Evaluating undercounts in epidemics

Michael Li, Jonathan Dushoff, David J. D. Earn, and Ben Bolker

09 July 2025

Introduction

Several papers[1–3] have promoted formulas that claim to provide bounds on the completeness of sampling of infectious disease cases, based only on case reports. We believe these approaches are fundamentally wrong, and that it is impossible to estimate undercounting from incidence data without a specialized sampling design or some kind of auxiliary information.

The authors' basic idea uses formulas developed by Chao to estimate bounds on true population sizes based on the numbers of individuals observed multiple times. For example, their estimator for the lower bound on true cases is $\Delta N(t)(\Delta N(t)-1)/(1+\Delta N(t-1))$, where $\Delta N(t)$ is the number of new cases observed per reporting period; an extended formula adjusts for mortality. The upper bound involves $\Delta N(t-2)$ as well[1,3].

Critique

This approach misuses the mark-recapture formulas. Cases identified at time $t-1$ are supposed to be representative of the number of cases counted twice: why? The fact that the same individual *could* be counted twice in the cumulative case report (for some sampling designs) is irrelevant. How can comparing yesterday's count to today's provide information about the completeness of sampling?

In principle, the number of unobserved individuals (hidden cases) can be estimated if individuals can be re-identified, or even with unmarked individuals given an appropriate sampling design[4]. In practice public health case reporting rarely uses such sampling designs. Case reporting is usually exclusive (i.e. someone who has been identified as a case will not be reported again later), or anonymized so that we cannot identify a particular infected individual as double-counted. Mark-recapture methods are sometimes used in public health, but "one needs at least two sources of information with individual case reporting and a unique personal identifier for each case"[5].

Exponential example

During the initial phase of an epidemic, various measures of current epidemic size (incidence, cumulative incidence, etc.) all grow geometrically. Suppose the incidence (true number of new infections: $I(t)$) grows at a rate λ per time step, i.e. $I(t)=I(0)\lambda^t$, and suppose a fraction a (the *ascertainment ratio*) of these cases are reported. The (lower bound on the) estimated ascertainment ratio $\hat{a}$ equals the number of reported cases, divided by the total number of estimated cases, i.e. (reported cases + estimated bound on hidden cases). Using the simpler, non-bias-corrected formula:

$$
\begin{aligned}
\hat{a} &= \frac{\Delta N(t)}{\Delta N(t) + H(t)} \\
&= \frac{\Delta N(t)}{\Delta N(t) + \frac{[\Delta N(t)]^2}{\Delta N(t-1)}} \\
&= (1 + \Delta N(t)/\Delta N(t-1))^{-1} \\
&= \left(1 + aI(0)\lambda^t/(aI(0)\lambda^{t-1}\right)^{-1} \\
&= 1/(1+\lambda).
\end{aligned}
$$

The estimated ascertainment ratio $\hat{a}$ depends only on the epidemic growth rate; it is independent of the true ascertainment ratio. As we show below, the bias-corrected formula has similar problems.

Simulation example (SIR model)

We simulated a discrete-time deterministic SIR epidemic, applied two different ascertainment ratios ($a = \{0.2, 0.8\}$) to derive time series of case reports, and used the `asymptor` package[6] (which computes bounds incorporating bias corrections) to estimate hidden cases (Figure 1). The estimated lower and upper bounds are largely independent of the true a; at the beginning and end of the epidemic, when the absolute number of cases is lower, the bias correction terms make the estimated bounds on the ascertainment ratio *higher* in the low-ascertainment scenario and *vice versa*.

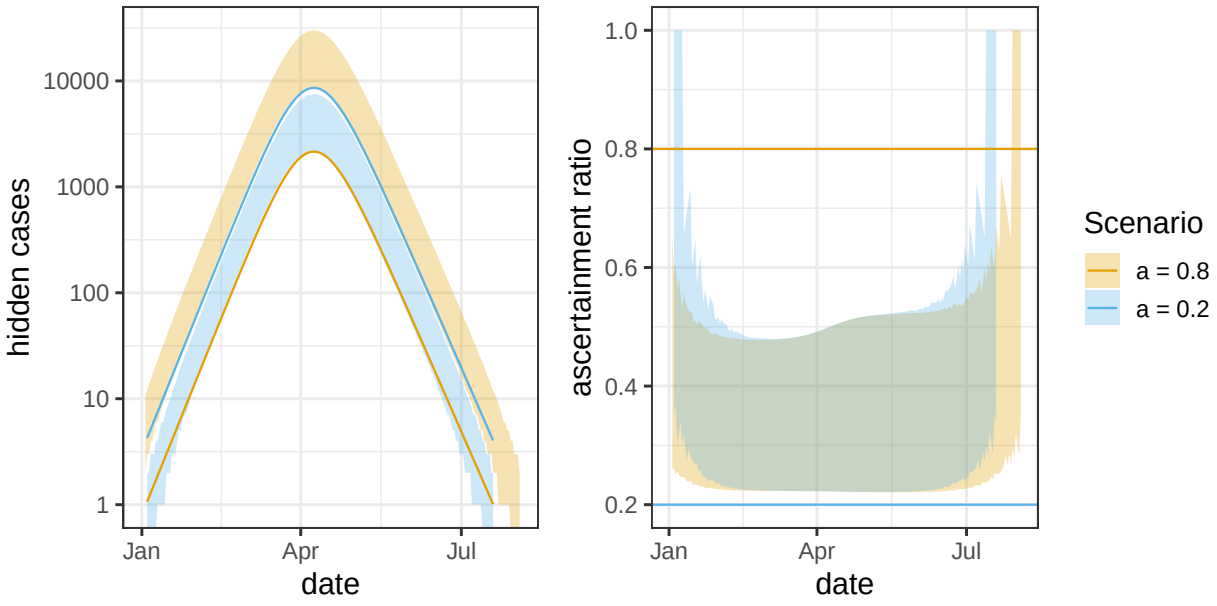

Figure 1: Estimates based on a simulated SIR epidemic with $\mathcal{R}_0 = 1.3$, infectious period of 3.33 days, $N = 10^6$. Left panel, true numbers of hidden cases (lines) with estimated bounds (regions). Right panel, true ascertainment ratios (lines) and estimated bounds (regions).

Simulation example (Richards curve)

We ran additional simulations using a Richards curve for the cumulative incidence of the epidemic[7]. We computed expected incidence by differencing the cumulative incidence, drew a random negative binomial deviate with this mean, and used a binomial sample with probability equal to the ascertainment ratio a to get the number of observed cases. Throughout, we used a shape parameter of $s = 2$ and a final epidemic size

of 10^5 for the Richards curve, and a negative binomial dispersion parameter $k = 5$. We varied the reporting period ($\Delta t = \{1, 7\}$); starting incidence ($I_0 = \{20, 40\}$); epidemic growth rate ($r = 0.01$ to 0.08 per day); and ascertainment ratio (a from 0.05 to 0.6). We ran each simulation for 100 days and used `asymptor` to compute bounds on the ascertainment ratio.

The authors indicated (pers. comm.) that they intended the estimator to be used at the beginning of an epidemic. Therefore we considered only sample points when the number of cases was between 5 and 500 (exclusive) and the lower bound estimator for hidden cases was greater than 1.

For each simulation run (80 in total), we computed the mean and confidence intervals for the estimated lower and upper bounds of $\hat{a}$ over time (Figure 2). As suggested by our SIR example, the bounds on $\hat{a}$ rarely include the true value, and are *largely independent of the true values of* a. The only noticeable signal arises from the bias-correction terms: simulations with lower overall case numbers (low r, low a, $\Delta t = 1$) have smaller upper bounds and larger lower bounds. In simulations without noise and with the simpler, non-bias-corrected expression for the lower bound (not shown), the lower-bound estimates of $\hat{a}$ are completely independent of a, and indeed of any parameters other than the epidemic growth rate.

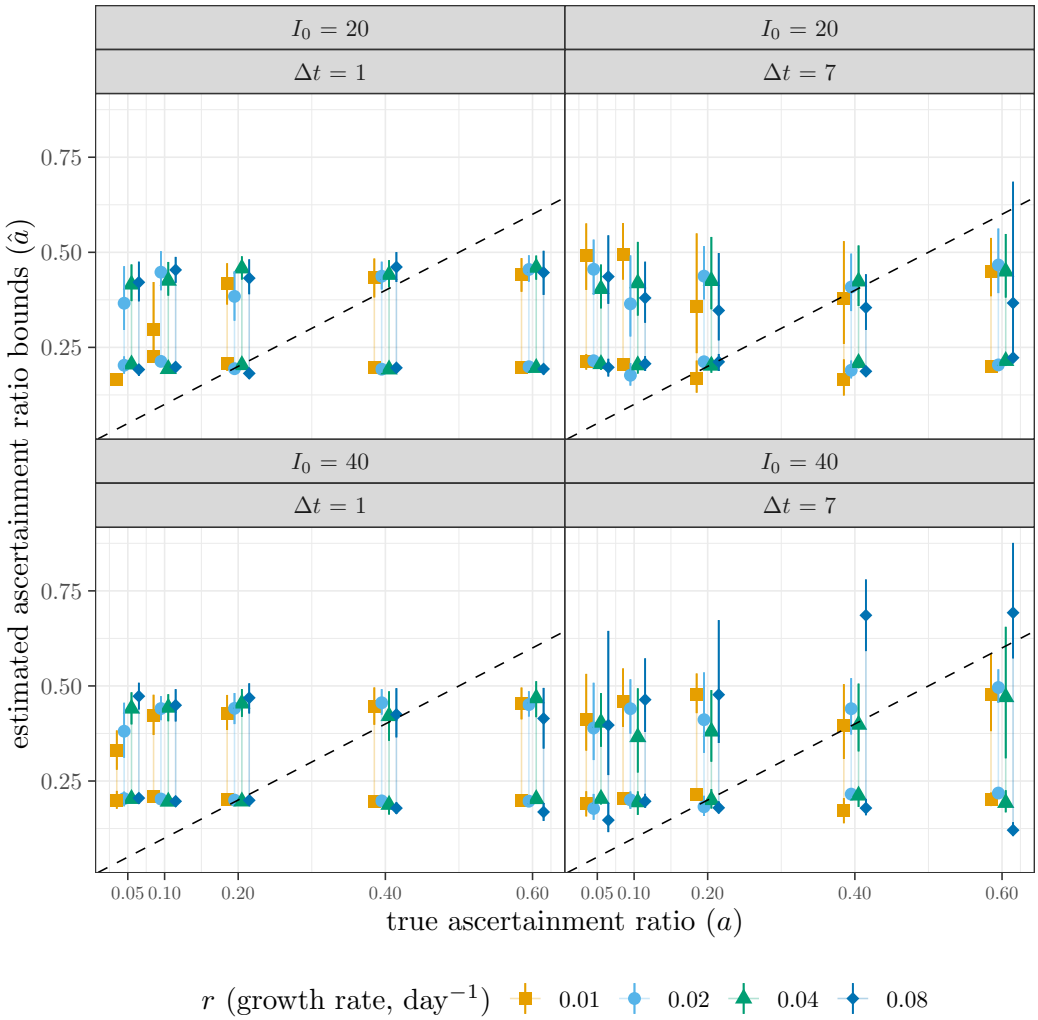

Figure 2: Ratio of lower/upper bounds to observed cases. $\Delta t = 1$, $I(0) = 40$ (results are similar for other choices of Δt and $I(0)$). Left-hand axis shows ratios of bounds to observed cases; right-hand axis shows the estimated ascertainment ratio ($1/(1 + \text{bounds ratio})$). Dashed line is the one-to-one line.

We conclude that the authors' formula appears to work well because it leads to plausible bounds on the ascertainment ratio (0.2 – 0.5) for realistic values of the epidemic growth rate, but that it is in fact nearly unrelated to the true ascertainment ratio and should not be applied to disease outbreak incidence data.

Source code for all examples is available at https://github.com/wzmli/undercount/.

References

1. Böhning, D., Rocchetti, I., Maruotti, A. & Holling, H. Estimating the undetected infections in the Covid-19 outbreak by harnessing capture–recapture methods. *International Journal of Infectious Diseases* **97**, 197–201 (2020).
2. Maruotti, A., Böhning, D., Rocchetti, I. & Ciccozzi, M. Estimating the undetected infections in the Monkeypox outbreak. *Journal of Medical Virology* 1–4 (2022) doi:10.1002/jmv.28099.
3. Rocchetti, I., Böhning, D., Holling, H. & Maruotti, A. Estimating the size of undetected cases of the COVID-19 outbreak in Europe: An upper bound estimator. *Epidemiologic Methods* **9**, (2020).
4. Royle, J. A. & Dorazio, R. M. *Hierarchical modeling and inference in ecology: The analysis of data from populations, metapopulations and communities*. (Academic Press, 2008).
5. Desenclos, J.-C. & Hubert, B. Limitations to the universal use of capture-recapture methods. *International Journal of Epidemiology* **23**, 1322–1323 (1994).
6. Gruson, H. asymptor: Estimate the lower and upper bound of asymptomatic cases in an epidemic using the capture/recapture methods (package version 1.0). (2020).
7. Ma, J., Dushoff, J., Bolker, B. M. & Earn, D. J. D. Estimating initial epidemic growth rates. *Bulletin of Mathematical Biology* **76**, 245–260 (2014).