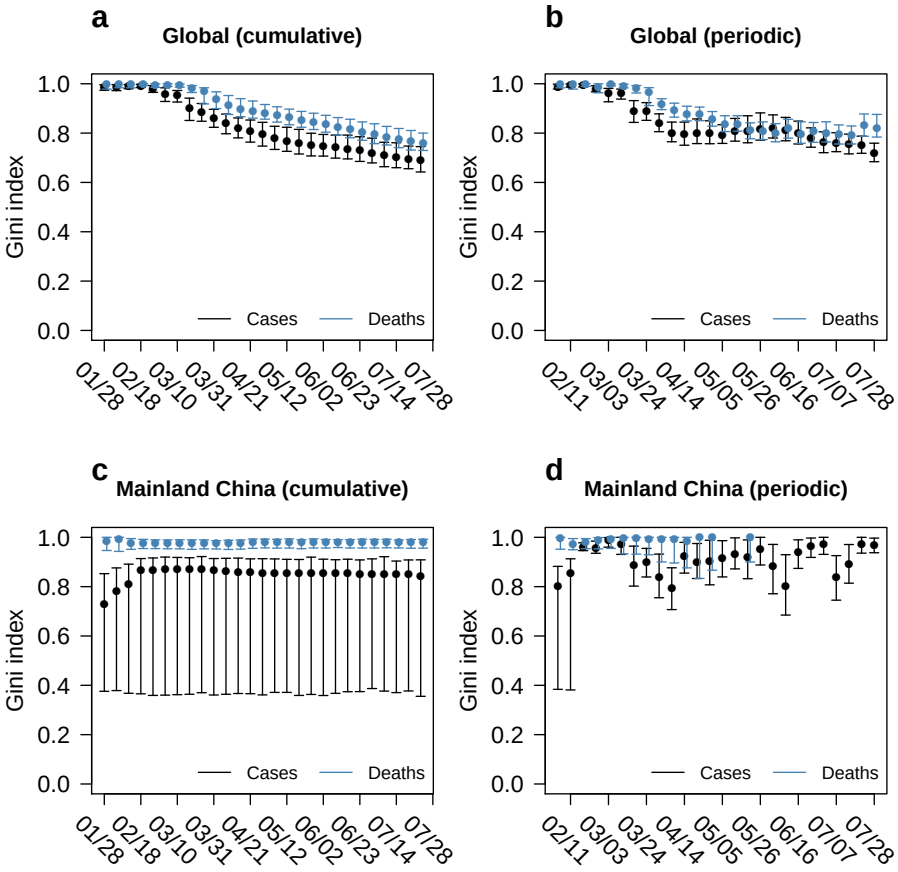

Figure 1. Time course of Gini indices describing incidence (black) and death data (blue) of the COVID-19 pandemic. The horizontal axis shows the calendar time between January 28th and July 28th, 2020. Shown are the global estimates (a and b) and estimates for mainland China (c and d) together with their 95% confidence intervals. The left panels show the cumulative and the right ones the periodic approach. For better visualization the death estimates are shown one day apart from the incidence estimates.

the *cumulative* and the *7-day* perspective is reflected in the relationship between both Gini indices $GI_{Incidence}$ and GI_{Death}. Therefore, CRFs may look quite heterogeneous when being calculated from the aggregated data, which is not so pronounced when CRFs are calculated for specific time periods. The CFRs for each province are shown in the Supplementary Table S1. Of note, since no deaths were reported for China in July 2020 in any province, the Gini index for the death rates could not be calculated. Theoretically, this situation corresponds to perfect equality ("nothing happened anywhere") but the Gini index is not defined as its calculation would require division by the value of zero (mean).

Figure 2 shows a comparison between the European Union (EU, panels a and b) and the United States (US, panels c and d) with left panels showing the cumulative and the right panels the periodic approach. The EU and US data show less heterogeneity than seen for the global data. In Figure 2a, the GI values for the EU are persistent over time while more dynamics seems to be

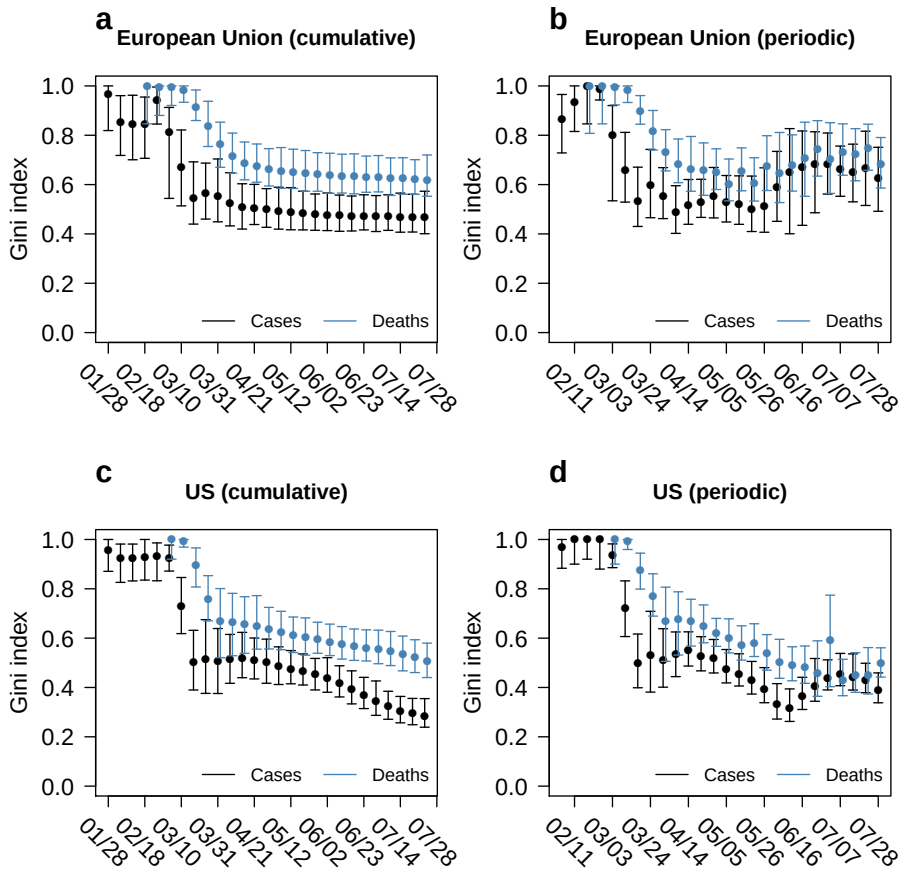

Figure 2. Time course of Gini indices describing incidence (black) and death data (blue) of the COVID-19 pandemic. The horizontal axis shows the calendar time between January 28th and July 28th, 2020. Shown are the estimates for the countries of the European Union (a and b) and for the states of the United States (US, c and d) together with their 95% confidence intervals. The left panels show the cumulative and the right ones the periodic approach. For better visualization the death estimates are shown one day apart from the incidence estimates.

present in the *7-day* approach (Figure 2b). The US (Figure 2 panels c and d) seems to have lower GIs compared to the EU in spite of the fact that incidences increase. The US states align in the bad. The corresponding case fatality rates for the countries of the EU and the US states are shown in the Supplementary Tables S2 and S3.

Additionally, the Gini index for different continents was evaluated. Both the *cumulative* view and the *7-day* approach are shown in Figure 3. The GIs were calculated for N COVID-19 affected countries per continent (as of July 28, 2020): Asia N=43, Europe N=49, America N=35, and Africa N=55. No Gini index is shown for Oceania due to the small number of affected countries (N=4). The data for Asia (panels a and b) show crossing curves for $GI_{Incidence}$ and GI_{Death}. It is the only situation in our data with $GI_{Incidence} > GI_{Death}$. The *cumulative* GIs seem to be similar to the *7-day* GIs. Overall, the heterogeneity remains at a high level ($GI \approx 0.7$). In Europe (panels c and d) the pandemic also started early but no deaths were reported during the first few weeks, so that the Gini index could not be calculated. A decrease in the heterogeneity for both incidences and

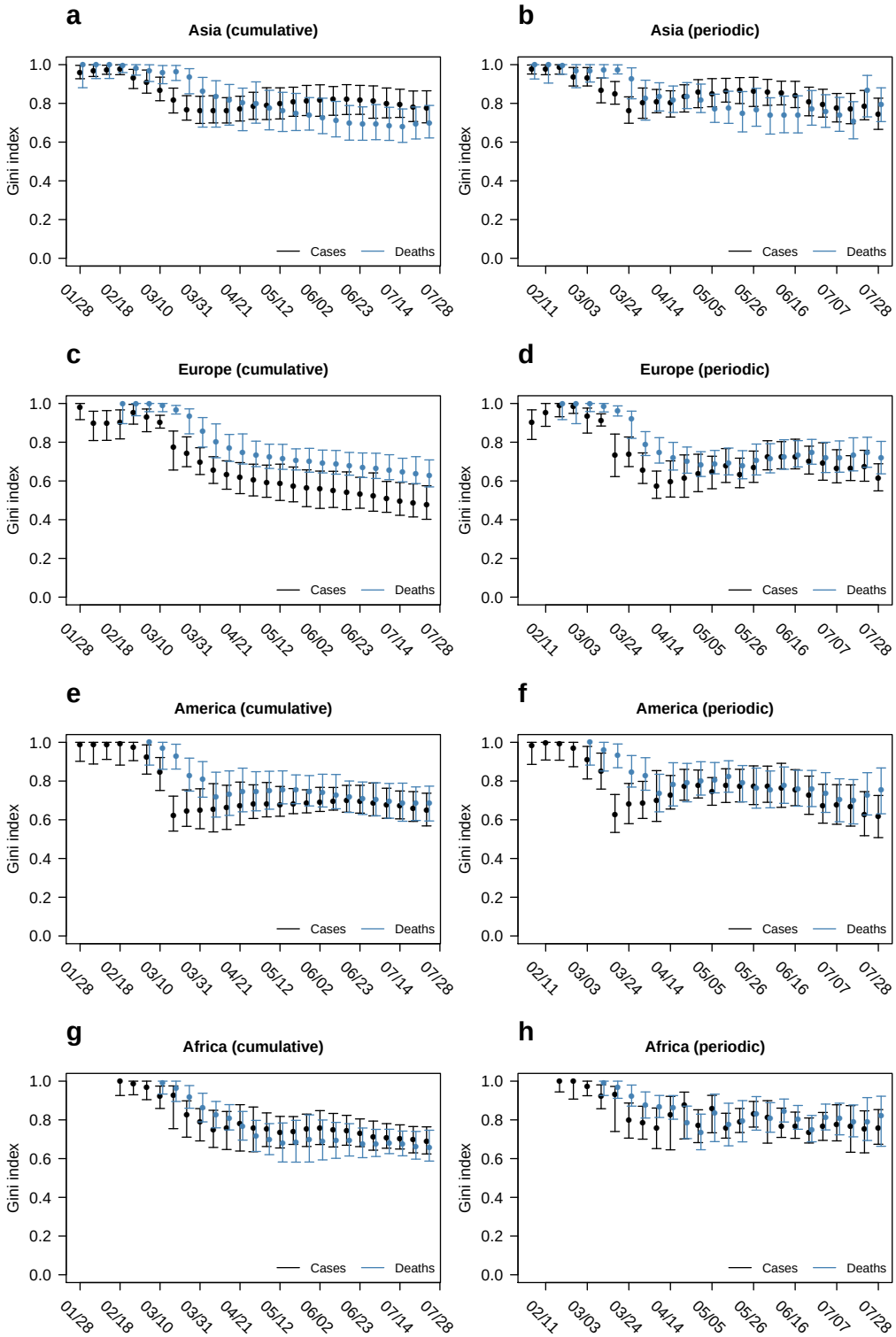

Figure 3. Time course of Gini indices describing incidence (black) and death data (blue) of the COVID-19 pandemic. The horizontal axis shows the calendar time between January 28th and July 28th, 2020. Shown are the estimates for the Asia (a, b), Europe (c, d), America (e, f) and Africa (g, h). The bars represent 95% confidence intervals. The left panels show the cumulative and the right ones the periodic approach. For better visualization the death estimates are shown one day apart from the incidence estimates.

deaths is observed and in the *cumulative* view and a plateau seems not be to reached by the end of July 2020. Countries like Italy reported high death rates at the beginning in comparison to other European countries. Additionally, policies to handle the pandemic differ within European countries. In spite of this fact, Europe seems to have the lowest GI values across the different continents. In the Americas, as shown in panels e and f, no deaths were reported until end of February. After an initial decrease a slight increase followed by a further decrease of the Gini index is observed. By the end of the observation time the heterogeneity for incidences and deaths is comparable. Both the *cumulative* and the *7 day* approaches seem similar. The spread of the pandemic started later in Africa (panels g and h) and the Gini index for both incidences and deaths remains similar. As seen for other continents, too, the heterogeneity remains high.

Gini index for Germany

The German data were analysed between April 16th and July 30th, 2020. Germany-wide a total of 206916 cases and 9128 deaths have been reported up to end of July, 2020. The federal states Bavaria and North Rhine-Westphalia (NRW), are two large federal states which are together responsible for 48% of all reported cases and deaths (Bavaria: 25% of cases, 29% of deaths, NRW: 23% of cases, 19% of deaths). The overall CFR is 4.4% for Germany, and 5.2% and 3.6% for Bavaria and NRW, respectively.

Figure 4 shows the Gini indices for incidence and death rates over time for Germany with different regional focus. Both the *cumulative* and the *7 day* Gini index are shown over the 16 federal states (panels a and b) and over the 401 administrative districts (*Landkreise*, panels c and d). The heterogeneity between the federal states is lower than for the international data seen

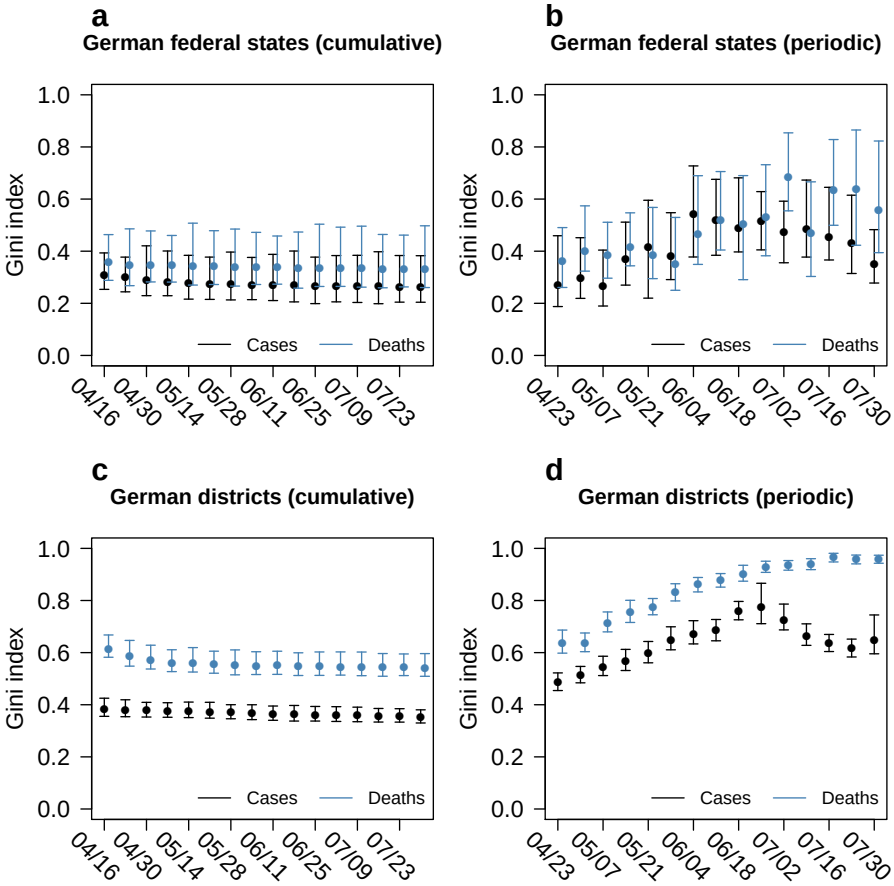

Figure 4. Time course of Gini indices describing incidence (black) and death data (blue) of the COVID-19 pandemic for Germany. The horizontal axis shows the calendar time between April 16th and July 30th, 2020. Shown are the estimates over the federal states (a and b) and over the administrative districts (c and d) both in *cumulative* (left) and *periodic* approach (right). The bars indicate 95% confidence intervals. For better visualization the death estimates are shown one day apart from the incidence estimates.

in Figures 1-3. The *cumulative* perspective appears to less dynamic as the *7-days* approach, where the latter probably reflects

effects of recent outbreaks in Bavaria and NRW. However, at the level of federal states, both $GI_{Incidence}$ and GI_{Death} seem to be similarly heterogeneous. When looking at the Gini index at the district level (panels c and d) for both approaches distinct differences between $GI_{Incidence}$ and GI_{Death} can be seen. The inequality between the administrative units appears larger as at the level of federal states. Local outbreaks make single *Landkreise* for a specific period very distinct from the remaining districts, which leads to high spatial heterogeneity and a high Gini index. Local outbreaks happened in nursing homes, factories, farms, and communal accommodation for refugees. In the *cumulative* approach we observe larger GI_{Death} (0.6) compared to $GI_{Incidence}$ (0.4).

Figure 5 shows the time course of the Gini index for Bavaria (panels a and b) and NRW (panels c and d). Bavaria has 96 and NRW has 53 administrative districts. Both states show high epidemic activity compared to other German federal states. An overview of the underlying data is provided in the Supplementary Table S4. When considering only data from Bavaria

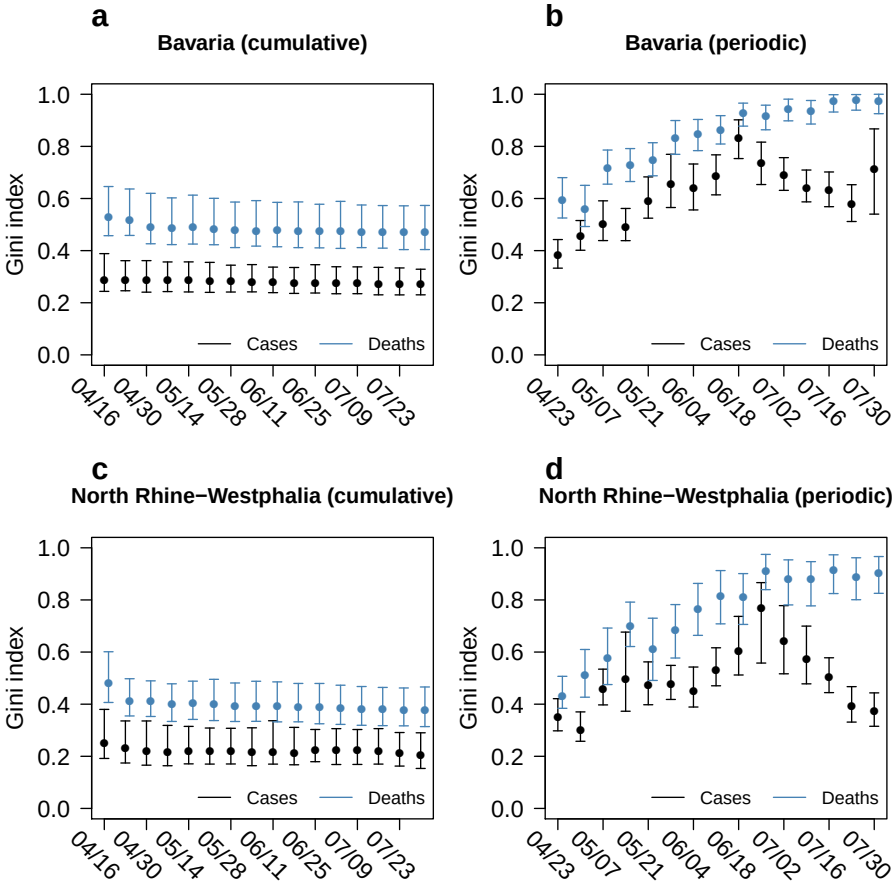

Figure 5. Time course of Gini indices describing incidence (black) and death data (blue) of the COVID-19 pandemic for Germany. The horizontal axis shows the calendar time between April 16th and July 30th, 2020. Shown are the estimates for the federal state of Bavaria (a and b) and North Rhine-Westphalia (c and d) both in *cumulative* (left) and *periodic* approach (right). The bars indicate 95% confidence intervals. For better visualization the death estimates are shown one day apart from the incidence estimates.

(panels a and b) the time course of the Gini index resembles the situation of German district level data (panels c and d of Figure 4): Deaths are more heterogeneously distributed than incidences but the confidence intervals are wider (smaller number of administrative districts, N=96, compared to N=401). The Gini index for incidences is about 0.3 and the one for deaths about 0.5 (*cumulative* approach). Of all German federal states Bavaria has the highest total number of infections and deaths per 100 000 inhabitants: The cumulative incidence rate per 100000 is 388 and death rate per 100000 is 20 as of July 30th, 2020 (for the rates also see Supplementary Table S4). Panel d shows the Gini index for the federal state of North Rhine - Westphalia. It is the largest federal state of Germany with nearly 18 million inhabitants. The inequality of incidence and death rates seems to resemble the situation of Bavaria: Both federal states seem to follow similar equality pattern reflected in similar time courses of $GI_{Incidence}$ and GI_{Death}.

Inhomogeneity of case fatality rates

By means of Poisson regression and likelihood ratio test (LRT) we investigated if the CFRs are homogeneous over the geographical areas. As indicated by Table 1 inhomogeneous CRFs for all regions were found.

Region	No. of observations	Type	p value (LRT)
World	186	countries	**<0.001**
Mainland China	31	provinces	**<0.001**
EU	27	countries	**<0.001**
US	51	states	**<0.001**
Asia	43	countries	**<0.001**
Europe	49	countries	**<0.001**
America	35	countries	**0.004**
Africa	55	countries	**<0.001**
Germany	16	federal states	**<0.001**
Germany	401	districts	**<0.001**
Bavaria	96	districts	**0.030**
North Rhine -Westphalia	53	districts	**0.016**

Table 1. Results (p value) from the likelihood ratio test (LRT) to test for homogeneity of the case fatality rates (CFR) between different geographical regions. For the calculation of the CFR the cumulated incidence and death data up to End of July 2020 was used. Significant p values are shown in bold. EU = European Union, US = United States.

Gini index and disease maps

Gini index can be used together with disease maps to summarize the information given in the maps. An example is shown in Figure 6, where an incidence and a death rate map are shown together with the corresponding graph of the Gini index (*cumulative* view). Both the maps and the Gini index use rates per 100 000. The heterogeneity seen in the maps is summarized by the Gini index. Of note, the maps correspond to the last point used for the calculation of the Gini index (July 28th, 2020). Whereas the maps are restricted to one point in time, showing the whole time course of the Gini index offers additional insights into the development of the heterogeneity over time.

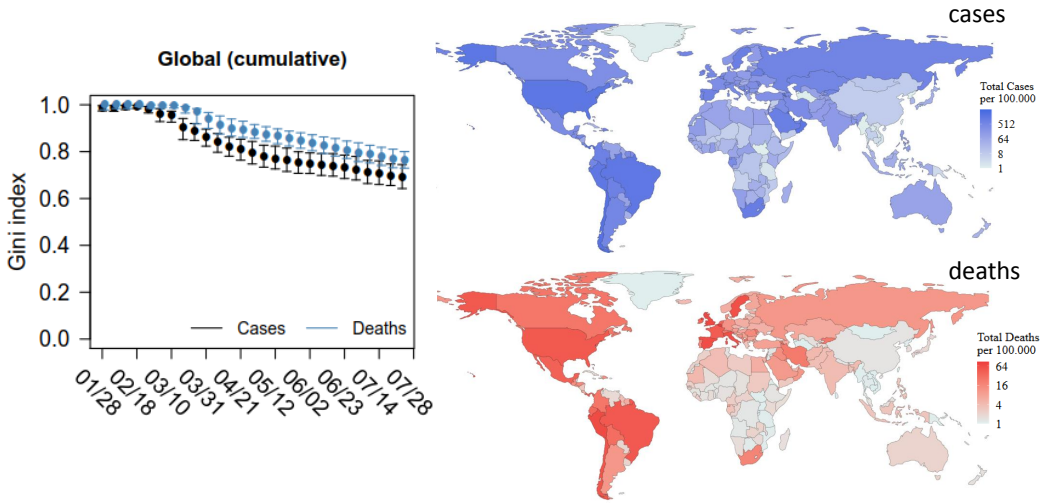

Figure 6. Time course of Gini indices describing global incidence (black) and death rates (blue) of the COVID-19 pandemic (left). The graph of the Gini index is accompanied by incidence rate map (blue-shaded upper map) and death rate map (red-shaded lower map). The maps show rates per 100 000, the same rates as used for the Gini index, too. The maps were extracted from[3].

Another example at a national level is shown in Figure 7 for Germany. The Gini index is shown in the *cumulative* view and the maps correspond to the last time point, where the Gini index was evaluated. By looking at the maps one might assume that

the heterogeneity in the death rates per district could be larger than the heterogeneity in the incidence rates. The graph of the Gini index gives more confidence for such an assumption, since the overall heterogeneity is presented as a number.

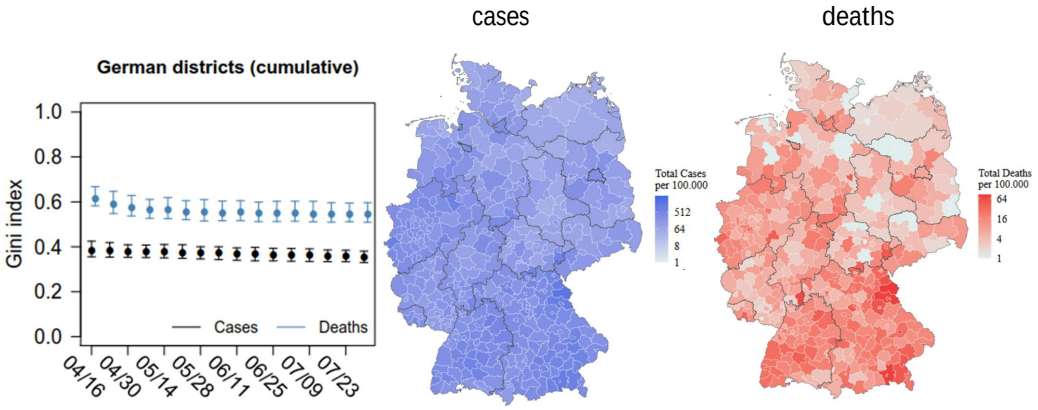

Figure 7. Time course of Gini indices for Germany at a district level for incidence (black) and death rates (blue) of the COVID-19 pandemic (left). The graph of the Gini index is accompanied by incidence rate map (blue-shaded map, middle) and death rate map (red-shaded map, right). The maps show rates per 100 000, the same rates as used for the Gini index, too. The maps were extracted from[3].

Discussion

The aim of this paper is to explore the relevance of the Gini index to allow for better understanding of inequality within the SARS-CoV-2 pandemic incidence and death data. The Gini index is a measure of *inequality* and may be helpful in understanding the information provided by spatial-temporal analyses. A specific interest of our research is also to study if larger spatial scales express more *heterogeneity* compared to analysis on a smaller spatial scale. We also provide two aspects of time course analyses over regions: Whereas the *cumulative* approach looks at the aggregated data over time, the *7-days* approach looks at periodically changing data over time. While in our data the *cumulative* view seems to mostly represent the difference between observational units at the beginning of the pandemic dynamics (it stabilizes during the time course), the *7-days* view is able to react on events during the time course such as the beginning of a second wave or changes in the policies to cope with the epidemic. Also, the cumulative approach delivers more robust disease measures by avoiding time-dependent variation within the data. It is regularly used to provide official disease information: Since start of the epidemic a total of xx persons were reported positive, yy persons died, zz persons recovered.

With regard to the incidences we observe that GIs decrease when the spatial scale of the analysis gets smaller. Inequality is more pronounced on the global scale compared to the situation within a nation or a continent. There is a complementary theoretical approach based on *heavy-tail* properties of distributions[4]. *Heavy-tail* distributions express more distinct inequalities within populations and can be seen as extreme mixtures of common distributions. They present challenges for the data analyst since they may not have estimable means or variances. Under these conditions, many familiar statistical approaches do not work any more. Heavy-tailed phenomena in relation to COVID-19 data have been recently investigated[5]. Heavy-tail observations force us to use distributional information like extreme quantiles to operationalize the information.

An interesting observation is that in many cases the heterogeneity in the distribution of deaths is higher than the heterogeneity in the distribution of cases. This was seen globally, for Europa (including the 27 countries of the EU) and for the US and also within Germany. A specific feature of the GI is its invariance with respect to a multiplication with a positive number. The number which compares incidences and deaths of an epidemic is the *case fatality rate* (CFR). A higher GI_{Death} compared to $GI_{Incidence}$ may reflect heterogeneity of CRFs between the regions compared. The difference is more pronounced in the *cumulative* view. We found inhomogeneous CFRs over the regions considered.

In this context, it is of interest to look at the German data. In panel a of Figure 4, the German federal states show the lowest GIs in our data (GI $\sim$ 0.3 which is comparable to the GI of the uniform distribution). The *7-day* approach shows more dynamics and some GI increase in May. There may be a relationship to the fact that Germany changed its policy by moving from a centrally guided approach to more regional (federal) policies. The federal GI measures for incidences and deaths seem not to be too different. This is different in panels c and d where we look at smaller administrative districts (*Landkreise* and *kreisfreie Städte*), where $GI_{Death} > GI_{Incidence}$ can be observed. One explanation for this might be the numerical artefact of the GI:

As the spectrum of the spread of the infection changes from nationwide spatial spread to more focused local outbreaks, as observed for Germany, there will be more districts with zero reported deaths and the Gini index will increase.

The comparison between Bavaria and NRW is of interest because of the political rivalry between the two most populated German federal states. Both states are similar in the *cumulative* as well as the *7-days* approach. The latter shows specific local outbreaks like in a large meat processing factory (NRW) or at a farm in Markt Inchenhofen (Bavaria), both in June 2020. Also the outbreak in Mamming (Bavaria) end of July 2020 is probably seen in the last data point for Bavaria. We also observe that in NRW the GIs for incidences recently decrease while the GIs for death do continuously increase. It would be interesting to see if the testing strategies introduced in August 2020 to screen holiday returnees may modify the corresponding GIs.

We showed how one could monitor inequality in the distribution of incidences or deaths at different levels using the Gini index. For this, we normalized the data to the respective population and used the incidence and fatality rates per 100000 inhabitants. We found high heterogeneity for both incidences and deaths at global and continent-wide level. This is probably related to different temporary onset of the epidemic, different quality and capacities of health care systems and the control measures implemented by the countries. Globally and in Europe the heterogeneity regarding fatalities across countries is larger than the one for incidences. In Asia the situation has reversed during the time course of the pandemic and in the other continents (America and Africa) incidences and deaths seem to be equally heterogeneously distributed. One possible explanation for the observation seen in the Asian data could be based on the standardization of the data: we used incidence and death rates, which were normalized to the respective population to gain a relative measure comparable across countries. Normalized rates for countries with low number of inhabitants can get large. For example the largest cumulative incidence rates in Asia at the end of July were reported for Qatar and Bahrain (3880 and 2432 per 100 000, respectively) and the largest death rate in Asia was reported for Kyrgyzstan with 21 per 100 000. All these countries have only low numbers of inhabitants for Asian countries (between 1.6 million and 6.4 million) but contribute a substantial amount to the relative heterogeneity, as measured by the Gini index in our study. It seems that the *relative* heterogeneity between incidence rates in Asia is currently higher than in death rates but it remains to be seen if the situation will reverse again.

Heterogeneity between countries can be a consequence of differences in diverse infrastructures, such as the reporting system, the health care system, or the effectivity of implemented control measures. In general, if countries would learn from each other and adapt the best strategy against the epidemic, one would expect a decreasing GI because of a general alignment in strategies. Higher death GIs reflect additional heterogeneity in death outcomes compared to incidences between the countries analysed. Especially, the data does not allow to adjust incidence and death rates for age and sex. It is well known that women and young people have lower risk to die compared to men and people above 70 years[6]. With regard to a set-up of an objective global surveillance system the exploration of heterogeneity contributed by national reporting systems may be an issue, which needs to be addressed.

We used rates instead of absolute numbers to be able to compare small and large countries. Due to the scale invariance property the GIs would not have changed, even if we would have used another rate than per 100 000 inhabitants. This approach is a strength and a weakness at the same time: We are able to compare differently populated countries against each other, but the question remains if the relative rates are the preferred ones which should be used for such comparisons. When using the Gini index for absolute numbers the heterogeneity would be mainly caused by differences in incidences and death numbers between small and large countries. We believe choosing a normalized approach and comparing the heterogeneity of rates is the preferred option to comparing absolute numbers.

Though the Gini index may be useful in quantifying heterogeneity related to COVID-19 burden, it has some shortcomings. The accuracy of the Gini index is dependent on reliably reported incidence and death data. Underreporting of cases and deaths might thus pose a problem. If the underreporting is similar across different countries or regions the Gini index should still be reliable. However, if the reporting rates between different countries or regions differ to a substantial extend, the true inequality might be overestimated by the Gini index. So, in a way, the Gini index shows the worst inequality scenario. Other way around, an underestimation of the true heterogeneity is not probable to occur. Another limitation is that very different incidence and death distributions can result in identical Gini indices. Because the Gini index attempts to summarize a two dimensional area (namely the gap between the Lorenz curve and the equality line) in to a single number, it masks information about the "shape" of inequality. The Gini index is also not able to capture demographic variations among subgroups within the distribution, such as the distribution of deaths across age. By using a dichotomized age variable (<60 years compared to $\geq$ 60 years old German population) we found out that the Gini indices for both groups over German districts were not equal and thus the heterogeneity can be probably partly explained by the age structure of the infected/deceased persons (data not shown). Further similar investigations remain left for future work.

Due to the properties of the Gini index it might not be straightforward to judge, if a decrease of heterogeneity in incidence/death rates over time should be attributed to success in fighting the disease or if it is related to a uniform increase of the disease activity across the geographical regions. Therefore, we propose to use disease maps/another tools for an overview of the disease activity together with the Gini index to summarize the data seen on the maps.

Methods

The Data

For the analyses we combined data from seven different sources. The global death and incidence data was retrieved from the COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University (JHU CSSE)[7]. The population data for each country was extracted from the European Centre for Disease Control and Prevention (ECDC) COVID-19 database[8], where a population estimate for the year 2019 is provided. The data regarding the continents was also extracted from the ECDC data. These data were used for the global analysis (Figure 1a, b) and the analyses per continent (Figure 3). Death and incidence data for mainland of China was taken from the JHU CSSE data and included data on Chinese 31 provinces. Data for Hong Kong and Macau were excluded. Population estimates for the provinces were obtained from the Chinese National Bureau of Statistics of China for the year 2019[9]. These data were used for Figure 1 (c, d). A subset of the JHU CSSE data was the basis for the analysis of the EU data, which is shown in Figure 2 (a,b). The list of the current member states of the European Union was retrieved from the official website of the EU[10]. The population data for each US state was obtained from the United States Census Bureau[11]. Puerto Rico was not included. A population estimate for the year 2019 was extracted. Combined with the data provided by the JHU CSSE Figure 2 (c, d) was created. The data for Germany was downloaded from the Robert Koch-Institute (RKI) COVID-19 dashboard[12]. We used the data provided on July 21th, 2020. For this analysis we summarized the daily incidences and deaths separately for each German federal state (16 in total) and for each German administrative district (401 in total). The population data for each German district was obtained from the German Federal Statistical Office Destatis[13]. The newest estimate for the year 2019 was used. The regional data for Bavaria and NRW were a subset of the nationwide German data and contains incidences and deaths for 96 districts on a daily basis.

For all analyses we excluded any cruise ship data. The data were analyzed using R Version 3.6.3[14].

The Gini index of inequality

The *Gini* index is widely used to characterize inequality in wealth and income[2]. The Gini index also quantifies inequality in the distribution.

The Gini index can be calculated as

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{2n^2\bar{x}},$$

where x is an observed value (such as number of deaths), n is the number of values observed and $\bar{x}$ is the mean value. If one sorts the values of x in ascending order, then G can be computed in a faster manner:

$$G = \frac{2}{n^2\bar{x}}\sum_{i=1}^{n} i(x_i - \bar{x}) = \frac{\sum_{i=1}^{n}(2i-n-1)x_i}{n\sum_{i=1}^{n}x_i}.$$

Here, each sorted x has rank i.

Since the observations used for the calculation of the Gini index are ordered, and thus not independent, is is difficult to find a reliable formula for the variance, which in turn is needed for the calculation of the confidence intervals. This is why the confidence intervals are calculated via bootstrap re-sampling methods[15].

We do apply the Gini index to incidences and death counts attributable to COVID-19. The Gini index ranges from zero to one, where the value of zero would mean that the incidences or deaths are evenly distributed among the affected countries. On the other hand, the value of one would mean that all burden of COVID-19 incidences or fatalities lies upon one single country. To make the counts more comparable across the countries, we did not use absolute numbers but incidence or death rates per 100000 persons. For this we simply divided the absolute counts by the population of the respective country and then multiplied this number by 100000. This way we are able to assess the relative heterogeneity/burden of disease between the countries. Otherwise, large heterogeneity between different countries would always be pronounced, since small countries could contribute only a small number of incidences and/or deaths, whereas large countries would seem heavier affected through larger absolute incidence or death numbers. For the longitudinal evaluation of inequality of COVID-19 incidences and deaths we calculated the Gini index every seven days, whereas the corresponding confidence intervals were obtained using bootstrapping. We plotted the Gini index of the fatalities one day apart from the Gini index of incidences for better visualization of the results. Of note, due to the small number of affected countries in Oceania (N=4: Australia, Fiji, New Zealand and Papua New Quinea) the Gini index was not calculated. The corresponding confidence intervals would have been very wide reaching from 0 to 1 for many data points resulting in no meaninful estimation of the GI. In addition to this, no cruise ship data was not used, since it could not clearly assigned to a continent.

Properties of Gini index

The Gini index is scale invariant. By multiplying all incidence/death rates by a constant factor, the value of the GI does not change. Intuitively, when all incidence rates scaled by a common factor, the cumulative distribution the rates does not change,

as a given fraction of population still holds the same fraction of incidences/deaths. The areas under the Lorenz Curve, therefore, do not change. Applying this idea to the ongoing pandemic leads to the following reasoning: Assume that the probability of dying is the same for each observational unit (such as district, (federal) state, or country). This means that of all reported incidences I a constant proportion p dies:

$$D = p \cdot I,$$

where D stays for the deaths. Due to the scale invariance introduced above the Gini indices for incidences and deaths over the observational units would be equal. This in turn implicates that if the both Gini indices are not equal, the risk of dying is not the same over the geographical areas. The constant factor is also called case fatality rate.

On the other hand, the Gini index is not translation invariant. If a constant amount is added (subtracted) to (from) each value the Gini index decreases (increases). In terms of the pandemic, if all districts/countries succeed in lowering the death rate by the same constant amount, the corresponding Gini index would increase.

The Lorenz Curve

The *Lorenz* curve is closely related to the Gini index and serves as a visualization of inequality within a distribution[16]. In our application, the proportion of affected countries is shown in the horizontal axis and the proportion of all COVID-19 deaths is shown in the vertical axis. A diagonal line presents the line of equality, along which x% of countries are responsible for x% of the fatalities (corresponding to Gini index of zero). The Gini index is equal to the ratio of the area between curve and the area below the diagonal. The Lorenz curve is calculated for one point in time.

Poisson regression for case fatality rates

In order to evaluate if the case-fatality rates between different regions are homogeneous we applied Poisson regression in the following way: The dependent variable was the absolute cumulative death count per observational unit (i.e. district, (federal) state or country) and the independent variable was the district itself. We used the absolute cumulative incidence counts per region as an offset. The p value from the likelihood ratio test between the model and an offset-adjusted intercept only model was used to decide if the case-fatality rates over the region was homogeneous or not. The case-fatality rates are shown in Supplementary Tables S1 -S4.

Data Availability

All datasets generated and/or analysed during the current study together with the R code are available in the OSF repository, `https://osf.io/c3ren/`.

References

1. Health 2020: the european policy for health and well-being. https://www.euro.who.int/en/health-topics/health-policy/health-2020-the-european-policy-for-health-and-well-being. Accessed: August 9th, 2020.
2. Gini, C. Sulla misura della concetrazione e della variability dei cartteri. *Atti del Reale Instituto Veneto di Scienze* **LXXII**, 1203–1248 (1914).
3. Lmu stablab. https://corona.stat.uni-muenchen.de/maps/. Accessed: August 6th, 2020.
4. Cirillo, P. & Taleb, N. N. Tail risk of contagious diseases. *Nat. Phys.* **16**, 606–613, DOI: https://doi.org/10.1038/s41567-020-0921-x (2020).
5. Manz, K. & Mansmann, U. Distributional challenges regarding data on death and incidences during the sars-cov-2 pandemic up to july 2020. *medRxiv* DOI: 10.1101/2020.07.24.20161257 (2020). https://www.medrxiv.org/content/early/2020/07/24/2020.07.24.20161257.full.pdf.
6. Williamson, E. J. *et al.* Factors associated with covid-19-related death using opensafely. *Nature* DOI: https://doi.org/10.1038/s41586-020-2521-4 (2020).
7. Covid-19 data repository. https://github.com/CSSEGISandData/COVID-19. Accessed: July 1st, 2020.
8. Covid-19 data repository. https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases. Accessed: June 30th, 2020.
9. National bureau of statistics of china (nbs). http://data.stats.gov.cn/english/easyquery.htm?cn=E0103. Accessed: July 29th, 2020.
10. Official website of the european union. https://europa.eu/european-union/about-eu/countries_en. Accessed: July 20th, 2020.

11. United states census bureau. https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html. Accessed: July 24th, 2020.

12. Robert koch institut covid-19 dashboard. https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bf1d4. Accessed: July 24th, 2020.

13. The german federal statistical office. https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/Administrativ/04-kreise.html. Accessed: July 22th, 2020.

14. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2020).

15. Efron, B. & Tibshirani, R. Improvements on cross-validation: The bootstrap method. *J. Am. Stat. Assoc.* **92**, 548–560 (1997).

16. Lorenz, M. O. Methods of measuring the concentration of wealth. *Publ. Am. Stat. Assoc.* **9**, 209–219 (1905).

Author contributions

U.M. conceptualized the idea. K.M. acquired and prepared the data. K.M. and U.M. analyzed and interpreted the data, drafted the manuscript and approved the submitted version.

Competing interests

The authors declare no competing interests.

Supplementary Information

The data sets

The raw data for the global data set was processed as follows: We extracted the total death count and the daily death counts per country between January 22 2020 and July 28, 2020. For the Netherlands, United Kingdom and France we excluded the data from the following oversea territories (Netherlands: Aruba, Curacao, Sint Maarten, Bonaire, Sint Eustatius and Saba; UK: Bermuda, Cayman Islands, Gibraltar, Montserrat, Anguilla, British Virgin Islands, Turks and Caicos Islands and Falkland Islands (Malvinas); France: French Guiana, French Polynesia, Guadeloupe, Mayotte, New Caledonia, Reunion, Saint Barthelemy, St Martin, Martinique, Saint Pierre and Miquelon).

Incidence and death rates in mainland China

Table S1 shows the cumulative number of incidences, deaths, incidence and death rates per 100000 inhabitants and the case-fatality rate for the 31 provinces of China as of July 28th, 2020.

Province	**Incidences**	**Deaths**	**Incidence rate per 100000**	**Death rate per 100000**	**Case-fatality rate (%)**
Hubei	68135	4512	114.957	7.6126	6.62
Hainan	171	6	1.810	0.0635	3.51
Beijing	932	9	4.327	0.0418	0.97
Heilongjiang	947	13	2.525	0.0347	1.37
Shanghai	744	7	3.064	0.0288	0.94
Henan	1276	22	1.324	0.0228	1.72
Chongqing	583	6	1.866	0.0192	1.03
Tianjin	204	3	1.306	0.0192	1.47
Xinjiang	400	3	1.585	0.0119	0.75
Anhui	991	6	1.557	0.0094	0.61
Hebei	349	6	0.460	0.0079	1.72
Shaanxi	323	3	0.833	0.0077	0.93
Gansu	167	2	0.631	0.0076	1.20
Jilin	157	2	0.583	0.0074	1.27
Shandong	799	7	0.793	0.0070	0.88
Guangdong	1674	8	1.453	0.0069	0.48
Hunan	1019	4	1.473	0.0058	0.39
Guizhou	147	2	0.406	0.0055	1.36
Liaoning	217	2	0.499	0.0046	0.92
Yunnan	190	2	0.391	0.0041	1.05
Guangxi	255	2	0.514	0.0040	0.78
Inner Mongolia	258	1	1.016	0.0039	0.39
Sichuan	604	3	0.721	0.0036	0.50
Fujian	366	1	0.921	0.0025	0.27
Jiangxi	932	1	1.997	0.0021	0.11
Zhejiang	1270	1	2.171	0.0017	0.08
Jiangsu	655	0	0.812	0.0000	0.00
Ningxia	75	0	1.079	0.0000	0.00
Qinghai	18	0	0.296	0.0000	0.00
Shanxi	201	0	0.539	0.0000	0.00
Tibet	1	0	0.028	0.0000	0.00

Table S1. Incidences, deaths, incidence and death rates per 100000 inhabitants and case-fatality rate in % in the 31 Chinese provinces as of 30th July 2020.

Incidence and death rates for EU

Table S2 shows the cumulative number of incidences, deaths, incidence and death rates per 100000 inhabitants and the case-fatality rate for the 27 current member states of the European Union as of July 28th, 2020.

Province	Incidences	Deaths	Incidence rate per 100000	Death rate per 100000	Case-fatality rate (%)
Belgium	66662	9833	581.92	85.8364	14.75
Spain	280610	28436	597.84	60.5833	10.13
Italy	246488	35123	408.37	58.1896	14.25
Sweden	79494	5702	777.05	55.7370	7.17
France	221077	30226	329.90	45.1048	13.67
Ireland	25929	1764	528.71	35.9689	6.80
Netherlands	53647	6164	310.42	35.6668	11.49
Luxembourg	6375	113	1038.45	18.4071	1.77
Portugal	50410	1722	490.53	16.7565	3.41
Romania	47053	2239	242.36	11.5326	4.76
Germany	207707	9131	250.19	10.9987	4.40
Denmark	13811	613	237.87	10.5579	4.44
Austria	20677	713	233.41	8.0485	3.45
Hungary	4456	596	45.60	6.0986	13.38
Finland	7404	329	134.18	5.9624	4.44
Slovenia	2101	117	100.97	5.6225	5.57
Estonia	2038	69	153.83	5.2083	3.39
Bulgaria	10871	355	155.30	5.0714	3.27
Poland	43904	1682	115.62	4.4295	3.83
Czechia	15799	374	148.35	3.5118	2.37
Croatia	4923	140	120.77	3.4345	2.84
Lithuania	2027	80	72.54	2.8631	3.95
Cyprus	1067	19	121.82	2.1692	1.78
Greece	4279	203	39.90	1.8928	4.74
Malta	708	9	143.45	1.8235	1.27
Latvia	1220	31	63.54	1.6146	2.54
Slovakia	2204	28	40.44	0.5137	1.27

Table S2. Incidences, deaths, incidence and death rates per 100000 inhabitants and case-fatality rate in % in the 27 countries of the European Union as of 28th July 2020.

Incidence and death rates for US

Table S3 shows the cumulative number of incidences, deaths, incidence and death rates per 100000 inhabitants and the case-fatality rate for the US states as of July 28th, 2020.

Province	Incidences	Deaths	Incidence rate per 100000	Death rate per 100000	Case-fatality rate (%)
New Jersey	180295	15825	2029.85	178.166	8.78
New York	412878	32653	2122.38	167.851	7.91
Massachusetts	116182	8551	1685.63	124.062	7.36
Connecticut	49077	4423	1376.52	124.057	9.01
Rhode Island	18725	1005	1767.57	94.869	5.37
District of Columbia	11945	583	1692.53	82.607	4.88
Louisiana	111038	3812	2388.53	82.000	3.43
Michigan	87958	6421	880.74	64.295	7.30
Illinois	174968	7638	1380.76	60.275	4.37
Delaware	14476	580	1486.60	59.563	4.01
Maryland	85524	3458	1414.63	57.198	4.04
Pennsylvania	114083	7154	891.13	55.882	6.27
Mississippi	54085	1540	1817.28	51.745	2.85
Arizona	165934	3408	2279.71	46.821	2.05
Indiana	63678	2924	945.87	43.433	4.59
Georgia	170843	3509	1609.08	33.049	2.05
Colorado	45296	1807	786.56	31.378	3.99
Alabama	82366	1491	1679.85	30.409	1.81
South Carolina	84109	1565	1633.59	30.396	1.86
New Hampshire	6500	409	478.04	30.080	6.29
New Mexico	19791	626	943.85	29.855	3.16
Ohio	86497	3382	739.98	28.933	3.91
Minnesota	52281	1620	927.03	28.725	3.10
Florida	441977	6117	2057.84	28.481	1.38
Iowa	42928	839	1360.60	26.592	1.95
Nevada	44936	759	1458.89	24.642	1.69
Virginia	86994	2095	1019.20	24.544	2.41
California	470762	8679	1191.43	21.965	1.84
Texas	408493	5913	1408.80	20.393	1.45
Washington	54205	1548	711.83	20.329	2.86
Missouri	45181	1232	736.16	20.074	2.73
North Carolina	116700	1860	1112.69	17.734	1.59
Nebraska	25157	321	1300.50	16.594	1.28
Kentucky	28111	719	629.21	16.093	2.56
Wisconsin	50179	906	861.82	15.561	1.81
Tennessee	99044	999	1450.31	14.628	1.01
Arkansas	40181	428	1331.46	14.182	1.07
South Dakota	8492	123	959.92	13.904	1.45
North Dakota	6141	100	805.84	13.122	1.63
Oklahoma	33775	509	853.56	12.863	1.51
Kansas	26058	340	894.45	11.671	1.30
Maine	3838	121	285.52	9.002	3.15
Vermont	1405	56	225.16	8.975	3.99
Idaho	19220	160	1075.51	8.953	0.83
Utah	38855	286	1211.96	8.921	0.74
Oregon	17408	303	412.73	7.184	1.74
West Virginia	6120	108	341.49	6.026	1.76
Montana	3474	51	325.04	4.772	1.47
Wyoming	2589	26	447.34	4.492	1.00
Alaska	2730	22	373.18	3.007	0.81
Hawaii	1757	26	124.09	1.836	1.48

Table S3. Incidences, deaths, incidence and death rates per 100000 inhabitants and case-fatality rate in the US states as of 28th July 2020.

Incidence and death rates in Germany

Table S4 shows the cumulative number of incidences, deaths, incidence and death rates per 100000 inhabitants and the case-fatality rate for the German federal states as of July 30th, 2020.

Federal State	**Incidences**	**Deaths**	**Incidence rate per 100000**	**Death rate per 100000**	**Case-fatality rate (%)**
Bavaria	50695	2621	387.674	20.0433	5.17
Saarland	2857	174	288.438	17.5667	6.09
Baden-Württemberg	37036	1843	334.576	16.6493	4.98
Hamburg	5338	261	289.923	14.1757	4.89
North Rhine - Westphalia	47962	1730	267.456	9.6472	3.61
Thuringia	3355	182	156.546	8.4922	5.42
Hesse	11805	518	188.403	8.2671	4.39
Lower Saxony	14305	649	179.206	8.1303	4.54
Bremen	1758	55	257.399	8.0529	3.13
Brandenburg	3531	168	140.570	6.6881	4.76
Berlin	9069	223	248.818	6.1183	2.46
Rhineland-Palatinate	7456	239	182.528	5.8509	3.21
Saxony	5525	225	135.485	5.5175	4.07
Schleswig-Holstein	3379	156	116.649	5.3854	4.62
Saxony-Anhalt	1996	64	90.385	2.8981	3.21
Mecklenburg-Western Pomerania	849	20	52.744	1.2425	2.36

Table S4. Incidences, deaths, incidence and death rates per 100000 inhabitants and case-fatality rate in the 16 German federal states as of 30th July 2020.