Heliyon 10 (2024) e23408

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

# Gaining further insights into the COVID-19 pandemic in Australia: Evidence using capture-recapture methods

Joanne Thandrayen [a,*], Bernard Baffour [b]

[a] National Centre for Epidemiology and Population Health, College of Health and Medicine, Australian National University, 54 Mills Road, Acton ACT 2601, Australia
[b] School of Demography, College of Arts and Social Sciences, Australian National University, 146 Ellery Cres, Acton ACT 2601, Australia

ABSTRACT

*Objective:* We re-examined the reported number of COVID-19 cases in Australia and across its states during the first wave of the pandemic. We provided estimates of the total number of cases, adjusted for under-reporting.
*Methods:* Publicly available data sourced from Australian governments at federal, state and territory levels included records on cumulative confirmed COVID-19 cases and cumulative deaths occurring in Australia and across its states on a daily basis. Lower bound and upper bound estimates of the total number of COVID-19 cases in Australia and across its states, that included the undetected cases that have not been recorded, were estimated.
*Results:* On January 25, 2020, Australia recorded its first 4 cases of COVID-19 and the first death occurred on March 3, 2020. On April 1, 2020, 4864 cases had been reported with 21 deaths. Our estimation showed that on April 1, 2020, the minimum and maximum number of COVID-19 cases in Australia were in fact 10,160 (95 % CI: 9781–10,538) and 21,748 (95 % CI: 21,607–22,014) respectively. We estimated that the total number of cases were at least twice and at most four times the observed cases recorded. These differences were also found at the state level where in New South Wales there was a minimum and maximum of 207 and 447 cases in total for every 100 reported cases, while in Victoria these figures were much lower at 157 and 265 respectively for every 100 reported cases.
*Conclusion:* Case ascertainment during the pandemic is known to have been underestimated due to difficulties in testing and contact tracing, amongst others. Capture-recapture methods provided a measure of the gap between the official number of cases recorded and the actual number during the first wave of the pandemic.

## 1. Introduction

Since its early stages, cases of the coronavirus disease 2019 (COVID-19) arising from the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) have been highly under-reported in official statistics [1–4]. As such, there remains a significant difference between the observed (reported) number of cases and the 'true' number of cases. A population consists of infected and uninfected people, infected people may be symptomatic or not, and their probability of being detected depends heavily on symptoms that some approaches to case ascertainment such as sero-prevalence surveys [5] or modelling based on Infection Fatality Ratios [6] refer only to the broader population of symptomatic infections. Evidence from numerous sources cite the difficulties in identification of asymptomatic cases, delayed symptoms, rapidly evolving variants and strains, efficacy of contract tracing, inchoate health systems, inadequacies in serological testing, amongst others [7–12]. For these reasons, the total number of cases is unknown, yet the information on the spread of the epidemic is routinely based on the observed cases, which have been shown to be a fraction of the total number of cases [10,13–15]. While the pandemic appears to be contained, in most jurisdictions, practically all countries have been implementing

* Corresponding author.
*E-mail address:* Joanne.Thandrayen@anu.edu.au (J. Thandrayen).

https://doi.org/10.1016/j.heliyon.2023.e23408
Received 9 October 2023; Received in revised form 26 November 2023; Accepted 4 December 2023
Available online 11 December 2023
2405-8440/

strategies against future outbreaks [16,17].

Capture-recapture methods, which originated in ecology with the aim of estimating the size of an unknown (possibly elusive) population [18], have now been extensively used in epidemiology and public health [19–23], and also the social sciences in measuring hard-to-count phenomena [24,25]. In public health settings, capture-recapture modelling has traditionally been used with two or more sources or lists. Another approach is based on a single source with repeated identifications (i.e. captures) of people in that source during a specified observational period [26–29]. In the latter case, the technique uses information on the number of people who have been identified once, twice, and so on during the study period to estimate the number of people not identified in the single source. Essentially the difference is that the former approach examines repeated identifications across multiple sources and the latter examines repeated identifications within a single source.

Since the number of undetected COVID-19 cases is unknown, capture-recapture methods have increasingly been applied across the pandemic [1,2,30] to estimate the true number of cases (taking into account both the observed and unobserved cases). In this paper, we apply the capture-recapture approach to estimate the total number of COVID-19 cases in Australia and provide a measure of the precision through calculating upper and lower bounds of these estimates. Knowing these figures is important for assessing the extent of the unreported cases and critical for policy planning and healthcare delivery.

While capture-recapture estimators have been applied in the estimation of COVID-19 prevalence, it is not a simple task to provide an accurate estimate of the true number of COVID-19 cases. Böhning et al., 2020 [1] set out an approach that provides an estimate of the total number of cases, including undetected cases. Their modelling framework is innovative in its use of cumulative time series distributions to incorporate differences in probabilities of being detected, and how this may vary over time. They also used a modified estimator proposed by Chao [31,32] to estimate the lower bound of the total number of cases, whose records may not have been collected. Rocchetti et al., 2020 [2] provided an extension which gives an upper bound of the total number of cases, accounting for the transmission and the rapid spread of the epidemic, and this has been important in designing a more effective public health response. Our analysis investigated both approaches to provide lower bound and upper bound estimates of the total number of cases. By doing so we can quantify the minimum and maximum levels of under-ascertainment of the recorded number of cases, which in turn enables a more reliable assessment of the true extent of the pandemic.

Australia has been one of the few developed countries in the world that has successfully slowed down the spread of COVID-19. Its relative geographic isolation and country's position as an island nation meant that it was effectively able to shut down its international borders, unlike in Europe [33]. Furthermore, the general population adherence to physical distancing measures, extended lockdown compliance and mask mandates have also been cited as contributing factors [34–37]. The political system in Australia devolved the response and management of the pandemic to the individual states (of which there are six – New South Wales, Victoria, Queensland, Western Australia, South Australia and Tasmania), and two territories (the Australian Capital Territory and Northern Territory). However, this meant that the respective states and territories varied in the implementation of public health protocols including state/territory border closures and entry/exit restrictions, quarantine rules for returning residents, lockdowns, social distancing rules and vaccination rollouts [33–37].

As a result, with a population of roughly 25 million, Australia has experienced lower infection and death rates than many comparable developed countries, with 23,129 deaths out of over 11 million confirmed cases of the disease, as of October 6, 2023 [38]. However, these figures are only those reported and collected by the state/territory official health systems, do not include the potentially large number of asymptomatic infections, and mask remarkable heterogeneity in data, recording practices and population behaviour and distribution, and as such can under-estimate the true magnitude of the pandemic. To our knowledge, estimation of the true number of COVID-19 cases, particularly including the number of undetected cases using capture-recapture methods has not been undertaken. Our study has been designed to meet this goal. In addition to the 'true' estimates, we also provide information about the precision and reliability of these estimates through computing lower and upper bounds, and their corresponding confidence intervals.

## 2. Methods

### 2.1. Data source

This study utilized data from public records sourced from Australian governments at federal, state and territory levels [39]. The data included records on cumulative counts of confirmed COVID-19 cases and deaths occurring in Australia on a daily basis. Similar information was available by states and territories. We analysed data over an observational period of 30 days starting from the day on which the first death was recorded. In Australia, the first 4 cases of COVID-19 were recorded on January 25, 2020 and the first death occurred on March 3, 2020. We calculated the estimates of a lower bound and upper bound and their respective confidence intervals for the total number of COVID-19 cases in Australia and across its states.

### 2.2. Statistical approach

We followed the capture-recapture framework described in Böhning et al. (2020) [1] and Rocchetti et al. (2020) [2] for the computation of the lower and upper bound estimators respectively. A summary of their mathematical formulation is provided below.

Using their notations, let $t_0$ and $t_m$ denote the start and end of an observational period respectively. Let $N(t)$ and $D(t)$ denote the cumulative count of observed cases and deaths respectively at day $t$ ($t = t_0, \cdots, t_m$). The number of new cases and deaths at day $t$ is thus $\Delta N(t) = N(t) - N(t-1)$ and $\Delta D(t) = D(t) - D(t-1)$ respectively.

The above situation can be related to a capture-recapture approach as follows. Each infected individual $i$ can be identified $x$ times

($x = 0, 1, 2, \ldots$) which is typically provided by the number of days that individual will remain infected. The corresponding frequencies are $f_0, f_1, f_2, \ldots$ for the whole distribution of cases (i.e. symptomatic and asymptomatic infections).

The unobserved frequency $f_0$ can be estimated by using a lower bound estimator [40,41] of the form:

$$\widehat{f}_0 = \frac{f_1^2}{f_2}$$

where $f_1$ and $f_2$ are the observed frequencies of those identified exactly once and twice respectively.

This result is a standard formula in capture-recapture literature and relies on the counts following a counting distribution such as a geometric distribution. This means that we can find the frequency of missing a case as the ratio of the squared frequency of identifying a case exactly once divided by the frequency of detecting a case twice. While, this result is reliant on the validity of the geometric distribution, it typically allows an arbitrary unknown distribution to reflect the varying probabilities of detection in the population [1, 40]. It is then a matter of connecting $f_1$ and $f_2$ with the data at hand. At day $t$, $f_1(t)$ is the number of infected people identified only once i.e. the new cases which is given by $\Delta N(t)$. Similarly, $f_2(t)$ is the number of infected people identified at day $(t-1)$ and those remaining infected at day $t$. This can be computed as $\Delta N(t-1) - \Delta D(t)$. In this instance, $f_0$ the frequency of unobserved cases, is still unknown and has to be estimated as $\widehat{f}_0$.

### 2.3. Lower bound estimator

Böhning et al. (2020) [1] provided a lower bound estimator for the number of undetected cases at day $t$ which is given by:

$$\widehat{f}_0^{LB}(t) = \frac{|\Delta N(t)|^2}{\Delta N(t-1) - \Delta D(t)} \tag{1}$$

The global estimate of undetected cases is then obtained by summing over all days in the observational period:

$$\widehat{f}_0^{LB} = \sum_{t=t_0+1}^{t_m} \frac{|\Delta N(t)|^2}{\Delta N(t-1) - \Delta D(t)} \tag{2}$$

In practice, a bias-corrected form of the above estimator [32] is used.

$$\widehat{f}_0^{LB} = \sum_{t=t_0+1}^{t_m} \frac{\Delta N(t)|\Delta N(t)-1|}{1 + \Delta N(t-1) - \Delta D(t)} \tag{3}$$

Note that $\Delta N(t-1) - \Delta D(t)$ is set to 0 if it becomes negative. The lower bound estimate of the total number of cases is given by:

$$N_{LB} = N(t_m) + \widehat{f}_0^{LB} \tag{4}$$

where $N(t_m)$ is the number of cases observed at the end of the observational period.

### 2.4. Upper bound estimator

Rocchetti et al. (2020) [2] provided an upper bound estimator for the number of undetected cases at day $t$ which is of the form:

$$\widehat{f}_0^{UB}(t) = n^*_{obs}(t) \times \frac{\widehat{\pi}_0^{UB}(t)}{1 - \widehat{\pi}_0^{UB}(t)} = \frac{n^*_{obs}(t)}{1 - \widehat{\pi}_0^{UB}(t)} - n^*_{obs}(t) \tag{5}$$

where $n^*_{obs}(t) = f_1(t) + f_2(t) + f_3(t)$; $f_1(t) = \Delta N(t)$; $f_2(t) = \Delta N(t-1) - \Delta D(t)$;

$$f_3(t) = \Delta N(t-2) - \Delta D(t-1) - \Delta D(t);$$

$$\widehat{\pi}_0^{UB}(t) = \frac{p_2(t) - p_1(t)}{\left(1 - \frac{\widehat{\pi}_1(t)}{\widehat{\pi}_2(t)}\right) + p_2(t) - p_1(t)};\ p_1(t) = \frac{f_1(t)}{n^*_{obs}(t)};\ p_2(t) = \frac{f_1(t) + f_2(t)}{n^*_{obs}(t)};$$

$$\widehat{\pi}_j(t) = \widehat{\pi}_0(t) + [1 - \widehat{\pi}_0(t)]\, p_j(t);\ \widehat{\pi}_0(t) = \frac{\widehat{f}_0(t)}{f_1(t) + f_2(t) + \widehat{f}_0(t)};$$

$\widehat{f}_0(t)$ is defined in its bias-corrected form (see above section on lower bound estimator).

The global estimate of undetected cases is then obtained by summing over all days in the observational period:

$$\widehat{f}_0^{UB} = \sum_{t=t_0+2}^{t_m} \frac{n^*_{obs}(t)}{1 - \widehat{\pi}_0^{UB}(t)} - n^*_{obs}(t) \tag{6}$$

The final upper bound estimate of the total number of cases is given by:

$$N_{UB} = N(t_m) + \widehat{f}_0^{UB} \tag{7}$$

where $N(t_m)$ is the number of cases observed at the end of the observational period.

We also summarised the computations described above by the following steps.

**Step 1**. Prepare a table of values as shown in Table 1. The date, N(t), and D(t) columns should be available in a database.

**Step 2**. Construct the columns t, ΔN(t), and ΔD(t). t is an incremental count with t = 0 as the starting value, ΔN(t) = N(t) - N(t-1), and ΔD(t) = D(t) - D(t-1).

**Step 3**. Calculate a lower bound estimator for the number of undetected cases at day t by using equation (1).

**Step 4**. Calculate a global estimate of undetected cases by summing over all days in the observational period by using equation (2).

**Step 5**. Calculate the bias-corrected form of that global estimator by using equation (3).

**Step 6**. Calculate the lower bound estimate of the total number of cases by using equation (4). To obtain the upper bound estimator, follow the same steps as those described above but using equations (5)–(7).

While these lower and upper bounds provide us with a better picture of the total number of cases, including the unobserved cases, we still need to address the uncertainty involved in those estimators. Böhning el al. (2020) [1] extended previous work by Niwitpong et al. (2013) [41] to formulate a closed form solution for the variance of the lower bound estimator and corresponding confidence intervals (see Böhning el al. (2020) [1] for details). Reduced bootstrap methods were used to estimate the variance of the upper bound estimator and corresponding confidence intervals (see Rocchetti et al. (2020) [2] for details).

We produce estimates of not only the total number of cases (adjusted for the unreported cases), but also provide upper and lower bounds of these estimates which characterise the relative reliability of the analysis. Taken together this allow us to make more robust inferences accounting for the underlying uncertainty of the results.

## 3. Results

The results showed that the estimates using only the observed number of COVID-19 cases vastly under-estimated the 'true' total number of cases (including the unreported number of cases). We found that this varied geographically (i.e. by state). Fig. 1 shows the

**Table 1**
Cumulative counts of confirmed COVID-19 cases (N(t)) and deaths (D(t)) from COVID-19 in Australia starting from $t_0$ = March 3, 2020 to $t_m$ = April 1, 2020.

| Date | t | N(t) | D(t) | ΔN(t) | ΔD(t) |
|---|---|---|---|---|---|
| March 3, 2020 | 0 | 40 | 1 | | |
| March 4, 2020 | 1 | 51 | 2 | 11 | 1 |
| March 5, 2020 | 2 | 59 | 2 | 8 | 0 |
| March 6, 2020 | 3 | 63 | 2 | 4 | 0 |
| March 7, 2020 | 4 | 73 | 2 | 10 | 0 |
| March 8, 2020 | 5 | 80 | 3 | 7 | 1 |
| March 9, 2020 | 6 | 92 | 3 | 12 | 0 |
| March 10, 2020 | 7 | 112 | 3 | 20 | 0 |
| March 11, 2020 | 8 | 127 | 3 | 15 | 0 |
| March 12, 2020 | 9 | 157 | 3 | 30 | 0 |
| March 13, 2020 | 10 | 198 | 4 | 41 | 1 |
| March 14, 2020 | 11 | 249 | 4 | 51 | 0 |
| March 15, 2020 | 12 | 297 | 6 | 48 | 2 |
| March 16, 2020 | 13 | 376 | 6 | 79 | 0 |
| March 17, 2020 | 14 | 454 | 6 | 78 | 0 |
| March 18, 2020 | 15 | 567 | 7 | 113 | 1 |
| March 19, 2020 | 16 | 709 | 7 | 142 | 0 |
| March 20, 2020 | 17 | 847 | 8 | 138 | 1 |
| March 21, 2020 | 18 | 1072 | 8 | 225 | 0 |
| March 22, 2020 | 19 | 1352 | 8 | 280 | 0 |
| March 23, 2020 | 20 | 1680 | 8 | 328 | 0 |
| March 24, 2020 | 21 | 2050 | 9 | 370 | 1 |
| March 25, 2020 | 22 | 2430 | 10 | 380 | 1 |
| March 26, 2020 | 23 | 2809 | 14 | 379 | 4 |
| March 27, 2020 | 24 | 3180 | 14 | 371 | 0 |
| March 28, 2020 | 25 | 3640 | 15 | 460 | 1 |
| March 29, 2020 | 26 | 3984 | 17 | 344 | 2 |
| March 30, 2020 | 27 | 4250 | 19 | 266 | 2 |
| March 31, 2020 | 28 | 4561 | 20 | 311 | 1 |
| April 1, 2020 | 29 | 4864 | 21 | 303 | 1 |

evolution of the pandemic in the first wave (March 2020 to May 2020) in Australia. There was a sharp increase in the cumulative counts of confirmed COVID-19 cases between end of March 2020 and start of April 2020. After that period, the increase appeared to plateau. This observation guided our choice for selecting the study period of March 3, 2020 to April 1, 2020. As an illustration, the cumulative counts of confirmed COVID-19 cases and deaths, and the number of new cases and deaths in Australia were provided for our chosen period of study (Table 1).

Table 2 shows the lower and upper bound estimates of the total number of COVID-19 cases respectively for Australia and five of its states: New South Wales; Victoria; Queensland; Western Australia; Tasmania. The number of deaths recorded in the observational period of 30 days (starting from the day on which the first death was recorded) was very small (under 5 deaths in total) in South Australia, Australian Capital Territory, and Northern Territory. We thus did not report on the estimates of undetected COVID-19 cases in the latter three jurisdictions.

In general, the total number of cases were at least twice (Table 2) and at most four times (Table 2) the observed cases recorded in Australia. A ratio (total/observed) of 2 and 4 implied that for every observed patient there were 2–4 undetected people respectively. These unseen cases might be those who were asymptomatic or experienced very mild symptoms of infection. There were differences at the state levels with New South Wales having the highest ratio and Victoria the lowest ratio (Fig. 2).

While point estimates provide a reasonable indication of the extent of COVID-19 outbreak in the chosen time period, there is some uncertainty associated with using only point estimates. We calculated 95 % Confidence intervals (CIs) for the lower bound estimator of the total number of cases (Table 2). These CIs provide a range of values that allow us to express how sure we are about the derived estimates. They measure the uncertainty in our estimates and provide a sense of reliability, with narrower CIs indicating better precision and wider CIs indicating poorer precision.

All CIs included their respective point estimates and were narrow. Bootstrapped 95 % CIs were also computed for the upper bound estimator of the total number of cases (Table 2). Similar patterns were observed in those latter CIs compared to those of the lower bound estimator. However, for smaller states, such as Tasmania, we had wider CIs. Our estimation of the total of cases of COVID-19 using both approaches was thus valid.

## 4. Discussion

The capture-recapture approach used in this study is a computationally easy and low cost method for estimating the true number of COVID-19 cases in Australia and across its states. This information can be useful in health policy planning for evaluating localised development of pandemics and allows for informed data driven decisions in pandemic response. In our study, we provided a lower bound and upper bound estimates of the true number of people having had COVID-19 in Australia during the first wave of the pandemic as well as reliable confidence intervals around those estimates. Our lower bound estimate for Australia is consistent with that of Phipps, Grafton and Kompas (2020) [42] who followed a backcasting approach. Their estimated number of people having had COVID-19 was 9000 (95 % CI: 6000–15,000) at March 28, 2020 whereas ours was 10,160 (95 % CI: (9781–10,538) at April 1, 2020. With these two estimates, public health policy makers are now able to (a) be aware of the minimum number of cases required for health

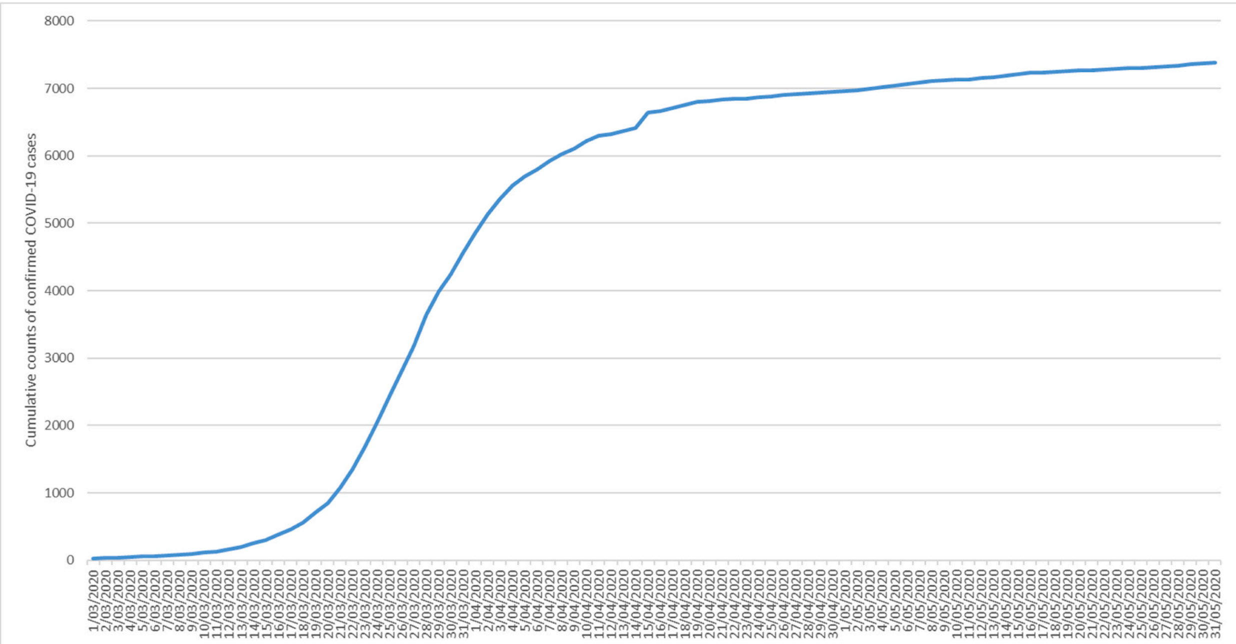


**Fig. 1.** Cumulative counts of confirmed COVID-19 cases during the first wave (March 2020 to April 2020) of the pandemic in Australia.

**Table 2**
Estimated total cases of COVID-19 for Australia and its states/territories, April 2020.

| | | | Lower bound | | Upper bound | |
|---|---|---|---|---|---|---|
| | Observational period[a] | Observed cases | Total cases (95 % CI) | Ratio of total/observed (95 % CI) | Total cases (95 % CI) | Ratio of total/observed (95 % CI) |
| Australia** | 03/03/20 to 01/04/20 | 4864 | 10160 (9781,10538) | 2.09 (2.01,2.17) | 21748 (21607,22014) | 4.47 (4.44,4.53) |
| NSW | 07/03/20 to 05/04/20 | 2580 | 5334 (5057,5611) | 2.07 (1.96,2.17) | 11524 (11406,11765) | 4.47 (4.42,4.56) |
| Vic | 28/03/20 to 26/04/20 | 1349 | 2113 (1820,2405) | 1.57 (1.35,1.78) | 3569 (3293,3903) | 2.65 (2.44,2.89) |
| Qld | 25/03/20 to 23/04/20 | 1024 | 1640 (1491,1790) | 1.60 (1.46,1.75) | 2957 (2886,3122) | 2.89 (2.82,3.05) |
| WA | 24/03/20 to 22/04/20 | 546 | 930 (795,1064) | 1.70 (1.46,1.95) | 1762 (1660,2069) | 3.23 (3.04,3.79) |
| Tas | 30/03/20 to 28/04/20 | 220 | 419 (304,535) | 1.91 (1.38,2.43) | 781 (694,985) | 3.55 (3.15,4.48) |

[a] 30 days from first death recorded; ** Australia included all states and territories (NSW, Vic, Qld, WA, SA, Tas, ACT, NT).

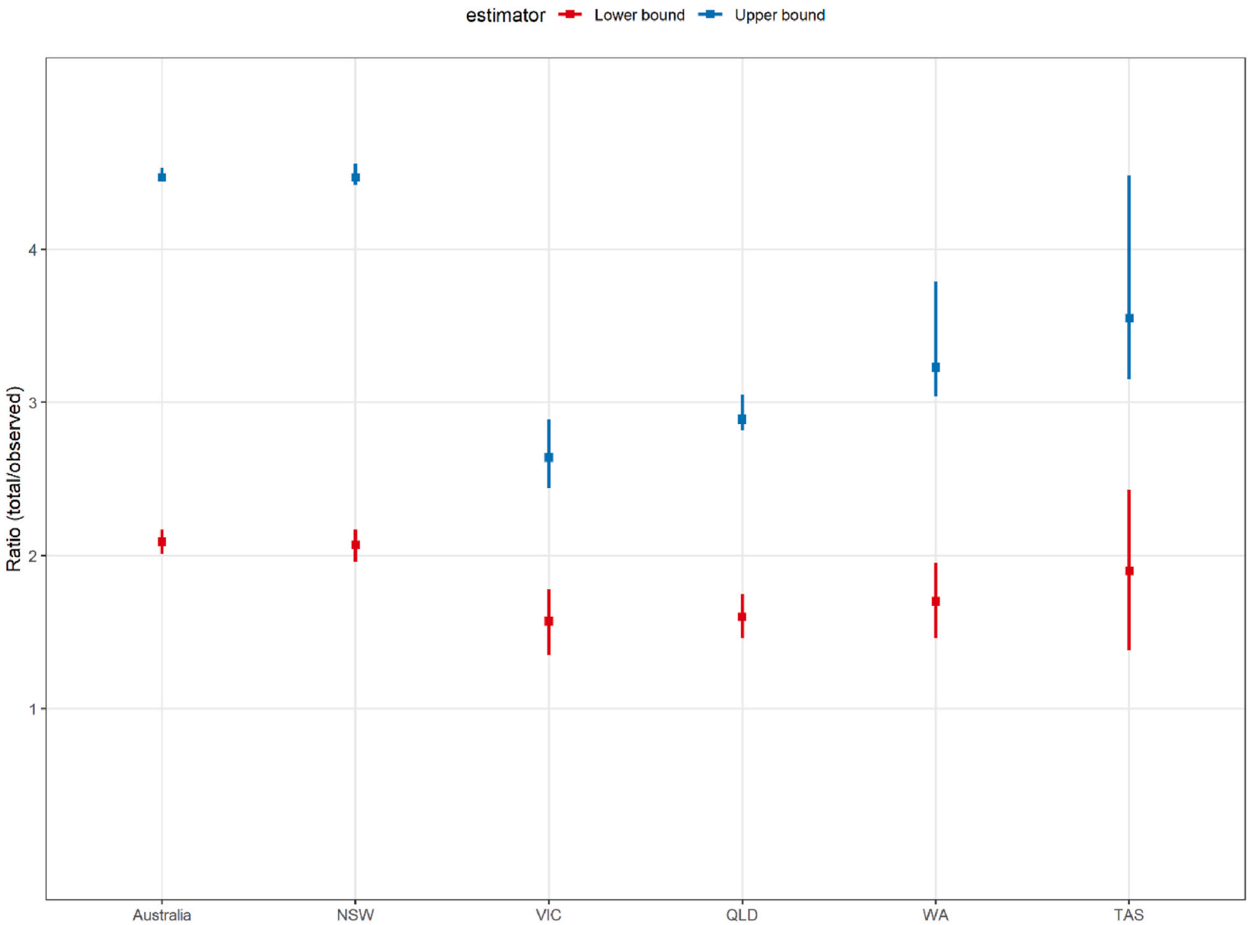


**Fig. 2.** Ratio (with confidence intervals) of total number of infections versus observed number of infections for Australia and its states, April 2020.

care demand, and (b) the possible true extent of the epidemic for the worst-case scenario planning.

We also provided estimates at the state-level to show regional variation across Australia given that states and territories have jurisdiction on reporting and health decision making. The state estimates showed that New South Wales had the highest number of undetected cases while Victoria had the lowest. This can be explained by differences in the measures implemented to curb the spread of the disease with Victoria having the strictest rules in place.

Our analysis utilized simple time series of cumulative data which are available to the general public or the academic sector to shed light on the undetected cases of COVID-19 across Australia. In contrast, individual data are not publicly available which thus did not allow us to provide age or sex-specific estimates. Future work could explore this aspect if individual data becomes available. Another advantage is that the estimators incorporate the heterogeneity in the detection probability, by using the time series information to adjust for variations in reporting. In addition, the true number of cases is bounded (in theory) by the total number of observed cases (at the minimum) and positive infinity (at the maximum), and as such the confidence intervals of the upper bound are rarely symmetric. A bootstrap routine (as implemented here) has the advantage of producing non-symmetric confidence intervals, unlike one using normalisation.

While capture-recapture methods are recognised for providing valuable insights, it should be acknowledged that they can potentially lead to incorrect conclusions, in particular when the modelling assumptions are unmet [43]. In addition, capture-recapture studies can be resource-intensive when the population of interest is large. Ethical and privacy concerns may also arise as the identification of cases relies on handling sensitive information such as names, date of births, and addresses. While our research provides a practical contribution to understanding the spread of newly developing infectious diseases, we are not stating that this approach can be implemented into other healthcare settings without due consideration being given to the research question of interest and the availability of reliable data. In our application of this approach to assessing the true number of COVID-19 cases in Australia during the first wave of the pandemic, we are assuming that all data collected by each state and territory in Australia are accurate.

Our study focused on the first wave of COVID-19 in Australia. The main result is the number of people with COVID-19 in the population at that time was clearly under-diagnosed. The undetected number of cases is in fact more important than the observed cases as they are those who are more likely to transmit the disease and should thus be in self-isolation. In this study, we did not attempt to undertake estimation in different waves of the pandemic. Future work could focus on this aspect and would employ time series techniques to smooth the data and subsequently apply the statistical approaches investigated in the current work. We also did not provide estimates for the states/territories that reported under 5 deaths (i.e., South Australia, Australian Capital Territory, and Northern Territory). Capture-recapture methods for population estimation with small counts [32] will be important in this regard, but are currently under-developed.

The present study demonstrated the need to investigate alternative methods to direct counting when a new pandemic that is unknown to the society emerges. Traditional methods to counting such as testing and contact tracing though useful are costly and time consuming. The capture-recapture method was low-cost, straightforward, and can be easily replicated in other countries, and at other timeframes. Public health systems would benefit to use this technique to inform the public of the true number of cases on a daily basis. This approach can be of significant value for surveillance and the management of infectious diseases in future. Measuring the gap between the recorded and actual number of cases is important to reduce uncertainty of numbers that feed into all modelling and predictions of the COVID-19 pandemic in Australia. Accurate predictions of how big or when the next wave will occur is necessary for government policy, planning, and the general public. Accurate case numbers would also assist in making informed decisions regarding which strategies to adopt during peaks such as vaccine booster campaign, mask rules as well as to assess whether those strategies will work. While the interest in COVID-19 research may have subsided recently, the importance of knowing the actual numbers could quickly arise if a new and more deadly variant emerges. By applying novel statistical methods, we might be able to provide sound information to address the unknown in a relatively short period, and perhaps save lives in the process.

## Ethics statement

Not applicable.


## Funding

The authors received no specific funding for this work.


## Data availability statement

Data included in article/supplementary material/referenced in article.

## CRediT authorship contribution statement

**Joanne Thandrayen:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Bernard Baffour:** Writing – review & editing, Methodology.

## Declaration of generative AI and AI-assisted technologies in the writing process

Not applicable.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

Not applicable.

## Abbreviations

| | |
|---|---|
| ACT | Australian Capital Territory |
| NSW | New South Wales |
| NT | Northern Territory |
| Qld | Queensland |
| SA | South Australia |
| Tas | Tasmania |
| Vic | Victoria |
| WA | Western Australia |

## References

[1] D. Böhning, I. Rocchetti, A. Maruotti, H. Holling, Estimating the undetected infections in the Covid-19 outbreak by harnessing capture–recapture methods, Int. J. Infect. Dis. 97 (2020) 197–201.
[2] I. Rocchetti, D. Böhning, H. Holling, A. Maruotti, Estimating the size of undetected cases of the COVID-19 outbreak in Europe: an upper bound estimator, Epidemiol. Methods 9 (1) (2020), 20200024.
[3] A.R. Tuite, V. Ng, E. Rees, D. Fisman, Estimation of COVID-19 outbreak size in Italy, Lancet Infect. Dis. 20 (5) (2020) 537.
[4] M. Yue, H.E. Clapham, A.R. Cook, Estimating the size of a COVID-19 epidemic from surveillance systems, Epidemiology 31 (4) (2020) 567–569.
[5] D. McConnell, C. Hickey, N. Bargary, L. Trela-Larsen, C. Walsh, M. Barry, et al., Understanding the challenges and uncertainties of seroprevalence studies for SARS-CoV-2, Int. J. Environ. Res. Publ. Health 18 (9) (2021).
[6] G. Luo, X. Zhang, H. Zheng, D. He, Infection fatality ratio and case fatality ratio of COVID-19, Int. J. Infect. Dis. 113 (2021) 43–46.
[7] R.W. Peeling, C.J. Wedderburn, P.J. Garcia, D. Boeras, N. Fongwen, J. Nkengasong, et al., Serology testing in the COVID-19 pandemic response, Lancet Infect. Dis. 20 (9) (2020) e245–e249.
[8] H. Nishiura, T. Kobayashi, T. Miyama, A. Suzuki, S.-M. Jung, K. Hayashi, et al., Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19), Int. J. Infect. Dis. 94 (2020) 154–155.
[9] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, et al., Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2), Science 368 (6490) (2020) 489–493.
[10] M. Day, Covid-19: identifying and isolating asymptomatic people helped eliminate virus in Italian village, BMJ 368 (2020) m1165.
[11] D. Blumenthal, E.J. Fowler, M. Abrams, S.R. Collins, Covid-19 — implications for the health care system, N. Engl. J. Med. 383 (15) (2020) 1483–1488.
[12] K. Sharun, R. Tiwari, K. Dhama, T.B. Emran, A.A. Rabaan, A. Al Mutair, Emerging SARS-CoV-2 variants: impact on vaccine efficacy and neutralizing antibodies, Hum. Vaccines Immunother. 17 (10) (2021) 3491–3494.
[13] M. Pollán, B. Pérez-Gómez, R. Pastor-Barriuso, J. Oteo, M.A. Hernán, M. Pérez-Olmeda, et al., Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study, Lancet (London, England) 396 (10250) (2020) 535–544.
[14] C. Rothe, M. Schunk, P. Sothmann, G. Bretzel, G. Froeschl, C. Wallrauch, et al., Transmission of 2019-nCoV infection from an asymptomatic contact in Germany, N. Engl. J. Med. 382 (10) (2020) 970–971.
[15] Y. Yu, Y.-R. Liu, F.-M. Luo, W.-W. Tu, D.-C. Zhan, G. Yu, et al., COVID-19 Asymptomatic Infection Estimation, medRxiv, 2020.
[16] S. Lim, M. Sohn, How to cope with emerging viral diseases: lessons from South Korea's strategy for COVID-19, and collateral damage to cardiometabolic health, Lancet Reg Health West Pac 30 (2023), 100581.
[17] Y. Assefa, C.F. Gilks, S. Reid, R. van de Pas, D.G. Gete, W. Van Damme, Analysis of the COVID-19 pandemic: lessons towards a more efective response to public health emergencies, Glob. Health 18 (10) (2022).
[18] C.C. Sekar, W.E. Deming, On a method of estimating birth and death rates and the extent of registration, J. Am. Stat. Assoc. 44 (245) (1949) 101–115.
[19] International Working Group for Disease Monitoring and Forecasting, Capture-recapture and multiple-record systems estimation I: history and theoretical development, Am. J. Epidemiol. 142 (10) (1995) 1047–1058.
[20] R.E. LaPorte, S.R. Dearwater, Y.F. Chang, T.J. Songer, D.J. Aaron, R.L. Anderson, et al., Efficiency and accuracy of disease monitoring systems: application of capture-recapture methods to injury monitoring, Am. J. Epidemiol. 142 (10) (1995) 1069–1077.
[21] G. Verlato, M. Muggeo, Capture-recapture method in the epidemiology of type 2 diabetes: a contribution from the Verona Diabetes Study, Diabetes Care 23 (6) (2000) 759–764.
[22] A. Chao, P.K. Tsay, S.H. Lin, W.Y. Shau, D.Y. Chao, The applications of capture-recapture models to epidemiological data, Stat. Med. 20 (20) (2001) 3123–3157.
[23] D. Böhning, P.G.M. van der Heijden, J. Bunge, Capture-Recapture Methods for the Social and Medical Sciences, CRC Press LLC, Philadelphia, PA, UNITED STATES, 2017.
[24] A.G. Davies, R.M. Cormack, A.M. Richardson, Estimation of injecting drug users in the City of Edinburgh, Scotland, and number infected with human immunodeficiency virus, Int. J. Epidemiol. 28 (1) (1999) 117–121.
[25] R.Q. Gurgel, J.D. da Fonseca, D. Neyra-Castañeda, G.V. Gill, L.E. Cuevas, Capture-recapture to estimate the number of street children in a city in Brazil, Arch. Dis. Child. 89 (3) (2004) 222–224.
[26] C. Rossi, F. Mascioli, Capture-recapture methods to estimate prevalence indicators for evaluating drug policies, Bull. Narc. 60 (2008).
[27] M.J. Cruyff, P.G.M. van der Heijden, Point and interval estimation of the population size using a zero-truncated negative binomial regression model, Biom. J. 50 (6) (2008) 1035–1050.
[28] D. Böhning, P.G.M. van der Heijden, A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations, Ann. Appl. Stat. 3 (2) (2009) 595–610.
[29] P.G.M. van der Heijden, G. Cruts, M.J. Cruyff, Methods for population size estimation of problem drug users using a single registration, Int. J. Drug Pol. 24 (6) (2013) 614–618.
[30] A. Sarría-Santamera, N. Abdukadyrov, N. Glushkova, D. Russell Peck, P. Colet, A. Yeskendir, et al., Towards an accurate estimation of COVID-19 cases in Kazakhstan: back-casting and capture-recapture approaches, Medicina 58 (2) (2022) 253.
[31] A. Chao, Estimating the population size for capture-recapture data with unequal catchability, Biometrics 43 (4) (1987) 783–791.
[32] A. Chao, Estimating population size for sparse data in capture-recapture experiments, Biometrics 45 (2) (1989) 427–438.

[33] M. Boyd, M.G. Baker, N. Wilson, Border closure for island nations? Analysis of pandemic and bioweapon-related threats suggests some scenarios warrant drastic action, Aust N Z J Public Health 44 (2) (2020) 89–91.
[34] B. Edwards, N. Biddle, M. Gray, K. Sollis, COVID-19 vaccine hesitancy and resistance: correlates in a nationally representative longitudinal survey of the Australian population, PLoS One 16 (3) (2021), e0248892.
[35] K. Murphy, H. Williamson, E. Sargeant, M. McCarthy, Why people comply with COVID-19 social distancing restrictions: self-interest or duty? Aust. N. Z. J. Criminol. 53 (4) (2020) 477–496.
[36] C.R. MacIntyre, P.-Y. Nguyen, A.A. Chughtai, M. Trent, B. Gerber, K. Steinhofel, et al., Mask use, risk-mitigation behaviours and pandemic fatigue during the COVID-19 pandemic in five cities in Australia, the UK and USA: a cross-sectional survey, Int. J. Infect. Dis. 106 (2021) 199–207.
[37] A.L. Quigley, M. Trent, H. Seale, A.A. Chughtai, C.R. MacIntyre, Cross-sectional survey of changes in knowledge, attitudes and practice of mask use in Sydney and Melbourne during the 2020 COVID-19 pandemic, BMJ Open 12 (6) (2022), e057860.
[38] World Health Organization. Coronavirus (COVID-19) Dashboard by country. https://covid19.who.int/region/wpro/country/au (viewed Oct 2023).
[39] O'Brien J. COVID-19 Data for Australia. https://github.com/M3IT/COVID-19_Data (viewed Aug 2021).
[40] A. Chao, R.K. Colwell, Thirty years of progeny from Chao's inequality: estimating and comparing richness with incidence data and incomplete sampling, SORT Stat Oper Res Trans 41 (1) (2017) 3–54.
[41] S.-A. Niwitpong, D. Böhning, P.G.M. van der Heijden, H. Holling, Capture–recapture estimation based upon the geometric distribution allowing for heterogeneity, Metrika 76 (4) (2013) 495–519.
[42] S.J. Phipps, R.Q. Grafton, T. Kompas, Robust estimates of the true (population) infection rate for COVID-19: a backcasting approach, R. Soc. Open Sci. 7 (11) (2020), 200909.
[43] J.C. Desenclos, B. Hubert, Limitations to the universal use of capture-recapture methods, Int. J. Epidemiol. 23 (1994) 1322–1323.