Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Inferring the true number of SARS-CoV-2 infections in Japan

Lauren McKenzie a, Affan Shoukat a,b,c, Kai On Wong a, Koju Itahashi a, Eiji Yasuda d, Alex Demarsh a, Kamran Khan a,e,f

a BlueDot, Toronto, Ontario, Canada
b Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada
c Medical Affairs Department, Meiji Seika Pharma Co., Ltd., Chuo-ku, Tokyo, Japan
d Digital Transformation Sect., Meiji Seika Pharma Co., Ltd., Chuo-ku, Tokyo, Japan
e Department of Medicine, University of Toronto, Toronto, Ontario, Canada
f Digital Transformation Sect., Meiji Seika Pharma Co., Ltd., Chuo-ku, Tokyo, Japan

Introduction: In Japan, as of December 31, 2021, more than 1.73 million laboratory-confirmed cases have been reported. However, the actual number of infections is likely to be under-ascertained due to the epidemiological characteristics such as mild and subclinical infections and limited testing availability in the early days of the pandemic. In this study, we infer the true number of infections in Japan between January 16, 2020, and December 31, 2021, using a statistical modelling framework that combines data on reported cases and fatalities.

Methods: We used reported COVID-19 deaths and age-specific infection fatality ratios (IFR) to impute the true number of infections. Estimates of IFR were informed from published studies and were adjusted to reflect the effects of pharmaceutical interventions, mass vaccination, and evolving variants. To account for the uncertainty in IFR, we sampled values from relevant distributions.

Results: We estimated that as of December 31, 2021, 3.07 million (CrI: 2.05–4.24 million) people had been infected in Japan, which is 1.77 times higher than the 1.73 million reported cases. Our meta-analysis confirmed that these findings were consistent with the intermittent seroprevalence studies conducted in Japan.

Conclusions: We have estimated that a substantial number of COVID-19 infections in Japan were unreported, particularly in adults. Our approach provides a more realistic assessment of the true underlying burden of COVID-19. The results of this study can be used as fundamental components to strengthen population health control and surveillance measures.

1. Introduction

Reliable estimates of the true number of SARS-CoV-2 infections in a population are necessary for evaluating the course of COVID-19, effectiveness of control strategies, and situational awareness. The number of confirmed cases, while readily available, tends to largely underestimate the true number of infections due to several characteristics of SARS-CoV-2 including mild and subclinical infections [1,2], high transmissibility [3-5], and an incubation period with a long-tailed distribution [6]. Additionally, limited availability of tests, imperfect test sensitivities, and delays in reporting further exacerbates the under-ascertainment of infections.

In Japan, as of December 31, 2021, more than 1.73 million laboratory-confirmed cases have been reported since the first identified case [7]. However, studies based on seroprevalence data in Japan have reported a high degree of under-ascertainment of cases [8-13]. For instance, the ascertainment rate of non-severe cases was estimated to be 0.44 at the start of the pandemic [12]. A more recent government report found that cumulative prevalence of infection in Osaka was 3.78% by end of 2021, while only 2.30% of infections were reported, with similar findings across four other prefectures [13]. Seroprevalence studies generally provide more accurate estimation of population prevalence. However, these studies are sparse in time and space and their use is not recommended in low prevalence settings as results could lead to a false sense of security regarding the extent of immunity in the population impeding mitigation strategies and interventions. In addition, tests conducted before seroconversion or after antibodies have waned can lead to false positives and false negatives [14].
To estimate an accurate level of case ascertainment in Japan, we developed a statistical modelling framework that combines data on reported cases and deaths. We posit that reported death counts are a more accurate indicator of disease burden than confirmed cases [15,16]. Countries with highly accessible healthcare and reliable collection of COVID-19 mortality data, including Japan, routinely and comprehensively capture and report disease-specific deaths [15,17]. Accordingly, we used age-specific infection fatality ratios (IFRs) along with reported age-stratified death counts to infer the true number of infections in the country. The IFR value represents the proportion of deaths among all individuals infected with COVID-19, including symptomatic and subclinical infections, thus providing a way to convert a population’s mortality rate into an estimate of infections. Consequently, our methodology enables us to remove selection bias due to individuals with mild symptoms or asymptomatic infection in case ascertainment. Our results can inform the efficacy of public health measures taken to protect the population from infection and estimate the underlying population immunity with higher accuracy and confidence than measures based on reported cases.

2. Methods

The total number of infections was estimated using daily reported COVID-19 deaths in Japan between January 16, 2020 and December 31, 2021 [7]. The number of deaths $D_t$ on each day $t$ was split into age-specific mortality data using the population age-structure in Japan [18]. For each age, the number of infections $N_t$ on day $t$ was then imputed using the formula

$$N_t = \frac{D_t}{IFR}.$$

Age-specific estimates of IFR were obtained from Ref. [19], which were calculated prior to vaccination, pharmaceutical measures, and evolution of less severe variants; however, recognizing these dynamics have a pronounced effect in reducing the IFR [20,21], we adjusted the estimates accordingly. Specifically, we linearly decreased the estimated IFR of each age to a maximum of 55% starting from February 17, 2021 (corresponding to the first vaccination in Japan) to the end of study period [22]. To account for the uncertainty in the IFR, we sampled values from log-normal distributions, with the mean parameters set to the adjusted IFR value. The standard deviation of the log-normal distribution was set to 1, relatively large compared to the mean estimates, to allow for a larger range of values to be sampled.

Next, we estimated the likely date of infection for each inferred infection, regardless of whether the patient recovered or died. Specifically, for each infection, we estimated the lag-time $t'$ such that the infection occurred $t - t'$ days before the reported death on day $t$. This lag-time was calculated as the sum of two independent intervals: the incubation period and the time-to-death following symptom onset. These two intervals were sampled for each inferred infection to account for uncertainty; in particular, the incubation period was sampled from a log-normal distribution with mean 5.7 days and standard deviation of 3.1 days [23] and the time-to-death was sampled from a log-normal distribution with mean of 14.5 days and a standard deviation of 6.7 days [24]. The total number of infections on any given day $t$ is then calculated by summing all infections associated to deaths at time $t > r$ (i.e., $N_t$) where the infection lag-time $t'$ is such that $t - t' = r$.

The variation in parameters was accounted for by running 1000 Monte Carlo simulations. The results were summarized by calculating the mean and associated 10% and 90% credible intervals of the inferred number of infections and were grouped into 10-year age groups: 1–10, 11–20, 21–30, 31–40, 41–50, 51–60, 61 to 70, 71–80, and 90+. All code and data files are available on https://github.com/BlueDot-global/true_infections.

To validate our estimates, we conducted a meta-analysis of multiple serology reports from January 1, 2020, to December 31, 2021 [13, 25–27]. Study eligibility criteria were population-based studies describing the prevalence of anti-SARS-CoV-2 (IgG and/or IgM) serum antibodies. Participants were from different socioeconomic backgrounds and age groups. A fixed-effects model was used to estimate pooled seroprevalence, and then extrapolated to the general population in June 2020, December 2020, and December 2021. Heterogeneity was assessed using I² statistics. Subgroup analyses were performed to explore potential sources of heterogeneity in the data.

3. Results

We estimate that as of December 31, 2021, 3.07 million (Credible Interval (CrI): 2.05–4.24 million) people have been infected in Japan, which is 1.77 times higher than the 1.73 million reported cases, resulting in a case ascertainment rate of 56.7% (CrI: 40.8–84.3%) (Fig. 1A). The estimated 3.07 million infections correspond to approximately 2.40% of the population of Japan. Stratified by age, we found the highest under-ascertainment in individuals less than 20 years of age (Table 1, Fig. 1B). Among this age group, we estimated a total of 861,097 infections (CrI: 777,153–997,231) as compared to 268,515 reported. For adults aged 20–60, we estimated 1.41 million infections (CrI: 1.21–1.64 million), corresponding to a case ascertainment rate of 83.8% (CrI: 72.1%–98.1%). Total number of infections among the elderly over 60 years of age was estimated to be 803,510 (CrI: 685,236–932,566), of which only 256,900 cases were reported. Fig. 1B shows the number of estimated number of infections by age group, overlaid with reported case counts in those age groups.

Our estimates of the total number of infections are consistent with multiple serology reports. Using a fixed-effects model, we estimated the pooled seroprevalence of the general population in June 2020, December 2020, and December 2021 to be 0.12% (95% confidence interval (CI): 0.06%–0.24%), 0.83% (95% CI: 0.68%–1.00%), and 2.34% (95% CI: 2.06%–2.76%) respectively. In comparison, the mean result of our estimation showed that 0.10% (CrI: 0.05%–0.15%) of the population was infected as of June 2020, 0.24% (CrI: 0.13%–0.37%) as of December 2020, and 2.39% (CrI: 1.6%–3.31%) by end of December 2021 showing significant agreement with clinically observed serology estimates. The I² statistic of the fixed-effects model, which represents the fraction of variance due to heterogeneity among studies, was 12% in June 2020, 87% in December 2020 and 88% December 2021. Due to the small number of studies included in each meta-analysis ($n \leq 5$), these values should be interpreted cautiously. We note that all studies included in the meta-analysis were conducted by the Ministry of Health, Labour and Welfare in Japan, and followed the same protocol.

4. Discussion

Understanding the extent of unreported infections in a country is crucial for effective policy and mitigation strategies and estimating the severity of the outbreak. In this study, we estimated the level of case ascertainment in Japan using a statistical framework that combines IFR estimates and publicly reported death counts. Our results indicate that a substantial number of COVID-19 infections in the country had not been reported, particularly in children less than 20 years of age and in adults aged 60 years of age and older. This under-ascertainment is explained by two crucial factors. First, early in the pandemic, Japan employed a RT-PCR testing policy in which individuals with mild infection were only tested available for asymptomatically infected individuals [30]. Second, by the end of 2021, the vaccination rate had exceeded 80% [31]. Since vaccines prevent COVID-19 disease but not SARS-CoV-2 infection, increasing coverage shifts symptomatic infections to asymptomatic infections [32], thereby exacerbating the under-ascertainment.
Reported cases, deaths, and estimated number of SARS-CoV-2 infections in Japan by age groups.

| Age Group | Reported number of infections | Number of deaths | Estimated mean number of infections | Credible interval |
|-----------|------------------------------|------------------|-------------------------------------|------------------|
| 1–10      | 94,046                       | 4                | 437,136                             | 373,733-506,191  |
| 11–20     | 174,469                      | 11               | 429,960                             | 363,419-491,039  |
| 21–30     | 424,964                      | 50               | 488,818                             | 417,488-567,790  |
| 31–40     | 281,203                      | 114              | 301,910                             | 258,155-351,404  |
| 41–50     | 265,446                      | 295              | 277,462                             | 237,703-322,406  |
| 51–60     | 210,861                      | 785              | 342,240                             | 291,682-398,232  |
| 61–70     | 104,272                      | 2051             | 323,871                             | 275,939-375,449  |
| 71–80     | 78,502                       | 4553             | 258,222                             | 221,098-299,242  |
| 81–90     | 53,922                       | 8636             | 202,618                             | 172,347-235,724  |
| 90+       | 20,204                       | 1894             | 18,798                              | 15,851-22,150    |

* The number of deaths for each age group are estimated (rounded) using Japan’s population age structure.

Despite the relatively less robust testing strategies, social distancing measures, and lockdown restrictions, the estimated prevalence of 2.29% by end of 2021 was lower than other countries in the Western Pacific as well as USA and the UK [33]. Japan’s high population density and early exposure to SARS-CoV-2 make this pattern more counterintuitive. This low prevalence can be explained by several reasons. First, due to the already established social norm, the public’s compliance to wearing face masks and coverings during COVID-19 was strong both early and late stages in the pandemic [34]. In addition, considering shortages of medical masks in the early days of the pandemic, the government sent over 100 million reusable face masks to its citizens in April 2020 ensuring strong compliance early in the pandemic [35]. Second, Japan established a robust network of contact tracers to track the spread of SARS-CoV-2 [36]. Considering that roughly 20% of SARS-CoV-2 infections are responsible for 80% of transmission [37], the country adopted cluster-focused retrospective contact-tracing methodology; this method looks backwards to identify when and where a case was originally infected to isolate COVID-19 transmission clusters. Published studies have shown that retrospective contact tracing results in identifying 2–3 times more cases than prospective contact tracing [38]. Finally, we presume that efficacious messaging on the risk factors of droplet and aerosol viral transmission early in the pandemic, along with the culture of shared responsibility, lead to strong self-regulation and awareness in the population.

Our estimates of case ascertainment should be considered in the context of several limitations. While death counts are a more accurate indicator of disease burden [15], the number of deaths attributable to COVID-19 may have been underestimated, especially during the first months of the pandemic, affecting our model estimates [16]. Furthermore, IFR values are subject to considerable change due to factors that affect severe outcomes such as increasing vaccination coverage, emerging new variants, and improved pharmaceutical interventions in the population. In our study, we accounted for these confounding factors by adjusting IFR values after the start of vaccination in Japan. In the absence of an observational study analyzing the time evolution of age-specific IFR values, we derived our adjustment factor based on a modelling study [22]. Nonetheless, our estimates are consistent with serological data from Japan, with model’s mean estimates considerably overlapping the 95% confidence intervals reported in the studies.

Effective vaccination strategies to achieve herd immunity and non-pharmaceutical control measures, coupled with robust surveillance systems, are needed to restrain the spread of SARS-2-CoV variants in the country. Sustainable and feasible long-term control of COVID-19 requires continuous review and understanding of the severity of the ongoing pandemic. Our method to estimate the true underlying cases could be a fundamental component to strengthen these population health control and surveillance measures.

Author contributions

AS, LM, AD, and KK conceived the idea. LM, AS, KW, KI, and EY collected and contributed data. LM, AS, and KW performed statistical analysis and code programming. All authors interpreted the results and contributed to the writing of the manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

KK is the founder of BlueDot, a social enterprise that develops digital technologies for public health. LM, AS, AD, KW, and KK are employed at BlueDot. KI and EY are employed at Meiji Seika Pharma.
