



OPEN

Reliably quantifying the evolving worldwide dynamic state of the COVID-19 outbreak from death records, clinical parametrization, and demographic data

Jose M. G. Vilar^{1,2}✉ & Leonor Saiz³✉

The dynamic characterization of the COVID-19 outbreak is critical to implement effective actions for its control and eradication but the information available at a global scale is not sufficiently reliable to be used directly. Here, we develop a quantitative approach to reliably quantify its temporal evolution and controllability through the integration of multiple data sources, including death records, clinical parametrization of the disease, and demographic data, and we explicitly apply it to countries worldwide, covering 97.4% of the human population, and to states within the United States (US). The validation of the approach shows that it can accurately reproduce the available prevalence data and that it can precisely infer the timing of nonpharmaceutical interventions. The results of the analysis identified general patterns of recession, stabilization, and resurgence. The diversity of dynamic behaviors of the outbreak across countries is paralleled by those of states and territories in the US, converging to remarkably similar global states in both cases. Our results offer precise insights into the dynamics of the outbreak and an efficient avenue for the estimation of the prevalence rates over time.

The global spread of the COVID-19 outbreak has had a major global impact. As of January 21, 2021, there have been 2.50 million reported deaths and 122.1 million confirmed cases¹. Massive travel restrictions, imposed quarantines, lockdowns, and other nonpharmaceutical interventions (NPIs) around the world have been able to slow down the progression of the outbreak, but their gradual lifting has already led to its deadly resurgence in multiple areas². Currently, it is still unclear how best to proceed to balance the economic, societal, ethical, and health trade-offs present. The most widespread quantification, based on the basic reproduction number, is insufficient to address the different trade-offs³. It indicates whether the outbreak is growing up or dying out, but it misses crucial information, including the timescales of the processes, the latent potential for a resurgence of the outbreak, and the possibility of actively controlling and tracing the infectious population.

Obtaining a detailed dynamic characterization of the outbreak, however, has proved to be particularly challenging and has been achieved only over small controlled populations. This characterization involved testing for the causative virus SARS-CoV-2, identifying contacts and the infection time, and following up the clinical evolution⁴⁻⁷. These analyses have led to a precise mathematical characterization of the clinical evolution of the disease, including the distributions of times from infection to symptom onset, from symptom onset to death, and from primary to secondary infections.

The available epidemiological information, however, has not been sufficiently reliable to be used directly and the detailed characterization is not feasible at a global scale. Methods based on compartmental models⁸ or Bayesian approaches⁹ have relied on location-specific information that is not readily available at a global scale.

¹Biofisika Institute (CSIC, UPV/EHU), University of the Basque Country (UPV/EHU), P.O. Box 644, 48080 Bilbao, Spain. ²Basque Foundation for Science, IKERBASQUE, 48011 Bilbao, Spain. ³Department of Biomedical Engineering, University of California, 451 E. Health Sciences Drive, Davis, CA 95616, USA. ✉email: j.vilar@ikerbasque.org; lsaiz@ucdavis.edu

These methods have typically pivoted toward using either the number of reported COVID-19 cases over time or the reported death counts over time.

Focusing on the number of cases is challenging because they represent only an undetermined fraction of the actual infections and this fraction depends on the testing capacity, which was remarkably low in the early stages of the pandemic, and on how the testing capacity evolves over time. In addition, there are a large fraction of asymptomatic infections, some of which are uncovered through contact tracing approaches while others remain undetected⁸. Moreover, it is not straightforward to determine when a positive testing person was infected, which depends on the location-specific testing procedures as well. Therefore, the connection between reported cases and the actual number of infections depends on the testing policies of each location, how they are implemented, and how they change over time.

An alternative approach is to use daily death counts. In this case, there is detailed statistical information on the disease progression over time through its different stages from infection to potential death, as well as on the age-stratified death ratios. The main drawbacks of this approach are on the mathematical side because it requires the solution of an inverse problem (finding the infectious population that would lead to the observed death curves) and on the fact that death counts are typically much lower than the number of infections, which make them prone to the inherent random fluctuations of the death process. These challenges are usually overcome by restricting the analyses to locations with large numbers of deaths and by imposing constraints on the reproduction number based on the information available about NPIs⁹.

Here, we provide a general quantitative approach for reliably quantifying the temporal evolution of the COVID-19 outbreak infectious and infected population utilizing multiple data sources, including daily death records, clinical parametrization of the disease, and location-specific demographic data. Explicitly, we use the customary infection-age structured mathematical description^{4,10–12}, which relies on the detailed statistical characterization of the disease progression over time, to develop an approach to obtain explicit estimates of the number of infectious and infected individuals in terms of the epidemiological death curves. We find that there is a general time delay between the infectious population and the daily deaths and a different time delay between the infected population and the cumulative number of deaths, which depend on the clinical parameters of the disease. The major location-specific contribution, besides the death counts, is the age structure of the population, which determines the infection fatality rate (IFR) and therefore the proportionality factors between infectious population and daily deaths and between infected population and cumulative deaths. We validated the approach with prevalence data of the infectious (PCR-RT testing) and infected (antibody testing) populations at a global scale and for states within the US, as well as with the timings of the peak infectiousness against the dates of the major country-wide lockdowns in Europe. To further quantify the temporal evolution and controllability of the COVID-19 outbreak we also obtained the temporal evolution of the growth rate of the infectious population. We consider explicitly countries in the world and states and territories within the United States (US) with at least 30 COVID-19 reported deaths as of January 21, 2021, which covers 97.4% of the world and 99.9% of the US population.

Results

Optimal dynamical constraints. The approach considers the dynamics of the infectious population, $n_I(t)$, at time t described through the expression

$$\frac{dn_I(t)}{dt} = k_G(t)n_I(t), \quad (1)$$

which establishes the definition of the (per capita) growth rate $k_G(t)$. As this expression results from the definition of the per capita growth rate, $k_G(t) \equiv \frac{1}{n_I(t)} \frac{dn_I(t)}{dt}$, it is completely general and independent of the underlying dynamics of the infection.

The underlying infection dynamics dictates the relationship of $k_G(t)$ and $n_I(t)$ with the different epidemiological quantities. Based on an infection-age structured mathematical description^{10,11} (See “Methods: Infection-age structured dynamics”), we have developed an approach to uncover the optimal dynamic constraints for these relationships in terms of delays and scaling factors (See “Methods: Dynamical constraints”).

We find that $n_I(t)$ is optimally related to the rate of increase of the expected cumulative deaths, n_D , at a later time $t + \tau_D$ according to

$$n_I(t) = \frac{\tau_G}{IFR} \frac{dn_D(t+\tau_D)}{dt}, \quad (2)$$

where IFR is the infection fatality rate and τ_G is the average generation time. Note that the first derivative of the expected cumulative number of deaths is obtained from the expected number of daily deaths as $\frac{dn_D(t)}{dt} = n_D(t) - n_D(t-1)$ (See “Methods: Dynamical constraints implementation with discrete time”) and that expected deaths are obtained from raw death counts reported by the Johns Hopkins University Center for Systems Science and Engineering¹ (See “Methods: Expected deaths”).

The preceding equation for $n_I(t)$ explicitly takes into account that, on average, an infectious individual within $n_I(t)$ has been infected at time $t - \frac{\tau_G + \sigma_G^2}{2\tau_G}$ and potentially dies with probability IFR at a time $\tau_I + \tau_{OD}$ after infection. Here, σ_G^2 is the variance of the generation time, and τ_I and τ_{OD} are the incubation and symptom onset-to-death average times, respectively. The values of these characteristic times have been estimated in days as $\tau_G = 6.5$, $\sigma_G = 4.2$, $\tau_I = 6.4$, and $\tau_{OD} = 17.8$ from precise follow up of specific groups of patients^{5,13,14}, which leads to $\tau_D = \tau_I + \tau_{OD} - \frac{\tau_G + \sigma_G^2}{2\tau_G} = 19.6$ (See “Methods: Dynamical constraints” and “Methods: Clinical parameters”).

Therefore, the value of τ_D can be interpreted as the average number of days from infection to death ($\tau_I + \tau_{OD}$) minus the average number of days that an individual remains infectious after infection ($\frac{\tau_G^2 + \sigma_G^2}{2\tau_G}$).

A general assumption of the approach is that the *IFR* for each age group remains the same for all locations⁵ and that the overall *IFR* for each location is obtained as the average over its specific population's age distribution (See "Methods: Infection fatality rate (*IFR*)").

The growth rate is obtained directly from Eqs. (1) and (2) as

$$k_G(t) = \frac{d}{dt} \ln \frac{dn_D(t+\tau_D)}{dt}, \quad (3)$$

which is related to the time-varying reproduction number, R_t , through the Euler–Lotka equation (See "Methods: Reproduction number").

The expected cumulative number of infected individuals at a time t , $n_T(t)$, follows from

$$n_T(t) = \frac{1}{IFR} n_D(t + \tau_I + \tau_{OD}), \quad (4)$$

which is obtained also from the dynamical constraints (See "Methods: Dynamical constraints").

To compare with prevalence studies, we used the dynamical constraints (See "Methods: Dynamical constraints") to obtain the relationship of the expected number of seropositive individuals, $n_{SP}(t)$, with the infected population,

$$n_T(t) = n_{SP}(t + \tau_{SP}), \quad (5)$$

where τ_{SP} is the average seroconversion time after infection. We also obtained the relationship of the expected number of positive reverse transcription polymerase chain reaction (RT-PCR) testing individuals, $n_{TP}(t)$, with the infectious population,

$$n_I(t) = \frac{\tau_G}{\Delta t_{TP}} n_{TP} \left(t + \tau_{TP} - \frac{\tau_G^2 + \sigma_G^2}{2\tau_G} \right), \quad (6)$$

where τ_{TP} is the average time for testing positive after infection and Δt_{TP} is the average number of days of positive testing. The values of these additional characteristic times have been estimated in days as $\tau_{SP} = 13.4$, $\tau_{TP} = 14.4$, and $\Delta t_{TP} = 20$ from clinical studies^{14–17} (See "Methods: Clinical parameters").

Implementation. Equations (1–4) completely characterize the dynamics of the outbreak from the expected cumulative deaths, the age structure of the population, and the general clinical parameters of the infection. We used a workflow (Fig. 1A) that incorporates explicitly the death counts compiled by the Johns Hopkins University Center for Systems Science and Engineering¹, the age structure reported by the United Nations for countries¹⁸ and the US Census for states and territories¹⁹, previously estimated aged-stratified *IFR*⁵, and other previously estimated clinical parameters^{5,13–17}. The expected number of daily deaths was inferred using density estimation from the death curves after preprocessing to minimize reporting artifacts (See "Methods: Inference and extrapolation"). Equations (5) and (6) were used to set the appropriate scale and delay to compare with seroprevalence studies.

Validation against prevalence data and NPIs. To validate the approach, we contrasted the estimated infectious and infected populations with the results from available antibody seroprevalence and RT-PCR testing studies for countries in the world and locations in the US with at least 30 reported deaths (Fig. 1B). The observed and estimated values agree with each other within the 95% credibility intervals (CrI), with a 1.56-fold accuracy over almost a 1,000-fold variation and a correlation coefficient (ρ) on a logarithmic scale of $\rho = 0.94$.

We combined into a separate analysis (Fig. 1C) the data of two additional comprehensive antibody seroprevalence studies within the US, which included 49²⁰ and 38²¹ states with non-zero prevalence values. For the states present in the two studies, we considered their average values. The estimated values agree with the observed prevalence with 1.61-fold accuracy and $\rho = 0.80$. The agreement for the combined data is better than for the data of each of the studies independently (1.65-fold accuracy and $\rho = 0.73$ for one study²⁰ and 1.74-fold accuracy and $\rho = 0.77$ for the other study²¹) and better than the agreement of both studies with each other ($\rho = 0.55$ for the 38 states common to both studies), indicating that the estimations of the approach fall within the observed variability of the prevalence studies. Indeed, the approach is effectively unbiased collectively, with the (geometric) average of the estimations being just a factor 1.12 (globally) and 1.13 (US) larger than that of the observations.

Overall, the validation of the approach shows that it can reproduce the available prevalence studies over a factor 615.0 between minimum and maximum values for countries and states with 72.0% (globally) and 77.6% (US) of the estimates within a factor 2 of the observed values, and with 100.0% (globally) and 93.9% (US) within a factor 3. There are no countries and only three states with values outside the factor 3 boundary. In the cases of Rhode Island and New Hampshire, the estimated infectious population is larger than the ones reported by one study²¹, which is consistent with the observed age-specific seroprevalence biased to older populations. In the case of Oregon, the estimated infectious population is smaller than the ones reported^{20,21}, which is consistent with the observed age-specific seroprevalence biased to younger populations²¹. In the case of Oregon, in addition, only 338 of the 1123 excess deaths during the outbreak before the collection of the specimens for analysis were attributed to COVID-19²².

We also validated the ability of the approach to capture the effects of NPIs (Fig. 1D). The inferred timings of the peak infectiousness (maximum infectious population) are concordant with the dates of the major early country-wide lockdowns in Europe⁹, with an overall average deviation of 0.0 days between lockdown and peak

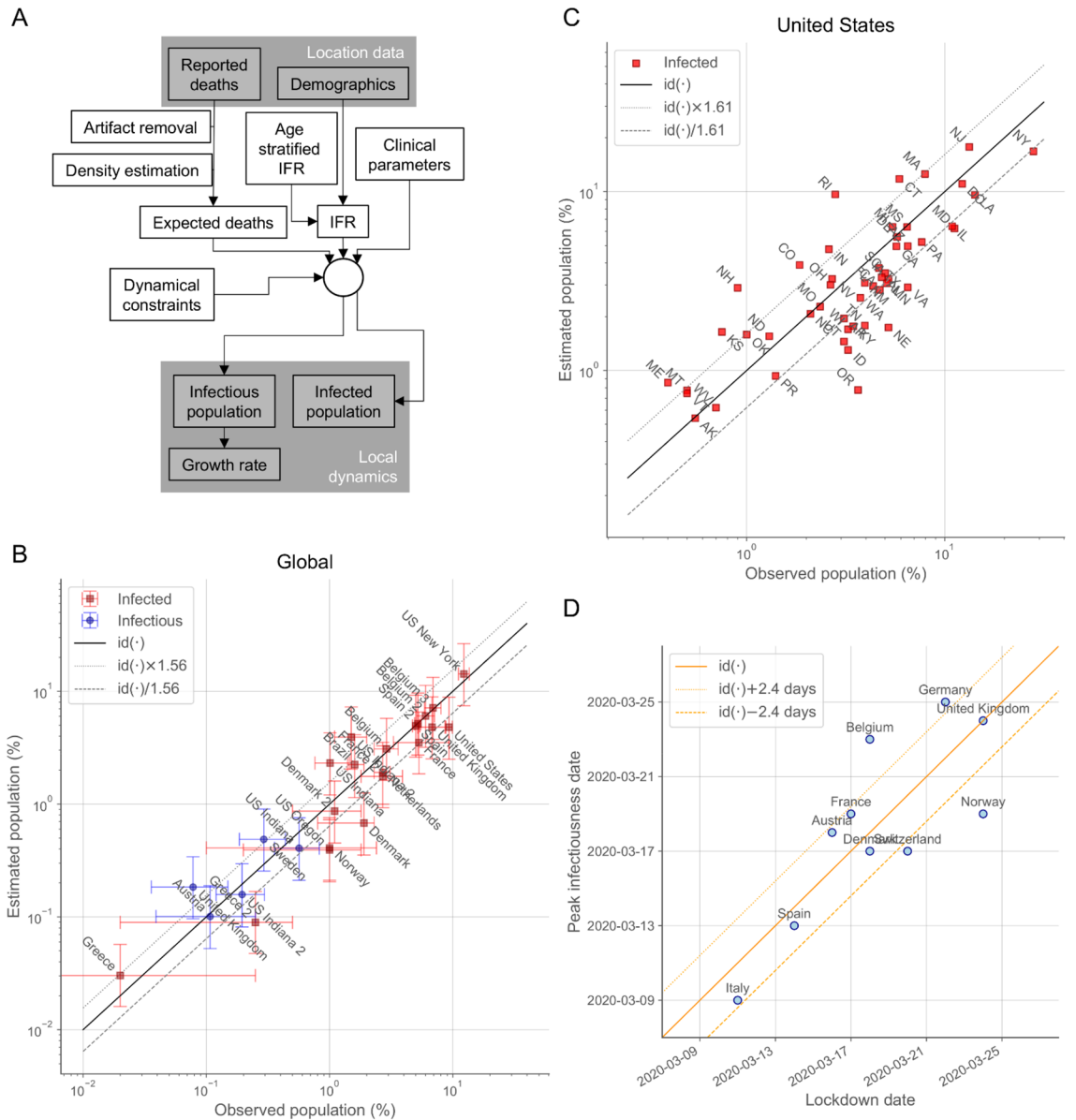


Figure 1. The approach consistently estimates prevalence and the timing of NPIs. The approach (A) has been validated with prevalence data of the infectious (PCR-RT testing) and infected (antibody testing) populations at a global scale (B) and for states within the US (C). The continuous black lines (B,C) represent the perfect prediction (identity function denoted by $id(\cdot)$) with the parallel dotted/dashed lines indicating the fold accuracy. Global data were obtained from sources described in Supplementary Table S1. Several locations have estimates for multiple date ranges. State data correspond to two studies with specimen samples taken primarily on the first two weeks of July²⁰ and of August, 2020²¹. The inferred timings of the peak infectiousness are plotted against the dates of the major country-wide lockdowns in Europe⁹ (D). The continuous orange line represents perfect concordance and the parallel dotted/dashed lines indicate the mean absolute error.

dates and a mean absolute deviation of 2.4 days. This concordance also indicates that there is only a small variability in the average timing between infection and death among those countries.

Outbreak progression motifs. We analyzed explicitly the time evolution of the growth rate, the infectious population, the cumulative number of infections, and the relationship between growth rate and infectious population for all countries and states with at least 30 deaths (Supplementary Fig. S1 and Table 1).

The characterization of the dynamics in terms of the growth rate and infectious population (Supplementary Fig. S1) shows that there are prototypical types of behavior, or motifs (highlighted with representative examples in Fig. 2). Locations either contained (Fig. 2A–C) or amplified (Fig. 2E,F) the outbreak in its initial stages. In the case of initial containment, the growth rate switched rapidly to negative values. Subsequently, the behavior branched into contained sporadic resurgences below the initial maximum infectiousness (Fig. 2A), uncontrolled resurgence of the infectious population over the initial maximum infectiousness (Fig. 2B), and sustained decrease

	World	United States
Highest local maximum of the infectious population	7.35 million on July 17, 2020	1.42 million on March 28, 2020
Current infectious population	14.46 million on January 21, 2021	2.42 million on January 21, 2021
Infected population	375.0 million as of January 21, 2021	52.0 million as of January 21, 2021
Underreporting factor	3.1 (only 122.1 million cases reported)	2.1 (only 24.6 million cases reported)
New infections per week	15.70 million	2.62 million
Growth rate: 10% quantile	-0.015/days (45.7-day half-life)	-0.020/days (34.5-day half-life)
Growth rate: median	0.000/days	0.000/days
Growth rate: 90% quantile	0.026/days (27.1-day doubling time)	0.011/days (64.5-day doubling time)
Population considered (locations with at least 30 deaths)	97.4%	99.9%
Number of locations considered	154	53
Number of locations with more than 10% of the population infected	44	43
Number of locations with more than 20% of the population infected	8	11
Number of locations with less than 7 actively infectious individuals per 100,000 population	22	0
Number of locations that reached 1,000 actively infectious individuals per 100,000 population	27	14
Correlation coefficient between infectious/infected population estimates and observations	0.94	0.80
Infectious/infected population estimates within a factor 2 of the observations	72.0%	77.6%
Infectious/infected population estimates within a factor 3 of the observations	100.0%	93.9%

Table 1. Key estimates of the outbreak.

of the outbreak (Fig. 2C). The specific behavior depended on the success of the measures implemented, e.g. targeted control and moderate lockdowns, and their subsequent relaxation²³. In the case of initial amplification, the dynamics proceeded in diverse ways, including an increasing infectious population converging to zero growth (Fig. 2D), a fast-evolving infectious population switching from positive to negative growth (Fig. 2E), and fast convergence to subsequent sustained residual growth (Fig. 2F). In general, the locations that reached a substantial negative growth rate are those that implemented long-term strict lockdowns, whereas zero or small positive growth rates correspond to intermediate measures with partial restrictions²³. In many cases, lifting restrictions has led to fast switching from negative to positive sustained growth, as illustrated by United Kingdom, and Italy (Fig. 2E), indicating the inability of these locations to contain the outbreak even at low values of the infectious population.

Global dynamics. At a global scale, the outbreak is characterized by two early local maxima of the overall worldwide infectious population on January 25, 2020 and March 26, 2020, coincidental with the regression of the outbreak in China, initially, and in Europe, afterward, reaching the highest local maximum of 7.35 million active infections (Fig. 3A,B) on July 17, 2020 with a subsequent increasing trend since September 22, 2020. In the US, there has been a local maximum of the infectious population on March 28, 2020 with 1.42 million active infections, a smaller local maximum on July 14, 2020, and a subsequent increasing trend since September 15, 2020 (Fig. 3D,E). The estimated infected population is 375.0 million, growing at a rate of 15.70 million new infections per week, for the World (Fig. 3C) and 52.0 million, growing at a rate of 2.62 million new infections per week, for the US (Fig. 3F), which is a factor 3.1 for countries in the World and 2.1 for locations in the US higher than the corresponding reported cases¹.

State of the outbreak across locations and its controllability. At a local scale, the per capita infectious population of countries and states has only exceptionally crossed the 1% value (27 out of 154 countries and 14 out of 53 states) (Fig. 4). The infected populations have surpassed the 10% of the total population only for 44 countries and 43 states (Fig. 4), with 8 countries and 11 states reaching the 20% value. These results confirm that relying on herd immunity is not a realistic option. Controlling the outbreak by actively tracing and quarantining newly and potentially infected individuals has been successfully implemented, until early October, 2020, in South Korea, with the use of large resources and with occasional outbreaks that required short-term extended human-interaction restrictions, and almost successfully in Japan, with voluntary business closures and other restrictions²³. These countries always remained below 7 actively infectious cases per 100,000 individuals (0.007% of their population) until early October, 2020, with average values of 2.2 (South Korea) and 2.6 (Japan) from March 1 to October 1, 2020. Only 0 of the states and 22 of the countries analyzed have had a per capita number of infectious individuals below the average value of South Korea since May 1, 2020, which indicates that 54.1% of the World (72.6% excluding China) and 97.9% of the US human population reside in countries or states that have not allowed targeted controllability of the outbreak with the resources used by South Korea. This inability

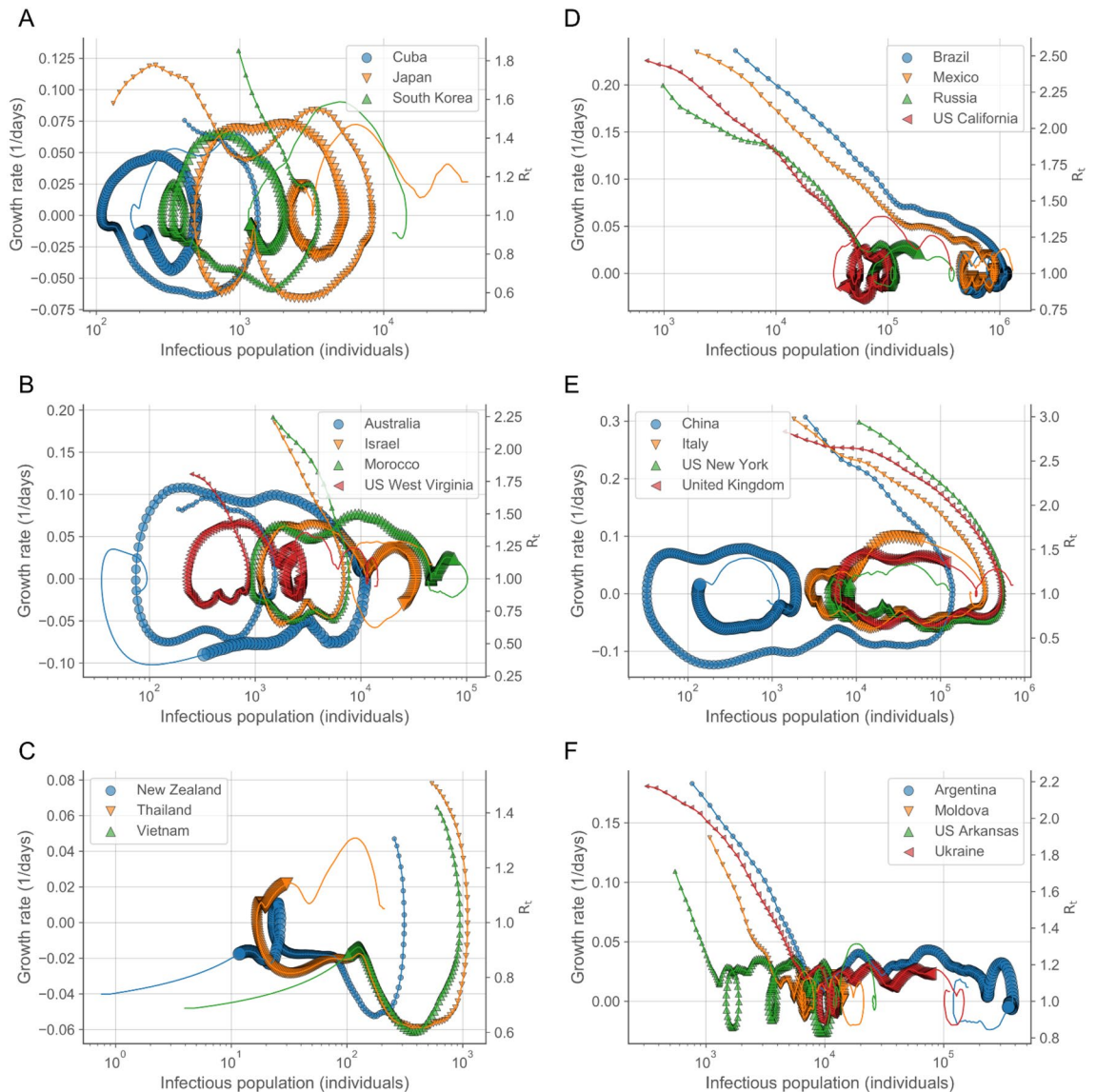


Figure 2. Trajectories in the growth rate-infectious population space followed prototypical types of behavior in the first-middle stages of the outbreak. Each day is indicated by a symbol increasing in size with time. The largest symbol corresponds to October 1, 2020. Complete trajectories, up to December 30, 2020, are indicated by the colored lines. The right axes indicate the growth rate in the scale of the reproduction number as $R_t = 1 + k_G(t)\tau_G$ (See “Methods: Reproduction number”). In all cases in the figure, the trajectory starts with positive growth rate ($R_t > 1$) at the smallest symbol and moves rightwards increasing the infectious population until the growth rate becomes negative ($R_t < 1$), at which point the infectious population starts to decrease and the trajectory moves leftwards. After the growth rate changes to negative values, the trajectories exhibit multiple types of differentiated behavior that can be classified in several groups depending on the extent of the initial and subsequent growth. If the growth rate reverts to positive values, the trajectory would move rightwards again. Common types of behavior include initial containment (A–C) with subsequent minor increases (A), amplified resurgence (B), and a sustained regression (C) of the outbreak; and initial amplification (D–F) with slow convergence to zero growth (D), convergence to negative growth (E), and fast convergence to initial sustained residual positive growth (F). The prefix “US” has been added to the name of the locations in the US. Confidence intervals for the growth rates and infectious population are provided in Supplementary Fig. S1.

to control the outbreak as soon as restrictions are lifted, even at low values of the infectious population but above the South Korea average value, is illustrated by United Kingdom and Italy (Fig. 2E).

From local to global dynamics. The collective properties of the individual local dynamics, as quantified by the distribution of growth rates across countries and states over time (Fig. 5), shows a progressive double stabilization of the outbreak. The double stabilization means that growth rates have mainly moved from large initial values towards zero values both at local scales and at a global scale. At local scales, growth rates for most

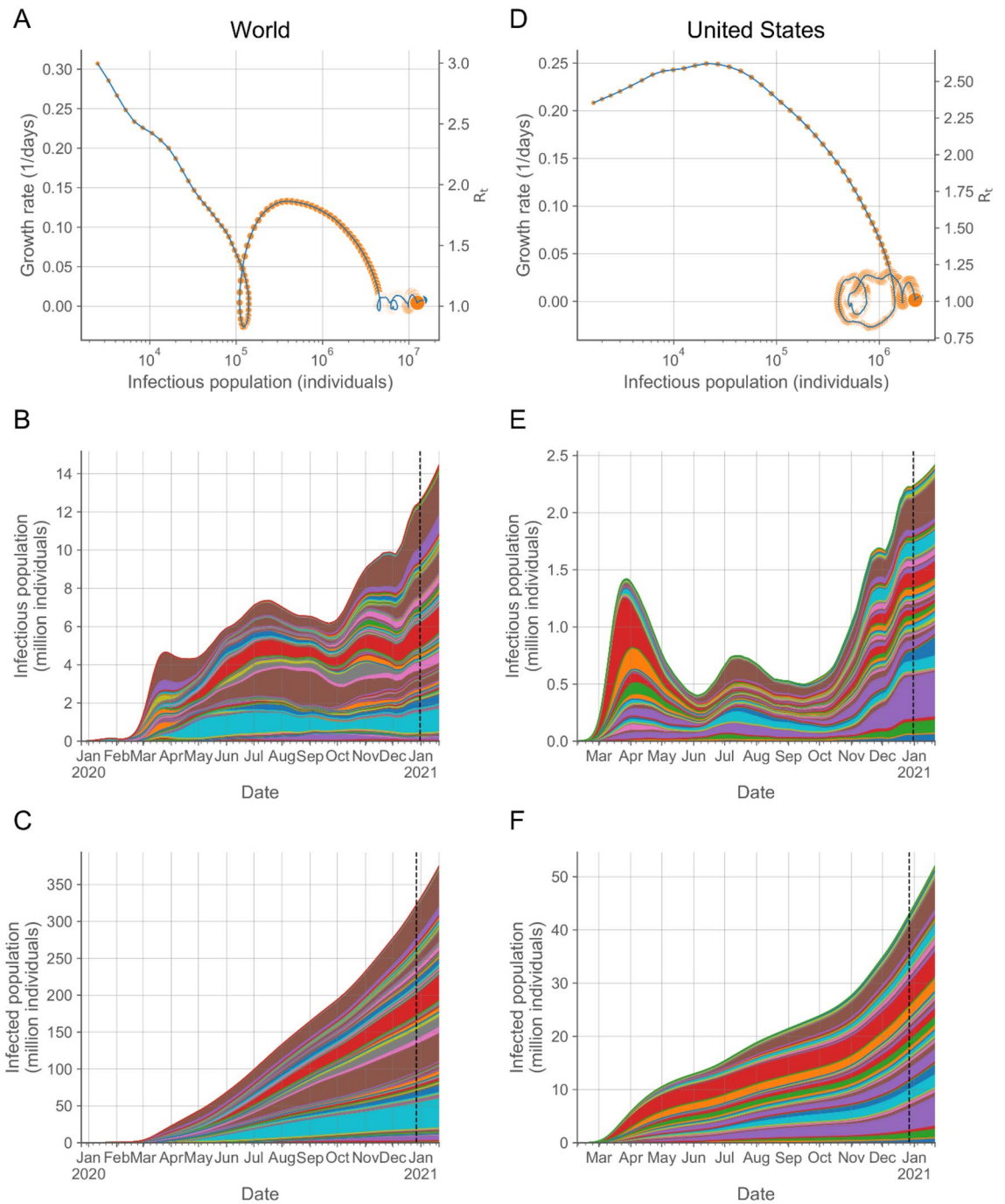


Figure 3. Characterization of the temporal evolution of the COVID-19 outbreak. Trajectory in the growth rate-infectious population space (A,D) and infectious (B,E) and infected (C,F) populations over time of the aggregate values for countries in the world (A–C) and locations in the US (D–F). In the trajectories (A,D), each day is indicated by a symbol increasing in size with time, ending with the largest symbol on December 30, 2020. The blue line at the end indicates the extrapolation to current time (January 21, 2021). The right axes indicate the growth rate in the scale of the reproduction number as $R_t = 1 + k_G(t)\tau_G$ (See “Methods: Reproduction number”). Each colored region in the area plots represents the contribution of a country (B,C) or a state (E,F) to the overall infectious and infected populations. Countries and states are arranged in alphabetical order from bottom to top. The vertical dashed lines (B,C,E,F) indicate the transition from inference to extrapolation. Individual data for all the countries and states are provided in Supplementary Fig. S1.

countries, whether positive or negative, decreased in absolute value, leading to a slowdown of the dynamics. At a global scale, positive and negative values have converged towards statistically compensating each other,

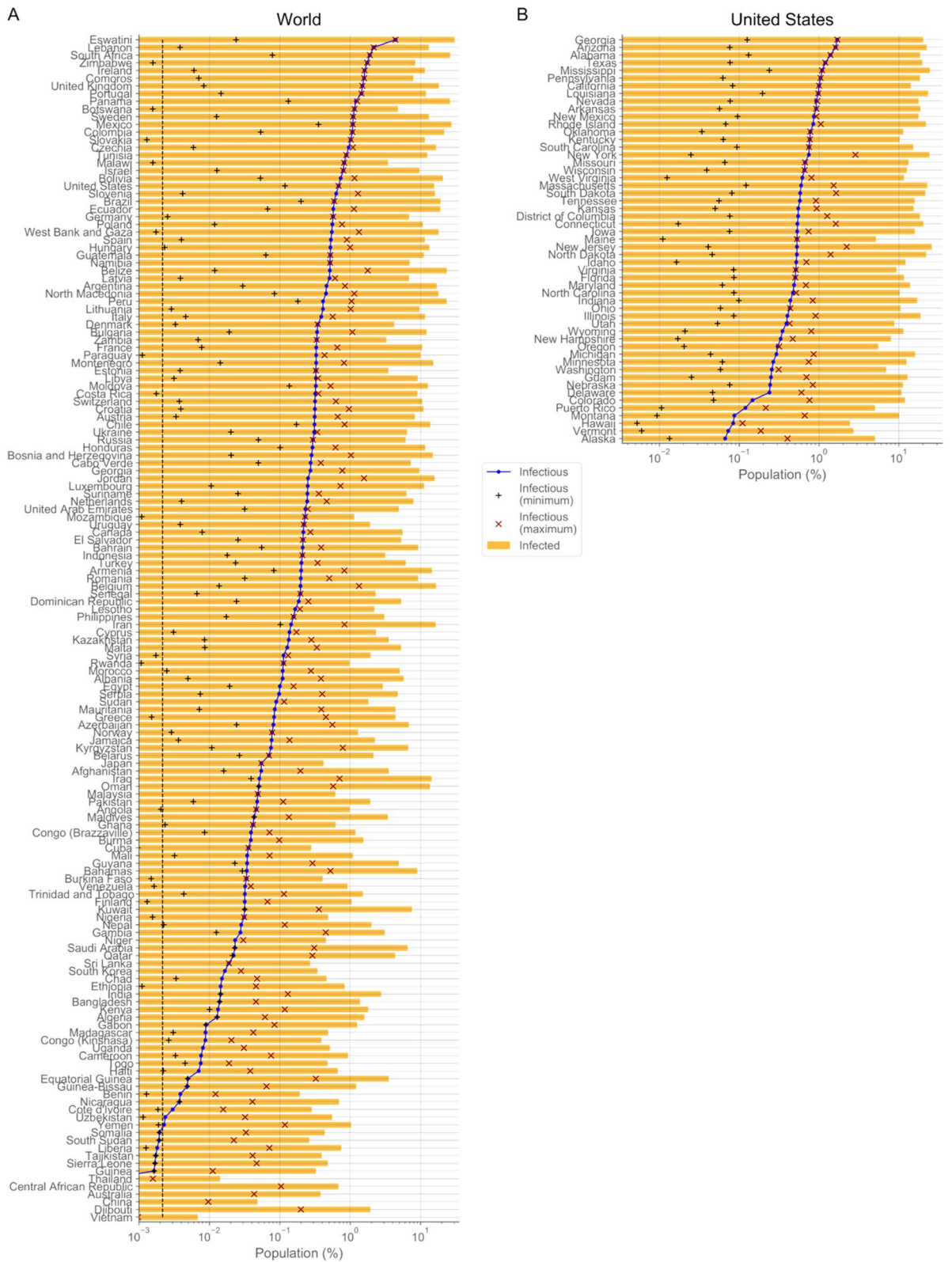


Figure 4. Per capita state of the outbreak. The percentage of infectious and infected populations on the last day estimated (January 21, 2021) and of the minimum and maximum infectious population reached are shown for countries in the world (A) and locations in the US (B). The minimum was computed for values after May 1, 2020. Countries and US locations have been sorted according to their per capita infectious population. The vertical dashed line (A) indicates the time-averaged infectious population of South Korea from March 1 to October 1, 2020.

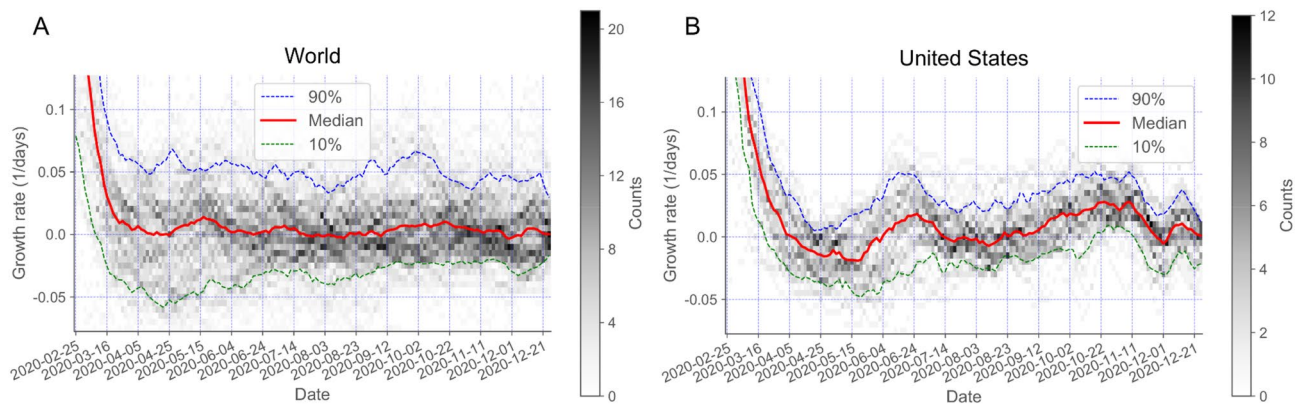


Figure 5. The temporal evolution of the distribution of growth rates shows a double stabilization of the outbreak. The temporal evolution of the median (red line), first decile (green line), and tenth decile (blue line) of the growth rates are plotted over the gray-coded histogram for countries in the world (A) and states and territories of the US (B). The median converges towards zero values and the deciles move closer to each other in a fluctuating manner for countries in the World and locations in the US. The double stabilization means that growth rates have mainly moved towards zero values both at local scales and at a global scale. At local scales, growth rates for most countries and states, whether positive or negative, decreased in absolute value, leading to a slowdown of the dynamics. At World and US global scales, positive and negative values for each of the locations have converged towards statistically compensating each other, decreasing further the overall net growth of the infectious population.

decreasing further the overall net growth of the infectious population. Specifically, the latest estimates of the growth rates on December 30, 2020 for countries in the world have a median value of 0.000/days, with 80% of the countries within the narrow range of values from -0.015 /days to 0.026 /days, which implies a very slow local dynamics with half-lives and doubling times of 45.7 and 27.1 days, respectively. Similarly, the median value of the growth rates for locations in the US is 0.000/days, with 80% of locations ranging from -0.020 /days (34.5 days half-life) to 0.011 /days (64.5 days doubling time).

Discussion

The dynamical constraints we have obtained through a detailed infection-age mathematical description of the outbreak allowed us to find the optimal time delays and scaling factors to connect the evolution of the reported death counts over time with those of the infectious and infected populations. Overall, integrating these constraints through a workflow with essential preprocessing steps showed that the approach can consistently infer the precise timing of NPIs and estimate prevalence data across countries in the world and locations in the US. A prominent feature of the approach is its ability to provide reliable results even for low death counts, which overcomes the major limitations of choosing between unreliable infection case data (highly dependent on testing rates) or noisy death counts as input to the inference problem⁹.

The approach assumes a general age-stratified *IFR*. In general, these quantities are expected to depend potentially on specific features of the population and the medical care facilities available. The available studies show a minimal variability among different countries and other locations that reported on prevalence²⁴. It also assumes an age-uniform exposure (attack rate), which is consistent with data for other respiratory diseases⁵ and holds to a large extent when there is information available for COVID-19^{20,25,26}. We have also assumed a constant generation interval typical of non-confinement locations, which has been observed to shorten in some cases by NPIs²⁷. Prevalence studies can also depend on the diminishing antibody levels after infection^{15,28}, collecting and processing specimens for analysis²⁹, and potential biases towards specific population groups²⁰. In addition, there might be a degree of under-reporting of COVID-19 deaths, as suggested by excess mortality not attributable to other causes than COVID-19²². Our results show that all of these potential deviations on the assumptions, on the data, and on prevalence studies collectively have only a restricted impact on the approach, with 72.0% (globally) and 77.6% (US) of the estimates within a factor 2 of the observed values and 100.0% (globally) and 93.9% (US), within a factor 3. This accuracy of the estimations is highly remarkable in the context of the observed prevalence spread over a factor 615.0 between minimum and maximum values, from 0.02 to 12.30%.

The analysis in terms of the growth rate–infectious population trajectories has revealed universal types of behavior of the outbreak for countries around the world and locations within the US. This information can be used to anticipate the response to enacting, modifying, or lifting NPIs. The most marked example is the response to strict lockdowns across countries in Europe (e.g. United Kingdom, Italy, Belgium, Spain, France, Germany, Austria, Netherlands, and Switzerland) and Northeast states in the US (e.g. New York, New Jersey, Massachusetts, Pennsylvania, and District of Columbia). They followed the same type of behavior (fast decrease of the growth rate from high to sustained negative values) until major restrictions were lifted in the European countries²³, turning the growth rate of their infectious populations into highly sustained positive values. Our results show that those European countries had small actively infectious populations but not as small as required for targeted controllability. They also show that most Northeast states in the US are in a similar resurgence path but at much

earlier stages, with many of their NPIs still in place and with markedly smaller growth rates, which makes their reaching as deadly a resurgence as in Europe still avoidable.

At a global scale, the outbreak has reached a net growth rate fluctuating near zero values but with a high infectious population. A similar state has also been reached in the US. This type of fluctuating states, with long stagnant overall infectious population periods and median growth rate close to zero, is expected of bounded unsynchronized fluctuating populations³⁰, such as those from uncoordinated locations aiming at just preventing an unrestricted expansion of the outbreak rather than at its eradication. This widespread feature is present for both countries in the world and locations within the US. Considering the NPIs implemented²³, our results show that there have been locations with interventions to move the growth rate towards zero values and that there have been locations switching on and off severe measures to decrease temporarily the active infectious population. Despite not growing substantially since reaching its highest local maximum of 7.35 million active infections on July 17, 2020, the high value of the global infectious population attained is currently leading to 15.70 million new infections per week that replace the same ballpark number of individuals that stop being infectious. This high turnover makes the control of any potential resurgence extremely costly.

At a local scale, our results show a highly variable temporal evolution of the infectious populations, both over time for each location and across locations. Having an up-to-date estimate of the infectiousness of populations would allow policymakers to better implement travel planning among locations. The approach has proven to accurately track the effects of local NPIs. We also expect it to play a fundamental role in evaluating the progress of vaccination efforts, especially considering the challenges present, such as waning immunity levels and pathogen evolution³¹.

Methods

Infection-age structured dynamics. For the description of the dynamics, we follow the customary infection-age structured approach (for details see for instance Refs.^{4,10–12}). Explicitly, we consider the infection-age structured dynamics of the number of individuals $u_I(t, \tau)$ at time t who were infected at time $t - \tau$ given by

$$\frac{\partial}{\partial t} u_I(t, \tau) + \frac{\partial}{\partial \tau} u_I(t, \tau) = 0 \quad (7)$$

with boundary condition

$$u_I(t, 0) = j(t). \quad (8)$$

Here, τ is the time elapsed after infection, referred to as infection age, and $j(t) = \int_0^\infty k_I(t, \tau) u_I(t, \tau) d\tau$ is the incidence, with $k_I(t, \tau)$ being the rate of secondary transmissions per single primary case.

The solution is obtained through the method of characteristics³² as

$$u_I(t, \tau) = j(t - \tau) \quad (9)$$

for $t \geq \tau$ and $u_I(t, \tau) = 0$ for $t < \tau$. The resulting renewal equation, $j(t) = \int_0^\infty k_I(t, \tau) j(t - \tau) d\tau$, is used as the basis for the definitions of the reproduction number $R_t = \int_0^\infty k_I(t, \tau) d\tau$ and the probability density of the generation time $f_{GT}(\tau) = \frac{k_I(t, \tau)}{R_t}$.

The infectious population is given by

$$n_I(t) = \int_0^\infty P_I(\tau) u_I(t, \tau) d\tau, \quad (10)$$

which considers that an individual remains potentially infectious after a time τ from infection with probability

$$P_I(\tau) = \int_\tau^\infty f_{GT}(l) dl. \quad (11)$$

Therefore, in terms of the incidence [substituting Eq. (9) in Eq. (10)], we have

$$n_I(t) = \int_0^\infty P_I(\tau) j(t - \tau) d\tau. \quad (12)$$

Additionally, we consider the expected cumulative number of infections, $n_T(t)$, expressed in terms of the overall accumulated incidence as

$$n_T(t) = \int_0^t j(s) ds, \quad (13)$$

and the dynamics of the expected cumulative deaths, $n_D(t)$,

$$\frac{d}{dt} n_D(t) = IFR \int_0^t f_{OD}(t-l) \int_0^l f_I(l-s) j(s) ds dl, \quad (14)$$

which takes into account that deaths occur with probability given by the infection fatality rate, IFR , at times after infection given by the convolution of the probability density functions of the incubation, f_I , and symptom onset-to-death, f_{OD} , times.

Similarly, the variation of the expected number of seropositive individuals at a time t , $n_{SP}(t)$, is expressed as

$$\frac{d}{dt} n_{SP}(t) = \int_0^t f_{SP}(t-s) j(s) ds, \quad (15)$$

where f_{SP} is the probability density function of the seroconversion time after infection, and the expected number of individuals with positive RT-PCR testing $n_{PT}(t)$, as

$$n_{TP}(t) = \int_0^\infty P_{TP}(\tau) u_I(t, \tau) d\tau, \quad (16)$$

where $P_{TP}(\tau)$ is the probability that an infected individual would test positive at a time τ after infection.

Dynamical constraints. To obtain a closed set of equations for the different epidemiological quantities, we developed an approach to optimally simplify the convolutions. Explicitly, for the expressions involving an integral $\int_0^\infty A(\tau) j(t - \tau) d\tau$ of a function A with the incidence j , we perform a series expansion of the incidence around the infection-age time τ_A ,

$$j(t - \tau) = j(t - \tau_A) + j'(t - \tau_A)(\tau_A - \tau) + O(j''), \quad (17)$$

with the value of τ_A chosen as

$$\tau_A = \frac{\int_0^\infty \tau A(\tau) d\tau}{\int_0^\infty A(\tau) d\tau}. \quad (18)$$

The specific value of τ_A leads directly to a first-order approximation,

$$\int_0^\infty A(\tau) j(t - \tau) d\tau = j(t - \tau_A) \int_0^\infty A(\tau) d\tau + O(j''), \quad (19)$$

because $\int_0^\infty A(\tau)(\tau_A - \tau) d\tau = 0$ by the definition of τ_A .

Using this approach, we obtain

$$n_I(t) = j\left(t - \frac{\tau_G^2 + \sigma_G^2}{2\tau_G}\right) \tau_G + O(j'') \quad (20)$$

from Eq. (12), where $\tau_G = \int_0^\infty \tau f_{GT}(\tau) d\tau$ and $\sigma_G^2 = \int_0^\infty (\tau - \tau_G)^2 f_{GT}(\tau) d\tau$ are the average and variance of the generation time, respectively, and

$$\frac{d}{dt} n_D(t) = IFR j(t - \tau_I - \tau_{OD}) + O(j'') \quad (21)$$

from Eq. (14), where $\tau_I = \int_0^\infty \tau f_I(\tau) d\tau$ and $\tau_{OD} = \int_0^\infty \tau f_{OD}(\tau) d\tau$ are the incubation and symptom onset-to-death average times, respectively. These expressions lead straightforwardly to

$$\frac{d}{dt} n_D(t) = \frac{IFR}{\tau_G} n_I\left(t + \frac{\tau_G^2 + \sigma_G^2}{2\tau_G} - \tau_I - \tau_{OD}\right) \quad (22)$$

and

$$n_D(t) = IFR n_T(t - \tau_I - \tau_{OD}), \quad (23)$$

up to $\mathcal{O}(j'')$. Note that we have used $\int_0^\infty 2\tau P_I(\tau) d\tau = \int_0^\infty d(\tau^2 P_I(\tau)) - \int_0^\infty \tau^2 dP_I(\tau) = \int_0^\infty \tau^2 f_{GT}(\tau) d\tau$ and $\int_0^\infty P_I(\tau) d\tau = \int_0^\infty d(\tau P_I(\tau)) - \int_0^\infty \tau dP_I(\tau) = \int_0^\infty \tau f_{GT}(\tau) d\tau$.

These expressions are used to estimate the infectious population $n_I(t)$ from the daily deaths, $\frac{d}{dt} n_D$, at time $t + \tau_I + \tau_{OD} - \frac{\tau_G^2 + \sigma_G^2}{2\tau_G}$ and the cumulative infected population $n_T(t)$ from the cumulative deaths, n_D , at time $t + \tau_I + \tau_{OD}$, leading to

$$n_I(t) = \frac{\tau_G}{IFR} \frac{d}{dt} n_D\left(t + \tau_I + \tau_{OD} - \frac{\tau_G^2 + \sigma_G^2}{2\tau_G}\right), \quad (24)$$

$$n_T(t) = \frac{1}{IFR} n_D(t + \tau_I + \tau_{OD}). \quad (25)$$

Similarly, we obtain

$$\frac{d}{dt} n_{SP}(t) = j(t - \tau_{SP}) + O(j'') \quad (26)$$

$$n_{TP}(t) = j(t - \tau_{TP}) \Delta t_{TP} + O(j''), \quad (27)$$

where τ_{SP} is the average seroconversion time after infection and Δt_{TP} is the average number of days an individual tests positive, which up to $\mathcal{O}(j'')$ leads to

$$n_{SP}(t) = n_T(t - \tau_{SP}), \quad (28)$$

$$\frac{d}{dt} n_D(t) = \frac{IFR}{\Delta t_{TP}} n_{TP}(t + \tau_{TP} - \tau_I - \tau_{OD}). \quad (29)$$

Combining Eqs. (28) and (29) with Eqs. (24) and (25) leads to

$$n_I(t) = \frac{\tau_G}{\Delta t_{TP}} n_{TP}\left(t + \tau_{TP} - \frac{\tau_G^2 + \sigma_G^2}{2\tau_G}\right), \quad (30)$$

$$n_T(t) = n_{SP}(t + \tau_{SP}), \quad (31)$$

which is used to validate the values of the estimated infectious population $n_I(t)$ from RT-PCR testing results, n_{TP} , at time $t + \tau_{TP} - \frac{\tau_G + \sigma_G}{2\tau_G}$ and the cumulative infected population $n_T(t)$ from seropositivity testing, n_{SP} , at time $t + \tau_{SP}$.

Expected deaths. The raw cumulative death counts over time, $n_W(t)$, are obtained from the Johns Hopkins University Center for Systems Science and Engineering¹ for countries and for US locations.

The daily death counts $\Delta n_W(t) = n_W(t) - n_W(t - 1)$ are considered to contain reporting artifacts if they are negative or if they are unrealistically large. This last condition is defined explicitly as larger than 4 times its previous 14-day average value plus 10 deaths, $\Delta n_W(t) > 10 + 4 \times \frac{1}{14}(n_W(t) - n_W(t - 14))$, from a non-sparse reporting schedule with at least 2 consecutive non-zero values before and after the time t , $\Delta n_W(t) \neq \frac{1}{5}(n_W(t + 2) - n_W(t - 3))$.

Reporting artifacts identified at time t are considered to be the result of previous miscounting. The excess or lack of deaths are imputed proportionally to previous death counts. Explicitly, death counts are updated as

$$n_W(t - 1 - i) \leftarrow n_W(t - 1 - i) \frac{n_W(t-1)_{estimated}}{n_W(t-1)} \quad (32)$$

with $n_W(t - 1)_{estimated} = n_W(t) - \frac{1}{7}(n_W(t - 1) - n_W(t - 8))$ for all $i \geq 0$. In this way, $\Delta n_W(t)$ is assigned its previous seven-day average value.

The expected daily deaths, $\Delta n_D(t)$, are obtained through a density estimation multiscale functional, f_{de} , as $\Delta n_D(t) = f_{de}(\Delta n_W(t))$, which leads to the estimation of the expected cumulative deaths at time t as $n_D(t) = n_W(t_0) + \sum_{s=t_0+1}^t \Delta n_D(s)$. Specifically,

$$f_{de}(\Delta n_W(t)) = (1 - r_1)dd_0 + r_1((1 - r_2)dd_1 + r_2dd_2) \quad (33)$$

with

$$r_1 = e^{-0.3dd_1}, \quad (34)$$

$$r_2 = e^{-3dd_2}, \quad (35)$$

$$dd_0 = ma_{14}(ma_{14}(\Delta n_W(t))), \quad (36)$$

$$dd_1 = rg_{12}(ma_{14}(\Delta n_W(t))), \quad (37)$$

$$dd_2 = rg_{48}(ma_{14}(\Delta n_W(t))), \quad (38)$$

where $ma_{14}(\cdot)$ is a centered moving average with window size of 14 days and $rg_{\sigma}(\cdot)$ is a centered rolling average through a Gaussian window with standard deviation σ . The specific value of the window size has been chosen to mitigate weekly reporting effects. The values of the standard deviations of the Gaussian windows have been selected to achieve a smooth representation of the expected death estimation for each country as shown in the bottom panels of Supplementary Fig. S1.

Reporting delays. We consider an average delay of two days between reporting a death and its occurrence. This value is obtained by comparing the daily death counts reported for Spain¹ and their actual values³³ from February 15 to March 31, 2020. The values of the root-mean-squared deviation between reported and actual deaths shifted by 0, 1, 2, 3, and 4 days are 77.9, 58.4, 38.5, 58.7, and 88.6 deaths respectively.

Infection fatality rate (IFR). The infection fatality rate is computed assuming homogeneous attack rate as

$$IFR = \frac{1}{\sum_a g_a} \sum_a IFR_a g_a, \quad (39)$$

where IFR_a is the previously estimated IFR for the age group a ⁵ and g_a is the population in the age group a as reported by the United Nations for countries¹⁸ and the US Census for states¹⁹.

Clinical parameters. We obtained the values of the average τ_G and standard deviation σ_G of the generation time from Ref.¹³, of the averages of the incubation τ_I and symptom onset-to-death τ_{OD} times from Refs.^{5,14}, and of the average number of days Δt_{TP} of positive testing by an infected individual from Refs.^{15,17}. The average time at which an individual tested positive after infection τ_{TP} was computed as $\tau_{TP} = \tau_I - 2 + \Delta t_{TP}/2$, where we have assumed that on average an individual started to test positive 2 days before symptom onset. The average seroconversion time after infection τ_{SP} was estimated as τ_I plus the 7 days of 50% seroconversion after symptom onset reported in Ref.¹⁶.

Dynamical constraints implementation with discrete time. We implemented the dynamical constraints to compute the infectious and infected population as outlined in the main text and as detailed in the previous section of this document, using days as time units. Time delays were rounded to days to assign daily values.

The first derivative of the cumulative number of deaths is computed as

$$\frac{dn_D(t)}{dt} = \Delta n_D(t), \quad (40)$$

with $\Delta n_D(t) = n_D(t) - n_D(t - 1)$.

The growth rate was computed explicitly from the discrete time series as the centered 7-day difference

$$k_G(t) = \frac{1}{7} (\ln(\Delta n_D(t + 4) + \Delta n_D(t + 3)) - \ln(\Delta n_D(t - 3) + \Delta n_D(t - 4))). \quad (41)$$

The 7-day value was chosen to mitigate reporting artifacts.

Confidence and credibility intervals. Confidence intervals associated with death counts were computed using bootstrapping with 10,000 realizations³⁴. These confidence intervals were combined with the credibility intervals of the *IFR* in infectious and infected populations assuming independence and additivity on a logarithmic scale.

Fold accuracy. The fold accuracy, F_A , is explicitly computed as

$$\log F_A = \frac{1}{N} \sum_{i=1}^N |\log x_i^{obs} - \log x_i^{est}|, \quad (42)$$

where $|\cdot|$ is the absolute value function, x_i^{obs} is the i^{th} observation, x_i^{est} is its corresponding estimation, and N is the total number of observations.

Inference and extrapolation. Because of the delay between infections and deaths, inference for the values of the growth rate and infectious populations ends on December 30, 2020 and for the values of the infected populations ends on December 26, 2020. Extrapolation to the current time (January 21, 2021) is carried out assuming the last growth rate computed.

Reproduction number. The quantities R_t and $k_G(t)$ are related to each other through the Euler–Lotka equation, $R_t^{-1} = \int_0^\infty f_{GT}(\tau) e^{-k_G(t)\tau} d\tau$, which considers $j(t - \tau) \simeq e^{-k_G(t)\tau} j(t)$ in the renewal equation $j(t) = \int_0^\infty k_I(t, \tau) j(t - \tau) d\tau$. Generation times can generally be described through a gamma distribution $f_{GT}(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau}$ with $\alpha = \tau_G^2 / \sigma_G^2$ and $\beta = \tau_G / \sigma_G^2$, which leads to $R_t = (1 + k_G(t) / \beta)^\alpha$ for $k_G(t) > -\beta$ and $R_t = 0$ for $k_G(t) \leq -\beta$. In the case of the exponentially distributed limit ($\alpha \simeq 1$) or small values of $k_G(t) / \beta$, it simplifies to $R_t = 1 + k_G(t) \tau_G$ for $k_G(t) > -1 / \tau_G$ and $R_t = 0$ for $k_G(t) \leq -1 / \tau_G$. Global prevalence data were obtained from multiple data sources^{35–42}, as described in Supplementary Table S1.

Received: 22 January 2021; Accepted: 14 September 2021

Published online: 07 October 2021

References

1. JHU CSSE COVID-19 Data, <<https://github.com/CSSEGISandData/COVID-19>> (2020).
2. Li, Y. *et al.* The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2: A modelling study across 131 countries. *Lancet. Infect. Dis.* [https://doi.org/10.1016/s1473-3099\(1020\)30785-30784](https://doi.org/10.1016/s1473-3099(1020)30785-30784) (2020).
3. Ma, J. Estimating epidemic exponential growth rate and basic reproduction number. *Infect. Dis. Model.* **5**, 129–141. <https://doi.org/10.1016/j.idm.2019.12.009> (2020).
4. Wu, J. T. *et al.* Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat. Med.* **26**, 506–510. <https://doi.org/10.1038/s41591-020-0822-7> (2020).
5. Verity, R. *et al.* Estimates of the severity of coronavirus disease 2019: A model-based analysis. *Lancet Infect. Dis.* **20**, 669–677. [https://doi.org/10.1016/S1473-3099\(20\)30243-7](https://doi.org/10.1016/S1473-3099(20)30243-7) (2020).
6. Rothe, C. *et al.* Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *N. Engl. J. Med.* **382**, 970–971. <https://doi.org/10.1056/NEJMc2001468> (2020).
7. Böhmer, M. M. *et al.* Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: A case series. *Lancet. Infect. Dis.* **20**, 920–928. [https://doi.org/10.1016/s1473-3099\(20\)30314-5](https://doi.org/10.1016/s1473-3099(20)30314-5) (2020).
8. Li, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493. <https://doi.org/10.1126/science.abb3221> (2020).
9. Flaxman, S. *et al.* Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261. <https://doi.org/10.1038/s41586-020-2405-7> (2020).
10. Grassly, N. C. & Fraser, C. Mathematical models of infectious disease transmission. *Nat. Rev. Microbiol.* **6**, 477–487. <https://doi.org/10.1038/nrmicro1845> (2008).
11. Hethcote, H. W. The mathematics of infectious diseases. *SIAM Rev.* **42**, 599–653 (2000).
12. Fraser, C., Riley, S., Anderson, R. M. & Ferguson, N. M. Factors that make an infectious disease outbreak controllable. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 6146–6151. <https://doi.org/10.1073/pnas.0307506101> (2004).
13. Bi, Q. *et al.* Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: A retrospective cohort study. *Lancet. Infect. Dis.* **20**, 911–919. [https://doi.org/10.1016/s1473-3099\(20\)30287-5](https://doi.org/10.1016/s1473-3099(20)30287-5) (2020).
14. Backer, J. A., Klinkenberg, D. & Wallinga, J. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Euro Surveill.* <https://doi.org/10.2807/1560-7917.ES.2020.25.5.2000062> (2020).
15. Long, Q.-X. *et al.* Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nat. Med.* **26**, 1200–1204. <https://doi.org/10.1038/s41591-020-0965-6> (2020).

16. Wolfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature* **581**, 465–469. <https://doi.org/10.1038/s41586-020-2196-x> (2020).
17. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *Lancet* **395**, 1054–1062. [https://doi.org/10.1016/S0140-6736\(20\)30566-3](https://doi.org/10.1016/S0140-6736(20)30566-3) (2020).
18. *UN World Population Prospects: Total Population - Both Sexes*, <<https://population.un.org/wpp/Download/Standard/Population>> (2020).
19. *US Census: ACS Demographic and Housing Estimates 2018*, <<https://data.census.gov/cedsci>> (2019).
20. Anand, S. *et al.* Prevalence of SARS-CoV-2 antibodies in a large nationwide sample of patients on dialysis in the USA: A cross-sectional study. *Lancet* [https://doi.org/10.1016/S0140-6736\(20\)32009-2](https://doi.org/10.1016/S0140-6736(20)32009-2) (2020).
21. *CDC COVID Data Tracker: Nationwide Commercial Laboratory Seroprevalence Survey*, <<https://covid.cdc.gov/covid-data-tracker/#national-lab>> (2020).
22. Woolf, S. H. *et al.* Excess deaths from COVID-19 and other causes, March–July 2020. *JAMA* **324**, 1562–1564. <https://doi.org/10.1001/jama.2020.19545> (2020).
23. Hale, T. *et al.* *Oxford COVID-19 Government Response Tracker*, Blavatnik School of Government, <www.bsg.ox.ac.uk/covidtracker> (2020).
24. Levin, A. T. *et al.* Assessing the age specificity of infection fatality rates for COVID-19: Systematic review, meta-analysis, and public policy implications. *medRxiv* <https://doi.org/10.1101/2020.07.23.20160895> (2020).
25. Pollan, M. *et al.* Prevalence of SARS-CoV-2 in Spain (ENE-COVID): A nationwide, population-based seroepidemiological study. *Lancet* **396**, 535–544. [https://doi.org/10.1016/S0140-6736\(20\)31483-5](https://doi.org/10.1016/S0140-6736(20)31483-5) (2020).
26. Erikstrup, C. *et al.* Estimation of SARS-CoV-2 infection fatality rate by real-time antibody screening of blood donors. *Clin. Infect. Dis.* <https://doi.org/10.1093/cid/ciaa849> (2020).
27. Ali, S. T. *et al.* Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. *Science* **369**, 1106–1109. <https://doi.org/10.1126/science.abc9004> (2020).
28. Herzog, S. *et al.* Seroprevalence of IgG antibodies against SARS coronavirus 2 in Belgium: A serial prospective cross-sectional nationwide study of residual samples. *MedRxiv* <https://doi.org/10.1101/2020.06.08.20125179> (2020).
29. Walker, A. S. *et al.* Viral load in community SARS-CoV-2 cases varies widely and temporally. *MedRxiv* <https://doi.org/10.1101/2020.10.25.20219048> (2020).
30. Vilar, J. M. G. & Rubi, J. M. Determinants of population responses to environmental fluctuations. *Sci. Rep.* **8**, 887. <https://doi.org/10.1038/s41598-017-18976-6> (2018).
31. Saad-Roy, C. M. *et al.* Immune life history, vaccination, and the dynamics of SARS-CoV-2 over the next 5 years. *Science* **370**, 811–818. <https://doi.org/10.1126/science.abd7343> (2020).
32. Murray, J. D. *Mathematical Biology* 2nd edn. (Springer, 1993).
33. *Ministerio de Sanidad. Fallecidos COVID-19*, <https://www.msbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov-China/documentos/Fallecidos_COVID19.xlsx> (2020).
34. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 2013).
35. Salje, H. *et al.* Estimating the burden of SARS-CoV-2 in France. *Science* **369**, 208–211. <https://doi.org/10.1126/science.abc3517> (2020).
36. Hallal, P. *et al.* Remarkable variability in SARS-CoV-2 antibodies across Brazilian regions: nationwide serological household survey in 27 states. *MedRxiv* <https://doi.org/10.1101/2020.05.30.20117531> (2020).
37. Bogogiannidou, Z. *et al.* Repeated leftover serosurvey of SARS-CoV-2 IgG antibodies, Greece, March and April 2020. *Eurosurveillance* <https://doi.org/10.2807/1560-7917.Es.2020.25.31.2001369> (2020).
38. Gallian, P. *et al.* Lower prevalence of antibodies neutralizing SARS-CoV-2 in group O French blood donors. *Antiviral Res.* **181**, 104880. <https://doi.org/10.1016/j.antiviral.2020.104880> (2020).
39. Menachemi, N. *et al.* Population point prevalence of SARS-CoV-2 infection based on a statewide random sample—Indiana, April 25–29, 2020. *MMWR Morb. Mortal. Wkly Rep.* **69**, 960–964. <https://doi.org/10.15585/mmwr.mm6929e1> (2020).
40. Tunheim, G. *et al.* *Seroprevalence of SARS-CoV-2 in the Norwegian population measured in residual sera collected in April/May 2020 and August 2019*. (Norwegian Institute of Public Health, ISBN (digital): 978–82–8406–109–2, 2020).
41. Sutton, M., Cieslak, P. & Linder, M. Notes from the field: Seroprevalence estimates of SARS-CoV-2 infection in convenience sample—Oregon, May 11–June 15, 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 1100–1101. <https://doi.org/10.15585/mmwr.mm6932a4> (2020).
42. *European Centre for Disease Prevention and Control. Daily update of new reported cases of COVID-19 by country worldwide*, <<https://www.ecdc.europa.eu/sites/default/files/documents/COVID-19-geographic-distribution-worldwide.xlsx>> (2020).

Acknowledgements

J.M.G.V. acknowledges support from Ministerio de Ciencia e Innovación under grant PGC2018-101282-B-I00 (MCI/AEI/FEDER, UE). L.S. acknowledges support from the University of California, Davis.

Author contributions

J.M.G.V. and L.S. designed research, performed research, analyzed data, and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-99273-1>.

Correspondence and requests for materials should be addressed to J.M.G.V. or L.S.

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

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



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

© The Author(s) 2021