Supplementary appendix for

Early detection of seasonality and second waves prediction in the COVID-19 pandemic

Márcio Watanabe

Correspondence to souza_marcio@id.uff.br

Table of contents

1. Methods

This section is a summary of the methodology applied in this study. Further details can be found in the next sections. The data we use was obtained from the John Hopkins University website and it was verified from the data released by national disease control centers, when this information was available. The data were analyzed using the statistical software R 64 bits version 4.0.0

We gathered daily data for the 50 countries with the greatest epidemics, as measured by the number of confirmed cases from March 1 to May 1 (or other period when cited). The inclusion criteria were: 1) have at least 1000 confirmed cases in this period; 2) having data from daily confirmed cases by either notification or onset of symptoms; 3) have at least 90% of its territory in the same hemisphere (this criteria was necessary to make clear for which hemisphere the country would be assigned. In this criterion, Colombia, Ecuador, Indonesia and Singapore were removed from the pool).

Then, when appropriate, we separate countries into 2 different groups, according to whether their majority territory is in the Northern or Southern hemispheres. We obtained a few countries for the Southern hemisphere group so we decided to include other countries with more than 1000 cases in the period that were among the top 100 epidemics. Using these criteria, Argentina and New Zealand were included in the Southern hemisphere group. For each country, we calculate the rate of confirmed cases per 100k inhabitants. For each day, we calculate the rate per 100k inhabitants for each selected country and calculate its simple mean. We study the global dynamics of the pandemic by analyzing the time series curves for average daily cases and rates.

We can interpret the time curves of the average rates as showing the average dynamics of the epidemic in each hemisphere from a country perspective. That is, it displays the expected dynamics if we randomly draw a country. An advantage of this curve is that it gives us an insight into the behavior of the pandemic itself, without being too much affected by a single country.

We define the expected moment of seasonality reversion as the expected week (day or month) when the transmission rate of a seasonal disease changes due to seasonal reason. This variable varies from country to country and even within a country. It can also vary from one disease to another. As COVID-19 is transmitted as a respiratory disease, we estimate the expected seasonality moment of reversal for COVID-19 from H1N1 2009 pandemic and along with data from other respiratory syndromes.

To quantify the change in confirmed case rates in each hemisphere, we perform a simple linear regression before and after the expected moment of seasonality reversal. To decrease the influence of other factors and a possible confounding effect, we consider the slopes for a short period of ten days before and two weeks after the expected seasonality moment of reversal.

The mean seasonal effect for the Northern hemisphere (MSEN) is calculated as the difference between the slopes for that hemisphere before and after the expected seasonality moment of reversal. We test the hypotheses that seasonality did not affect COVID-19 transmission with the null hypotheses H0: MSEN=0. Likewise, we perform definition and testing for the Southern hemisphere.

We calculate the slope variation for each hemisphere at time point t as the difference between the slope of the curve in a short period of days (ten days or two weeks) after t minus the slope of the curve in a short period of days before t. The 10day (2-week) extension is chosen because a shorter time interval would be subject to a great variability due to a small number of time points in its estimation. On the other hand, a longer period of time would be subject to the influence of other factors, generating possible confounding factors.

To confirm that seasonality effect is due to seasonality and not to confounding factors, we run a multiple linear regression analysis. We take as response variable Y the seasonality effect (A-B) for each country. The explanatory variables are the seasonal factor X_{HP}, the social distancing factor X_{SD} and the income factor X_{IC}. We set X_{HP} as the indicator variable if the country belongs to the Northern hemisphere, X_{SD} as the discrete score varying from 0 to 2 according to the level of social distancing measures adopted (low/none, moderate, high/national lockdown) and X_{IC} as the country growth domestic product per capita (GDP).

To assess the variability of seasonal effect between different countries, we display the box-plots of seasonal effects (slope differences before and after the expected seasonality moment of reversal) for countries from both hemispheres separately.

2. Social distancing effect

In this section we display more details about the analysis of the effects of the social distancing measures.

In table 1 below, we display the top 50 countries with the largest number of confirmed cases from 2020-03-01 to 2020-05-01 plus Argentina and New Zealand. We also show the intervention dates when social distancing measures started in each country. For those countries that adopted different social distancing measures on different dates, we considered only the first one. Belarus, Japan, South Korea and Sweden have not adopted social distancing measures in this time period and thus were excluded from this specific analysis.

Table 1. Social distancing measures starting dates

Country	Intervention date
Argentina	2020-03-20
Australia	2020-03-23
Austria	2020-03-16
Bangladesh	2020-03-26
Belarus	-
Belgium	2020-03-18
Brazil	2020-03-24
Canada	2020-03-23
Chile	2020-03-19
Colombia	2020-03-25
Czech Republic	2020-03-16
Denmark	2020-03-13
Dominican Republic	2020-03-19
Ecuador	2020-03-16
Egypt	2020-03-25
Finland	2020-03-16
France	2020-03-17
Germany	2020-03-23
India	2020-03-25
Indonesia	2020-03-28
Iran	2020-03-14
Ireland	2020-03-12
Israel	2020-04-02
Italy	2020-03-09
Japan	-
Malaysia	2020-03-18
Mexico	2020-03-23
Netherlands	2020-03-15
New Zealand	2020-03-10
Norway	2020-03-12
Pakistan	2020-03-24
Panama	2020-03-25
Peru	2020-03-16
Philippines	2020-03-15
Poland	2020-03-13
Portugal	2020-03-19

Qatar	2020-03-11
Romania	2020-03-25
Russia	2020-03-28
Saudi Arabia	2020-03-09
Serbia	2020-03-15
Singapore	2020-04-07
South Africa	2020-03-26
South Korea	-
Spain	2020-03-14
Sweden	-
Switzerland	2020-03-16
Turkey	2020-04-03
Ukraine	2020-03-17
United Arab Emirates	2020-03-26
United Kingdom	2020-03-23
United States	2020-03-19
Northern hemisphere mean	2020-03-19
Southern hemisphere mean	2020-03-19
Global mean	2020-03-19

The Global mean included all countries from the table, except Belarus, Japan, South Korea and Sweden. The Southern hemisphere mean represents the average of intervention dates from Argentina, Australia, Bolivia, Brazil, Chile, Peru, South Africa and New Zealand. The Northern hemisphere mean was obtained from the Global group excluding countries from Southern hemisphere group, Colombia, Ecuador, Indonesia and Singapore which were not assigned neither to Northern nor to Southern hemisphere groups according to one of the inclusion criteria.

Both Northern and Southern hemisphere groups have the same truncated intervention mean date 2020-03-19 (Southern: mean = 19.7 standard deviation = 5.4 Northern: mean = 19.5 standard deviation = 6.3). The Global truncated intervention mean date, which also includes the four countries close to equator line, is 2020-03-19 (mean = 19.8 standard deviation = 6.6).

In the main text, we have shown the global mean daily rates. Next in figure S1 below we show the mean rates for Northern and Southern hemispheres groups from 2020-03-05 to 2020-04-15. It also shows the regression lines before and immediately after the expected effect of social distancing interventions. We estimate a rough 7 day delay from the mean intervention date 2020-03-19 and a possible effect in the pooled rate due to time to symptom onset plus testing results. Thus, the left regression lines (before intervention) use data from March 17 to March 26 and the right regression lines use data from March 27 to April 5.

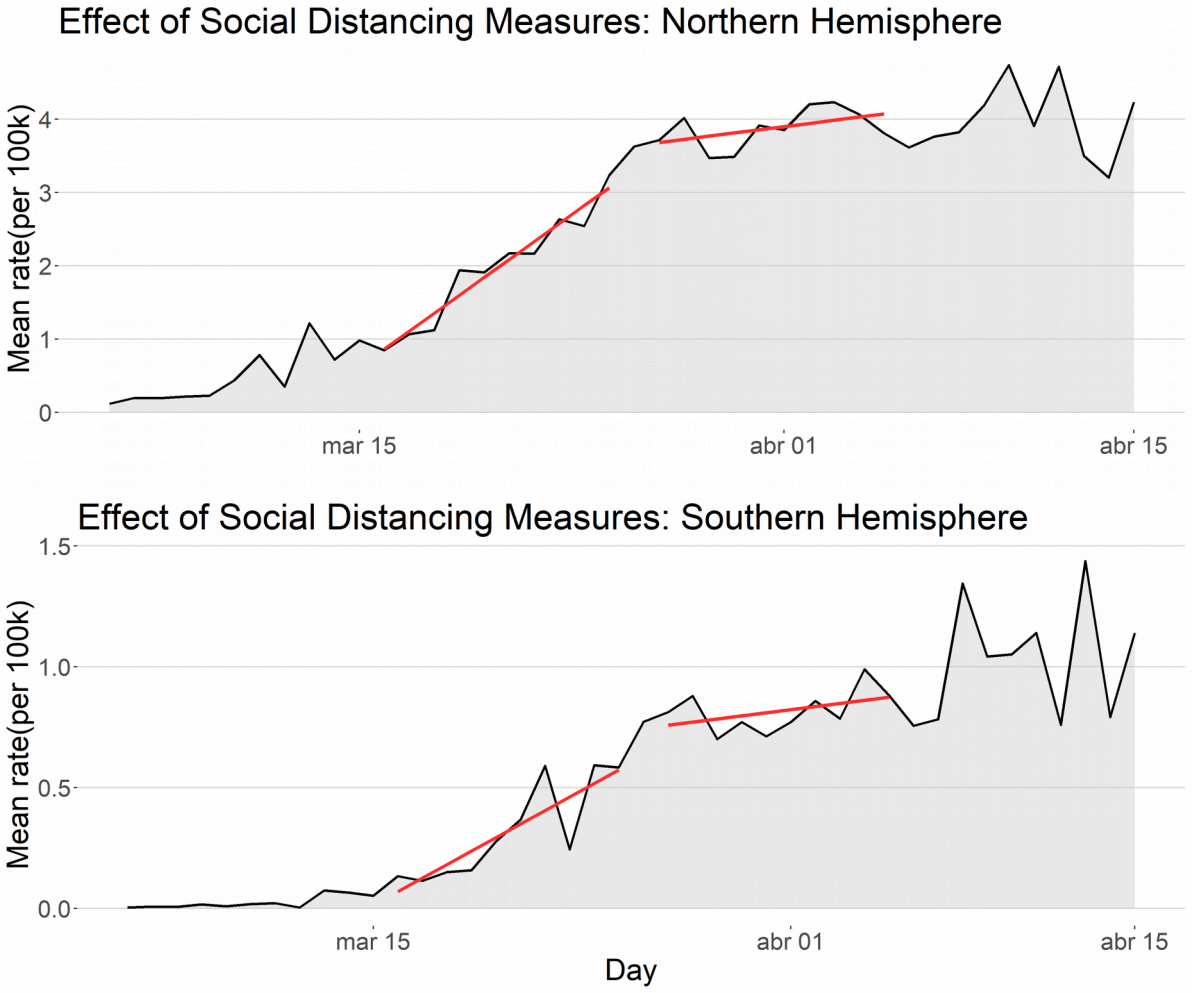

Figure S1 - Social distancing effects: The black curves show the mean rates of cases per 100k inhabitants in Northern and Southern Hemispheres receptively. The left red lines are the linear regression lines for a period of 10 days before the effects of social distancing measures appear. The right red lines are the linear regression line for a 10 days period starting one week after March 19, the average start date for social distancing measures.

We observe similar qualitative behavior in both hemispheres. Both hemispheres curves had a decrease in their growth speeds measured by the difference between slopes from right and left regression lines. To measure the quantitative effect, we obtain slope estimates for both regression lines before (B) and after (A) and 95% confidence intervals.

For the Northern hemisphere we have B=0.2610 (CI= [0.2048, 0.3172]), A=0.0431 (CI= [-0.0193, 0.1055]). MESD= -0.2178, which represents a relative reduction of 83.5%. Note that A and B are not independent. The closer the regression lines are from each other we can suppose more positively correlated A and B are. As the upper limit 0.1055 for A is less than the lower limit 0.2048 for B, we reject the hypothesis that A = B at 95% confidence level. Thus, there is sufficient statistical evidence that the social distancing measures have decreased,

at least for the short term, the Northern average growth rate of COVID-19's cases with an estimated relative reduction of 83.5% in the speed of growth.

For the Southern hemisphere we have B=0.0681 (CI= [0.0393, 0.0969]), A=0.0130 (CI= [-0.0081, 0.0341]). MESD= -0.0552, which represents a relative reduction of 81.0%. As the upper limit 0.0341 for A is less than the lower limit 0.0393 for B, we reject the hypothesis that A = B at 95% confidence level. Thus, as before there is sufficient statistical evidence that the social distancing measures have decreased, at least for the short term, the Southern average growth rate of COVID-19's cases with an estimated relative reduction of 81.0% in the speed of growth. Note that the relative reductions in both hemispheres are very similar.

3. Total number of confirmed cases

In the main text data analysis was based on rates (number of cases per 100k inhabitants). Rates provide a perspective for the distribution of the epidemics from a country point of view, in the sense that each country has the same weight in mean calculation. Here we repeat the analysis of the seasonality effect but for the total number of cases instead for the mean of rates. This provides a population perspective to the analysis, in the sense that each person has the same weight in calculations. It can be interpreted as if we have only two countries in the world, the Northern and the Southern hemispheres and we were comparing their epidemics. One disadvantage of the next analysis is that outliers have a stronger influence here than in the mean rate analysis.

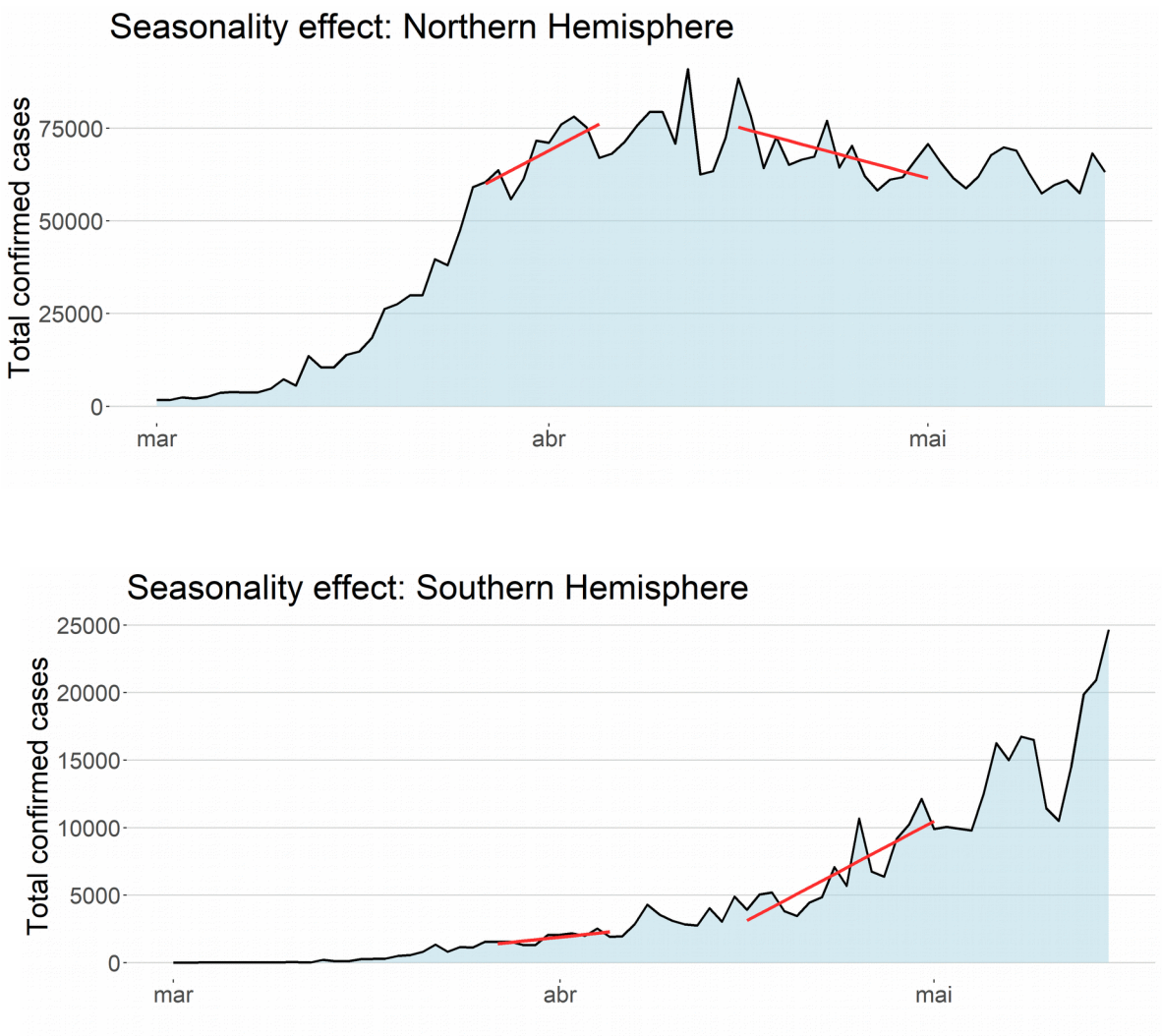

Figure S2 - Seasonality effect: black curves show the total number of confirmed cases in the Northern and Southern hemispheres, respectively. Left red lines are the linear regression lines immediate before the expected seasonal reversal moment and right red lines are linear regression lines immediate after the expected seasonal reversal moment for both hemispheres.

Figure S2 above gives a clear picture of how the total number of cases moved into different directions in Northern and Southern hemispheres just after the estimated moment of seasonal reversal. It has the same qualitative behavior of the mean rates for both hemispheres.

To quantify this difference, we obtain the estimates of the slopes of the Northern hemisphere and 95% confidence intervals given by B=1787.7 (CI=[374.5, 3201.0]), A= -914.1 (CI=[-1677.6, -150.6]). MSEN= -2701.8, which represents a relative reduction of 151.1%. We interpret a relative reduction greater than 100% as a reduction which changes a positive slope to a negative one. Remind that A and B are not independent. The closer the regression lines are from each other we can suppose more positively correlated A and B are. As the upper limit -150.6 for A is less than the lower limit 374.5 for B, we reject the hypothesis that A = B at 95% confidence level.

For the Southern hemisphere, slope estimates and 95% confidence intervals are given by B=97.3 (CI=[24.3, 170.2), A= 490.5 (CI=[310.6, 670.4]). MSES= 393.2, which represents a relative increase of 404.1%. As the upper limit 170.2 for B is less than the lower limit 310.6 for A we reject the hypothesis that A = B with 95% confidence level.

These results confirm that in the second half of April a statistical significant change in the speed of variation of cases occurred in both hemispheres, with an increase in the growth speed of the number of cases in the Southern hemisphere and a decrease in the growth speed of the number of cases in the Northern hemisphere. This reinforces the evidence of a consistent seasonal influence in the transmission of COVID-19.

4. Confirmatory tendency of seasonal effect

4.1 Seasonal mean rate time series

Seasonality period can vary from one country to another: in the first semester the seasonal period may start from March to July, depending on the region. Thus, to measure seasonality in subgroups of countries we need a method more robust than the Seasonal effect, which is based on slopes for daily rates of confirmed cases. In addition, when $R0_{min} > 1$, the difference caused by seasonality in the slopes from both hemispheres may not be as appealing as it is when $R0_{min} < 1$, since the rates would keep growing in both hemispheres. Here, we introduce the Seasonal rate which we define to be the daily difference between the mean rate of cases from the Northern hemisphere and the mean rate of cases from the Southern hemisphere.

We display the Seasonal rate = mean rate of the Northern hemisphere – mean rate of the Southern hemisphere in figure S3 below. There are some advantages in measuring seasonal effect through this time series. The seasonal effects from both hemispheres are summed up, which has as a number of interesting consequences. The first is that we have the entire data from both hemispheres integrated in a single curve. The second, is that the seasonal effect is more robust to changes in the time interval used in the regression line after the expected left limit from the seasonal period measured (the seasonal reversal moment). The third is that the form of this curve does not depend whether $R0_{min}$ is greater or lower than one.

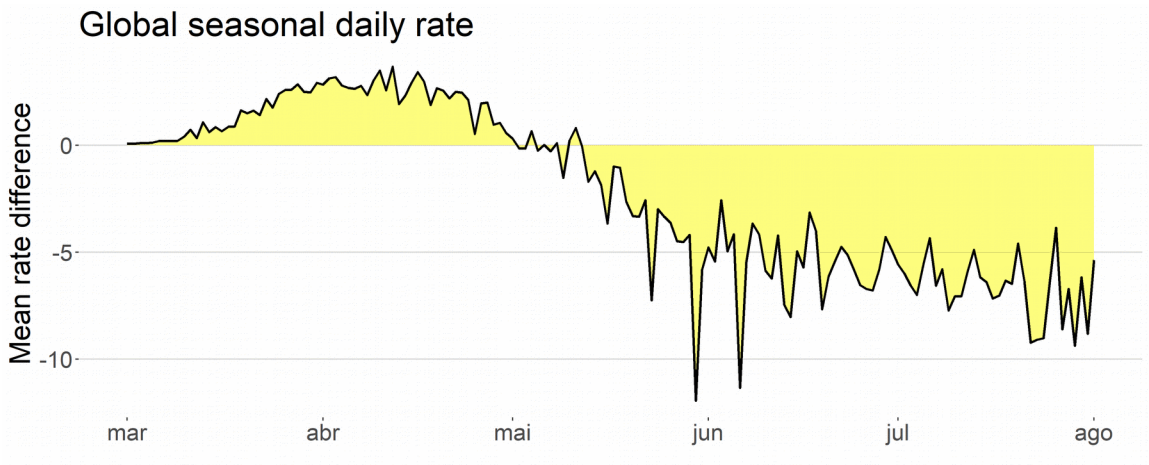

Figure S3 – Global seasonal rate: Larger epidemics alternates between Northern (y>0) and Southern (y<0) hemispheres. The reversal seasonal moments coincide with the moment when the slope of the seasonal rate curve equals zero. In the COVID-19 pandemic data above, this happened in the mid-April. Positive slopes (increasing seasonal rate) are expected in the high season for the Northern hemisphere and negative slopes (decreasing seasonal rate) are expected in the high season for the Southern hemisphere.

When the seasonal rate is positive it means that the mean rate in the Northern hemisphere is greater than the mean rate of the Southern hemisphere. On the other hand, when the seasonal rate is negative we have that the mean rate of cases in the Southern is the greatest. Thus, from figure S3 we see that until May the Northern

hemisphere was the epicenter of COVID-19 cases and in the beginning of May the Southern hemisphere became the epicenter of the pandemic. Until the end of March, the difference increased in favor of the Northern hemisphere. In April, there is a turnaround of events and the difference starts to decrease in favor of the Southern cases. This coincides with the beginning of the high season of respiratory seasonal infections in the Southern hemisphere. From April until August this tendency of decreasing of the seasonal rate remains consistent. A new reversal of seasonal effect is expected for the beginning of October.

4.2 Subgroup analyses: low income countries

We repeat the previous analysis but include only low income countries in the top 50 epidemics. The inclusion criteria was to have a gross domestic product (GDP) per capita lower than 20.000 dollars. For this analysis, Bolivia is included in the low income Southern group.

The motivation for this analysis is to verify the hypothesis whether the difference in the behavior of the rate of cases between both hemispheres could be explained by economic aspects since except for Australia and New Zealand all Southern hemisphere countries are low income.

We list the countries included below with its respective GDP per capita in dollars in 2018 (source: World bank).

Southern hemisphere: Argentina (11683), Bolivia (3548), Brazil (8920), Chile (15923), Peru (6941) and South Africa (6374). Mean = 8898 and standard deviation = 4380.

Northern hemisphere: Bangladesh (1698), Belarus (6289), Dominican Republic (8050), Egypt (2549), India (2009), Iran (5627), Malaysia (11373), Mexico (9673), Panama (15575), Pakistan (1482), Philippines (3102), Poland (15420), Romania (12301), Russia (11288), Serbia (7246), Turkey (9370) and Ukraine (3095). Mean = 7420 and standard deviation = 4724.

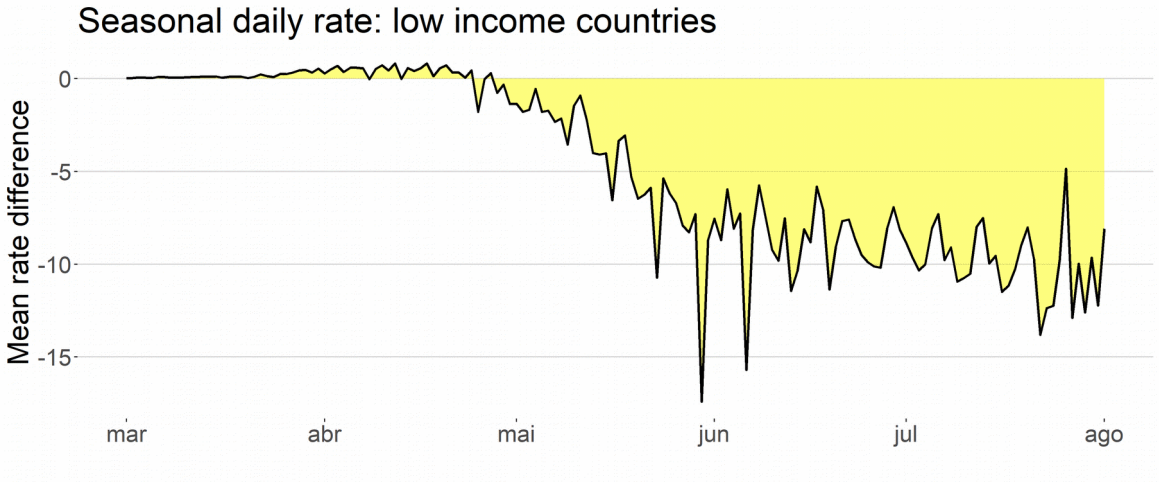

Figure S4 – Seasonal rate for low income countries: Larger epidemics alternates between Northern (y>0) and Southern (y<0) hemisphere groups. The reversal seasonal moments coincide with the moment when the

slope of the seasonal rate curve equals zero. In the COVID-19 pandemic data above, this happened in the mid-April. The difference in favor of the group is more pronounced than in the general case of figure S3.

We observe a qualitative behavior very similar with the previous comparison between all countries from both hemisphere groups. The Northern hemisphere groups present a higher mean rate of cases until April when we observe a fast change. In May there is an inversion, with the average rate of cases in the Southern hemisphere group presenting more cases. This tendency maintains during the entire seasonal period which is going to finish just in September.

Quantitatively, the seasonal differences between low income countries from both hemispheres are even more pronounced than when included mid and high income countries.

Therefore, this gives data based evidence that the seasonal effect is even more pronounced in low income countries. One possible explanation is that social distancing effect was higher in high income countries contributing to a decrease in the rate of cases in high income countries independently of the geographical location. Consequently, we expect a decrease in the seasonal rate, which measures the differences between rates from Northern and Southern hemispheres since both rates would be smaller, softening the natural dynamics of the pandemic.

4.3 Multiple regression analyses

To confirm that the difference between the growth speeds from the end of April and the end of March is due to seasonality and not to confounding effects, we run simple and multiple linear regression analysis. We take as response variable Y the seasonality effect (A-B) for each country. The explanatory variables are the seasonal factor X_{HP}, the social distancing factor X_{SD} and the income factor X_{IC}. We set X_{HP} as the indicator variable if the country belongs to the Northern hemisphere, X_{SD} as the discrete score varying from 0 to 2 according to the level of social distancing measures adopted (low/none, moderate, high/national lockdown) and X_{IC} as the country growth domestic product per capita (GDP).

The simple linear regression lines for each explanatory variables are given by $Y = 0.096 - 0.202X_{HP}$, $Y = -0.060 - 0.019X_{SD}$, $Y = 0.004 - 0.094X_{IC}$ and the only statistically significant model at 95% confidence level was the first one for the seasonal factor, with p-value = 0.04. The interpretation for the seasonal factor model is that the mean effect Y = 0.096 for the Southern hemisphere, which means we have a mean increase of 0.096 for the growth rate of the countries from the Southern hemisphere. In contrast, the mean effect for countries from the Northern hemisphere was Y = 0.096 -0.202, that is a reduction of 0.106 in the growth rate.

For the multiple linear analysis, we make a simple re-scaling of the factors X_{SD} and $X_{IC.}$. To do this we just multiply each variable by the average value of X_{HP} and divide them by their respective means. After re-scaling variables to the X_{HP} scale we are able to compare their coefficient estimates and also compare them to the estimates obtained in the simple linear regression for X_{HP}. Thus, we obtain for the additive model the estimate $Y = 0.202 - 0.184X_{HP} - 0.057X_{SD} - 0.085X_{IC}$. All three factors contribute to a reduction on Y, but the seasonal factor X_{HP} has the largest absolute coefficient 0.184 which is similar to the 0.202 obtained in the simple regression.

To assess dependence between factors, we run a multiple linear regression model with interaction terms obtaining for re-scaled variables the estimate $Y = 0.108 - 0.162X_{HP} + 0.034X_{SD} + 0.009X_{IC} - 0.011X_{HP}X_{SD} - 0.014X_{HP}X_{IC} - 0.093X_{SD}X_{IC}$. We observe that the individual coefficient of the seasonal factor X_{HP} remains the largest and its estimate is very close to the estimate obtained in the additive model. The interaction coefficients between seasonal factors and the other two factors are considerably small which shows that the seasonal factor affected all countries independently of the levels of social distancing and income factors. The individual coefficients for social distancing and income factors changed from negative to positive. The reduction measured previously in the additive model is contained in the interaction between social distancing factor and income factor. This implies that the social distancing effectiveness is highly correlated with income and that its impact was bigger in high income countries than in low income countries.

We finish this section showing a sensitivity analysis for the time interval chosen for the slopes calculations. We calculate mean seasonal effects and the coefficients b for simple linear regression lines $Y = a + bX_{HP}$ to assess the impact of the time interval change. The originally chosen pair of intervals were [2020-03-27, 2020-04-05] and [2020-04-16, 2020-05-01]. The mean effect was -0.077 and the regression coefficient was -0.202. Alternative pairs of time intervals are ([2020-03-24, 2020-04-02], [2020-04-13, 2020-04-28]) with mean -0.118 and coefficient -0.186 and ([2020-03-30, 2020-04-08], [2020-04-19, 2020-05-01]) with mean -0.026 and coefficient -0.159. All cases presented negative means and coefficients, with the original pair having the greatest absolute value for the regression coefficient.

5. Modeling seasonality effect

The seasonal (or time-inhomogeneous) SEIR model is a deterministic model typically used to describe the dynamics of an epidemic where the transmission rate changes periodically. Suppose we have a population of a fixed size N. The number of people in this population susceptible to the disease at time t is denoted by S(t). The number of people exposed to the infection at time t, but not capable yet of transmitting the disease is denoted by E(t). The number of people which is infective at time t is denoted by I(t) and the number of people which is immune to the infection is denoted by R(t). For convenience, we omit time in notation and write just S, E, I and R, whenever this is possible. Observe that S + E + I + R = N for all t. Let β(t) denote the rate of transmission at time t, α denote the mean time an infected person is non-infective and γ denote the mean recovery time. The dynamics of the epidemics are given by the following differential equations:

$$\frac{dS}{dt} = -\beta(t) I \frac{S}{N}$$
$$\frac{dE}{dt} = \beta(t) I \frac{S}{N} - \alpha E$$
$$\frac{dI}{dt} = \alpha E - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

The classical SEIR model is obtained when β(t) is a constant. The above model makes the simplifying assumption that the transmission is homogeneously mixed in the population, that is, each infective person transmits the infection at the same rate β(t)S/N. In addition, observe that once an individual is immune to the disease, he remains immune for all t.

The transmission function β(t) can be chosen according to the case of interest. Here, we choose the simplest possible seasonal function, the so-called term-time function

$$\beta(t) = \beta_{max}, \text{if } t \in \text{high seasonal period}$$
$$\beta_{min}, \text{if } t \in \text{low seasonal period ¿}$$

where $\beta_{max} > \beta_{min} > 0$. This function simplifies the analysis of the model, which makes possible understand the dynamics in a clear way. Besides its simplicity, it reproduces qualitatively the natural dynamics from seasonal diseases in a short period of few years. For longer periods, one can use a variation with birth and death rates. Our intention here is to make a qualitative comparison of this model to the classical SEIR in order to understand the effect of seasonality in the qualitative behavior of the dynamics of seasonal diseases. Alternative models, such as the TSIR model, can also be used to model seasonal effect in epidemiological dynamics [5].

We generate a simulated dynamics for the time period from 2020-02-01 to 2021-07-01 for two similar cities, one in the Northern hemisphere and the other in the Southern hemisphere. Population sizes are taken to be N = 30000 and initial conditions E=10, I=10 and R=0. We estimate α = 4.6 and the mean recovery time γ = 6.75

(*3,4)*. The high seasonal period is assumed to be from April to September in the Southern hemisphere and from October to March in the Northern hemisphere. The low seasonal periods are the complementary periods of the year.

For the seasonal SEIR with term-time transmission function, there are three possible scenarios with different qualitative behavior according to the pair of reproduction numbers ($R0_{min}$, $R0_{max}$). The first case is when $R0_{max} > 1$ and $R0_{min} < 1$. In this case, the epidemics obey the seasonal pattern clearly, with larger outbreaks in the high seasons and small number of cases in the low seasons. The second case is when $R0_{min} > 1$. Here, one can expect outbreaks starting at any moment of the year, but in the high seasonal period when $\beta = \beta_{max}$, the size and intensity of the outbreaks are larger than in the low seasonal period when $\beta = \beta_{min}$. The last case is when $R0_{max}$ is less than or equal to 1. In this case, there is no tendency to occur large epidemics at any time, thus we restrict our analysis to the first two cases and do not show any results on this last situation. For the simulations below, we assume $R0_{max} = 1.5$, $R0_{min} = 0.6$ in the first case and $R0_{max} = 1.5$, $R0_{min} = 1.1$ in the second case.

Figures S5 and S6 below show the dynamics for Northern and Southern hemispheres differs considerably from each other on the seasonal SEIR, independently from $R0_{min}$ be bigger or smaller than 1. On the contrary, Figure S7 shows that similar cities will have similar dynamics in both hemispheres in the classical (non-seasonal) SEIR. When $R0_{min} < 1$, epidemic waves alternate from one hemisphere to the other. In this case, the Northern hemisphere has a small, interrupted first epidemic wave, and a bigger second wave is expected in the next high season. Meanwhile, many cities in the Southern hemisphere will present a very big first wave and a very small second wave. This resembles the COVID-19 situation of many cities, particularly in Europe.

In case $R0_{min} > 1$, figure S6 the situation from the Southern hemisphere does not change qualitatively. On the contrary, the Northern hemisphere city would have a different dynamic. A continuous flatted first wave is observed during the low season period until the moment when the high season comes and the number of cases grows faster but the herd immunity is achieved in a softer manner in comparison to the Southern curve. This dynamic resembles the COVID-19 outbreak from India, for example.

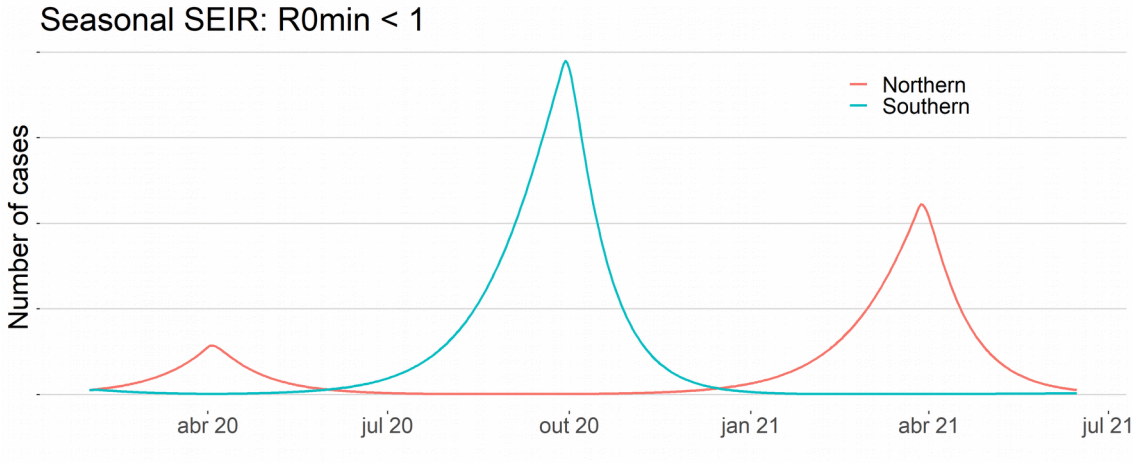

Figure S5 – Seasonal SEIR: $R0_{min} < 1$. Larger epidemics occur in alternated periodic pattern in the Northern and the Southern hemispheres. Since the starting date was in the end of a north seasonal period, we can expect that the Northern first wave to be smaller than its second wave (left and right red curves).

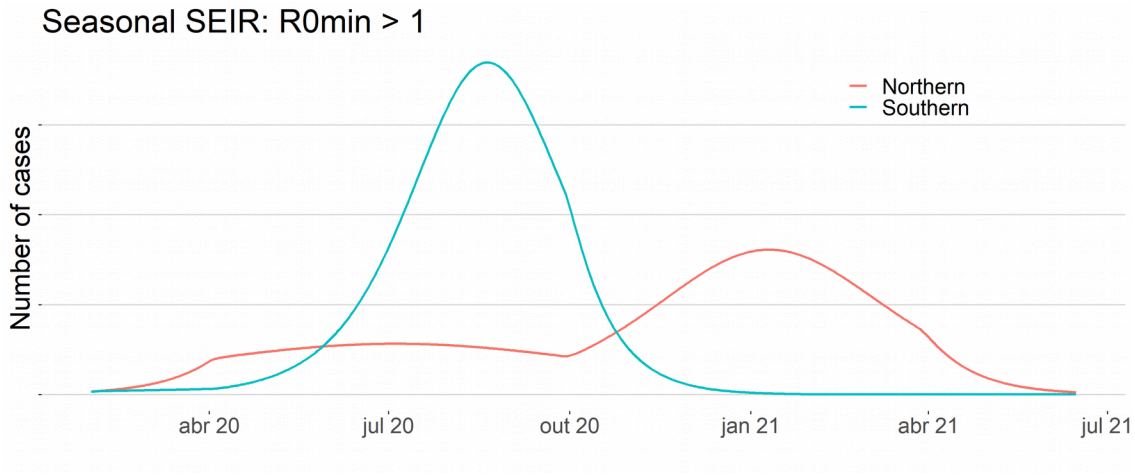

Figure S6 – Seasonal SEIR: $R0_{min} > 1$. Larger epidemics occur in alternated seasonal periods in Northern and Southern hemispheres. Here we observe a long epidemic on the Northern city, with varying growth speed: Slower increase in the low season from May to September and a faster increase after October.

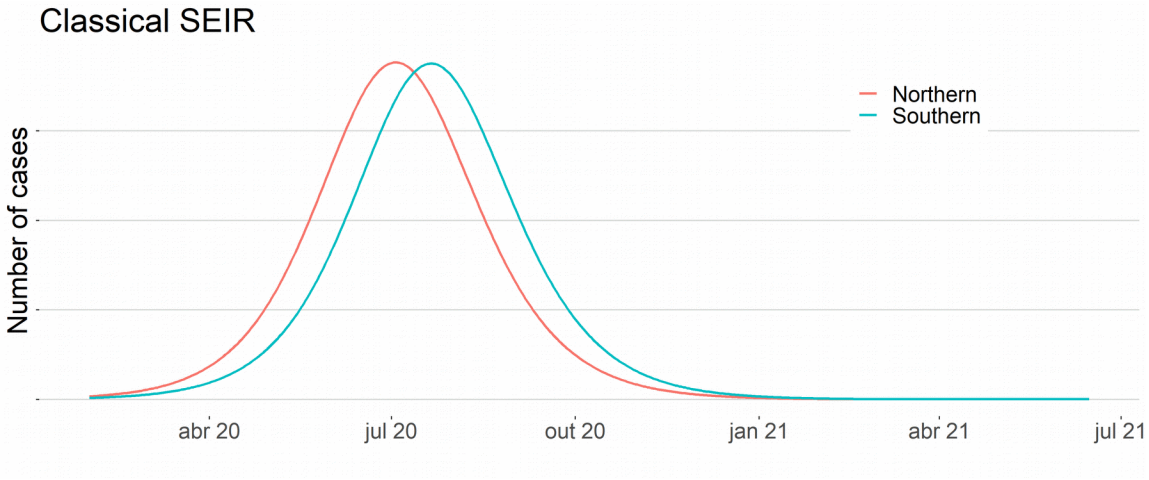

Figure S7 – Classical SEIR: Without seasonality, a similar behavior is expected in similar cities from the Northern and Southern hemispheres. Large single waves occur in similar periods in both hemispheres and herd immunity controls the moment when the speed of growth changes from positive to negative.

Figure S8 below displays the seasonal rate for simulated seasonal SEIR, when $R0_{min} < 1$. We observe alternated waves from both hemispheres, with the slopes of the seasonal rate changing signs close to the reversal seasonal moments, when a high seasonal period finishes in one hemisphere and stars in the other (vertical lines). We expect a similar dynamics for COVID-19 seasonal rate.

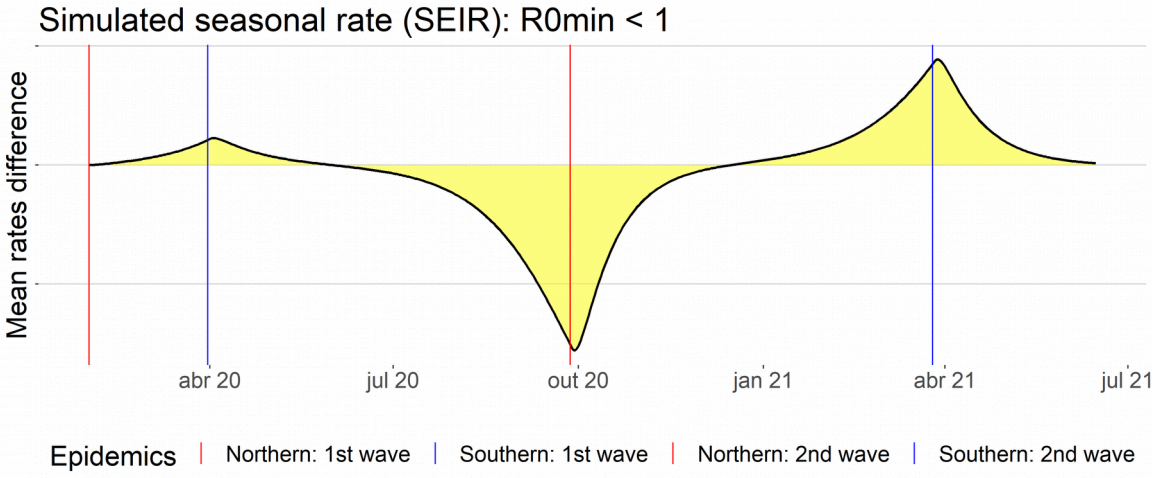

Figure S8 – Simulated seasonal rate: Larger epidemics alternates between Northern (y>0) and Southern (y<0) hemispheres. The reversal seasonal moments (vertical lines) coincide with the moments when the slope of the seasonal rate curve equals zero, except from a one week delay. Northern waves start at red vertical lines (positive slopes) and Southern waves start at blue vertical lines (negative slopes). The Northern second wave is bigger than its first wave, which was interrupted by seasonal forces. Southern first wave is much larger than its second wave, where herd immunity is achieved in a short period.

6. Data limitations

Reported cases have many limitations as a database. There is a significant amount of unreported cases and a large variability of the quality and reliability of the data from country to country. As we used polled data methods to analyze the data, we have minimized the variability problem. In addition, we focused on data from the larger reported epidemics. There are many countries with weak developed epidemiological surveillance with very small reported epidemics. By including only the largest epidemics we avoid the inclusion of this less reliable data.

However, there were still some particular limitations which deserved some attention. In this section we examine two of these limitations.

6.1 Reported cases from China

The analyses of this article rely on a time series perspective. We used data from daily reported confirmed cases, thus it is necessary that each country follows a daily report rule. We have chosen to restrict our pool to the top 50 countries with the largest number of confirmed cases from 2020-03-01 to 2020-05-01. Another possibility, maybe more natural, is to restrict to a period starting from the beginning of the pandemic until May 1. The difference between the two groups obtained is just the inclusion of China in the last ranking in the place of Finland.

However, China confirmed cases did not follow a daily basis report. A significant part of the cases was reported in a single day, as we can see in figure S5 below.. For this reason, we have decided to exclude it from the pool and include Finland instead.

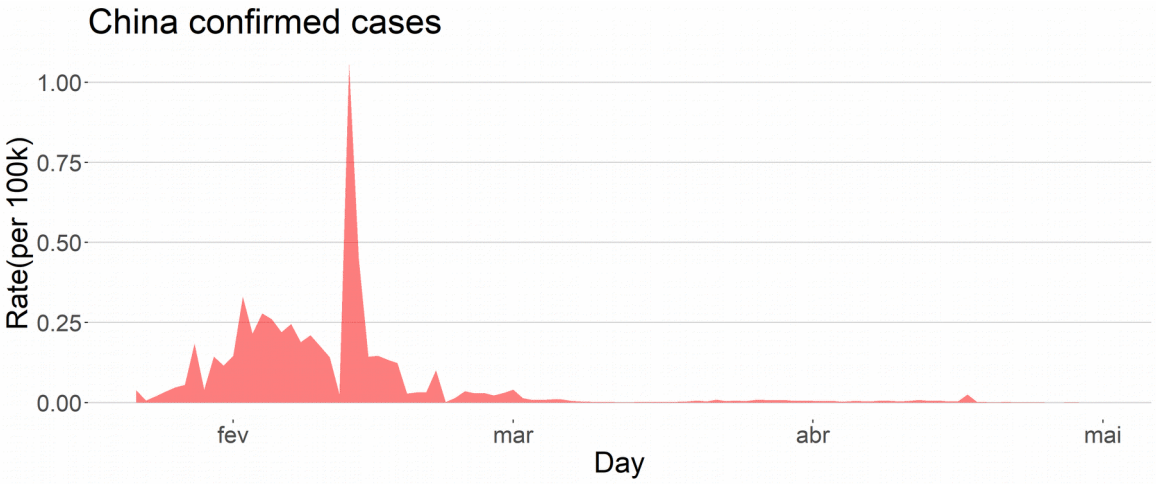

Figure S9 – Daily rate of confirmed cases of COVID-19 from China: A single day concentrates many of the reported cases.

6.2 Case definitions

The main information released concerning the dynamics of the COVID-19 pandemic in each country is the number of confirmed cases. This fundamental epidemiological surveillance data is used by governments, scientists and society to make projections and measure the burden of the disease.

In most countries, the definition of a confirmed case follows the World Health Organization-WHO guidance. The official definition states that a person becomes a confirmed case only in the presence of a laboratory confirmation (*1*). A recent study shows different definitions can significantly impact important epidemiological variables (*2*).

The crucial issue here is that laboratory tests are intended to detect the specific virus with high precision, such as the molecular RT-PCR test. On the contrary, serology tests are imprecise and are not included in this definition, as recommended by WHO itself (*3*). A significant number of serology tests produce false positive results, typically in the range 0,7% to 7,5% (*4*).

In addition, there is also a high rate of false negatives. In order to measure the possible impact this error could cause, note the Brazilian Ministry of Health has requested more than 20 million serology test kits, which implies that the number of confirmed cases could increase by up to 1.5 million due to false positives (*5*). Until June 23, Brazil officially had 1.145.906 confirmed cases, which represents about 0,5% of the total population.

Cases with clinical diagnosis or with only a positive serology test are considered only probable cases by WHO.[1] The European Union, the United States, Canada, Argentina, Chile, South Korea, Japan are some of the many countries which follow these definitions (*6-12*).

Brazil has its own definition for a confirmed case, which includes, in addition to laboratory confirmed cases, clinical suspicious cases that had contact with a confirmed case and cases that tested positive in serology tests (*13*). According to WHO, these last two should be considered probable cases. Until May 9, the proportion of confirmed cases that had a test confirmation by molecular RT-PCR was 67.6%, while the proportion of cases confirmed by serology tests was 32.4%. In addition, 1.4% of all confirmed cases were clinical cases (*14*). On May 29, the proportion of cases confirmed by RT-PCR reduced to only 56% (*15*). Thus, up to this date, almost one in two tested confirmed cases in Brazil should not be considered a confirmed case (*3,16*).

To access the impact of serology tests on reported cases, we fix the rate of clinically confirmed cases at 1.4%. Since may 9, we estimate that 47.3% of confirmed cases were confirmed through serology tests. Thus, 48.7% of the new cases should be classified as probable cases. We estimate that 51.3% of the cases reported after 9 may in Brazil are laboratory confirmed cases. This estimate agrees with the proportion of 54.3% of molecular tests in Brazil, which has requested 22 million serology tests and 24.6 million molecular RT-PCR tests (*5*). The number of officially confirmed cases on May 9 was 155.939 and on May 29 was 465.166. Therefore, the corrected number of confirmed cases in Brazil can be estimated by the formula:

$$Estimated\,confirmed\,cases = 103.939 + 0.513 \times (officialy\,confirmed\,cases - 155.939)$$

On June 23 the number of confirmed cases was 1.145.906. Applying the formula above we obtain the estimate 611.792 for the corrected number of confirmed cases and 534.114 for the number of probable cases.

In Brazil, molecular RT-PCR tests are used mainly in hospitalized patients and deaths. Thus, we consider the number of officially confirmed deaths as an approximation to the estimated number of confirmed deaths. We estimate the corrected lethality rate by the formula:

$$Corrected\,lethality\,rate = \frac{(officially\,confirmed\,deaths)}{(Estimated\,confirmed\,cases)}$$

On June 23 the number of confirmed deaths was 52.645. Applying the formula above we obtain the corrected lethality rate of 8.6% in contrast to the official rate of 4.6%, which implies a relative increase of 87.5%. Hence, it is necessary caution to compare Brazilian reported cases with reported cases from other countries.

In addition to Brazil, the United States also deviates from the standards in disclosing its results. Despite following WHO definitions, the United States Centers of Disease Control and Prevention (CDC) discloses the number of total cases, which includes confirmed cases plus probable cases (*1,7,8*). However, the percentage of laboratory confirmed cases in the United States correspond to around 96,6% of the total cases reported (*17*). Therefore, the impact of probable cases in the total reported cases in the United States' data is considerably smaller when compared to Brazil's data.

Corresponding author: Marcio Watanabe
email: souza_marcio@id.uff.br

Instituto de Matemática e Estatística,
Universidade Federal Fluminense, Brazil.

References

[1] World Bank website.
https://data.worldbank.org/
Date: 2020.
Date accessed: August 21, 2020.

[2] Our world in data website.
COVID-19: Government Response Stringency Index.
https://ourworldindata.org/
Date: 2020.
Date accessed: August 20, 2020

[3]Bi Q, Wu Y, Mei S, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study [published correction appears in Lancet Infect Dis. 2020 Jul;20(7):e148]. *Lancet Infect Dis*. 2020;20(8):911-919. doi:10.1016/S1473-3099(20)30287-5

[4] PGT Walker, C Whittaker, OJ Watson, et al. The impact of COVID-19 and strategies for Mitigation and suppression in low- and middle-income countries. Science (2020) published online June 12. https://doi.org/10.1126/science.abc0035

[5] World Health Organization.
Global Surveillance for human infection with coronavirus disease.
https://www.who.int/publications-detail/global-surveillance-for-human-infection-with-novel-coronavirus-(2019-ncov)
Date: 2020.
Date accessed: May 25, 2020.

[6] Tsang TK, Wu P, Lin Y, Lau EHY, Leung GM, Cowling BJ.
Effect of changing case definitions for Covid-19 on the epidemic curve and transmission parameters in mainland China: a modelling study. The Lancet Public Health 2020; 5: e289–96.

[7] World Health Organization.
Advice on the use of point-of-care immunodiagnostic tests for Covid-19.

https://www.who.int/news-room/commentaries/detail/advice-on-the-use-of-point-of-care-immunodiagnostic-tests-for-covid-19
Date: 2020.
Date accessed: May 25, 2020.

[8] Coronavirus kit evaluation program (in Portuguese)
https://testecovid19.org/avaliacoes/
Date: 2020.

Date accessed: May 25, 2020.

[9] Brazilian Ministry of Health.
https://www.saude.gov.br/noticias/agencia-saude/46760-ministerio-da-saude-amplia-para-46-2-milhoes-aquisicao-de-testes
Date: 2020.
Date accessed: May 25, 2020.

[10] European Centre for Disease Prevention and Control.
Case definition for coronavirus disease.
https://www.ecdc.europa.eu/en/case-definition-and-european-surveillance-human-infection-novel-coronavirus-2019-ncov
Date: 2020.
Date accessed: May 25, 2020.

[11] United States Centers for Disease Control and Prevention.
Coronavirus disease 2019.
https://wwwn.cdc.gov/nndss/conditions/coronavirus-disease-2019-covid-19/case-definition/2020/
Date: 2020.
Date accessed: May 25, 2020.

[12] United States Centers for Disease Control and Prevention.
Coronavirus disease 2019: cases in the U.S.

https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html
Date: 2020.
Date accessed: May 25, 2020.

[13] Canada government.
Interim national case definition.
https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/health-professionals/national-case-definition.html
Date: 2020.
Date accessed: May 25, 2020.

[14] South Korea government.
Case definition.
http://ncov.mohw.go.kr/en/baroView.do?brdId=11&brdGubun=112&dataGubun=&ncvContSeq=&contSeq=&board_id=&gubun=
Date: 2020.
Date accessed: May 25, 2020.

[15] Argentine government.
Case definition.
https://www.argentina.gob.ar/salud/coronavirus-COVID-19/definicion-de-caso

Date: 2020.
Date accessed: May 25, 2020.

[16] Chile government.
Case definition.
https://www.minsal.cl/wp-content/uploads/2020/04/Ord.-B51-N%C2%BA933.pdf
Date: 2020.
Date accessed: May 25, 2020.

[17] Brazilian Ministry of Health.

Case definition and notification.
https://coronavirus.saude.gov.br/definicao-de-caso-e-notificacao
Date: 2020.
Date accessed: May 25, 2020.

[18] Brazilian Ministry of Health.
Epidemiological bulletin 16.
https://portalarquivos.saude.gov.br/images/pdf/2020/May/21/2020-05-19---BEE16---Boletim-do-COE-13h.pdf
Date: 2020.
Date accessed: May 25, 2020.

[19] CNN Brazil website: data from Brazilian Ministry of Health (in Portuguese).
https://www.cnnbrasil.com.br/saude/2020/06/04/ate-maio-27-8-testes-rt-pcr-para-covid-19-deram-positivo
Date: 2020.
Date accessed: June 04, 2020.

[20] United States Centers for Disease Control and Prevention.
Evaluating and testing persons for coronavirus disease.
https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-criteria.html
Date: 2020.
Date accessed: May 25, 2020.

[21] United States Centers for Disease Control and Prevention.
CDC COVID data tracker.
https://www.cdc.gov/covid-data-tracker/index.html#cases
Date: 2020.
Date accessed: June 16, 2020