# Evaluating undercounts in epidemics

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

19 September 2022

## Introduction

Two papers (Böhning et al. (2020) and Maruotti et al. (2022)) have promoted a formula that claims to estimate the completeness of sampling of an infectious disease epidemic, using only case reporting counts, based on a mark-recapture formula developed by Chao. We believe this formula is fundamentally wrong, and that it is impossible to estimate undercounting without a specialized sampling design or some kind of auxiliary information.

From Böhning et al. (2020):

> The idea is to apply [the Chao estimator] day-wise. We take an arbitrary day $t$. At this day we have $\Delta N(t)$ new infections. This will be viewed as $f_1$, the infected people identified just once. If we look at $\Delta N(t-1)$, then this is the count of new infections the day before. But these will still be infected at day $t$ unless they decease. So, $f_2$ [the number of infections counted twice] corresponds to $\Delta N(t-1) - \Delta D(t)$. We can ignore the number of recoveries as we are looking at infections which are very recent (notified at day $t$ or $t-1$). Hence we are able to give the estimate for the number of hidden infections at day $t$ as $H(t) = \frac{[\Delta N(t)]^2}{\Delta N(t-1) - \Delta D(t)}$ ...

Or from Maruotti et al. (2022):

> We will denote with $N(t)$ the cumulative count of infections at week $t$ where $t = t_0, \ldots, t_m$. Hence $\Delta N(t) = N(t) - N(t-1)$ are the number of new infections at week $t$ where $t = t_0, t_1, \ldots t_m$ ... Here $\Delta N(t)$ corresponds to the infected people identified just once, and $\Delta N(t-1)$ is the number of those identified twice.

This latter paper does not use a death term, and gives the full bias-corrected expression (their eq 1) as

$$\frac{\Delta N(t)(\Delta N(t)-1)}{1+\Delta N(t-1)}.$$

## Critique

This approach is not an appropriate application of these mark-recapture formulas. Why are cases identified at time $t-1$ representative of the number of cases counted twice? The fact that the same individual *could* be counted twice in the cumulative case report (for some sampling designs) is not relevant. How can comparing yesterday's count to today's provide information about the completeness of sampling, or anything other than the rate of growth of the epidemic?

In principle, it is possible to estimate the number of unobserved individuals (hidden cases) if individuals can be re-identified when observed a second time, or even with unmarked individuals given an appropriate sampling design (Royle and Dorazio 2008). 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) and/or anonymized so that we cannot identify whether a particular case/infected individual was double-counted or not. 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" (Desenclos and Hubert 1994).

**Exponential example**

Suppose an epidemic is growing exponentially. In the epidemic phase, the various measures of current epidemic size (incidence, prevalence, cumulative incidence, etc.) all grow exponentially, or geometrically if we sample in discrete time. Suppose $I(t)$ ("new cases" or incidence) is growing at a geometric rate $\lambda$ per time step, i.e. $I(t) = I(0)\lambda^t$, and suppose we are sampling a fraction $a$ (for "ascertainment") of the incidence. For simplicity, we use the simple (non-bias-corrected) estimate, without death or recovery.

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

This shows that the estimated number of hidden counts depends on the growth rate $\lambda$ as well as the ascertainment fraction $a$ — not what we wanted. The "underreporting ratio" $H(t)/\Delta N(t) = H(t)/(aI(t)) = \lambda$ is independent of the ascertainment fraction. The bias-corrected formula, which we use in our simulations below, has similar problems.

More generally, looking at the form of the expression $[\Delta N(t)]^2/\Delta N(t-1)$, we can see that the estimate of underreporting will always be equal to the rate of acceleration of observed cases ($\Delta N(t)/\Delta N(t-1)$) times the number of new cases $\Delta N(t)$, so underreporting will be assumed to be larger (proportional to observed cases) when disease incidence is increasing, and smaller when incidence is decreasing.

**Simulation example (SIR model)**

We simulate a deterministic SIR epidemic in discrete time, apply two different ascertainment fractions to derive a time series of case reports, and apply the authors' formula (including bias corrections) to estimate the number of hidden cases (Figure 1). At the peak of the epidemic, when the number of cases is large, the method predicts the same ratio of hidden cases to observed case reports regardless of the ascertainment fraction; at the beginning and end of the epidemic, the bias correction terms make underreporting appear lower in the lower-ascertainment scenario, when the number of reported cases is lower. In the high-ascertainment case ($a = 0.8$) the true ratio of hidden to reported cases is 0.25, but the average estimated ratio is 0.904 (range 0.037 – 1.093); in the low-ascertainment case ($a = 0.2$) the true ratio is 4 and the mean estimate is 0.847 (range 0.004 – 1.085).

## Simulation example (Richards equation)

Based on correspondence with one of the authors of the original papers, we ran an additional set of simulations that used a Richards curve

$$\text{cuminc}(s) = K/(1 + s\exp(-sr(t - t_{\text{infl}})))^{(1/s)}$$

to model the cumulative growth of the epidemic (Ma et al. 2014). (We reparameterized the model using the initial incidence, i.e. the derivative of the curve at $t = 0$, rather than the inflection point parameter $t_{\text{infl}}$.) We then computed the incidence by differencing the cumulative incidence, and used the ascertainment ratio $a$ to get the number of hidden cases and observed cases (multiplying by $1 - a$ and $a$ respectively). Throughout, we used a shape parameter of $s = 2$ and a final size of $10^5$ in all simulations; we varied the reporting period ($\Delta t = \{1, 7\}$); starting incidence ($I_0 = \{20, 40\}$); epidemic growth rate ($r = \{0.01, 0.02, 0.04, 0.08\}$ per day);

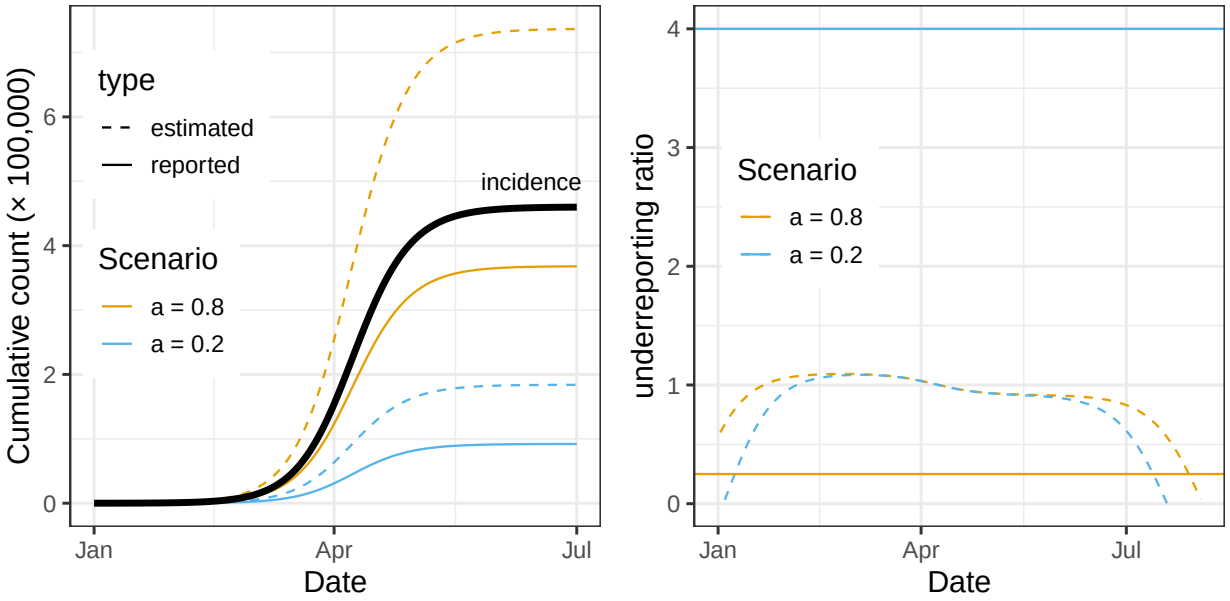


Figure 1: Simulated SIR epidemic with $\mathcal{R}_0 = 1.3$, infectious period of 3.33 days, $N = 10^6$. Left panel, cumulative numbers of reported and estimated-hidden cases along with true incidence (only one line appears because both scenarios have the same true incidence). Right panel, ratio of estimated (dashed) or true (solid) hidden cases to reported cases for each scenario.

and ascertainment ratio ($a = \{0.05, 0.1, 0.2, 0.4, 0.6\}$). We ran each simulation for 100 days. We used the `asymptor` package (Gruson 2020), which implements the lower-bound formula from Böhning et al. (2020) and the upper-bound formula from Rocchetti et al. (2020), as an independent check to make sure we had not incorrectly coded any of the formulas.

The authors indicated (pers. comm.) that they intended the estimator to be used at the beginning of an epidemic. To make sure that we were only computing bounds during reasonable times, we restricted our conclusions to periods when the number of cases was between 5 and 500 (exclusive) and when the lower bound estimator was greater than 1.

We ran deterministic simulations, to simplify comparisons. While omitting noise is unrealistic, adding (e.g.) negative binomial error for incidence and/or binomial error for ascertainment would not change any of our qualitative conclusions.

For each of the simulation runs (80 in total), we first evaluated whether the number of hidden cases estimated by `asymptor` in fact remained within the computed upper and lower bounds (Figure 2). Regardless of reporting period and initial incidence, the method always underestimated the number of hidden cases (i.e., the true number of hidden cases was above the estimated upper bound for at least one time period) when the ascertainment ratio was low ($a$ from 0.05 to 0.2) and overestimated the number of hidden cases (the true number of hidden cases was below the lower bound for at least one time period) when the ascertainment ratio was high ($a = 0.6$).

If we instead examine the estimated upper and lower bounds scaled by number of reported cases (Figure 3), we see that the scaled bounds are always between about 1 and 3. These values correspond to lower bounds on the ascertainment ratio of 0.2 to 0.25 and upper bounds of 0.5 to 0.6, *largely independent of the true values of the ascertainment ratio.* (In fact, the estimated lower bound on the ascertainment ratio is lower when the true ascertainment ratio is higher, because more cases are reported overall and thus the bias-correction terms have less effect.)

Checking the simpler, non-bias-corrected expression for the lower bound ($\Delta N(t)^2/\Delta N(t-1)$, not shown)

confirms that most of the observed variation in the estimated bounds is driven by the bias-correction terms. When the bias-correction terms are ignored, the estimated lower bound ratio during the early/exponential phase of the epidemic always corresponds to the discrete-time growth rate ($\lambda = \exp(r\Delta t)$); these estimated bounds are completely independent of the ascertainment ratio, and indeed of any parameters other than the epidemic growth rate.

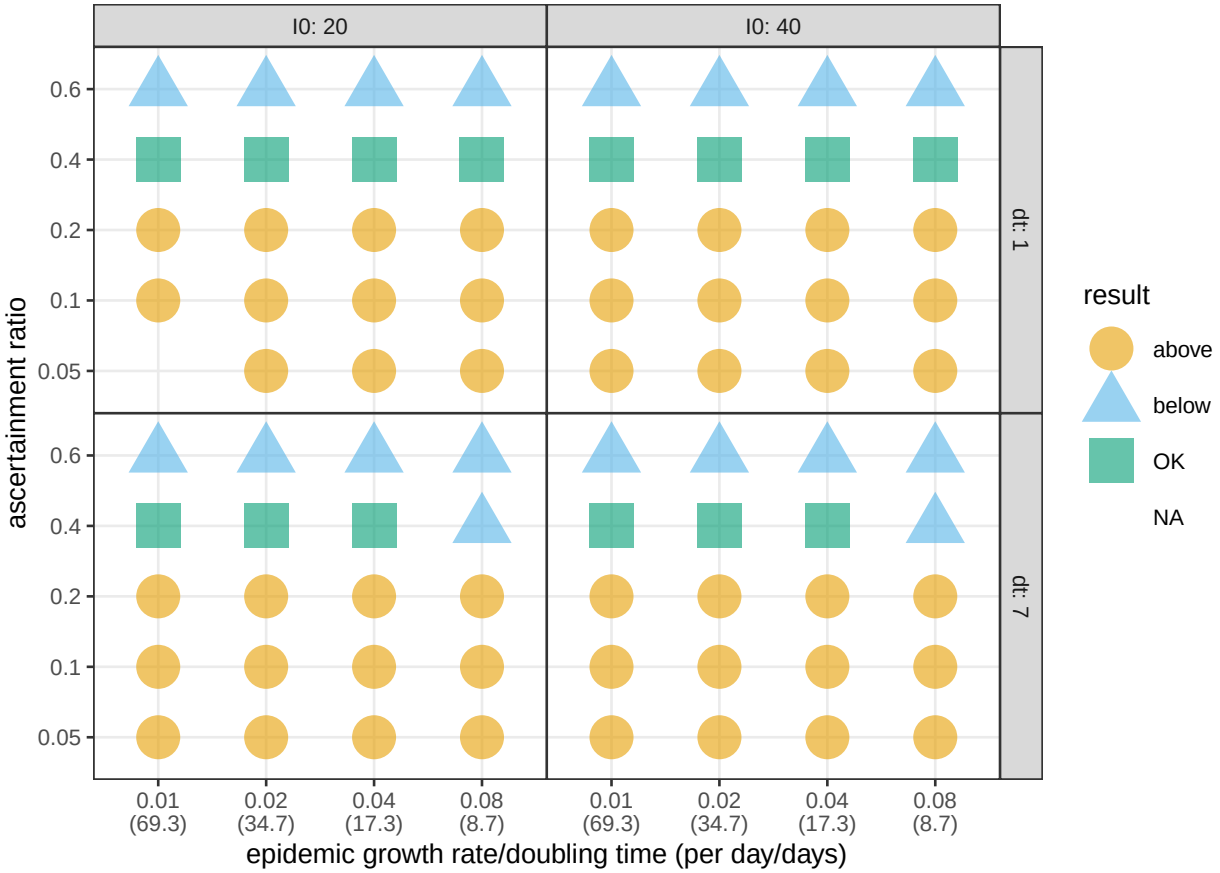


Figure 2: Outcome of lower/upper bound estimation for different parameter combinations. 'Above' (orange circle): true hidden cases > upper bound for at least one time point. 'Below' (blue triangle): true hidden cases < lower bound for at least one time point. Blank/NA: no time points met the filtering criterion. 'OK' (green square): true hidden cases were always within the estimated bounds.

We conclude that the authors' formula appears to work well because it leads to plausible bounds on the ascertainment ratio (0.25 – 0.5) for realistic values of the epidemic growth rate, but that it is in fact largely unrelated to the true ascertainment ratio and should not be applied to disease outbreak incidence data.

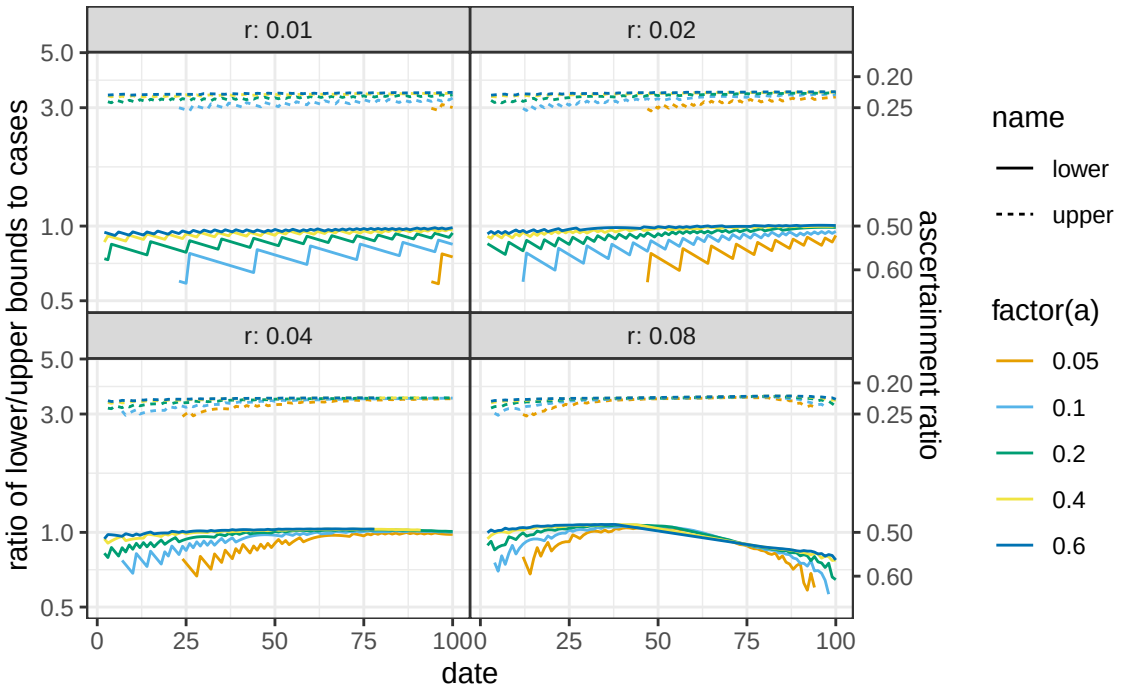


Figure 3: Ratio of lower/upper bounds to observed cases. $\Delta t = 1$, $I(0) = 40$ (results are similar for other choices of $\Delta t$ and $I(0)$). Left-hand axis shows ratios of bounds to observed cases; right-hand axis shows the estimated ascertainment ratio $(1/(1 + \text{bounds ratio}))$.

## References


Böhning, Dankmar, Irene Rocchetti, Antonello Maruotti, and Heinz Holling. 2020. "Estimating the Undetected Infections in the Covid-19 Outbreak by Harnessing Capture–Recapture Methods." *International Journal of Infectious Diseases* 97 (August): 197–201. https://doi.org/10.1016/j.ijid.2020.06.009.

Desenclos, Jean-Claude, and Bruno Hubert. 1994. "Limitations to the Universal Use of Capture-Recapture Methods." *International Journal of Epidemiology* 23 (6): 1322–3. https://doi.org/10.1093/ije/23.6.1322.

Gruson, Hugo. 2020. "asymptor: Estimate the Lower and Upper Bound of Asymptomatic Cases in an Epidemic Using the Capture/Recapture Methods (Package Version 1.0)." https://CRAN.R-project.org/package=asymptor.

Ma, Junling, Jonathan Dushoff, Benjamin M. Bolker, and David J. D. Earn. 2014. "Estimating Initial Epidemic Growth Rates." *Bulletin of Mathematical Biology* 76 (1): 245–60. https://doi.org/10.1007/s11538-013-9918-2.

Maruotti, Antonello, Dankmar Böhning, Irene Rocchetti, and Massimo Ciccozzi. 2022. "Estimating the Undetected Infections in the Monkeypox Outbreak." *Journal of Medical Virology* n/a (n/a). https://doi.org/10.1002/jmv.28099.

Rocchetti, Irene, Dankmar Böhning, Heinz Holling, and Antonello Maruotti. 2020. "Estimating the Size of Undetected Cases of the COVID-19 Outbreak in Europe: An Upper Bound Estimator." *Epidemiologic Methods* 9 (s1). https://doi.org/10.1515/em-2020-0024.

Royle, J. Andrew, and Robert M. Dorazio. 2008. *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities.* Academic Press.