Estimating global, regional, and national daily and cumulative infections with SARS-CoV-2 through Nov 14, 2021: a statistical analysis

*COVID-19 Cumulative Infection Collaborators**

Lancet 2022; 399: 2351–80

Published Online
April 8, 2022
https://doi.org/10.1016/S0140-6736(22)00484-6

See Comment page 2326

*Collaborators listed at the end of the Article

Correspondence to:
Ryan M Barber, Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA 98195, USA
rmbarber@uw.edu

Summary

Background Timely, accurate, and comprehensive estimates of SARS-CoV-2 daily infection rates, cumulative infections, the proportion of the population that has been infected at least once, and the effective reproductive number ($R_{effective}$) are essential for understanding the determinants of past infection, current transmission patterns, and a population's susceptibility to future infection with the same variant. Although several studies have estimated cumulative SARS-CoV-2 infections in select locations at specific points in time, all of these analyses have relied on biased data inputs that were not adequately corrected for. In this study, we aimed to provide a novel approach to estimating past SARS-CoV-2 daily infections, cumulative infections, and the proportion of the population infected, for 190 countries and territories from the start of the pandemic to Nov 14, 2021. This approach combines data from reported cases, reported deaths, excess deaths attributable to COVID-19, hospitalisations, and seroprevalence surveys to produce more robust estimates that minimise constituent biases.

Methods We produced a comprehensive set of global and location-specific estimates of daily and cumulative SARS-CoV-2 infections through Nov 14, 2021, using data largely from Johns Hopkins University (Baltimore, MD, USA) and national databases for reported cases, hospital admissions, and reported deaths, as well as seroprevalence surveys identified through previous reviews, SeroTracker, and governmental organisations. We corrected these data for known biases such as lags in reporting, accounted for under-reporting of deaths by use of a statistical model of the proportion of excess mortality attributable to SARS-CoV-2, and adjusted seroprevalence surveys for waning antibody sensitivity, vaccinations, and reinfection from SARS-CoV-2 escape variants. We then created an empirical database of infection–detection ratios (IDRs), infection–hospitalisation ratios (IHRs), and infection–fatality ratios (IFRs). To estimate a complete time series for each location, we developed statistical models to predict the IDR, IHR, and IFR by location and day, testing a set of predictors justified through published systematic reviews. Next, we combined three series of estimates of daily infections (cases divided by IDR, hospitalisations divided by IHR, and deaths divided by IFR), into a more robust estimate of daily infections. We then used daily infections to estimate cumulative infections and the cumulative proportion of the population with one or more infections, and we then calculated posterior estimates of cumulative IDR, IHR, and IFR using cumulative infections and the corrected data on reported cases, hospitalisations, and deaths. Finally, we converted daily infections into a historical time series of $R_{effective}$ by location and day based on assumptions of duration from infection to infectiousness and time an individual spent being infectious. For each of these quantities, we estimated a distribution based on an ensemble framework that captured uncertainty in data sources, model design, and parameter assumptions.

Findings Global daily SARS-CoV-2 infections fluctuated between 3 million and 17 million new infections per day between April, 2020, and October, 2021, peaking in mid-April, 2021, primarily as a result of surges in India. Between the start of the pandemic and Nov 14, 2021, there were an estimated 3·80 billion (95% uncertainty interval 3·44–4·08) total SARS-CoV-2 infections and reinfections combined, and an estimated 3·39 billion (3·08–3·63) individuals, or 43·9% (39·9–46·9) of the global population, had been infected one or more times. 1·34 billion (1·20–1·49) of these infections occurred in south Asia, the highest among the seven super-regions, although the sub-Saharan Africa super-region had the highest infection rate (79·3 per 100 population [69·0–86·4]). The high-income super-region had the fewest infections (239 million [226–252]), and southeast Asia, east Asia, and Oceania had the lowest infection rate (13·0 per 100 population [8·4–17·7]). The cumulative proportion of the population ever infected varied greatly between countries and territories, with rates higher than 70% in 40 countries and lower than 20% in 39 countries. There was no discernible relationship between $R_{effective}$ and total immunity, and even at total immunity levels of 80%, we observed no indication of an abrupt drop in $R_{effective}$, indicating that there is not a clear herd immunity threshold observed in the data.

Interpretation COVID-19 has already had a staggering impact on the world up to the beginning of the omicron (B.1.1.529) wave, with over 40% of the global population infected at least once by Nov 14, 2021. The vast differences in cumulative proportion of the population infected across locations could help policy makers identify the transmission-prevention strategies that have been most effective, as well as the populations at greatest risk for future infection.

This information might also be useful for targeted transmission-prevention interventions, including vaccine prioritisation. Our statistical approach to estimating SARS-CoV-2 infection allows estimates to be updated and disseminated rapidly on the basis of newly available data, which has and will be crucially important for timely COVID-19 research, science, and policy responses.

Funding Bill & Melinda Gates Foundation, J Stanton, T Gillespie, and J and E Nordstrom.

Introduction

Measuring SARS-CoV-2's daily infection rate, cumulative infections, and the proportion of the population with one or more infections is essential for understanding the determinants of past transmission, identifying ongoing inequities, predicting future trajectories of the COVID-19 pandemic, and, in theory, prioritising vaccination allocations. Daily infections are also the crucial input into measuring the changing effective reproductive number ($R_{effective}$, the number of subsequent infections caused by a new infection).[1–3] A robust assessment of $R_{effective}$ by day in each location is useful to help evaluate the effect of the wide range of non-pharmaceutical interventions that have been deployed during the pandemic. The $R_{effective}$ over time is also a crucial input into future forecasts of COVID-19.[4] Cumulative infections can help us identify

Research in context

Evidence before this study

This study was conceptualised and developed from the start of the pandemic to fill a void in the provision of timely estimates of SARS-CoV-2 infections for tracking the pandemic and to provide inputs to epidemiological models of transmission. Several research groups have estimated SARS-CoV-2 daily or cumulative infections in select locations at specific points in time. For example, the US Centers for Disease Control and Prevention estimates cumulative infections by approximating the infection–detection ratio (IDR) using assumptions about the portion of the population who will seek care. The Serotracker project reports on the universe of seroprevalence surveys and some attributes of these surveys, but it does not make estimates of cumulative infections based on these data. Noh and Danuser (2021) used reported deaths and published estimates of the infection–fatality ratio (IFR) to estimate cumulative infections for US states and select countries. To our knowledge, however, no source has provided estimates, either periodic or regularly updated, of global daily and cumulative SARS-CoV-2 infections at this resolution (399 administrative units).

Added value of this study

This study is the first comprehensive analysis of global daily and cumulative SARS-CoV-2 infections to date and improves upon previous infection estimation strategies in several important ways. First, we combined three approaches that have been used to estimate daily infections: cases divided by the IDR, hospitalisations divided by the infection–hospitalisation ratio (IHR), and deaths divided by the IFR. Combining these estimates gave us a more robust estimate of daily infections that was less susceptible to biases within and between each type of measure. Second, estimates of total COVID-19 deaths derived from a comprehensive assessment of excess mortality and a statistical estimate of the portion of excess mortality directly due to COVID-19 allowed for more meaningful interpretation of spatial heterogeneity in total COVID-19 mortality rates. Third, we used a systematic analysis of available seroprevalence data matched in space and time to cases, hospitalisations, and deaths to empirically estimate the IDR, IHR, and IFR. Because the IHR and IFR are profoundly age related, we also estimated age-standardised ratios for these quantities. Fourth, for locations without seroprevalence surveys, we used statistical models based on the available empirical data and the testing of a wide range of covariates to predict the IDR, IHR, and IFR. Fifth, we used daily infections to estimate cumulative infections and, with assumptions on cross-variant immunity, the cumulative number of individuals with one or more infections, as well as posterior estimates of cumulative IDR, IHR, and IFR. Sixth, we incorporated corrections to the primary data into the analysis to deal with known biases such as waning antibody test sensitivity. Seventh, our ensemble model reflects the uncertainty of the data sources, model design, and parameter assumptions included in the analysis. Finally, the methods developed to triangulate on daily infections, cumulative infections, and the proportion of the population infected once or more than once have been developed into easily applied statistical code, so estimates can be shared and updated rapidly and iteratively on the basis of the frequency of newly reported data.

Implications of all the available evidence

SARS-CoV-2 has been extremely widespread, causing 3·80 billion (95% uncertainty interval 3·44–4·08) infections and reinfections as of Nov 14, 2021, infecting 43·9% (39·9–46·9) of the world's population. The proportion of the population infected has varied greatly across countries, suggesting that host immunity characteristics and national and local policies play a crucial role in determining patterns of transmission. Our comprehensive modelling approach provides a database of daily infections and effective reproductive number by location from the beginning of the pandemic to Nov 14, 2021, which can be used to develop insights into the determinants of transmission, identify ongoing inequities, establish standards for vaccine prioritisation, and more.

which nations and communities have been able to keep transmission at lower levels, potentially creating the opportunity to learn from these success stories. Finally, a sound measurement of the proportion of the population ever infected could help to identify which communities are at greater risk of future transmission and might be a factor that should be considered in vaccine prioritisation.[5]

Several studies have estimated cumulative infections in select countries at specific points in time.[6–9] Some of these studies have used seroprevalence surveys, while others have made estimates of infections by assuming a particular infection–detection ratio (IDR).[7,10–12] One study estimated infections in the USA and other select countries,[13] and other studies have done multinational systematic reviews and meta-analyses of seroprevalence surveys.[14,15] The fundamental problem in all of these analyses is that each of the data series observed has potential biases: reported cases capture only a portion of infections, and this portion will be a function of the availability of testing; reported deaths capture only a subset of total COVID-19 deaths, and the infection–fatality ratio (IFR) can vary widely over time and across locations;[16–19] the proportion of patients with an infection who are admitted to hospital can also vary over time and location; and seroprevalence surveys can be influenced by sampling design, waning of sensitivity of antibody tests, and vaccination rates. Few studies have combined data from reported cases, reported deaths, hospitalisations, and seroprevalence surveys to triangulate daily infections, and WHO only routinely reports confirmed cases, not estimated infections.[20] The use of such sources of incomplete, biased, and heterogeneous case data uncritically in research, science, and policy will result in inferences confounded to unknown levels by these known problems.

In this study, we present an approach to estimating past SARS-CoV-2 daily infections, cumulative infections through Nov 14, 2021, and the proportion of the population with one or more infections on the basis of reported cases, total deaths attributable to COVID-19, hospitalisations, and seroprevalence surveys. This approach attempts to deal with the biases in each of these measures and use them all to triangulate daily infections. With this statistical approach to the fusion of these data streams, we aimed to provide a method that can be applied on a rapid and ongoing basis, so that these estimates remain maximally relevant for research, science, and policy and can be immediately and freely available. Importantly, we incorporated various sources of uncertainty in daily infections into the analysis to help informed assessment of the variation in space and time of the fidelity of the estimates.

Methods

Overview

We derived comprehensive global estimates of daily and cumulative SARS-CoV-2 infections for the duration of the COVID-19 pandemic, using the heterogeneous universe of reported epidemiological data (iteratively curated, corrected, and calibrated into an internally complete and consistent time series at national and subnational levels) to further timely research, discovery, and policy inference. Our approach can be divided into seven steps, which are applied by use of an ensemble model framework. First, we developed a dataset of reported COVID-19 cases, total COVID-19 deaths, and hospitalisations (where available), corrected for known biases such as lags in reporting. Second, we identified representative SARS-CoV-2 seroprevalence surveys that could be used to create a database of cumulative infections and adjusted them for waning antibody sensitivity, vaccinations, and reinfection from escape variants. Third, using adjusted seroprevalence survey data matched to cases, hospitalisations, and deaths, we created an empirical database of IDRs, infection–hospitalisation ratios (IHRs), and IFRs. Fourth, for locations without seroprevalence surveys and to estimate a complete time series for each location, we developed statistical models to predict the IDR, IHR, and IFR by location and day, as a function of a wide range of covariates. Fifth, three series of estimates of daily infections (cases divided by IDR, hospitalisations divided by IHR, and deaths divided by IFR) were combined into a more robust estimate of daily infections. Sixth, we used the combined time series of daily infections to estimate cumulative infections and the cumulative proportion of the population with one or more infections, and calculate posterior estimates of cumulative IDR, IHR, and IFR. Seventh, we converted daily infections into a historical time series of $R_{\text{effective}}$ by location and day, on the basis of assumptions of duration of the period from infection to infectiousness and time an individual spent being infectious. Estimates are given for all ages and both sexes combined for 190 countries and territories, and for subnational locations in ten of those countries, aggregated into 21 regions, seven super-regions,[21] and globally, from the start of the COVID-19 pandemic through Nov 14, 2021.

This study complies with the Guidelines for Accurate and Transparent Health Estimates Reporting recommendations (appendix 1, section 2).[22] All code used in the analysis can be found online.

See **Online** for appendix 1

For the **analysis code** see https://github.com/ihmeuw/covid-historical-model and https://github.com/ihmeuw/covid-model-infections

Ensemble framework

Our model system includes many component parts that are inherently uncertain, ranging from input data sources and parameter assumptions to model specification. To account for this, we developed an ensemble framework wherein we varied the data and model settings across 100 iterations of the analysis, which were then run independently to yield 100 estimates of infections. These sources of uncertainty include seroprevalence survey error; bootstrapped samples of our seroprevalence database; estimates of seroreversion rates; estimates of total COVID-19 mortality; parameterisation of cross-variant immunity, increased risk of hospitalisation and death from non-ancestral SARS-CoV-2 variants, and durations associated with COVID-19 natural history;

covariate selection and specification of statistical models of the IDR, IHR, and IFR; and triangulation of infections on the basis of cases, hospitalisations, and deaths (more details regarding the ensemble framework in appendix 1, section 9).

Data inputs and corrections

Data of reported cases were obtained largely from Johns Hopkins University (Baltimore, MD, USA),[23] with exceptions and additions noted in appendix 1 (section 4.1) and appendix 2 (section 4). Hospital admissions were largely sourced from national databases such as that of the Department of Health and Human Services (HHS) in the USA and the *Secretaria de Vigilância em Saúde* in Brazil (for an exhaustive list see appendix 2, section 1). Deaths were based on reported deaths data from Johns Hopkins University[23] and various national sources from locations where data inconsistencies were evident in the Johns Hopkins University datasets (more details in appendix 1, section 4.3, and appendix 2, section 2). To account for the prevalent issue of under-reporting in COVID-19 deaths, we applied a scalar of reported to total COVID-19 deaths in our analysis. Total COVID-19 deaths, as defined by WHO, are all deaths where the deceased individuals were actively infected with SARS-CoV-2 at the time of the death. Estimates of total COVID-19 mortality were constructed with use of the statistical model developed by the COVID-19 Excess Mortality Collaborators to predict the excess mortality rate for all locations between Jan 1, 2020, and Nov 14, 2021.[16] To estimate total COVID-19 mortality, we predicted a counterfactual excess mortality rate due to COVID-19 in which the IDR was set to the maximum observed values among all locations. The predicted excess mortality rate from this counterfactual analysis, corrected for under-reporting, resulted from insufficient testing and changes in mortality driven by behaviours such as deferred health care during periods of lockdown. We used the ratio of this counterfactual excess mortality rate and the prediction for the same period as a proxy for the proportion of excess mortality that is total COVID-19 mortality. Subsequently, a scalar of reported COVID-19 deaths to total COVID-19 deaths can be derived (more details in appendix 1, section 9.4). We identified seroprevalence surveys through a search protocol that leveraged previous reviews,[24,25] SeroTracker,[26] and routine inclusion of national and subnational surveys undertaken by governmental organisations. Studies that focused on specific subsets of the population—either a specific subpopulation such as health-care workers or specific locations such as specific cities—were typically excluded as a result of not being representative. In total, we identified 2817 seroprevalence survey datapoints (of 6420 reviewed) for inclusion in this analysis.

See Online for appendix 2

Although most data streams for daily cases, deaths, and hospitalisations are indexed by date of report, some are indexed by date of event; in these instances, lags in reporting create misleading trends in the most recent days of data. These trends are gradually corrected over time as reporting systems catch up but, to prevent this occurrence from influencing our models, we needed to evaluate each individual data source and determine an appropriate number of days to exclude in any iteration of the analyses.

Some hospital admissions data series only became available starting from weeks or months after the beginning of the COVID-19 pandemic—for example, the HHS database began in July, 2020. However, total cumulative hospitalisations are required to create our empirical estimate of IHR. In these instances, we leveraged information from the metrics that did have complete time coverage (cases and deaths) to impute the earlier portion of the admissions time series (appendix 1, section 4.2).

Seroprevalence survey adjustments

Seroprevalence surveys were corrected for vaccination, because vaccination generates a positive anti-spike antibody test in most individuals who receive the vaccine.[27] In locations where vaccination rates have increased over time, population levels of anti-spike antibodies will be elevated. To correct for this, we adjusted seroprevalence estimates downward on the basis of vaccination rates in adults in every location, accounting for vaccination of previously infected individuals (appendix 1, section 5.1).

Seroprevalence surveys provide an estimate of the number of individuals who have been infected with SARS-CoV-2 one or more times; these surveys do not detect repeat infections in a single individual. Because reinfection can be common in settings where escape variants such as beta (B.1.351), gamma (P.1), and delta (B.1.617.2) are present,[28–30] we had to adjust seroprevalence data to estimate the cumulative number of infections—that is, to include both first and any subsequent infections. We used a level of cross-variant immunity of 30% to 70% between escape variants and ancestral variants and alpha (B.1.1.7), on the basis of an empirical analysis conducted by the COVID-19 Forecasting Team (unpublished). This estimate did not take into account that some individuals could have been infected more than once with ancestral variants.[31] A detailed explanation of how we adjusted for escape variant prevalence is given in appendix 1 (section 5.2).

Lastly, seroprevalence surveys were corrected for waning sensitivity of antibody tests. We identified eight categories of antibody tests; for each of these, we used a reported curve of sensitivity over time.[32–34] To implement the correction based on waning, we used initial estimates of the timing of infection based on reported deaths. We did not adjust for specificity, as reported specificity for all available commercial assays included in the analysis is over 95% and mostly over 98% (more details in appendix 1, sections 5.3 and 9.3).[35]