Evaluating undercounts in epidemics: response to Maruotti *et al.* 2022

Michael Li[1,2], Jonathan Dushoff[3], David J. D. Earn[2], and Benjamin M. Bolker[2,3]

[1]Public Health Risk Science Division, National Microbiology Laboratory, Public Health Agency of Canada
[2]Department of Mathematics & Statistics, McMaster University
[3]Department of Biology, McMaster University

November 15, 2022

Abstract

1 Abstract

Maruotti *et al.* 2022 used a mark-recapture approach to estimate bounds on the true number of monkeypox infections in various countries. These approaches are fundamentally flawed; it is impossible to estimate undercounting based solely on a single stream of reported cases. Simulations based on a Richards curve for cumulative incidence show that, for reasonable epidemic parameters, the proposed methods estimate bounds on the ascertainment ratio of $\approx 0.2 - 0.5$ roughly *independently* of the true ascertainment ratio. These methods should not be used to estimate undercounting or ascertainment ratios.

2 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 flawed, and that it is impossible to estimate undercounting from incidence data without a specialized sampling design or some kind of auxiliary information.

The authors use mark-recapture formulas developed by Chao[4] and others[5] to estimate bounds on true population sizes based on the numbers of individuals observed multiple times. For example, the proposed estimator for the lower bound

on unobserved individuals (hidden cases) is ~~$\Delta N(t)(\Delta N(t) - 1)/(1 + \Delta N(t - 1))$~~ $\hat{H}(t) = \Delta N(t)(\Delta N(t) - 1)/(1 +$ where $\Delta N(t)$ is the number of new cases observed per reporting period; extended formulas adjust for mortality and recovery. The upper bound also involves $\Delta N(t-2)$.[1,3]

3 Critique

3.1 Logical argument

This approach misuses the mark-recapture formulas. Cases identified at time $t-1$ are claimed 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 ~~hidden cases can~~ (hidden) cases could be estimated if cases can be re-identified, or even with unmarked/unidentified cases given an appropriate sampling design.[6] 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 can provide valuable public health information in specific scenarios such as contact-tracing studies, but "one needs at least two sources of information with individual case reporting and a unique personal identifier for each case".[7] This limitation is fundamental to mark-recapture methods; standard case-reporting time series, which do not identifiably re-sample the same individuals, provide no information with which we could estimate the fraction of the population observed.

3.2 Mathematical argument

The simplest mathematical illustration of the problems with the method occurs during the exponential-growth phase of the epidemic (when the authors have suggested that their method is most appropriate). During this phase the incidence (true number of new infections: $I(t)$) grows at a rate λ per time step, i.e. $I(t) = I(0)\lambda^t$. Suppose a fraction a (the *ascertainment ratio*) of these cases are reported (i.e. a is the ratio of reported cases to the true incidence). An estimated lower bound on the number of hidden cases $\hat{H}$ can be converted to an upper bound on the estimated ascertainment ratio, $\hat{a}$. Using the simpler, non-bias-corrected formula:

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

The estimated upper bound of the ascertainment ratio $\hat{a}$ thus depends only on the epidemic growth rate; it is independent of the true ascertainment ratio. Furthermore, epidemics typically grow at rates of a few percent per reporting period; an epidemic with more than 20% growth per reporting period ($\lambda > 1.2$) would be catastrophic. Thus, the upper bound on the ascertainment ratio during the exponential phase would typically range only from about 0.45 to 0.5.

Applying a bias correction decreases the lower bound on the number of hidden cases, thus increasing the upper bound on $\hat{a}$. The results also depend on the overall number of reported cases, so the pattern is more complicated, but as we show below the estimated upper and lower bounds are still largely independent of the true ascertainment ratio.

3.3 Simulation example

We ran simulations using a Richards curve for the cumulative incidence of the epidemic~~.[8]~~;[8] this is a widely used phenomenological model for epidemic curves[9,10] and, according to the authors, is the same method they used to test their approach (pers. comm.). We computed expected incidence by differencing the cumulative incidence, drew a random negative binomial deviate with mean equal to the expected incidence, 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). These ranges encompass typical parameters of epidemic outbreaks (SARS-CoV-1, COVID-19, monkeypox, etc.), but we argue that the precise numerical values are not very important. The key aspects of a simulation are the epidemic growth rate ($\lambda = \exp(r\Delta t)$), which is the primary determinant of the ascertainment ratio bounds computed according to Maruotti *et al*'s method, and the typical number of cases reported per period, which determines the effects of the bias correction terms.

We ran each simulation for 100 days and used the R package `asymptor`[9][11] 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 1). The bounds on $\hat{a}$ rarely overlap 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 larger lower bounds and smaller upper bounds. The relationship between $\hat{a}$ and the growth rate r is barely visible as increasing values of the upper bound with r for the cases with $\Delta t = 1$ and low true ascertainment ratio; otherwise, this pattern is swamped by the effects of noise and bias correction. 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~~; some algebra shows that during the exponential growth phase of an epidemic, the (simplified) lower bound on $\hat{a}$ is exactly equal to $1/(1 + \exp(r\Delta t))$~~, as expected from the mathematical argument given above.

We conclude that the authors' formulas appear to work well because they lead to plausible bounds on the ascertainment ratio ($\approx 0.2 - 0.5$) for realistic values of the epidemic growth rate, but that they are in fact nearly unrelated to the true ascertainment ratio and should not be applied to estimate ascertainment ratios from disease outbreak incidence data.

Further details, and source code for all examples, are 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. Chao, A. Estimating population size for sparse data in capture-recapture experiments. *Biometrics* **45**, 427 (1989).

5. Alfò, M., Böhning, D. & Rocchetti, I. Upper bound estimators of the population size based on ordinal models for capture-recapture experiments. *Biometrics* **77**, 237–248 (2021).

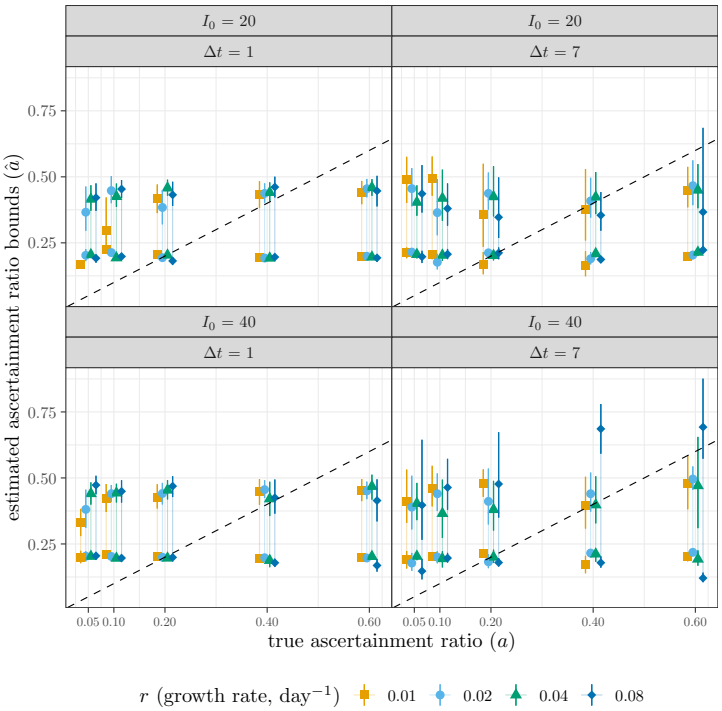

Figure 1: Comparison of true ascertainment ratio (a) to estimated lower and upper bounds of ascertainment ratio ($\hat{a}$). Dashed line is the one-to-one line (estimated = true).

6. Royle, J. A. & Dorazio, R. M. *Hierarchical modeling and inference in ecology: The analysis of data from populations, metapopulations and communities.* (Academic Press, 2008).

7. Desenclos, J.-C. & Hubert, B. Limitations to the universal use of capture-recapture methods. *International Journal of Epidemiology* **23**, 1322–1323 (1994).

8. Ma, J., Dushoff, J., Bolker, B. M. & Earn, D. J. D. Estimating initial epidemic growth rates. *Bulletin of Mathematical Biology* **76**, 245–260 (2014).

9. Chowell, G. *et al.* Using phenomenological models to characterize transmissibility and forecast patterns and final burden of Zika epidemics. *PLoS Currents* **8**, ecurrents.outbreaks.f14b2217c902f453d9320a43a35b9583 (2016).

10. Mingione, M., Ciccozzi, M., Falcone, M. & Maruotti, A. Short-term forecasts of Monkeypox cases in multiple countries: Keep calm and don't panic. *Journal of Medical Virology* (2022) doi:10.1002/jmv.28159.

11. 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).