medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

COVID-19 prevalence in 161 countries and over time

Stilianos Louca[1,2,*]

[1]*Department of Biology, University of Oregon, Eugene, USA*
[2]*Institute of Ecology and Evolution, University of Oregon, Eugene, USA*
[*]Corresponding author. www.loucalab.com

Abstract

Effectively evaluating, controlling and predicting the course of the COVID-19 pandemic requires knowledge of the true number of infections in the population. This number, however, generally differs substantially from the number of confirmed cases due to a large fraction of asymptomatic infections as well as geographically and temporally variable testing effort and strategies. Here I use age-stratified death count statistics, published age-dependent infection fatality risks and stochastic modeling to estimate the true prevalence and growth of COVID-19 infections among adults (age $\geq$ 20 years) in 161 countries, from early 2020 until November 1, 2020. My predictions are largely consistent with data from multiple previous nationwide seroprevalence surveys. As of November 1, 2020, the nationwide cumulative COVID-19 prevalence (past and current infections relative to the population size) is estimated at 31% (95%-CI 22-50) for Peru, 27% (17–41) for Mexico, 22% (14–34) for Brazil, 12% (7.2-20) for the US, 11% (6.4–18) for the United Kingdom, 8.2% (5.2–15) for France, 7.4% (4.9–13) for Sweden, 4.2% (2.5–6.8) for Canada, 1.8% (1.2–3) for Germany and 0.12% (0.074–0.26) for Japan. These time-resolved estimates expand the possibilities to evaluate the factors influencing the pandemic's progression and to assess vaccination needs around the world. Periodically updated estimates are available at: www.loucalab.com/archive/COVID19prevalence

Keywords: *COVID-19; SARS-CoV-2; infection fatality risk; exponential growth rate*

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Accurate estimates of the true prevalence of COVID-19 in a population are needed for evaluating (and optimizing) disease control policies and testing strategies, determining seasonal effects, predicting future disease spread, assessing the risk of foreign travel and determining vaccination needs [1]. Further, parameter-rich and potentially underdetermined epidemiological models [2, 3] generally benefit from independently obtained estimates of disease prevalence. Due to the existence of a large fraction of asymptomatic cases, as well as variation in reporting, testing effort and testing strategies (e.g., random vs symptom-triggered), confirmed case counts cannot be directly converted to infection counts and a comparison of confirmed case counts between countries is generally of limited informative value [4]. While large-scale seroprevalence surveys (e.g., using antibody tests) can yield information on the disease's true prevalence in a population, such surveys involve substantial financial and logistical challenges and only yield prevalence estimates at a specific time point.

In contrast to case reports, COVID-19-related death counts are generally regarded as less sensitive to testing effort and strategy [5, 6], and fortunately most countries have established nationwide continuous reporting mechanisms for death counts. Hence, in principle, knowing the infection fatality risk (IFR, the probability of death following infection) should permit a conversion of death counts to infection counts [5, 6]. The IFR of COVID-19, however, depends strongly on the patient's age, and hence the effective IFR of the entire population depends on the population's age structure as well as the disease's age distribution [7]. Indeed, it was shown that the age-dependency of the IFR, the age-dependency of COVID-19 prevalence, and the age structure of the population are largely sufficient to explain variation in the effective IFR between countries [8]. This suggests that age-stratified death counts can (and must) be used with age-dependent IFR estimates in order to obtain an accurate estimate of infection counts. This approach has been successfully used to estimate COVID-19 prevalence over time in Europe until May 4, 2020 [6].

Unfortunately, the ongoing pandemic necessitates continuously updated prevalence estimates. Moreover, age-stratified and time-resolved death statistics are not readily available for many countries with insufficiently comprehensive reporting, thus preventing a direct adoption of the above approach [6, 9]. In cases where only total death counts are available (e.g., as disseminated by the World Health Organization) one needs to somehow independently determine the likely age distribution of infections in order to convert total death counts to infection counts. Here I address this challenge by leveraging information on the age distribution of COVID-19 infections from multiple countries with available age-stratified death reports, to estimate the likely age-distribution of COVID-19 in other countries, while accounting for each country's age structure. Based on these calibrations, I estimate the prevalence of COVID-19 (cumulative number of infections, weekly new infections and exponential growth rate) over time in 161 countries up until November 1, 2020, among adults aged 20 years or more. My predictions are largely consistent with data from multiple previously published nationwide seroprevalence surveys.

Calibrating the age distribution of COVID-19 prevalence

In order to calculate infection counts solely from total (i.e., non-age-stratified) death counts, while accounting for the age-dependency of the IFR and each country's population age structure, independent estimates of the ratios of infection risks between age groups (i.e., the risk of infection in any one age group relative to any other age group) are needed. To determine the general distribution of age-specific infection risk ratios, I analyzed weekly age-stratified COVID-19-related death reports from 20 countries around the world using a probabilistic model of Poisson-distributed time-delayed death counts (see Methods for details). Briefly, for any given country c, any given week w, and any given age group g, I assumed that the number of new

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

infections during that week ($I_{c,w,g}$) is approximately equal to $\alpha_{c,g} I_{c,w,r} N_{c,g}/N_{c,r}$, where r represents some fixed reference age group, $N_{c,g}$ is the population size of age group g, and $\alpha_{c,g}$ is the relative risk of an individual in age group g being infected compared to that of an individual in age group r. The expected number of deaths in each age group 4 weeks later (roughly the average time lag between infection and death [10]), denoted $D_{c,w+4,g}$, was assumed to be $I_{c,w,g} R_g$, where R_g is the IFR for that age group. Age-specific IFRs were calculated beforehand by taking the average over multiple IFR estimates reported in the literature [8, 9, 11–14]. This model thus accounts for the age-structure of each country, the age-distribution of the disease in each country and the age-dependency of the IFR. A critical assumption of the model is that, in any given country, nationwide age-specific infection risks co-vary linearly between age groups over time, i.e., an increase of disease prevalence in one age group coincides with a proportional increase of prevalence in any other age group. This assumption is motivated by the observation that nationwide death rates generally covary strongly linearly between age groups (Fig. 1A and Supplemental Fig. S1); the adequacy of this model is also confirmed in retrospect (see below). For each country, I fitted the infection risk ratios $\alpha_{c,g}$ (for all $g \neq r$) as well as the weekly infections in the reference age-group $I_{c,w,r}$ (one per week) to the age-stratified weekly death counts using a maximum-likelihood approach and assuming that weekly death counts follow a Poisson distribution. This stochastic model explained the data generally well, with observed weekly death counts almost always falling within the 95% confidence interval of the model's predictions (Supplemental Fig. S2). This supports the initial assumption that infection risks co-vary approximately linearly between age groups over time and suggests that country-specific but time-independent infection risk ratios are largely sufficient for describing the age-distribution of COVID-19 infections in a country and over time. For any given age group g, the fitted infection risk ratios $\alpha_{c,g}$ differed between countries but were generally within the same order of magnitude (Fig. 1B). On the basis of this observation, and as explained in the next section, it thus seems possible to approximately estimate the number of infections in any other country based on total death counts, the population's age structure and the *ensemble* of infection risk ratios $\alpha_{c,g}$ fitted above.

Estimating infection counts over time

Based on the ensemble of fitted infection risk ratios and total (non-age-stratified) COVID-19-related death count reports disseminated by the WHO, I estimated the true weekly infection counts over time in each of 161 countries (details in Methods). Briefly, for any given country c, week w and any given set of relative infection risks $\alpha_1, \alpha_2, ..$, the total number of deaths 4 weeks later was assumed to be Poisson-distributed with expectation equal to:

$$I_{c,w,r} \sum_g R_g \alpha_g \frac{N_{c,g}}{N_{c,r}}, \tag{1}$$

consistent with the previously described model, where as before R_g is the IFR for age group g, $N_{c,g}$ is the population size of age group g and $I_{c,w,r}$ is the (a priori unknown) number of new infections in the reference age group r during week w. For the sum in Eq. (1), I considered only age groups at 20 years or older (in 5-year intervals), because estimates of the infection risk ratios α_g were unreliable for younger ages (due to low death counts) and because deaths among less-than-20-year olds were numerically negligible compared to the total number of deaths reported. For each week, the unknown $I_{c,w,r}$ was estimated via maximum-likelihood based on the total deaths reported 4 weeks later. The total number of new infections among $\geq$20-year olds during that week was estimated as $I_{c,w} = I_{c,w,r} \sum_g \alpha_g N_{c,g}/N_{c,r}$. Cumulative (i.e., past and current) infection counts were calculated as incremental sums of the weekly infection count estimates. The pandemic's exponential growth rate over time was subsequently calculated from the estimated weekly infection counts based on a Poisson distribution model and using a sliding-window approach.

Depending on the particular choice of infection risk ratios, this yielded different estimates for the weekly

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

nationwide infection counts, the cumulative infection counts and the exponential growth rates over time. Uncertainty in the true infection risk ratios in any particular country was accounted for by randomly sampling from the full distribution of fitted infection risk ratios multiple times, and calculating confidence intervals of the predictions based on the obtained distribution of estimates. Estimated weekly and cumulative infection fractions (i.e., relative to population size) and exponential growth rates over time are shown for a selection of countries in Fig. 2 and Supplemental Figs. S3 and S4. A comprehensive and periodically updated report of estimates for all 161 countries is being made available at www.loucalab.com/archive/COVID19prevalence. Global color-maps of the latest estimates for all countries are shown in Fig. 3.

To assess the accuracy of the above approach, I compared the estimated cumulative infection fractions to previously published nationwide antibody-based seroprevalence surveys across 12 countries (Supplemental Table S2). Only surveys attempting to estimate nationwide seroprevalence in the general population (in particular, either using geographically or demographically stratified sampling or adjusting for sample demographics) were included. Agreement between model estimates and seroprevalence estimates was generally good, with seroprevalence point-estimates for 8 out of 9 countries (all except Brazil) being included in the model's 95%-confidence intervals (Figs. 2K,L,N,O and Supplemental Fig. S3). Apart from potentially erroneous model predictions (discussed below), deviations from seroprevalence-based estimates may also be due to the fact that antibody concentrations in infected individuals (especially asymptomatic ones) can drop over time, rendering many of them seronegative [15–17]. Thus, previously infected individuals may not all be recognized as such. Further, sensitivity and specificity estimates for antibody tests performed in the laboratory or claimed by manufactures need not always apply in a community setting [17], thus introducing biases in seroprevalence estimates despite adjustments for sensitivity and specificity.

Case counts alone can yield wrong impressions

Estimates of the true COVID-19 prevalence in a population can yield insight into the pandemic's growth dynamics that may not have been possible from reported case counts alone. Indeed, according to the present estimates, in most countries case counts initially severely underestimated the number of true infections and often did not properly reflect the progression of the pandemic, although in many countries more recent case reports do capture a much larger fraction of infections and more closely reflect the pandemic's dynamics (Figs. 2A–E and Supplemental Fig. S4). For example, in the US, France, Sweden, Belgium, Spain, United Kingdom and many other European countries reported cases only reflected a small fraction of infections occurring in Spring 2020, while the majority of infections occurring in Summer and Fall 2020 have been successfully detected. Nevertheless, in multiple countries even recent case counts do not correctly reflect the actual dynamics of the pandemic, sometimes even suggesting an opposite trend in its growth. For example, recent case counts in Turkey, Iran, Egypt and Afghanistan severely underestimate the disease's rapid ongoing growth (Supplemental Fig. S5). Future investigations, enabled by the infection rate estimates presented here, might be able to identify the main factors (e.g., political, financial, organizational) driving the discrepancies between infections and detected cases and suggest concrete steps to eliminate them or correct for them.

As infection counts do not depend on testing effort and strategies, they are arguably more suitable for comparing the pandemic's progression between countries. Future investigations, enabled by the estimates presented here, might be able to identify concrete political, environmental and socioeconomic factors influencing the pandemic's growth. For example, my results indicate that as of November 1, 2020 Sweden — often criticized for it's reluctance to impose strong restrictions on it's citizens — was experiencing a slower increase of infections (relative to its population size) than many other European countries such as France, Spain, Italy and the United Kingdom (Supplemental Fig. S4). A similar observation can be made for the US, also frequently pointed out as a particularly severely affected country: As of November 1, 2020 the weekly

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

fraction of newly infected individuals (relative to population size) appears to be substantially lower in the US than in many European countries (Fig. 2F and Supplemental Fig. S4), while the cumulative fraction of infected individuals in the US is comparable to some European countries (e.g., Belgium, United Kingdom, Spain) and much lower than many South American countries (Supplemental Fig. S6). These observations highlight the importance of considering actual infection counts (and of course death counts) relative to population size when evaluating policy differences between countries.

Caveats

The predictions presented here are subject to some important caveats. First, incomplete, erroneous or age-biased reporting of COVID-19-related deaths will have a direct impact on the estimated infection counts. This caveat is particularly important for countries with less developed medical or reporting infrastructure, as well as for countries were reports may be censored or modified for political reasons. Comparisons of results between countries should thus be done with care. Second, the age-specific infection risk ratios ($\alpha_{c,g}$) were calibrated based on available age-stratified death statistics from a limited number of countries, and may not apply to all other countries (for example due to strong cultural differences). Uncertainty associated with this extrapolation is partly accounted for by considering infection risk ratios calibrated to multiple alternative countries (see Methods). Third, age-specific IFRs were obtained from studies in only a few countries (mostly western) and often based on a small subset of closely monitored cases (e.g., from the Diamond Princess cruise ship). These IFR estimates may not be accurate for all countries, especially countries with a very different medical infrastructure, different sex ratios in the population or a different prevalence of pre-existing health conditions (e.g., diabetes), all of which can affect the IFR. That said, estimated trends over time within any given country, in particular exponential growth rates (e.g., Figs. 2P–T), are unlikely to be substantially affected by such biases. To nevertheless examine the robustness of my estimates against variations in the IFR, I repeated the above analyses by considering for each age group an ensemble of IFRs, i.e., randomly sampling from the set of previously reported IFRs [8, 9, 11–14] rather than considering their mean. Median model predictions remained nearly unchanged, however unsurprisingly the uncertainty (i.e., confidence intervals) of the estimates increased (examples in Supplemental Fig. S7).

Conclusion

I have presented estimates of the true nationwide prevalence and growth rate of COVID-19 infections over time in 161 countries around the world, based on official COVID-19-related death reports, age-specific infection fatality risks and each country's population age structure. My estimates are largely consistent with data from nationwide general-population seroprevalence surveys. My findings suggest that while in many countries the detection of infections has greatly improved, there are also examples where even recent reported case counts do not properly reflect the pandemic's dynamics. In particular, comparisons between countries based on infection counts can yield very different conclusions than comparisons merely based on confirmed case counts. My estimates thus enable more precise assessments of the disease's progression, evaluation and improvement of public interventions and testing strategies, and estimation of nationwide vaccination needs.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

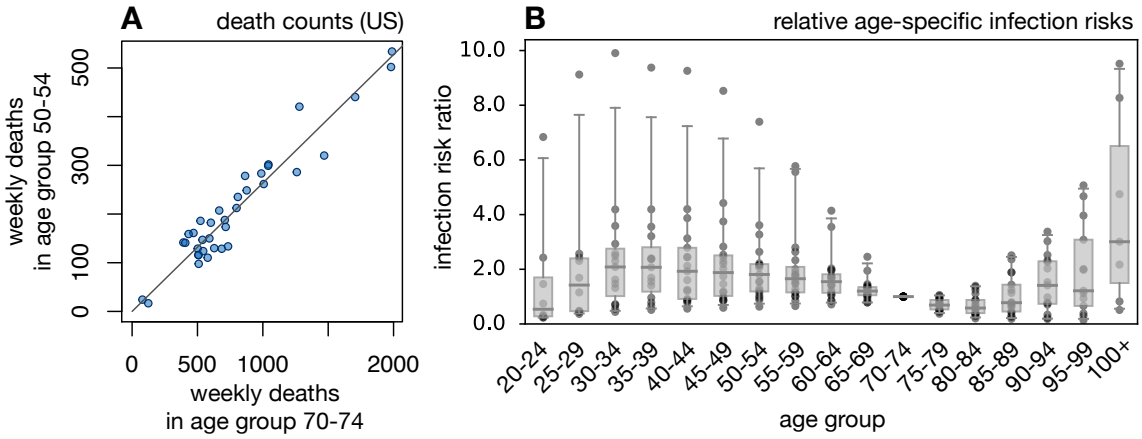

Figure 1: **Infection and death rates covary linearly between age-groups.** (A) Weekly reported COVID-19-related death counts in the US, in age group 70–74 (horizontal axis) and age group 50–54 (vertical axis). Each point corresponds to a different week (defined here as a 7-day period). The linear regression line is shown for reference. For additional age groups and countries see Supplemental Fig. S1. The strong co-linearity of death rates between age-groups suggests that infection risks also covary linearly between age-groups. (B) Relative infection risk ratios (relative to age group 70–74) for different countries, estimated based on death-stratified COVID-19-related death counts. Each column represents a different age group, and in each column each point represents a distinct country. Horizontal bars represent medians and boxes span 50%-percentiles of the data.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

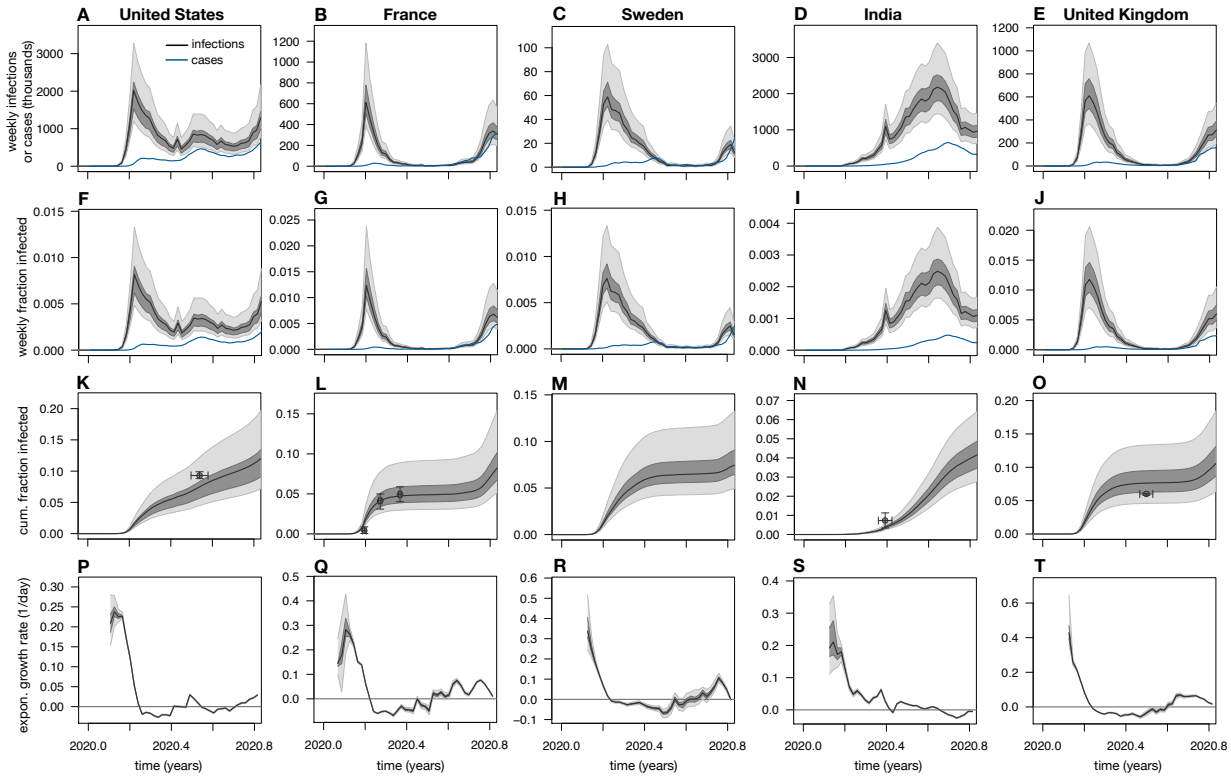

Figure 2: **Estimated nationwide infection rates (adults aged ≥20 years).** (A–E) Estimated nationwide weekly number of infections over time, for various countries (black curves show prediction medians, dark and light shades show 50 % and 95 % percentiles of predictions, respectively), compared to weekly reported cases (blue curves). Note that cases are shown 1 week earlier than actually reported (corresponding roughly to the average incubation time [10]) for easier comparison with infection counts. (F–J) Estimated nationwide weekly fraction of new infections (relative to population size), for the same countries as in A–E. (K–O) Estimated nationwide cumulative fraction of infections (compared to population size), for the same countries as in A–E. Small circles show empirical nationwide prevalence estimates from published seroprevalence surveys for comparison (horizontal error bars denote survey date ranges, vertical error bars denote 95%-confidence intervals as reported by the original publications; references in Supplemental Table S2). (P–T) Estimated exponential growth rate based on weekly infection counts, for the same countries as in A–E. Horizontal lines are shown for reference. Each column shows estimates for a different country. All model estimates refer to adults aged ≥20 years, while reported case counts (blue curves) refer to the entire population. Periodically updated estimates for all 161 countries are available at: www.loucalab.com/archive/COVID19prevalence

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

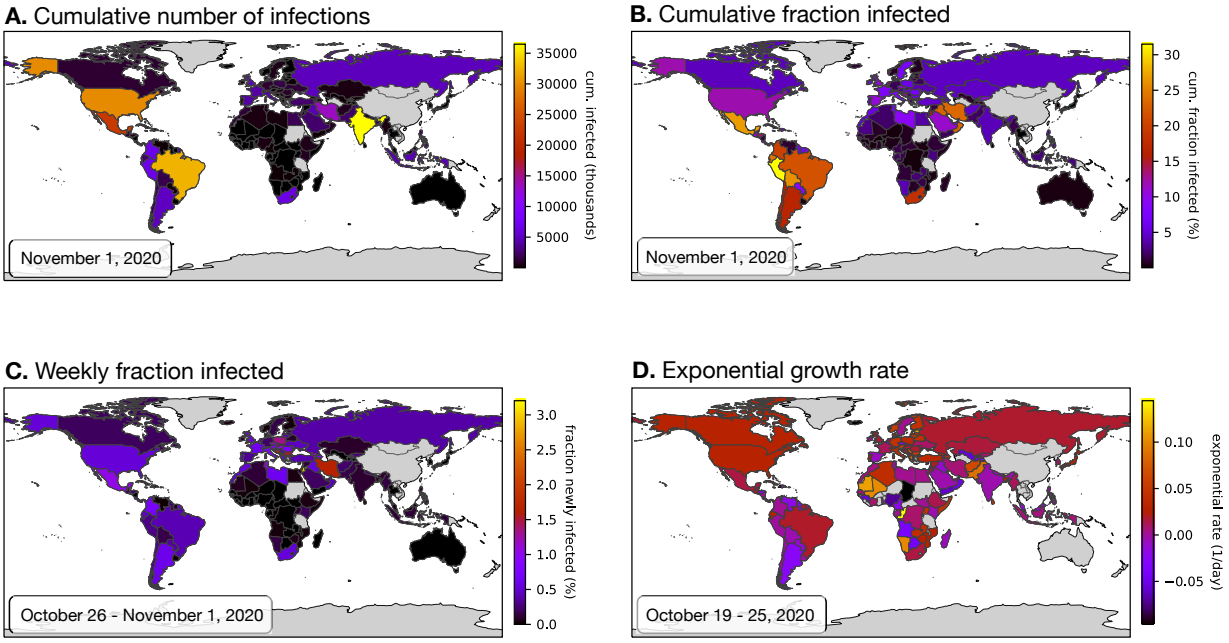

Figure 3: **Worldwide overview of latest estimates (adults aged ≥20 years).** Global map of the latest estimated nationwide (A) cumulative (past and current) number of infections, (B) cumulative fraction of infection (infections relative to population size), (C) weekly fraction of new infections (relative to population size) and (D) current exponential growth rate. Dates of the estimations are given in the lower-right corner of each figure. Countries for which an estimation was not performed (e.g., due to insufficient data) are shown in grey. Note that infection count estimates based on death statistics necessarily lag at least ∼4 weeks behind the current date.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Methods

Age-specific infection fatality risks

Age-specific infection fatality risks (IFRs) were calculated based on the following literature: Table 1 in [12], Supplementary Appendix Q in [8], Table S2 in [13], Table 2 in [11], Table S4 in [9], and Eq. (1) in [14]. For each considered age group, the average IFR across all of the aforementioned published IFRs was used, after linearly interpolating where necessary (Supplemental Table S1).

Calibrating age-specific infection risk ratios

Age-specific population sizes for each country (status 2019) were downloaded from the United Nations website (`https://population.un.org/wpp/Download/Standard/CSV`) on October 23, 2020 [18]. Time series of nationwide cumulative COVID-19-related death counts grouped by 5-year age intervals were downloaded on November 30, 2020 from COVerAGE-DB (`https://osf.io/7tnfh`), which is a database that gathers and curates official death count statistics from multiple official sources [19]. For each country included in COVerAGE-DB, and separately for each age-group, I ensured that cumulative death counts are non-decreasing over time by linearly re-interpolating death counts at problematic time points. The resulting time series were then linearly interpolated onto a regular weekly time grid, i.e., in which adjacent time points are 7 days apart (no extrapolation was performed, i.e., only dates covered by the original time series were included). The weekly number of new deaths in each age group were calculated as the difference of cumulative deaths between consecutive time points on the weekly grid. To ensure a high accuracy in the estimated infection risk ratios, in the following analysis I only considered countries for which COVerAGE-DB covered at least 10 weeks with at least 100 reported deaths each. The following 20 countries were thus considered: Argentina, Bangladesh, Belgium, Brazil, Chile, Colombia, Germany, Ecuador, France, United Kingdom, Indonesia, India, Italy, Mexico, Netherlands, Peru, Philippines, Sweden, Turkey and United States.

For each considered country c, I chose as "reference" age group r the age group that had the highest cumulative number of deaths. For each other age group g, I estimated the infection risk ratio $\alpha_{c,g}$, i.e., the probability of an individual in group g being infected relative to the probability of an individual in group r being infected, using a probabilistic model according to which the number of deaths in group g during week w (denoted $D_{c,w,g}$) was Poisson distributed with expectation:

$$D_{c,w,r} \cdot \alpha_{c,g} \cdot \frac{N_{c,g}}{N_{c,r}} \cdot \frac{R_g}{R_r}. \tag{2}$$

Here, $N_{c,g}$ is the population size of age group g in country c and R_g is the IFR for age group g. Under this model, the maximum-likelihood estimate for $\alpha_{c,g}$, i.e. given the weekly death count time series, is given by:

$$\hat{\alpha}_{c,g} = \frac{\sum_w D_{c,w,g}}{\sum_w D_{c,w,r}} \frac{N_{c,r}}{N_{c,g}} \cdot \frac{R_r}{R_g}. \tag{3}$$

To avoid errors due to sampling noise, only weeks with at least 100 reported deaths were considered in the sums in Eq. (3). I mention that $\alpha_{c,g}$ might also alternatively be estimated as the slope of the linear regression:

$$D_{c,w,g} \sim \alpha_{c,g} D_{c,w,r} \cdot \frac{N_{c,g}}{N_{c,r}} \cdot \frac{R_{c,g}}{R_{c,r}}. \tag{4}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Estimates obtained via linear regression were nearly identical to those obtained using the aforementioned Poissonian model, and were thus not considered further.

For purposes of evaluating the model's adequacy (explained below), I also estimated the weekly number of infections in the reference age group, $I_{c,w,r}$, via maximum-likelihood based on a probabilistic model in which $D_{c,w,g}$ was Poisson-distributed with expectation:

$$\mathbb{E}\{D_{c,w,g}\} = R_g I_{c,w-4,r} \hat{\alpha}_{c,g} \frac{N_{c,g}}{N_{c,r}}. \tag{5}$$

Under this model, the maximum-likelihood estimate for $I_{c,w-4,r}$ is given by:

$$\hat{I}_{c,w-4,r} = \frac{N_{c,r} \sum_g D_{c,w,g}}{\sum_g \hat{\alpha}_{c,g} R_g N_{c,g}}. \tag{6}$$

To evaluate the adequacy of the above stochastic model in explaining the original death count data, I simulated multiple hypothetical weekly death counts for each age group and compared the distribution of simulated death counts to the true death counts. Specifically, for each country c, week w and age group g, I drew 100 random death counts ($\tilde{D}_{c,w,g}$) from a Poisson distribution with expectation:

$$\mathbb{E}\{\tilde{D}_{c,w,g}\} = R_g \hat{I}_{c,w-4,r} \hat{\alpha}_{c,g} \frac{N_{c,g}}{N_{c,r}}. \tag{7}$$

Median simulated death counts and 50 % and 95% confidence intervals, along with the original death counts, are shown for a representative selection of countries and age groups in Supplemental Fig. S3. As can be seen in that figure, the model's simulated time series are largely consistent with the original data.

Estimating infection counts from total death counts

Time series of total (non-age-stratified) nationwide cumulative reported death and case counts were downloaded from the website of the World Health Organization (`https://covid19.who.int/table`) on November 30, 2020. Cumulative death and case counts were made non-decreasing and interpolated onto a weekly time grid as described above. Only countries that reported at least one death per week for at least 10 weeks were included in the analysis below. For each country c, week w and any particular choice of age-specific infection risk ratios $\alpha_1, \alpha_2, ..$, the number of infections was estimated as follows. Let r denote some fixed reference age group with respect to which infection risk ratios are defined, i.e., such that $\alpha_r = 1$ (here, ages 70–74 were used as reference). Let $I_{c,w,r}$ be the (a priori unknown) number of new infections occurring during that week in the reference age group. The number of deaths occurring 4 weeks later in any age group g, $D_{c,w+4,g}$, was assumed to be Poisson-distributed with expectation equal to:

$$\mathbb{E}\left\{D_{c,w+4,g}\right\} = R_g \alpha_g I_{c,w,r} \frac{N_{c,g}}{N_{c,r}}. \tag{8}$$

The total number of weekly deaths, $D_{c,w+4}$, is thus Poisson-distributed with expectation:

$$\mathbb{E}\left\{D_{c,w+4}\right\} = \sum_g R_g \alpha_g I_{c,w,r} \frac{N_{c,g}}{N_{c,r}}. \tag{9}$$

As explained in the main text, only age groups $\geq$20 years were included because infection risk ratios could not be reliably estimated for younger ages and because the contribution of younger ages to total death counts

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

can be considered numerically negligible. Under the above model, the maximum-likelihood estimate for $I_{c,w,r}$ is given by:

$$\hat{I}_{c,w,r} = \frac{D_{c,w+4} N_{c,r}}{\sum_g R_g \alpha_g N_{c,g}}. \tag{10}$$

The total number of weekly infections, $I_{c,w}$, can thus be estimated as:

$$\hat{I}_{c,w} = \hat{I}_{c,w,r} \sum_g \alpha_g \frac{N_{c,g}}{N_{c,r}}. \tag{11}$$

The cumulative number of total infections up until any given week can be estimated by summing the weekly infection counts.

Exponential growth rates over time were estimated from the weekly infection counts using a sliding-window approach, as follows. In every sliding window (spanning 4 consecutive weeks), an exponential function of the form $I(t) = Ae^{t\lambda}$ was fitted, where t denotes time in days and A and λ are unknown parameters (in particular, λ is the exponential growth rate in that window). The parameters A and λ were fitted via maximum likelihood, assuming that the total number of weekly infections, $I_{c,w}$, was Poisson distributed with expectation $Ae^{t_w \lambda}$. Under this model, the log-likelihood of the data (more precisely, of the previously estimated weekly infection counts) is:

$$\ln L = \sum_w \left[\hat{I}_{c,w} \ln A + \hat{I}_{c,w} \lambda t_w - Ae^{\lambda t_w} - \ln(\hat{I}_{c,w}!) \right], \tag{12}$$

where w iterates over all weeks in the specific sliding window. The maximum-likelihood estimates of A and λ are obtained by solving $\partial \ln L / \partial \lambda = 0$ and $\partial \ln L / \partial A = 0$, which quickly leads to the condition:

$$\frac{\sum_w e^{\hat{\lambda} t_w}}{\sum_g t_w e^{\hat{\lambda} t_w}} \cdot \sum_w t_w \hat{I}_{c,w} = \sum_w \hat{I}_{c,w}. \tag{13}$$

Equation (13) was solved numerically to obtain the maximum-likelihood estimate $\hat{\lambda}$.

To assess estimation uncertainties stemming from sampling stochasticity and uncertainties in the infection risk ratios, I repeated the above estimations 100 times using alternative infection risk ratios (for each age group drawn randomly from the set of infection risk ratios previously fitted to various countries) and replacing in Eq. (10) the death counts $D_{c,w+4}$ with values drawn from a Poisson distribution with mean $D_{c,w+4}$. Hence, rather than point-estimates, all predictions are reported in the form of medians and confidence intervals. Only infection risk ratios for which the corresponding linear curve (Eq. 4) achieved a coefficient of determination (R^2) greater than 0.5 were used (shown in Fig. 1), to avoid less accurately estimated infection risk ratios (typically obtained from countries with low death rates). Tables of all estimates for all considered countries up until November 1, 2020 are provided as Supplemental File 1; periodically updated estimates and visual summaries can be found at: www.loucalab.com/archive/COVID19prevalence

Data availability

All data used in this manuscript are publicly available at the locations described in the Methods section.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Competing interests

The author declares no conflict of interest.

Acknowledgements

S.L. was supported by a US National Science Foundation RAPID grant #2028986.

References

[1] Pearce, N., Vandenbroucke, J.P., VanderWeele, T.J. & Greenland, S. Accurate statistics on COVID-19 are essential for policy guidance and decisions. *American Journal of Public Health* **110**, 949–951 (2020).

[2] Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A.J. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences* **110**, 228–233 (2013).

[3] MacPherson, A., Louca, S., McLaughlin, A., Joy, J.B. & Pennell, M.W. A general birth-death-sampling model for epidemiology and macroevolution. *bioRxiv* 2020.10.10.334383 (2020).

[4] Lachmann, A., Jagodnik, K.M., Giorgi, F.M. & Ray, F. Correcting under-reported COVID-19 case numbers: estimating the true scale of the pandemic. *medRxiv* 2020.03.14.20036178 (2020).

[5] Lu, F.S. *et al.* Estimating the cumulative incidence of COVID-19 in the United States using four complementary approaches. *medRxiv* (2020).

[6] Flaxman, S. *et al.* Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).

[7] Dowd, J.B. *et al.* Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proceedings of the National Academy of Sciences* **117**, 9696–9698 (2020).

[8] Levin, A.T. *et al.* Assessing the age specificity of infection fatality rates for COVID-19: Meta-analysis & public policy implications. Working Paper 27597, National Bureau of Economic Research (2020).

[9] O'Driscoll, M. *et al.* Age-specific mortality and immunity patterns of SARS-CoV-2 infection in 45 countries. *medRxiv* 2020.08.24.20180851 (2020).

[10] Linton, N.M. *et al.* Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine* **9** (2020).

[11] Rinaldi, G. & Paradisi, M. An empirical estimate of the infection fatality rate of COVID-19 from the first Italian outbreak. *medRxiv* (2020).

[12] Pastor-Barriuso, R. *et al.* Infection fatality risk for SARS-CoV-2: a nationwide seroepidemiological study in the non-institutionalized population of Spain. *medRxiv* (2020).

[13] Salje, H. *et al.* Estimating the burden of SARS-CoV-2 in France. *Science* **369**, 208–211 (2020).

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

[14] Linden, M. *et al.* The foreshadow of a second wave: An analysis of current COVID-19 fatalities in Germany. *arXiv* (2020).

[15] Long, Q.X. *et al.* Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nature Medicine* **26**, 1200–1204 (2020).

[16] Bolotin, S. *et al.* SARS-CoV-2 seroprevalence survey estimates are affected by anti-nucleocapsid antibody decline. *medRxiv* (2020).

[17] La Marca, A. *et al.* Testing for SARS-CoV-2 (COVID-19): a systematic review and clinical guide to molecular and serological in-vitro diagnostic assays. *Reproductive BioMedicine Online* **41**, 483–499 (2020).

[18] DESA, U. World population prospects 2019, online edition. rev. Tech. Rep., United Nations, Department of Economic and Social Affairs, Population Division (2019).

[19] Riffe, T. *et al.* Coverage-db: A database of age-structured covid-19 cases and deaths. *medRxiv* (2020).

[20] Hallal, P. *et al.* Remarkable variability in sars-cov-2 antibodies across brazilian regions: nationwide serological household survey in 27 states. *medRxiv* (2020).

[21] Hallal, P.C. *et al.* SARS-CoV-2 antibody prevalence in Brazil: results from two successive nationwide serological household surveys. *The Lancet Global Health* **8**, e1390–e1398 (2020).

[22] Vu, S.L. *et al.* Prevalence of SARS-CoV-2 antibodies in France: results from nationwide serological surveillance. *medRxiv* (2020).

[23] Bogogiannidou, Z. *et al.* Repeated leftover serosurvey of SARS-CoV-2 IgG antibodies, Greece, March and April 2020. *Eurosurveillance* **25** (2020).

[24] Merkely, B. *et al.* Novel coronavirus epidemic in the Hungarian population, a cross-sectional nationwide survey to support the exit policy in Hungary. *GeroScience* **42**, 1063–1074 (2020).

[25] Murhekar, M. *et al.* Prevalence of SARS-CoV-2 infection in India: Findings from the national serosurvey, May-June 2020. *Indian Journal of Medical Research* **152**, 48–60 (2020).

[26] Snoeck, C.J. *et al.* Prevalence of SARS-CoV-2 infection in the Luxembourgish population: the CONVINCE study. *medRxiv* (2020).

[27] Slot, E. *et al.* Herd immunity is not a realistic exit strategy during a COVID-19 outbreak. *in review* (2020).

[28] Pollán, M. *et al.* Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. *The Lancet* **396**, 535–544 (2020).

[29] Ward, H. *et al.* Antibody prevalence for SARS-CoV-2 in England following first peak of the pandemic: REACT2 study in 100,000 adults. *medRxiv* (2020).

[30] 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. *The Lancet* **396**, 1335–1344 (2020).

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Code availability

All software used in this paper have been described in the Methods and are freely available online.

Data availability

All data are available as supplementary material and on public repositories described in the Methods.

Acknowledgements

S.L. was supported by National Science Foundation RAPID grant 2028986.

Competing interests

The authors declare that they have no competing interests.

Materials & Correspondence

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

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Information

Table S1: Age-specific infection fatality risks used in this study, obtained by averaging IFRs reported in multiple previous studies (see Methods for details).

age group (years)	IFR (%)
0–4	0.0015
5–9	0.00257
10–14	0.0102
15–19	0.0102
20–24	0.0115
25–29	0.0143
30–34	0.0221
35–39	0.0384
40–44	0.0641
45–49	0.118
50–54	0.209
55–59	0.403
60–64	0.739
65–69	1.41
70–74	2.56
75–79	5.55
80–84	9.47
85–89	10.77
$\geq$90	11.99

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table S2: Previously published nationwide seroprevalence estimates, considered in this study for comparison (Supplemental Fig. S3).

country	start date	end date	seroprevalence (95% CI)	reference
Brazil	2020-05-01	2020-05-15	0.014 (0.013–0.016)	[20]
Brazil	2020-05-14	2020-05-21	0.019 (0.017–0.022)	[21]
Brazil	2020-06-04	2020-06-07	0.031 (0.028–0.034)	[21]
France	2020-03-09	2020-03-15	0.0041 (0.0005–0.0088)	[22]
France	2020-04-06	2020-04-12	0.0414 (0.031–0.0499)	[22]
France	2020-05-11	2020-05-17	0.0493 (0.0402–0.0589)	[22]
Greece	2020-03-01	2020-03-31	0.0002 (0.00–0.00025)	[23]
Greece	2020-04-01	2020-04-30	0.0025 (0.0002–0.005)	[23]
Hungary	2020-05-01	2020-05-16	0.0068 (0.005–0.0086)	[24]
India	2020-05-11	2020-06-04	0.0073 (0.0034–0.0113)	[25]
Luxemburg	2020-04-16	2020-05-05	0.020586 (0.0134–0.0277)	[26]
Netherlands	2020-04-01	2020-04-15	0.027	[27]
Spain	2020-04-27	2020-05-11	0.046 (0.043–0.050)	[28]
United Kingdom	2020-06-20	2020-07-13	0.060 (0.058–0.061)	[29]
United States	2020-07-01	2020-07-31	0.093 (0.088–0.099)	[30]

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

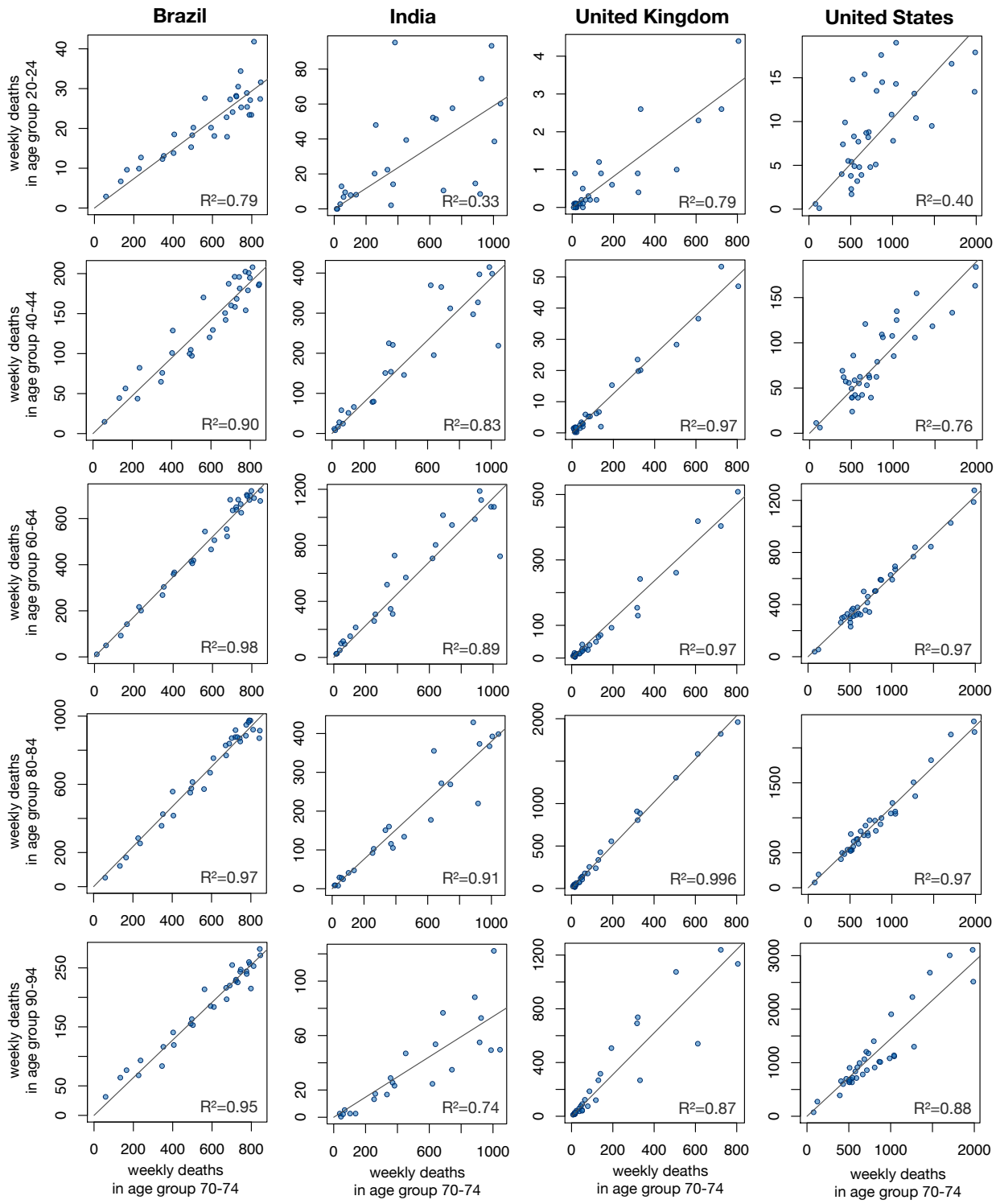

Figure S1: **Comparison of weekly death counts between age groups.** Weekly COVID-19-related death counts per age group (vertical axes) compared to death counts in the same week in age group 70–74 (horizontal axes), in various countries with particularly high reported death rates. Each column corresponds to a different country, each row to a different age group, and each point to a specific week. Regression lines are shown for reference; the corresponding fractions of explained variance (R^2) are shown in the figures.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

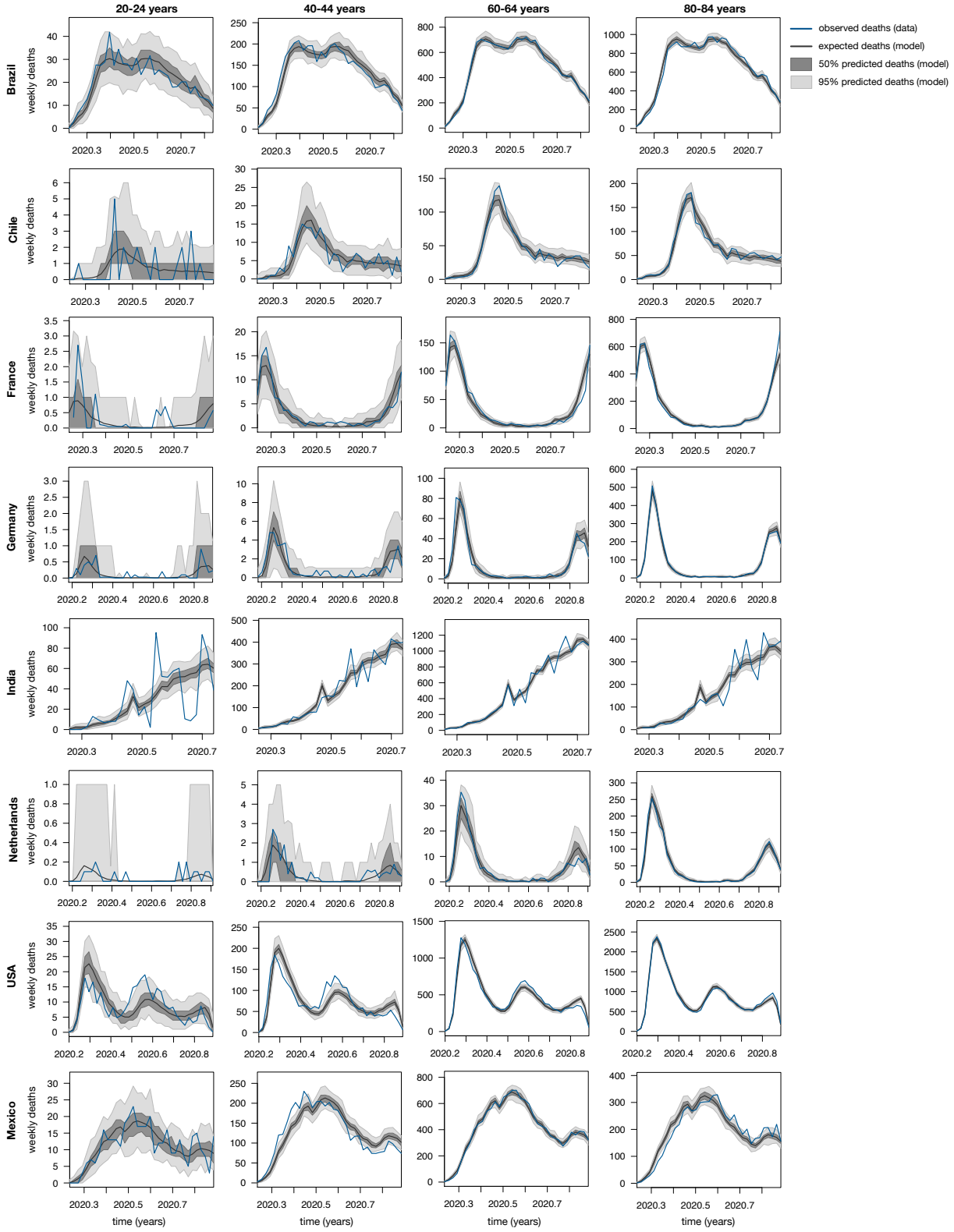

Figure S2: Reported weekly death counts for various countries and selected age groups (blue curves, source: COVerAGE database), compared to death counts predicted by the stochastic model fit to age-stratified death counts (black curves are expectations, dark shades denote 50% and light shades denote 95% confidence intervals). Each sub-figure shows a distinct age group for a distinct country.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

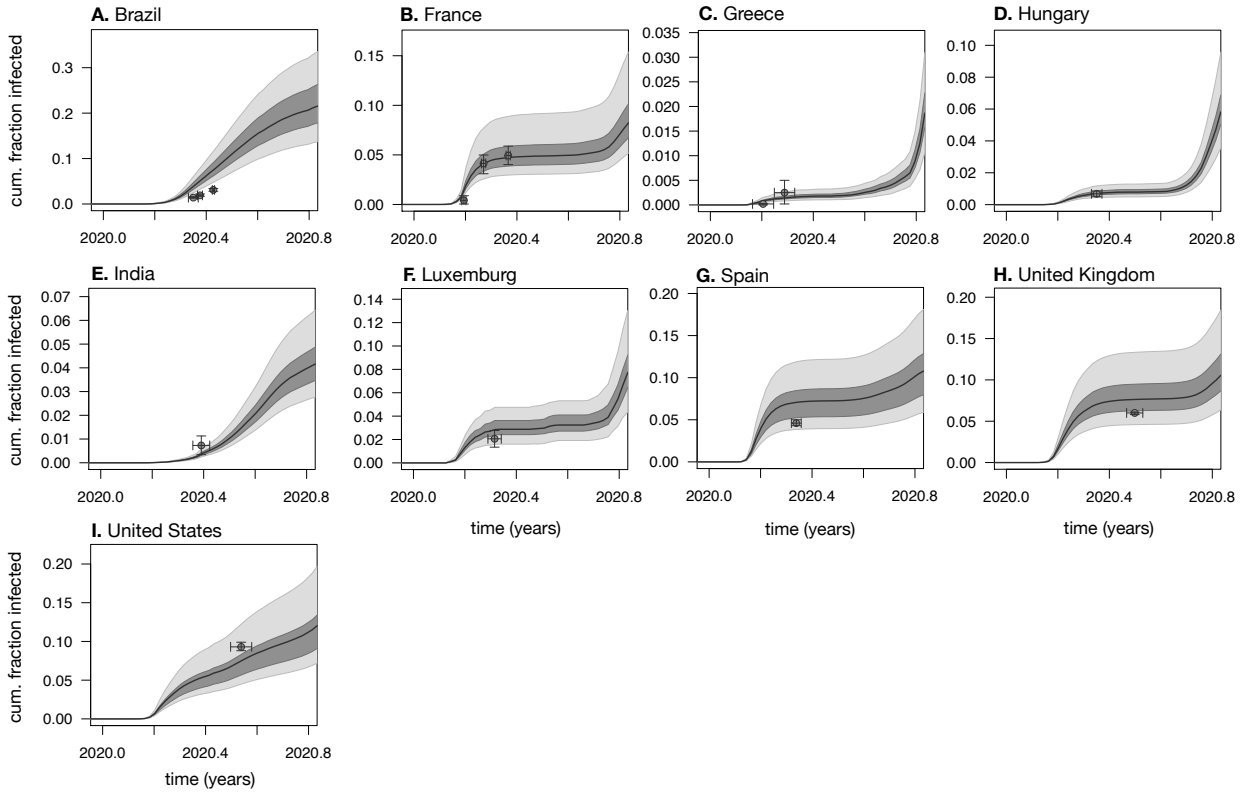

Figure S3: **Comparison of predicted infection fractions to seroprevalence data.** Predicted cumulative infection fractions relative to population size (aka. prevalence) in various countries (among adults aged $\geq$20 years). Black curves show prediction medians, dark and light shades show 50 % and 95 % confidence intervals, respectively. Small circles show empirical nationwide prevalence estimates from published seroprevalence surveys for comparison (horizontal error bars denote survey date ranges, vertical error bars denote 95%-confidence intervals as reported by the original publications). Seroprevalence data sources are listed in Supplemental Table S2.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

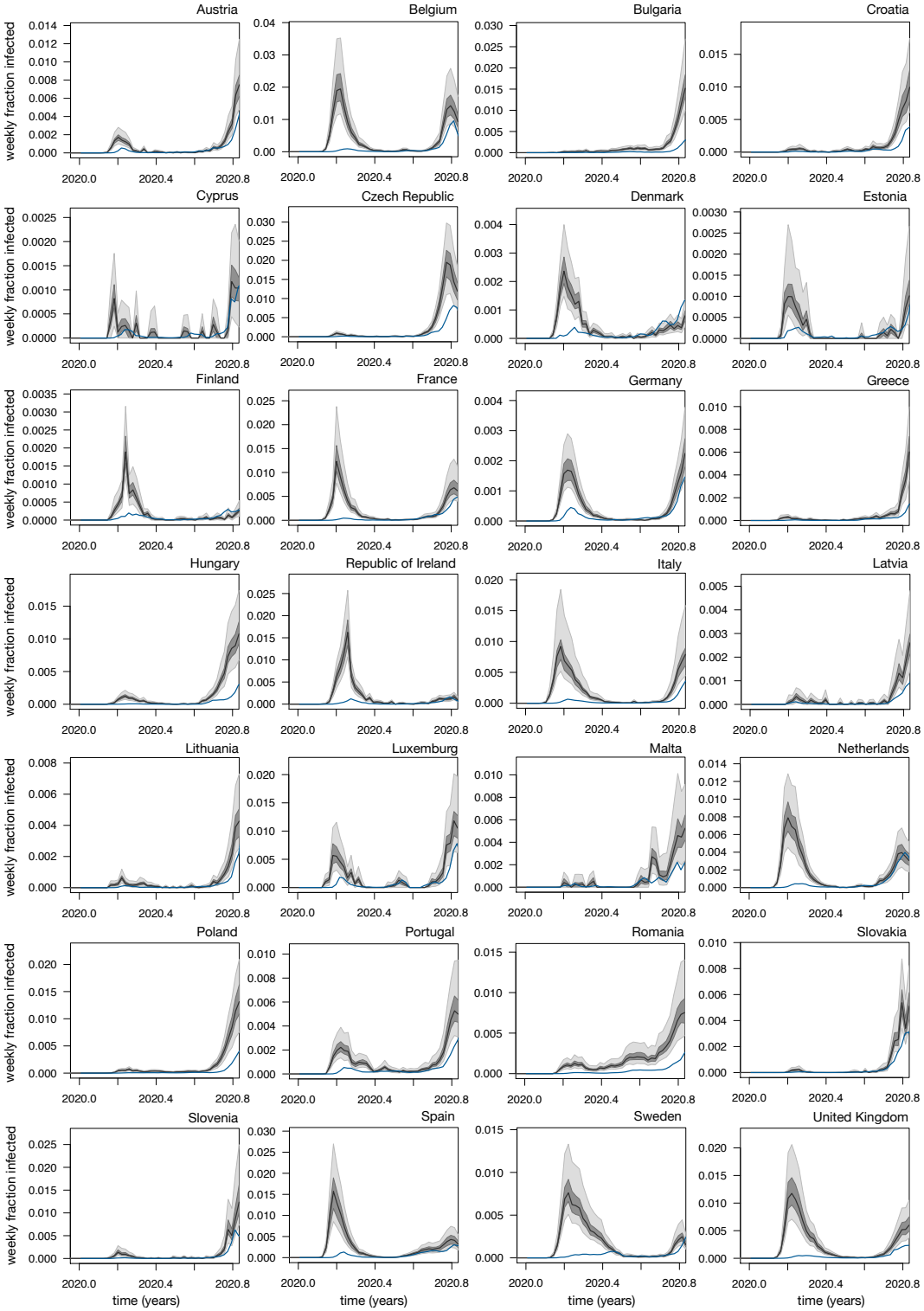

Figure S4: **Weekly infection fractions in Europe.** Estimated weekly nationwide fraction of new infections (relative to population size) over time, in member countries of the European Union as well as the United Kingdom, among adults aged $\geq$20 years. Black curves show prediction medians, dark and bright shades show 50 % and 95 % confidence intervals, respectively. Blue curves show weekly reported case fractions (all ages). Note that cases are shown 1 week earlier than actually reported (corresponding roughly to the average incubation time [10]) for easier comparison with infection counts.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

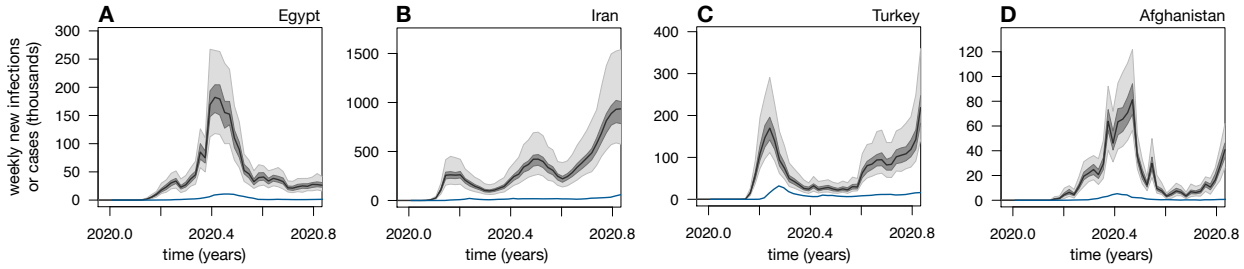

Figure S5: **Estimated infection counts contradict case counts.** Nationwide predicted weekly number of new infections (black curves and shades, among adults aged ≥20 years) and weekly reported cases (blue curves, all ages) over time in Egypt, Iran, Turkey and Afghanistan. Black curves show prediction medians, dark and bright shades show 50 % and 95 % confidence intervals, respectively.

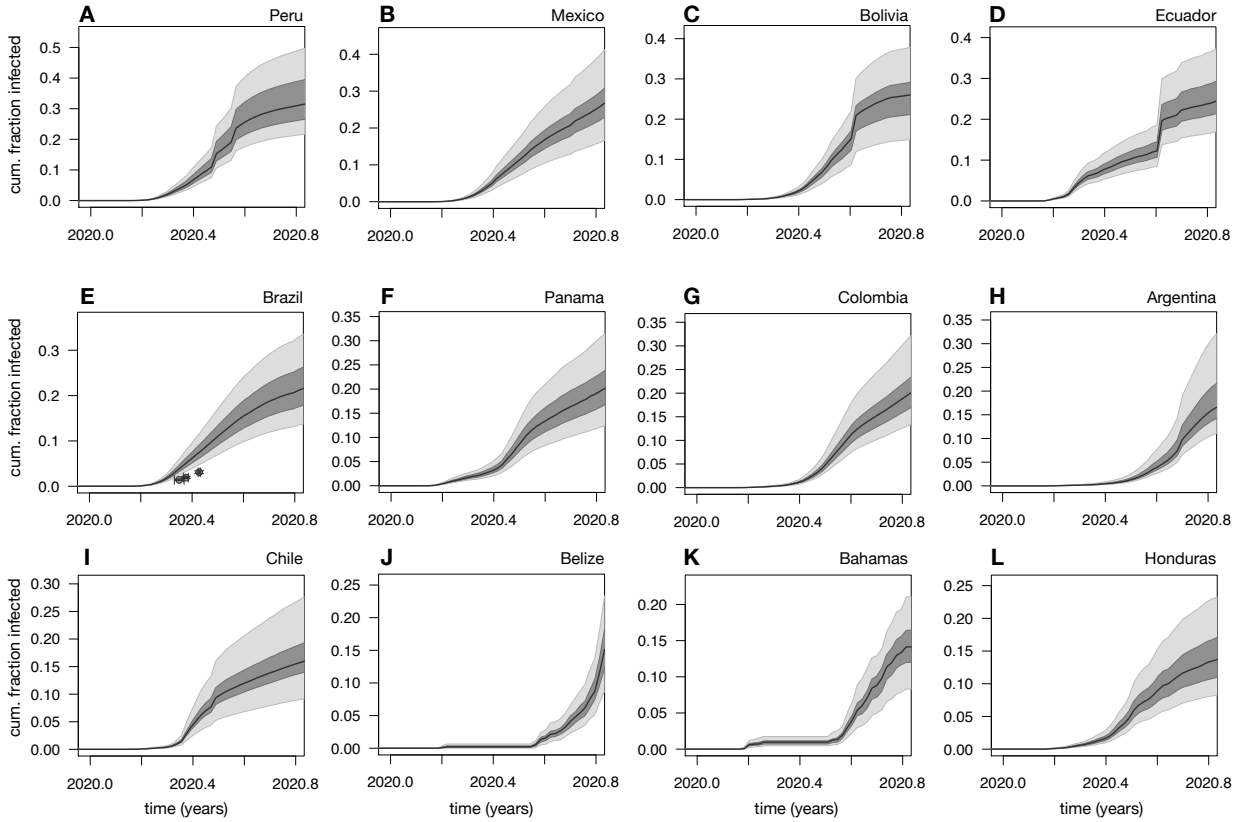

Figure S6: **Cumulative infection fractions in the Americas.** Estimated nationwide cumulative fraction of infections (relative to population size) over time, among adults aged ≥20 years, in countries of the Americas with particularly high estimates of infections. Black curves show prediction medians, dark and bright shades show 50 % and 95 % confidence intervals, respectively.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.01.20241539; this version posted December 2, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

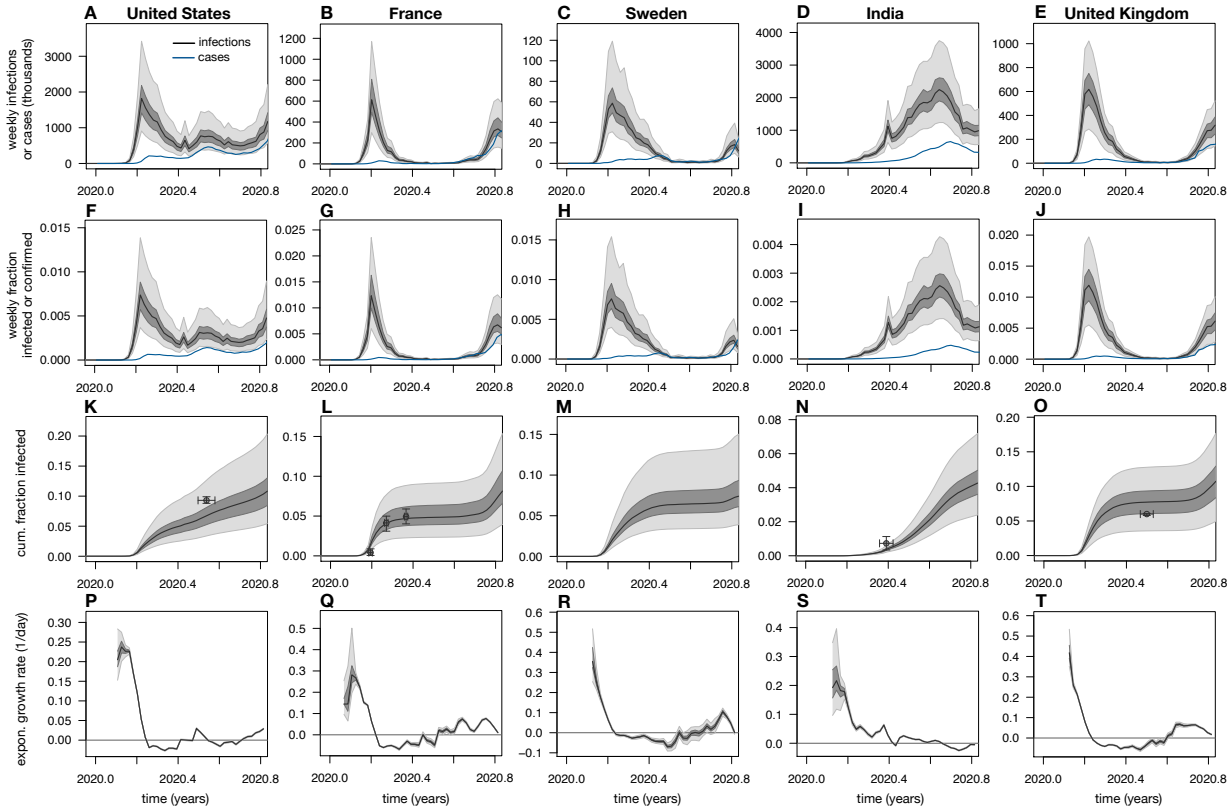

Figure S7: **Estimated nationwide infection counts (using the IFR ensemble).** Estimated nationwide weekly infection counts, infection fractions (relative to population), cumulative fractions infected and exponential growth rates, among adults aged $\geq$20 years.