Vienna Yearbook of Population Research 2022 (Vol. 20), pp. 143–166

The mathematics of the reproduction number *R* for Covid-19: A primer for demographers

Luis Rosero-Bixby[1,*] ***and Tim Miller***[2]

Abstract

The reproduction number *R* is a key indicator used to monitor the dynamics of Covid-19 and to assess the effects of infection control strategies that frequently have high social and economic costs. Despite having an analog in demography's "net reproduction rate" that has been routinely computed for a century, demographers may not be familiar with the concept and measurement of *R* in the context of Covid-19. This article is intended to be a primer for understanding and estimating *R* in demography. We show that *R* can be estimated as a ratio between the numbers of new cases today divided by the weighted average of cases in previous days. We present two alternative derivations for these weights based on how risks have changed over time: constant vs. exponential decay. We then provide estimates of these weights, and demonstrate their use in calculating *R* to trace the course of the first pandemic year in 53 countries.

Keywords: Covid-19; reproductive number *R*; demographic methods; net reproduction rate

1 Introduction

Health professionals and world leaders are talking more and more about the numbers *R* and R_0 (*R*-naught), the basic reproduction number.

Angela Merkel, a rare head of state with a scientific background, explained the trajectory of the Covid-19 pandemic on April 16, 2020, as follows:

> "*We are now at about a reproduction number of 1, so one person is infecting another one. . . . If we get to the point where everybody infects*

[1]University of Costa Rica, San Pedro, SJ, Costa Rica

[2]Department of Economic and Social Affairs, United Nations

[*]Correspondence to: Luis Rosero-Bixby, Lrosero@mac.com

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the United Nations and the University of Costa Rica.

https://doi.org/10.1553/populationyearbook2022.res1.3

> *1.1 people, then by October we will reach the capacity of our health care system with the assumed number of hospital beds. If we get to 1.2, so that everyone is infecting 20% more – out of five people, one infects two and the rest one. Then, we will reach the limits of our health care system in July. And if it is up to 1.3 people, then in June we will reach the limits of our health care system. So that's where we can see how little the margin is". (The Guardian News, 2020).*

The R was explicitly defined for the first time by the epidemiologist Klaus Dietz in 1975[1] (Dietz, 1975) as the expected number of infections (secondary cases) generated by a typical infected individual. If this occurs in a population in which everyone is susceptible (that is, at the beginning of an epidemic; hence the subscript zero), this number is R_0, or the basic reproduction number. In later stages of an epidemic, epidemiologists usually call the R the "effective" reproductive number, which is often represented as $R(t)$. This number can, in turn, be a cohort (longitudinal) R, which is called in some texts the "*case* reproductive number;" or a period (cross-sectional) indicator, which is sometimes called the "*instantaneous R*" (Gostic et al., 2020). This article focuses on the instantaneous, effective reproductive number, the $R(t)$, which we usually refer to simply as R.

R is considered to be an important indicator for monitoring the Covid-19 pandemic, and particularly for assessing the effects of infection control measures that frequently have high social and economic costs. R is also an important input for projecting future scenarios of disease spread. Moreover, knowing R_0 allows us to identify the threshold for *herd immunity*: i.e., the proportion of individuals in a completely susceptible population who need to become immune (naturally or by vaccination) in order to stop the growth of the epidemic curve. This threshold occurs at $(R_0 - 1)/R_0$ in homogeneous populations (Fine et al., 2011).

The demand for information about R for Covid-19 is so great that several websites provide estimates of R at the national and subnational levels, as well as the tools for producing estimates with user-provided data. The website https://shiny.dide.imperial.ac.uk/epiestim/ is an example of the latter (Cori et al., 2013). A systematic review of the Covid-19 literature up to September 2020 found 524 studies that reported R estimates, including 49 that explained the method and the data they used (Billah et al., 2020).

Although the concept of R is clear, the logic for its calculation in epidemiology is not easy to follow, as it usually requires the use of mathematical models and complex algorithms (Bettencourt and Ribeiro, 2008; Dietz, 1993; Nikbakht et al., 2019; Wallinga and Lipsitch, 2007). In addition, the results may vary substantially depending on the method used in the estimate (Billah et al., 2020). Hence, there is a demand for transparent and reasonable estimates of R.

[1] Earlier epidemiology in the field of malaria transmission used the concept of R in an effort to identify critical thresholds of population densities of mosquitos per human for stopping the spread of infection (Heesterbeek, 2002).

The purpose of this article is to use the toolbox of demographers to understand R, and to provide a straightforward procedure for estimating it. We seek to demystify the complexities of estimating this important indicator by following well-known procedures in demography, a discipline in which an analog of R – the net reproduction rate (NRR) – has been routinely computed for more than a century. The approach to estimating R we present in this article is similar to an approach that was recently developed in epidemiology by Cori and colleagues (Cori et al., 2013).

2 Simple (but not useful) formulas

In an ideal world in which we had access to perfect data, the reproduction number R could be calculated for each generation of infected individuals as the simple average of the number of infections generated by each member of the cohort. For example, the cohort of the first two infected persons in Costa Rica (March 6, 2020) had an $R = 4.5$, since, according to press reports, one case was a tourist who infected his spouse and the other was a doctor who infected eight people: $R = (1 + 8)/2 = 4.5$. However, this type of information is not available for the subsequent cohorts of individuals who were infected in the days that followed. Moreover, this information is not perfect, as it is possible that there were additional people who were infected by these two initial cases, but whose infections were not reported.

Another way to estimate R is the approach that has been used in demography since around 1880 (Lewes, 1984), and that was formally developed by Alfred Lotka, the father of mathematical demography (Dublin and Lotka, 1925). Lotka defined the NRR as the ratio of total births of daughters[2] in two successive generations, expressed as:

$$NRR = R = \int_u^v b(a)p(a)\,da \tag{1}$$

Where $b(a)$ is the fertility rate of women at age a and $p(a)$ is the probability of reaching this age alive (both variables refer only to females and female offspring), and the limits of the integral include the reproductive age range of women, which is, in practice, from $u = 15$ to $v = 49$ years.

If instead of applying the formula to population growth, we apply it to the reproduction of an outbreak – that is, to a cohort of individuals infected on the same date – the number of days elapsed since each cohort member was infected would be represented by a (the "age", defined as the days since infected); $b(a)$ would become the transmission rate of the infection at that "age" of a days, or the average number of people infected on day a; and $p(a)$ would become the probability of still being able to spread the disease after a days. The limits of the integral would be from u, or the first day when an individual achieves a sufficiently high viral load to become

[2] Lotka originally defined the NRR for generations of men and sons. However, for practical reasons, demographers compute it for women and daughters.

infectious; to the maximum number of days v that an infected person can still be infectious. Hence, R becomes the NRR of infected individuals or the reproduction number R in the lexicon of epidemiologists.

However, to use this formula as is customary in demography, it would be necessary to have data on daily counts of new cases of infected persons tabulated by the time-since-infection (duration of infection) of the person who infected them. The newly recovered cases,[3] as well as the deceased cases, should also be tabulated by the duration of the infection. Given that these data usually do not exist, it is necessary to make assumptions about the functional form of $b(a)$ and $p(a)$ to be able to estimate the reproduction number R indirectly given the lack of data disaggregated by duration a.

In the following sections, we present two approaches or models for estimating the reproduction number R using widely available data. To simplify the presentation, we assume no demographic change; i.e., a process with no births, deaths or migrants. In the discussion section, we address the robustness of the method to violations of these and other assumptions.[4]

3 A simple model with constant rates

Two heroic assumptions that can be used to simplify the estimation process are that the effective transmission rates and the recovery rates (or, more broadly to include deaths, the "removal rates") are constant throughout a person's infectious period; that is, that the rates are invariant with respect to a, days since infection.

If $b(a)$ is invariant with respect to a over the interval from u to v, then $b(a) = b$, which can be placed outside of the integral:

$$R = b \int_u^v p(a)\, da \tag{2}$$

Where b is the daily rate of effective transmission or the average number of people infected per day.

The probability of continuing to be infectious – or survival function $p(a)$ – is driven by the removal rate $g(a)$. In survival time analysis, this is the "hazard", "failure" or "mortality" rate. The following identity relates the survival function to the failure rate, which, in turn, nicely simplifies into a negative-exponential function

[3] Recovered cases are those of individuals who are no longer able to produce replication-competent virus.

[4] The acquired immunity of recovered individuals means that R declines over time because the pool of susceptible individuals is depleted. This dynamic of epidemics does not occur in the NRR of demography, as giving birth is a renewable process. The effect of a naturally declining R is, however, nil on the few days that individuals are sick with Covid-19, and can thus be omitted from the models used to estimate R in this article.

when the recovery rate $g(a)$ is invariant with respect to a (Keyfitz, 1968):

$$p(a) = e^{-\int_0^a g(x)\,d(x)} = e^{-ga} \tag{3}$$

The integral of $p(a)$ is well-known in demography and in survival time analysis: it is the area under the survival curve, which defines life expectancy; or, in this case, the number of days lived while infectious during the interval between u and v, which we call E.[5] Here, "surviving" means to continue in the infectious state.

Recalling Equation (2), the equation for the reproduction number R therefore simplifies into:

$$R = b \cdot E \tag{4}$$

b is the daily effective transmission rate of the infection (new infections per day), and
E is the mean number of days infectious.

This simple identity is useful to show that the reproduction number R has two components: the rate at which the infection is transmitted from one person to another and the mean duration of the infectious period. For example, if the daily transmission rate is $b = 0.2$ and the mean duration of the infectious period is $E = 10$ days, the reproduction number of the epidemic would be $R = 2.0$. Each case would produce two infections on average, under the two assumptions of invariance noted above.

4 Estimation of the effective transmission rate in a real population

The expected length of the infectious period E, and the recovery rate from the disease g that determines it, can reasonably be considered universal parameters determined by the biology of the infectious agent, which, in practice, vary little over time and from one population to another, at least as long as there is no treatment to speed recovery. Early data for Covid-19 suggest that the virus has an average infectious period of between eight and 15 days (Anastassopoulou et al., 2020; WHO, 2020; You et al., 2020). If an exogenous value of E is used, estimating R is a question of determining the specific transmission rate b of the population at each time t. The average transmission rate (under the aforementioned assumption of constancy over the infectious period) can be estimated as:

$$b(t) = c(t)/A(t) \tag{5}$$

$c(t)$ is the number of new cases on day t, and

[5] Solving the definite integral of $p(a)$ in Equation (3) yields the expected number of days a person remains infectious on average: $E = [p(u) - p(v)]/g$. If a person is infectious over the entire disease period, E is simply the inverse of g.

$A(t)$ is the number of currently active cases (infected people who are still spreading the disease) as of day t.

The number of new cases each day is a widely available statistic that is usually published in a timely fashion. However, the number of currently active cases needs to be estimated, which can be done using the data series of new cases in the previous days. Borrowing a basic relationship in demography (Lotka, 1998), which defines the size of a population based on the number of past births and the survival function, the number of active cases in the infective period u to v can be estimated with:

$$A(t) = \int_u^v c(t-a)p(a)\,da \tag{6}$$

Recalling from (2) that $R = b\int_u^v p(a)\,da$, the reproduction number R at time t is:

$$R(t) = \frac{c(t)}{\int_u^v c(t-a)p(a)\,da} \cdot \int_u^v p(a)\,da$$

Dividing both the numerator and the denominator by $\int_u^v p(a)\,da$ gives an expression with a clearer interpretation,

$$R(t) = \frac{c(t)}{\int_u^v c(t-a)\left[\frac{p(a)}{\int_u^v p(a)\,da}\right]da} \tag{7}$$

The numerator of this quotient is the number of new cases counted on day t, while the denominator is the weighted average of the cases reported during the previous u to v days. The weights used to obtain this average are represented by the term in square brackets, which we will call $w(a)$.[6] The weighting term is none other than the distribution of the "survival" function for the infectious state; that is, the proportion of people who continue to be infectious $(t-a)$ days after they first became infected. As previously shown in Equation (3), this is a simple negative exponential distribution under the assumption that the recovery rate is independent of the time elapsed since infected.

Moving on to the discrete version in which we solve the integral and simplify the fraction, we arrive at the following handy formula for estimating $R(t)$, which also assumes a fixed lag of six days between the date the infection occurs and the date the case is reported:

$$R(t-6) = c(t)\Big/\sum_{a=u}^{a=v} c(t-a)w(a) \tag{8}$$

The weights $w(a)$ are the aforementioned distribution of the survival function $p(a)$ evaluated over the interval u to v, which is determined by the following formula

[6] A quick and rough estimate of the denominator can be obtained by calculating the simple average – without weighting – of the cases in a period of at least 14 previous days.

(see Footnote 5):

$$w(a) = ge^{-ga}/(e^{-gu} - e^{-gv}) \tag{9}$$

Plausible parameters for estimating these factors are:

- Infectious interval: $u = 2$ and $v = 30$ days, and
- Daily recovery rate $g = 1/10$, which implies:
  - Mean duration of illness = 10 days and
  - Mean duration of infectiousness $E = 6$ days.

We took these parameters from early reports of the epidemiology of Covid-19 as observed mostly in the Hubei province in China (Anastassopoulou et al., 2020; Park et al., 2020; WHO, 2020). As knowledge of this disease progresses, different parameters may be favored in the future.

5 A (more realistic) model with exponential rates

Although epidemiology models of Covid-19 often assume that transmission and recovery rates are constant during the illness period, it is useful to explore alternative specifications of these two functions to better approximate the rates that have been observed during the first few months of the pandemic.

Regarding the transmission rate $b(a)$, initial data on the outbreak and measurements of the viral load while infected with the disease suggest a distribution with an early peak at two or three days followed by a sharp decline (He et al., 2020; Prakash, 2020). To keep the math simple, we assume a negative exponential function that declines quickly from the peak day of infection, which is also assumed to be the first day of infectiousness u:

$$b(a) = B_0 e^{-B_1(a-u)} \tag{10}$$

B_0 parameter representing the peak transmission rate on the initial day u, and
B_1 parameter indicating the speed of the decline in the transmission rate.

Regarding the removal rate $g(a)$, we did not find any estimates of its distribution for the novel Covid-19 disease in the literature. However, it seems reasonable to assume that the chance of recovery of an infected individual increases with time. The Gompertz model is a well-known function (and is convenient for integration purposes) for representing this behavior. It assumes that the rate of interest increases with duration time at a constant speed, which is a pattern observed for failure rates in most biological and mechanical entities (Keyfitz, 1968; Pollard, 1991):

$$g(a) = G_0 e^{G_1 a} \tag{11}$$

G_0 parameter representing the recovery at the beginning of the disease, and
G_1 parameter measuring the speed of increase in the recovery rate per unit of a.

The proportion of individuals who are still infectious after a days, or the survival function, is obtained by solving the integral in the formula below, which results in a double exponential function:

$$p(a) = e^{-\int_0^a g(x)\,d(x)} = e^{[(G_0/G_1)(1-e^{G_1 a})]} \tag{12}$$

Determining the effective reproduction number $R(t)$ with the functions $b(a)$ and $p(a)$ would entail estimation at each time t of the parameters defining these functions; most importantly, those of the transmission rate $b(a)$. The data required to do this are not available. Instead, we propose following a procedure that is well-known in demographic analysis: indirect standardization (Shryock and Siegel, 1976). In the first step of the procedure, we estimate the expected number of cases consistent with a reproductive number $R = 1$ with plausible distributions of $b(a)$ and $p(a)$, given the composition by duration a of active (currently infected) cases at time t.

The following relation estimates the expected number of cases given that $R = 1$:

$$c(t, R = 1) = \int_u^v c(t-a)[b(a)p(a)]\,da \tag{13}$$

In a second step, the $R(t)$ factor is estimated as a quotient between the observed and the expected cases:

$$R(t) \approx \frac{c(t)}{\int_u^v c(t-a)[b(a)p(a)]\,da} \tag{14}$$

Note that the denominator is, like in the model of constant rates (Equation (7)), a weighted average of the series of cases in the previous days, with the term in rectangular brackets as the weighting factor we have called $w(a)$.

Given the assumed functions for $b(a)$ and $p(a)$, and with the aforementioned lag of six days between infection and diagnosis, we arrive at the following formula in discrete terms for computing an estimate of the effective reproduction number $R(t)$ under the model we call "exponential rates":

$$R(t-6) = c(t) \Big/ \sum_{a=u}^{a=v} c(t-a)w(a) \tag{15}$$

This is the same formula as the one with the constant rates model (Equation (8)), but with a different set of weighting factors $w(a)$:

$$w(a) = B_0 e^{[-B_1(a-u)+(G_0/G_1)(1-e^{G_1 a})]} \tag{16}$$

These weighting factors $w(a)$ are the distributions derived by multiplying $b(a)$ times $p(a)$, starting with the day $a = u$ when infectiousness begins, which we are also assuming is the peak day of Covid-19 infectiousness.

Plausible parameters for estimating the set of weighting factors are:

- Infectious interval: $u = 2$ and $v = 30$ days (however, the upper limit is irrelevant, since the weighting factors reach zero by day 22);

- Parameters for the survival function $p(a)$ chosen to conveniently reproduce a 10-day mean duration of illness:
  $G_0 = 0.0169$ and
  $G_1 = 0.220$;
- Parameters for the effective transmission function $b(a)$ chosen to reproduce, in conjunction with $p(a)$, a convenient reproduction number $R = 1$:
  $B_0 = 0.157$ and
  $B_1 = 0.0508$.

As before, we chose the parameters on the basis of early knowledge of the Covid-19 epidemiology, mostly from the Chinese province of Hubei (Anastassopoulou et al., 2020; He et al., 2020; Park et al., 2020; Prakash, 2020; WHO, 2020).

6 Weighting factors, generation time and growth

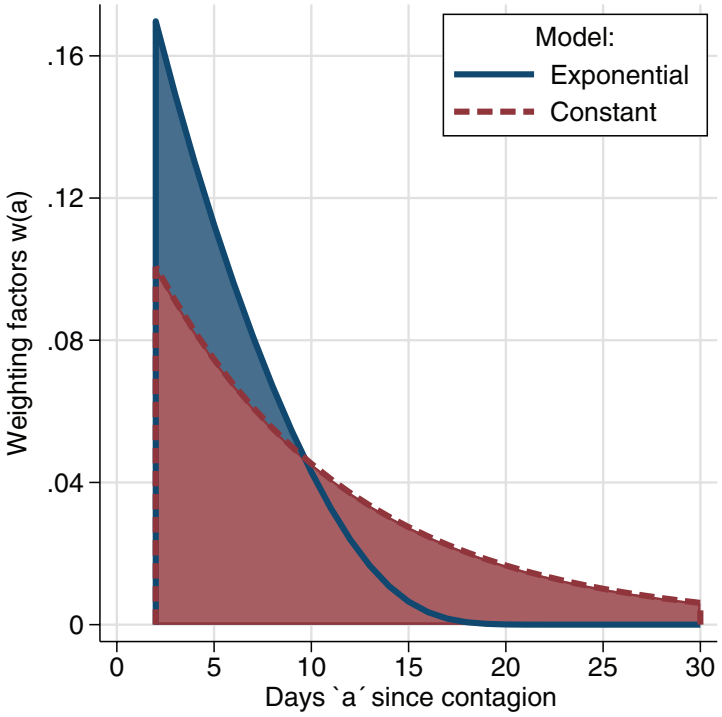
With two different sets of assumptions, we have arrived at the same relationship for estimating $R(t)$ as the quotient between the numbers of new cases in day t divided by the weighted average of cases in the previous days. Therefore, the choice of the correct set of weighting factors $w(a)$ becomes a key issue in estimating R. Figure 1 compares the $w(a)$ distributions in the previously presented constant and exponential models (the functions $b(a)$ and $p(a)$ behind the weighting factors are shown in Figure A.1 in the Appendix).

The constant rates model gives more weight to cases that occurred farther back in the past, while the exponential rates model gives more importance to more recent cases. If the number of new cases has changed little in the past, the $R(t)$ estimated with the two models will be similar. Remembering that these factors are in the denominator of the $R(t)$ formula, the constant rates model will result in higher $R(t)$ when the number of daily cases is increasing. The reverse will happen in later stages of the epidemic, when the number of daily cases is declining: i.e., the $R(t)$ estimates with the constant model will be lower. Therefore, the constant rates model and, in general, wider distributions will exaggerate extreme values of $R(t)$ estimates.

The two models can be considered archetypes for the choice of a weighting distribution for the indirect estimation of the reproduction number $R(t)$. Choosing a narrow distribution, as in the exponential model, gives more weight to recent cases, while a wider distribution, as in the constant model, gives more weight to older cases.

The shape of the $w(a)$ distribution is mostly driven by the shape of the transmission rate curve $b(a)$. To understand the transmission pattern of Covid-19, it is useful to look to evidence from recent outbreaks of other respiratory infections, such as: (1) the seasonal influenza curve with a high and narrow concentration in the first few days of illness; and (2) the SARS-2003 coronavirus outbreak with a wider and later distribution, which is somewhat similar to our rectangle of constant $b(a)$ (see Figure A.1 in the Appendix). Emerging data and estimates for the novel Covid-19 virus suggest that its transmission pattern resembles that of seasonal influenza,

Figure 1:
Weighting factors distributions



rather than of SARS, with a high concentration in days two to four (He et al., 2020); as in our exponential model.

The *generation time*[7] or length is an important indicator that epidemiologists often use to summarize the time it takes for an infected person to pass on the infection to others. It is a key input element in many epidemiological models that estimate the reproduction number R. This indicator is the mean duration a in our $w(a)$ distribution of weighting factors, which we call T:

$$T = \int_u^v aw(a)\,da \tag{17}$$

[7] The epidemiologic literature often uses the "serial interval" as an estimate of the "generation time". The generation time is the interval between the onset of infection for the "parent-child" cases. The serial interval is the observed period of the onset of symptoms between the infector and the infectee. The onset of infection and the onset of symptoms are separated by the "incubation period".

Since the integral in the exponential model does not have a simple analytical solution, we use numerical integration to derive the generation time (see Table A.1 in the Appendix) with:

$T = 10.20$ days in the model of constant rates, and

$T = 6.06$ days in the model of exponential rates.

Four review papers have identified nearly 40 articles on Covid-19 with estimates of T ranging from four to eight days (Billah et al., 2020; Griffin et al., 2020; Hussein et al., 2021; Park et al., 2020; Rai et al., 2021). As the estimates of our exponential model fall in the middle of this range, it appears that this model better represents the current state of knowledge about Covid-19 transmission than our constant model. An example of a set of $R(t)$ estimates with a shorter generation time of 3.6 days is from the Centre for the Mathematical Modeling of Infectious Diseases (CMMID) at the London School of Hygiene and Tropical Medicine (Abbott et al., 2020). As expected, these estimates result in smaller extreme figures; or, in other words, the estimates are very close to $R(t) = 1$ at all times.

The equivalent of the generation time in demographic analysis is the "mean interval between two consecutive generations," which Alfred Lotka, in his 1934 book *Analytical Theory of Biological Associations,* used to identify a relationship between the net reproduction rate R and a key indicator of the multiplication capacity of a population: the "intrinsic rate of growth" (Lotka, 1969). The relationship is:

$$R = e^{\rho T} \quad \text{or} \quad \rho = \ln(R)/T \tag{18}$$

In the context of Covid-19, ρ is the "intrinsic" or underlying rate of growth of the number of infectious individuals. Note that this growth rate may differ from the observed or real rate usually represented by lowercase r. In Lotka's words: "the ρ exposes the fundamental capacity of multiplication . . . while the r does not give us the true measure of that capacity since it is influenced by past factors we could call adventitious. The ρ is an asymptotic value to which the observed r will approach when those fundamental conditions remain the same" (Lotka, 1969, pp. 126–127). The observed growth r of Covid-19 cases is determined by both the fundamental conditions of its infectiousness and the momentum in the pool of individuals who are the source of infection. The intrinsic ρ is a rate free of momentum effects.

It is worth noting that several epidemiological studies have developed estimation procedures of R that start from this relationship and use observed growth rates as input and borrow T from models.[8] However, those studies usually do not make the distinction between the observed little r and the intrinsic ρ.

[8] Indeed, estimating the intrinsic growth rate directly from observed population data is a well-known approach in demography. In stable populations, births, deaths and population numbers are all growing at the intrinsic growth rate. In non-stable populations, Preston has shown that the growth rate of the population segment below the mean length of a generation is a good approximation of the intrinsic growth rate (Preston, 1986). Ediev, in generalizing the work of Fisher on reproductive value, has provided a method for estimating the intrinsic growth rate based on the dynamics of the population age structure (Ediev, 2007).

7 Estimates of $R(t)$ for Covid-19 in the real world

In this section, we analyze our estimates of $R(t)$ during the first year of the pandemic for 53 European and Latin American countries.[9] Figure 2 shows the results for Chile and Costa Rica, two Latin American countries known for maintaining good-quality health statistics. The figure illustrates the effect on R estimates of using the two different weighting distributions $w(a)$ corresponding to our constant and exponential assumptions. The figure also shows the relationship between the behavior of $R(t)$ and the epidemic curve of incidence over time.

The two proposed models produce approximately similar time-trend curves. They tell similar stories about when the reproduction number in each country is ascending, declining and crossing the $R = 1$ threshold; and about the speed of change in this indicator. However, at specific points in time, the level of the estimate may differ substantially, especially at extremely high or low levels. As expected, the model assuming constant rates exaggerates extreme values, tending toward higher values at high levels and lower values at low levels. This is in part because the mean generation time in the constant model is wider (10 days vs. six days). However, it is also because new cases tend to be increasing when $R > 1$ and to be decreasing when $R < 1$ (see the epidemic curves in the lower part of the figure), which, as we explained above, pulls the estimate up or down due to the greater weight assigned to older cases in the constant rates model.

As we noted in the previous section, our model of choice is the one that assumes exponential rates of removal and transmission of the Covid-19 disease. The "constant model" was developed for didactic purposes only.

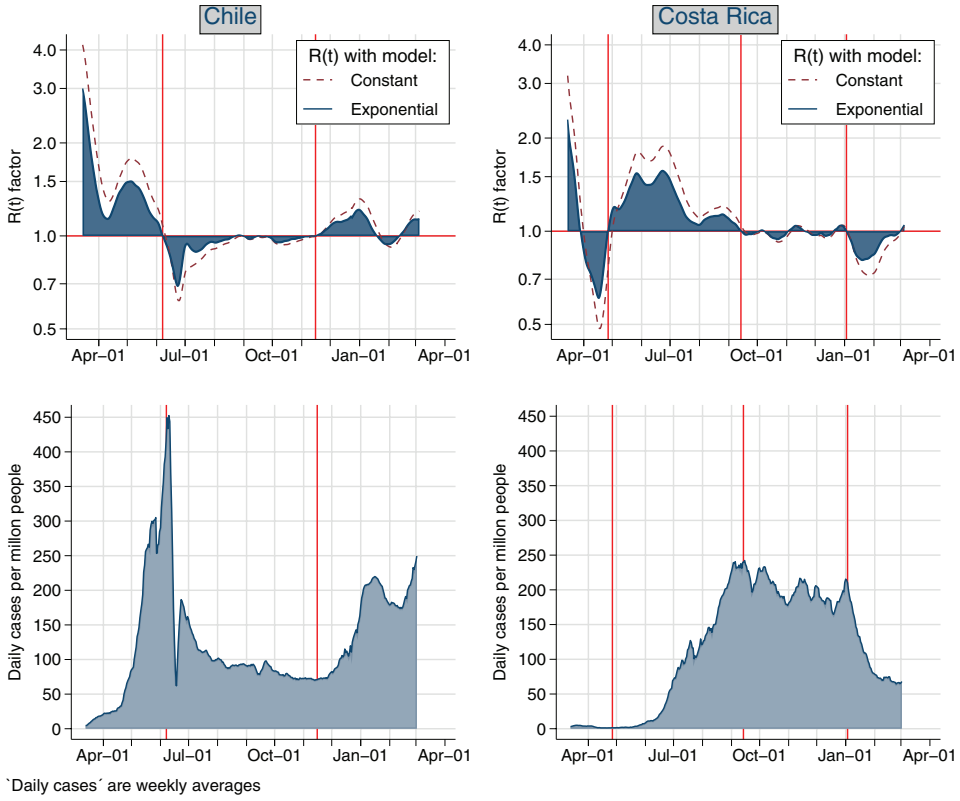
Figure 2 also illustrates the relationship between $R(t)$ and the epidemic curve of incidence. In periods when $R > 1$, the epidemic curve increases; and in periods when $R < 1$, the curve declines. When R is hovering around one, the number of new cases plateaus. This can occur at high levels, such as in Costa Rica from September to December; or at moderate levels, such as in Chile from August to November.

The points in time when $R(t)$ falls below the threshold of one are approximately the peak times of the pandemic waves: i.e., early July and early January in Chile and mid-September and January 1, 2021, in Costa Rica. $R(t)$ also shows the distinct phases or waves of the epidemic, delimited by the red vertical lines of Figure 2.

The $R(t)$ curves observed in these countries demonstrate the importance of taking aggressive action to contain the pandemic in its very early stages. Costa Rica

[9] We used the daily national series of confirmed Covid-19 cases from the "Our World in Data" website (Ritchie, 2020), accessed on March 10, 2021. The raw curves of cumulative cases were first smoothed out with local regression as implemented in the Stata software, command "lowess" (StataCorp, 2017). Clean daily numbers of cases were obtained by the difference in the smoothed cumulative curve, and were used as the input data in the estimation. Countries with populations of less than one million or unreliable data were excluded, along with the period before there were 100 accumulated cases. Our final analytical data file for Figures 2 and 3 is included as supplementary material in Excel and Stata-17 formats (available at https://doi.org/10.1553/populationyearbook2022.res1.3).

Figure 2:
***R*(*t*) and the incidence curve during the first year of the Covid-19 pandemic in Chile and Costa Rica**



Source: Daily national series of confirmed Covid-19 cases from the website "Our World in Data" (Ritchie, 2020), accessed on March 10, 2021.

employed that strategy by implementing aggressive contact tracing and testing programs, as well as drastic lockdown measures that essentially paralyzed the country from March 15 to April 15 (Rosero-Bixby and Jiménez-Fontana, 2021). Consequently, in Costa Rica, the $R(t)$ factor fell well below one, and the number of infections was contained at levels close to zero. In contrast, Chile did not reduce its R to the threshold of one or lower in April, and paid dearly for this failure with a devastating surge in infections in the following period. After the first month of the pandemic, both countries had rising R, but because the increase started at very different baselines, the results were vastly different. By June 15, the pandemic was exploding in Chile, at 260 daily cases per million population; whereas in Costa Rica, just 20 daily cases per million population were being reported.

The effects of Costa Rica's initial success in containing the virus were still apparent as long as one year after the start of the pandemic. As of March 10, 2021, the cumulative mortality caused by Covid-19 was 561 deaths per million residents in Costa Rica, compared to 1,117 deaths per million residents in Chile.

In general, subtle differences in the trajectory of the $R(t)$ resulted in two substantially different epidemic curves of incidence in Chile and Costa Rica. This is an obvious point from a demographic perspective: the absolute increase in population size is driven by both the reproduction rate and the initial population size. By the same token, both the R factor and the number of actively contagious individuals drive the incidence curve.

Broadening the scope of our analysis to 18 countries in Latin America and 35 countries in Europe, Figure 3 shows the results of our R estimates (exponential model), with weekly boxes displaying the distribution of countries by R. The box's hinges indicate the interquartile interval, and each box's central line indicates the median value of R for that week.

Epidemiologists pay special attention to the R_0 factor – the basic reproduction number – to characterize and model epidemic outbreaks. The level of $R(t)$ – the effective reproduction number – in the first days of an outbreak is an approximation of this basic R_0. The first boxes in the figure thus suggest that Covid-19 R_0 was in the interquartile range of 1.9 to 2.8 in European populations, whereas it was in the range of 2.3 to 2.5 in Latin American populations.

On both continents, the initial R declined sharply in the first few weeks, though more so in Europe than in America. In the European countries, R leveled out at around $R = 0.8$ in May, while in the Latin American countries, R leveled out at around $R = 1.15$. This means that in Europe, the first pandemic wave peaked (R crossed one) in early April, with the incidence of Covid-19 falling sharply thereafter. By contrast, in Latin America as a whole, the peak ($R = 1$) of the first pandemic wave seems to have occurred much later, in early August.

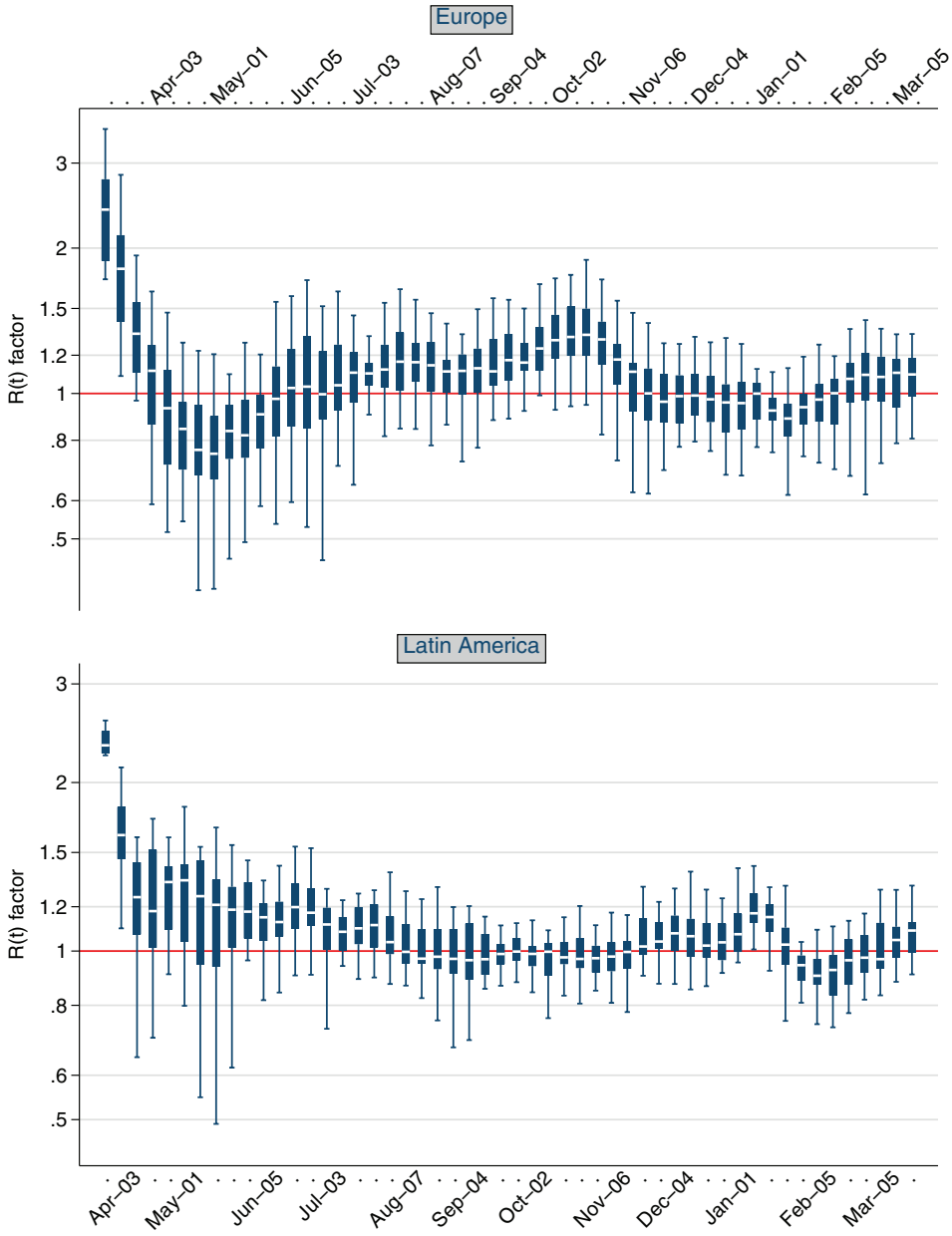
In Latin America, R hovered around $R = 1$ from August to December. Thus, the first wave did not really end, but instead plateaued at high levels of incidence.

In Europe, the Covid-19 pandemic has followed a trajectory of three well-defined waves: the initial wave peaked in April 2020; the second wave peaked in November 2020; and the third wave had not yet peaked by March 5, 2021.

One year after the start of the pandemic, the described trajectories of $R(t)$ resulted in a mortality toll that was 16% higher in Latin America, with 1,325 deaths per million people, than it was in Europe, with 1,139 deaths per million people.

The data from the 18 Latin American countries confirm our previous observation that the very early containment of R correlates with a less severe pandemic in the following months. In these countries, the correlation coefficient between the national level of R two weeks after case 100 was diagnosed and the death toll in the first year of the pandemic is strong, at 81%. However, this association is not observed in Europe, where the correlation coefficient is weak, at 5%. Figure A.2 in the Appendix shows the scatter plots behind these correlations.

Figure 3:
Weekly distribution by $R(t)$ of countries in Europe and Latin America



Source: Daily national series of confirmed Covid-19 cases from the website "Our World in Data" (Ritchie, 2020), accessed on March 10, 2021.

8 Discussion

The reproduction number R is a key indicator that has been used to characterize the dynamics of the Covid-19 pandemic, and to assess the effects of pandemic-related policy interventions. Unfortunately, the available statistics do not allow us to calculate this factor unequivocally. Instead, R must be estimated using indirect methods based on theoretical models and assumptions about the behavior of this novel disease. This article provides an approach for estimating R using methods and models developed a century ago in demography. The strengths of the proposed approach are the transparency of the assumptions from the point of view of demographers and the simplicity of the procedure.

The simple relationship used to estimate $R(t)$ on a daily basis is a quotient between the current number of new cases divided by a weighted average of the number of cases in the previous 20 or 30 days. We suggest using a set of weighting factors derived from assuming that: (1) the transmission rate of an infected individual declines sharply from a peak at day 2 of the illness following a negative exponential function; and (2) the recovery rate from the disease follows the Gompertz law of exponential growth with disease duration. A mean generation time of six days summarizes this suggested set of weighting factors. Early estimates of this interval, mostly for outbreaks in China's provinces, range from four to eight days. A weighting factors distribution with shorter generation times will result in $R(t)$ values that are closer to one; i.e., with less extreme values. We have shown that during stages of the outbreak when the number of new cases is increasing, shorter generation times (narrower distributions) result in lower $R(t)$ estimates; whereas during stages of the outbreak when the number of new cases is decreasing, shorter generation times result in higher (closer to one) estimates. In spite of these differences, the general time trend in $R(t)$ does not change meaningfully when different distributions are chosen. As our knowledge about this novel coronavirus improves, researchers will have more information that will enable them to make better informed choices about the distribution of the weighting factors used to estimate $R(t)$.

The strategy proposed in this article for estimating R is not new in epidemiology. A similar equation was proposed by Wallinga and Lipsitch (2007, Equation 4.2), and was implemented through web-based tools by Cori et al. (2013). The distribution $w(a)$, or the set of weighting factors of cases that occurred in previous days $t - a$, is called the "*infectivity profile*" by these authors, which is also the distribution of the generation time. Epidemiology studies assume a mathematical function for the $w(a)$ distribution, with the gamma function being the most commonly used (Knight and Mishra, 2020).

Using the computer tool provided by Cori et al. (2013), we were able to reproduce very closely our R estimates with the gamma function for a mean generation time of six days and a standard deviation of three. One study has recommended using the Cori et al. approach to estimate R after comparing it with two other epidemiological

methods applied to a simulated Covid-19 epidemic in which the true R is known (Gostic et al., 2020).

The main contribution of this article is that we demonstrated how the problem of estimating R can be approached with demographic thinking. The key set of weighting factors $w(a)$ is seen here not as a black box of a mathematical function, but as the product of two well-known demographic concepts: a survival function and a birth function, which could be defined analytically or with discrete observed distributions.

Our model assumes the absence of demographic change, meaning that births, deaths and migrations do not exist. Given that the time horizon involved in $R(t)$ estimates is short (30 days or less), including or excluding demographic change is unlikely to change the results in a meaningful way. Potential exceptions to this general observation are the arrival of imported cases of Covid-19 and mortality caused by Covid-19 itself.

Imported cases should not be counted in the numerator if the information is available, even though they must be included in the denominator. However, imported cases are statistically important only when the outbreak is at very low levels, and is in its initial stages.

Covid-19 deaths can be included by broadening the concept of the recovery rate $g(a)$ to a "removal rate" that would include both recovery and death as means of exiting the population of the infected. However, this correction would change the estimates very little, since the case fatality rate of Covid-19 has an order of magnitude of 0.01 (Worldometer, 2020), which, along with a mean period of illness of 15 days, is equivalent to a daily mortality rate of less than 0.001. Given that the mean daily recovery rate of Covid-19 is around 0.1, the correction would thus be about 1%. Such a small correction may well be omitted.

A weakness in all of the estimates on the numbers of reported cases is that this statistic is just the tip of the iceberg of all Covid-19 infections. But this does not necessarily invalidate the estimate. The estimated R would be valid insofar as these known observations are representative of the whole. Regardless of what proportion of cases is known and what proportion of cases is unknown, the important thing is that the known cases reflect the characteristics of the whole, and that this proportion does not change rapidly on the scale we are using to measure R. It is worth noting that given this weakness in the available input data, it might be pointless to use more intricate models to estimate R, which would seem to support the use of the simple approach this article proposes.

The R number is probably the best indicator for monitoring the dynamics in the propagation of an epidemic, and for taking action to contain it. It is like the speedometer in a car that tells us how quickly an epidemic is moving, and it does so in a more timely manner and with less contamination than its cousins; i.e., the rates of variation in the curves of incidence, hospitalizations or deaths. For example, in late January and early February 2021 in Costa Rica, the epidemic curve of incidence was declining, whereas R was clearly increasing (Figure 2). Thus, the

former indicator was misleading, while the *R* estimates reinforced the need to keep public health restrictions in place.

However, the *R* number tells only a partial story of an epidemic and its drivers. It does not, for example, tell us about the severity of an outbreak, which is better described by the incidence of diagnoses, the prevalence of hospitalizations or the mortality rate. In addition, because it is just an average, *R* can miss several important dimensions of reproduction, particularly in heterogeneous populations. For example, the existence of super-spreader individuals or clusters, which can be crucial in an outbreak, is totally hidden in this average. As a long tradition of demographic research has shown us, estimating the reproduction rate and assessing its meaning is just a first step in an ongoing quest to grasp the complexities of human behavior and the conditions that drive it.

Supplementary material

Available online at https://doi.org/10.1553/populationyearbook2022.res1.3

Supplementary file 1. Data in Excel format
Supplementary file 2. Data in Stata 17 format

ORCID

Luis Rosero-Bixby https://orcid.org/0000-0002-3063-3111

References

Abbott, S., Hellewell, J., Thompson, R. N., Sherratt, K., Gibbs, H. P., Bosse, N. I., … Chun, J. Y. (2020). Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Research, 5*(112), 112. https://doi.org/10.12688/wellcomeopenres.16006.2

Anastassopoulou, C., Russo, L., Tsakris, A., and Siettos, C. (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLoS ONE, 15*(3), e0230405. https://doi.org/10.1371/journal.pone.0230405

Bettencourt, L. M., and Ribeiro, R. M. (2008). Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS ONE, 3*(5), e2185. https://doi.org/10.1371/journal.pone.0002185

Billah, M. A., Miah, M. M., and Khan, M. N. (2020). Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence. *PLoS ONE, 15*(11), e0242128. https://doi.org/10.1371/journal.pone.0242128

Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology, 178*(9), 1505–1512. https://doi.org/10.1093/aje/kwt133

Dietz, K. (1975). Transmission and control of arbovirus diseases. *Epidemiology, 104*, 104–121.

Dietz, K. (1993). The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research, 2*(1), 23–41. https://doi.org/10.1177/096228029300200103

Dublin, L. I., and Lotka, A. J. (1925). On the true rate of natural increase: As exemplified by the population of the United States, 1920. *Journal of the American Statistical Association, 20*(151), 305–339.

Ediev, D. M. (2007). On an extension of RA Fisher's result on the dynamics of the reproductive value. *Theoretical Population Biology, 72*(4), 480–484. https://doi.org/10.1016/j.tpb.2007.03.001

Fine, P., Eames, K., and Heymann, D. L. (2011). "Herd Immunity": A rough guide. *Clinical Infectious Diseases, 52*(7), 911–916. https://doi.org/10.1093/cid/cir007

Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., ... De Salazar, P. M. (2020). Practical considerations for measuring the effective reproductive number, Rt. *PLoS Computational Biology, 16*(12), e1008409. https://doi.org/10.1371/journal.pcbi.1008409

Griffin, J., Casey, M., Collins, Á., Hunt, K., McEvoy, D., Byrne, A., ... More, S. (2020). Rapid review of available evidence on the serial interval and generation time of COVID-19. *BMJ Open, 10*(11), e040263. https://doi.org/10.1136/bmjopen-2020-040263

He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., ... Tan, X. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine, 26*(5), 672–675. https://doi.org/10.1038/s41591-020-0869-5

Heesterbeek, J. A. P. (2002). A brief history of R0 and a recipe for its calculation. *Acta Biotheoretica, 50*(3), 189–204. https://doi.org/10.1023/A:1016599411804

Hussein, M., Toraih, E., Elshazli, R., Fawzy, M., Houghton, A., Tatum, D., ... Duchesne, J. (2021). Meta-analysis on serial intervals and reproductive rates for SARS-CoV-2. *Annals of Surgery, 273*(3), 416–423. https://doi.org/10.1097/sla.0000000000004400

Keyfitz, N. (1968). *Introduction to the mathematics of population*. London: Addison-Wesley Publishing Co.

Knight, J., and Mishra, S. (2020). Estimating effective reproduction number using generation time versus serial interval, with application to COVID-19 in the Greater Toronto Area, Canada. *Infectious Disease Modelling, 5*, 889–896. https://doi.org/10.1016/j.idm.2020.10.009

Lewes, F. M. (1984). A note on the origin of the net reproduction ratio. *Population Studies, 38*(2), 321–324.

Lotka, A. J. (1969). *Teoría Analítica de las Asociaciones Biológicas* (G. A. Maccio, Trans.). Santiago: CELADE.

Lotka, A. J. (1998). *Analytical theory of biological populations* (D. P. Smith and H. Rossert, Trans.). New York: Plenum Press.

Nikbakht, R., Baneshi, M. R., Bahrampour, A., and Hosseinnataj, A. (2019). Comparison of methods to estimate basic reproduction number (R0) of influenza, using Canada 2009 and

2017–18 a (H1N1) data. *Journal of Research in Medical Sciences: The Official Journal of Isfahan University of Medical Sciences, 24*. https://doi.org/10.4103/jrms.JRMS_888_18

Park, M., Cook, A. R., Lim, J. T., Sun, Y., and Dickens, B. L. (2020). A systematic review of COVID-19 epidemiology based on current evidence. *Journal of Clinical Medicine, 9*(4), 967. https://doi.org/10.3390/jcm9040967

Pollard, J. P. (1991). Fun with Gompertz. *Genus, 47*(1/2), 1–20.

Prakash, M. K. (2020). *Quantitative COVID-19 infectiousness estimate correlating with viral shedding and culturability suggests 68% pre-symptomatic transmissions.* MedRxiv. https://doi.org/10.1101/2020.05.07.20094789

Preston, S. H. (1986). The relation between actual and intrinsic growth rates. *Population Studies, 40*(3), 343–351.

Rai, B., Shukla, A., and Dwivedi, L. K. (2021). Estimates of serial interval for COVID-19: A systematic review and meta-analysis. *Clinical Epidemiology and Global Health, 9*, 157–161. https://doi.org/10.1016/j.cegh.2020.08.007

Ritchie, H. (2020). Our World in Data, Coronavirus Source Data. Data on COVID-19 (coronavirus) by Our World in Data Retrieved 7 August 2020, from University of Oxford, Oxford Martin Programme on Global Development. https://ourworldindata.org/coronavirus-source-data

Rosero-Bixby, L., and Jiménez-Fontana, P. (2021). *Crónica de la pandemia de Covid-19 en Costa Rica*. Programa Estado de la Nación (PEN), Repositorio Institucional CONARE. https://repositorio.conare.ac.cr/handle/20.500.12337/8250

Shryock, H. S., and Siegel, J. S. (1976). *The methods and materials of demography*. New York: Academic Press.

StataCorp. (2017). *Stata Statistical Software: Release 15*. College Station, TX: StataCorpo LP.

The Guardian News (Producer). (2020). Angela Merkel uses science background in coronavirus explainer. 16 April 2020. https://www.youtube.com/watch?v=22SQVZ4CeXA

Wallinga, J., and Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences, 274*(1609), 599–604. https://doi.org/10.1098/rspb.2006.3754

WHO. (2020). *Report of the WHO-China Joint Mission on coronavirus disease 2019 (COVID-19).* World Health Organization, Geneva, Switzerland.

Worldometer (2020). COVID-19 coronavirus/death rate. Retrieved 23 July 2020, from https://www.worldometers.info/coronavirus/coronavirus-death-rate/

You, C., Deng, Y., Hu, W., Sun, J., Lin, Q., Zhou, F., . . . Zhou, X.-H. (2020). Estimation of the time-varying reproduction number of COVID-19 outbreak in China. *International Journal of Hygiene and Environmental Health*, Article 113555. https://doi.org/10.1016/j.ijheh.2020.113555

Appendix

Figure A.1:
Transmission rate $b(a)$ and "survival" function $p(a)$ in the constant and exponential rates models

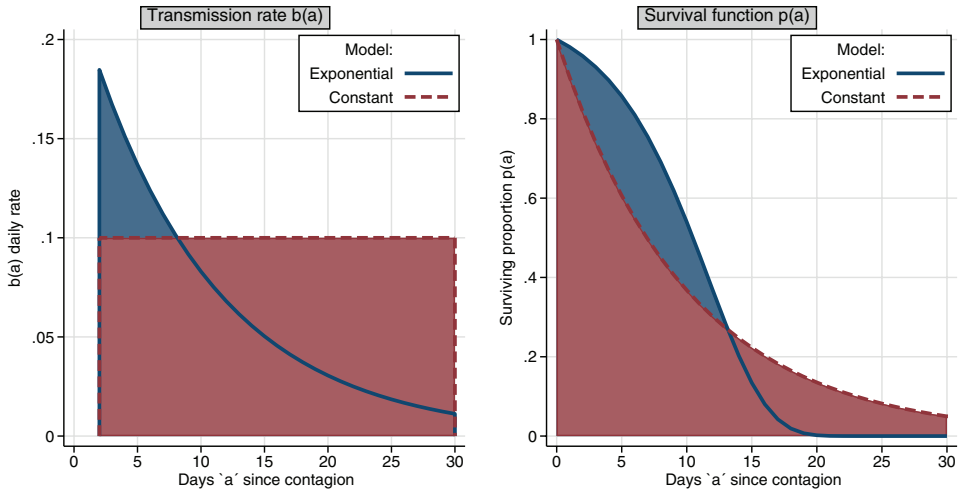
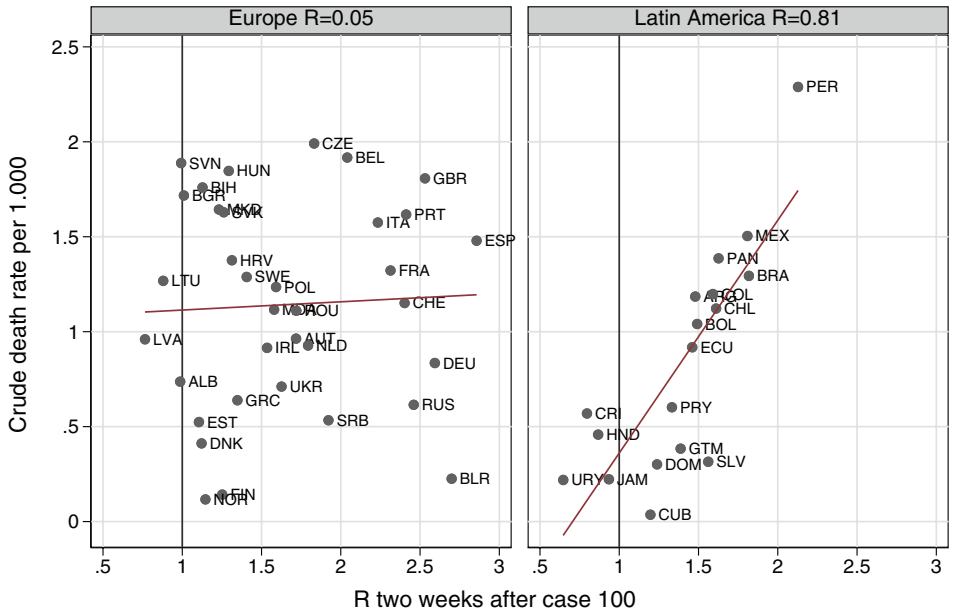


Figure A.2:
Correlation between the early level of *R* and the Covid-19 crude death rate in the first year



Note: Countries are identified by their ISO alpha-3 codes. United Nations Statistics Division, "*Standard Country or Area Codes for Statistical Use*" (M49 standard) https://unstats.un.org/unsd/methodology/m49/ Accessed on March 15, 2021.
Source: Daily national series of confirmed Covid-19 cases and deaths from the website "Our World in Data" (Ritchie, 2020), accessed on March 10, 2021.

Table A.1:
Weighting factors $w(a)$ to estimate $R(t)$ with two models

Days	**Constant rates model**			**Exponential rates model**			
a	$g(a)$	$p(a)$	$w(a)$	$g(a)$	$p(a)$	$b(a)$	$w(a)$
0.5	0	0.9512	0	0.0189	0.9911	0	0
1.5	0	0.8607	0	0.0235	0.9704	0	0
2.5	0.10	0.7788	0.1013	0.0293	0.9452	0.1531	0.1746
3.5	0.10	0.7047	0.0917	0.0365	0.9147	0.1455	0.1529
4.5	0.10	0.6376	0.0830	0.0455	0.8781	0.1383	0.1328
5.5	0.10	0.5769	0.0751	0.0568	0.8345	0.1315	0.1142
6.5	0.10	0.5220	0.0679	0.0708	0.7831	0.1249	0.0970
7.5	0.10	0.4724	0.0615	0.0882	0.7235	0.1188	0.0811
8.5	0.10	0.4274	0.0556	0.1099	0.6555	0.1129	0.0665
9.5	0.10	0.3867	0.0503	0.1370	0.5797	0.1073	0.0532
10.5	0.10	0.3499	0.0455	0.1708	0.4973	0.1020	0.0413
11.5	0.10	0.3166	0.0412	0.2129	0.4108	0.0969	0.0309
12.5	0.10	0.2865	0.0373	0.2654	0.3237	0.0921	0.0220
13.5	0.10	0.2592	0.0337	0.3308	0.2406	0.0876	0.0148
14.5	0.10	0.2346	0.0305	0.4123	0.1662	0.0832	0.0092
15.5	0.10	0.2122	0.0276	0.5139	0.1048	0.0791	0.0053
16.5	0.10	0.1920	0.0250	0.6405	0.0590	0.0752	0.0027
17.5	0.10	0.1738	0.0226	0.7984	0.0288	0.0715	0.0012
18.5	0.10	0.1572	0.0205	0.9951	0.0118	0.0679	0.0004
19.5	0.10	0.1423	0.0185	1.2404	0.0039	0.0645	0.0001
20.5	0.10	0.1287	0.0167	1.5461	0.0010	0.0614	0.0000
21.5	0.10	0.1165	0.0152	1.9271	0.0002	0.0583	0.0000
22.5	0.10	0.1054	0.0137	2.4020	0.0000	0.0554	0.0000
23.5	0.10	0.0954	0.0124	2.9940	0.0000	0.0527	0.0000
24.5	0.10	0.0863	0.0112	3.7319	0.0000	0.0501	0.0000
25.5	0.10	0.0781	0.0102	4.6516	0.0000	0.0476	0.0000
26.5	0.10	0.0707	0.0092	5.7980	0.0000	0.0452	0.0000
27.5	0.10	0.0639	0.0083	7.2269	0.0000	0.0430	0.0000
28.5	0.10	0.0578	0.0075	9.0080	0.0000	0.0409	0.0000
29.5	0.10	0.0523	0.0068	11.2280	0.0000	0.0388	0.0000
Sum			1.0000				1.0000
T			10.20				6.06