Using next generation matrices to estimate the proportion of cases that are not detected in an outbreak

H Juliette T Unwin[1], Anne Cori[1], Natsuko Imai[1], Katy A. M. Gaythorpe[1], Sangeeta Bhatia[1], Lorenzo Cattarino[1], Christl A. Donnelly[1,2], Neil M. Ferguson[1] and Marc Baguelin[1,3]

[1] MRC Centre for Global Infectious Disease Analysis; and the Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), School of Public Health, Imperial College London.
[2] Department of Statistics, University of Oxford (CAD)
[3] Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK (MB)

Keywords: Mathematical modelling, Disease Outbreaks; Disease Transmission, Infectious; Epidemiological Methods

Abbreviations: New Zealand, NZ; next generation matrices, NGM; basic reproduction number, R_0; severe acute respiratory syndrome coronavirus 2, SARS-CoV-2.

Contact tracing, where exposed individuals are followed up to break ongoing transmission chains, is a key pillar of outbreak response for many infectious disease outbreaks, such as Ebola and SARS-CoV-2. Unfortunately, these systems are not fully effective, and cases can still go undetected as people may not know or remember all of their contacts or contacts may not be able to be traced. A large proportion of undetected cases suggests poor contact tracing and surveillance systems, which could be a potential area of improvement for a disease response. In this paper, we present a novel method for estimating the proportion of cases that are not detected during an outbreak. Our method uses next generation matrices that are parameterized by linked contact tracing and case line-lists. We use this method to investigate the proportion of undetected cases in two case studies: the SARS-CoV-2 outbreak in New Zealand during 2020 and the West African Ebola outbreak in Guinea during 2014. We estimate that only 6% of SARS-CoV-2 cases were not detected in New

Zealand (95% credible interval: 1.31 – 16.7%), but over 60% of Ebola cases were not detected in Guinea (95% credible interval: 15 - 90%).

There are many non-pharmaceutical interventions for controlling infectious disease epidemics. Some control measures, such as case isolation and safe and dignified burials for Ebola, avoid secondary cases but others, such as contact tracing, avoid tertiary cases. Measures, which avoid secondary cases, are most effective when tertiary cases are also avoided and all (or nearly all) cases are identified so that interventions can be targeted to those who are infected and those who have been exposed (1). If contact tracing is implemented well, contacts of known cases can take precautions to reduce onward transmission by limiting their contacts and isolating quickly on symptom onset (2–4). However, if many cases are not detected, outbreaks can grow rapidly as undetected cases usually infect more people than detected cases (5).

Cases or deaths may not be reported for a variety of reasons (6). Poor availability of tests at the start of an outbreak of an emerging pathogen, such as SARS-CoV-2, may mean that those with symptoms cannot be diagnosed (7). Asymptomatic individuals may also not know they are infected unless tested for other reasons, such as through contact tracing or for routine medical care (8). Undetected cases are not unique to COVID-19 and under-reporting is common in Ebola outbreaks due to barriers to accessing health care and limited hospital capacity meaning many cases are not hospitalized and are less likely to be reported (9). Many patients may not seek health care due to mistrust and if they die, may be buried without notification, leading again to those cases being missed from official lists (10).

Infectious disease analysis and modelling are important tools for managing epidemics and can help provide quantitative evidence and situational awareness to government and public health responses (11). The importance of such analyses has been highlighted by the response to the COVID-19 pandemic, which has been, to a large extent, informed by epidemic modelling e.g. (12–14). However, these models often require robust case data to

make accurate transmission predictions. Over time attempts have been made to account for under-reporting in models. Some models assume perfect reporting (15,16), however, this can lead to an underestimation of the infection rate (6). Other methods assume a constant under-reporting rate (17) or use data augmentation techniques (6). More recently, many models have switched to using death data, which was believed to be more reliable than case data, because it is more likely consistent over time and between countries (13). This is especially important for methods which are robust to constant under-reporting.

We propose using a quasi-Bayesian next generation matrix (NGM) approach in this paper to estimate the proportion of cases that are not detected in an outbreak. This method is not disease specific and is simple to implement from contact tracing and surveillance data. The calculation can also be repeated throughout the outbreak to provide time varying estimates. We present two applications of our method: the SARS-CoV-2 outbreak in New Zealand in 2020 and the 2014 Ebola epidemic in Guinea.

METHODS

NGMs are often used to calculate the basic reproduction number (the average number of secondary infections generated by a primary infection in a large fully susceptible population), R_0, from a finite number of discrete categories that are based on epidemiologically relevant traits in the population, such as infected individuals at different stages of infection (e.g. exposed and infectious) or with different characteristics (e.g. age). The NGM is a matrix which quantifies the number of secondary infections generated in each category by an infected individual in a given category. R_0 is defined as the dominant eigenvector of this matrix (18,19). Here, we stratify infected individuals using information about their contact tracing status and whether they were being followed up at the time of symptom onset to assign infection pathways and construct our NGM. We identify three types of cases: i) cases that are not detected (ND), ii) cases that are detected but not under active surveillance (NAS), and (iii) cases that are detected and under active surveillance

(AS). Contact follow-up or surveillance takes different forms for different diseases; for Ebola, a contact under active surveillance would be undergoing in-person follow-up for 21 days after their last interaction with the case (20), whereas for COVID-19, a contact under active surveillance would have been notified by contact tracers, or through a mobile phone application, and asked to self-isolate for up to 10 days (21,22).

Formulation of the NGM

For contact tracing to be fully effective, the parent (or primary) case needs to be diagnosed and, if positive, all their contacts placed under active surveillance. The parent case therefore needs to know and remember everyone they have been in close contact with whilst they have been infectious and for these contacts to be contacted. Despite a contact being recalled and reported, they may not be under active surveillance if they cannot be identified due to missing or incorrect contact details or evasion from contact tracers. We assume in our model that: i) cases that are not detected and those cases detected but not under active surveillance have the same effective reproduction number (R) and therefore on average, infect the same number of secondary cases; and ii) contacts under surveillance who become cases have a lower effective reproduction number (scaled by α) because they are rapidly isolated after the onset of symptoms. We define ϕ as the proportion of contacts recalled, γ as the proportion of contacts actively under surveillance, and π as the proportion of cases detected or "re-captured" by community surveillance

We identify 12 pathways through which individuals can become infected by three different types of cases (Figure 1). These pathways are described as follows:

1. A case that was detected (with probability π), who was infected by a case that was not detected and was therefore not under active surveillance.
2. A case that was not detected (with probability 1-π), who was infected by a case that was not detected and was therefore not under active surveillance.

3. A case that was detected (with probability π), who was infected by a case that was detected but not under surveillance, was correctly recalled as a contact (with probability ϕ) and was under active surveillance (with probability γ).
4. A case that was detected (with probability π), who was infected by a case that was detected but that was not under surveillance, was correctly recalled as a contact (with probability ϕ) but was not under surveillance (with probability 1- γ).
5. A case that was not detected (with probability 1-π) case, who was infected by a case that was detected but not under surveillance, was correctly recalled (with probability ϕ) but was not under surveillance (with probability 1- γ).
6. A case that was detected (with probability π) case, who was infected by a case that was detected but not under surveillance, that was not recalled (probability 1-ϕ).
7. A case that was not detected (with probability 1-π) case, who was infected by a case that was detected but not under surveillance, that was not recalled (probability 1-ϕ).
8. A case that was detected (with probability π), who was infected by a case that was detected and under surveillance, was correctly recalled (with probability ϕ) and was under surveillance (with probability γ).
9. A case that was detected (with probability π) case, who was infected by a case that was detected and under surveillance, was correctly recalled (with probability ϕ) but was not under surveillance (with probability 1- γ).
10. A case that was not detected (with probability 1-π), who was infected by a case that was detected and under surveillance, was correctly recalled (with probability ϕ) but was not under surveillance (with probability 1- γ).
11. A case that was detected (with probability π), who was infected by a case that was detected and under surveillance, that was not recalled (with probability 1-ϕ).
12. A case that was not detected (with probability 1-π) case, who was infected by a case that was detected and under surveillance, that was not recalled (with probability 1-ϕ).

Seven of our twelve pathways result in detected cases. The cases from pathways 3, 4, 8, and 9 are individuals on contact lists who are detected as cases whereas, the cases from pathways 1, 6, and 11 are de novo cases that are not on any contact tracing list, but which are detected via other routes such as attending a health care unit. The cases from pathways 3 and 8 are contacts who were under surveillance at the time of symptom onset, while those from pathways 4 and 9 were not under surveillance at onset. The cases resulting from the pathways 2, 5, 7, 10 and 12 are not detected by the surveillance system. We use the notation FX to denote the probability of a case stemming from pathway X, for example F1 equals $R\pi$.

If $Z_n = [ND_n, NAS_n, AS_n]^T$ is a vector of the number of each type of case for generation n, the dynamics of the model is given by:

$$Z_{n+1} = AZ_n, \qquad (1)$$

where A is our NGM that represent the potential transitions from one generation of cases to the next

$$A = R\begin{bmatrix} 1-\pi & (1-\pi)(1-\gamma\phi) & \alpha(1-\pi)(1-\gamma\phi) \\ \pi & \pi(1-\gamma\phi) & \alpha\pi(1-\gamma\phi) \\ 0 & \gamma\phi & \alpha\gamma\phi \end{bmatrix}. \qquad (2)$$

From the eigenvalues of this NGM, we can calculate the proportion of each of the three types cases (*ND, NAS* and *AS*), see Supplementary Information (SI) A. In the limit as n goes to infinity, an equilibrium is reach and the proportion of cases that are not detected, μ_{ND}, can be calculated as:

$$\begin{aligned} \mu_{ND} &= \lim_{n\to\infty} \frac{ND_n}{ND_n + NAS_n + AS_n} \\ &= \frac{(-1+\pi)\left(1+\alpha(-2+\gamma\phi)-\pi\gamma\phi+\sqrt{-2\pi(1+\alpha(-2+\gamma\phi))\gamma\phi+\pi^2\gamma\phi]^2+(-1+\alpha\gamma\phi)^2}\right)}{2(\alpha-1)}. \end{aligned} \qquad (3)$$

Linking our model to contact tracing and surveillance system data

Cases are often recorded in line-lists during disease outbreaks, where dates of testing, symptom onset and hospitalization are recorded alongside information about the age and sex of the patient. When case lists are linked to contact lists, we can derive two ratios with which we parameterize our NGM. We define r_1 as the ratio of cases who were contacts but not under surveillance versus the cases who were contacts and under surveillance and r_2 as the ratio of de novo cases (cases that were not known contacts) versus detected cases that were contacts and under surveillance.

Following the pathways in Figure 1, we expand r_1 (the ratio of cases who were contacts but not under surveillance versus the cases who were contacts and under surveillance) as $\left[\frac{F4+F9}{F3+F8}\right]$. At the equilibrium of the surveillance process (SIA), we have $ND_n = \mu_{ND} C_n$, $NAS_n = \mu_{NAS} C_n$ and $AS_n = \mu_{AS} C_n$, where $C_n = ND_n + NAS_n + AS_n$ is the total number of cases at generation n, μ_{NAS} is the proportion of cases not under active surveillance and μ_{AS} is the proportion of cases under active surveillance. Therefore,

$$\begin{aligned} r_1 &= \frac{R\phi\pi(1-\gamma)\mu_{NAS}S_n + \alpha R\phi\pi(1-\gamma)\mu_{AS}S_n}{R\phi\gamma\mu_{NAS}S_n + \alpha R\phi\gamma\mu_{AS}S_n} \\ &= \frac{(1-\gamma)\pi}{\gamma}. \end{aligned} \qquad (4)$$

We re-write this as

$$\gamma = \frac{\pi}{r_1 + \pi}. \qquad (5)$$

We also expand r_2 (the ratio of de novo cases versus detected cases that were contacts and under surveillance) as $\left[\frac{F1+F6+F11}{F3+F8}\right]$. Therefore,

$$\begin{aligned} r_2 &= \frac{R\pi(\mu_{ND}S_n + (1-\phi)\mu_{NAS}S_n + \alpha(1-\phi)\mu_{AS}S_n)}{\mathrm{R}(\phi\gamma\mu_{NAS}S_n + \alpha\phi\gamma\mu_{AS}S_n)} \\ &= \frac{\pi(\beta + (1-\phi))}{\phi\gamma}, \end{aligned} \qquad (6)$$

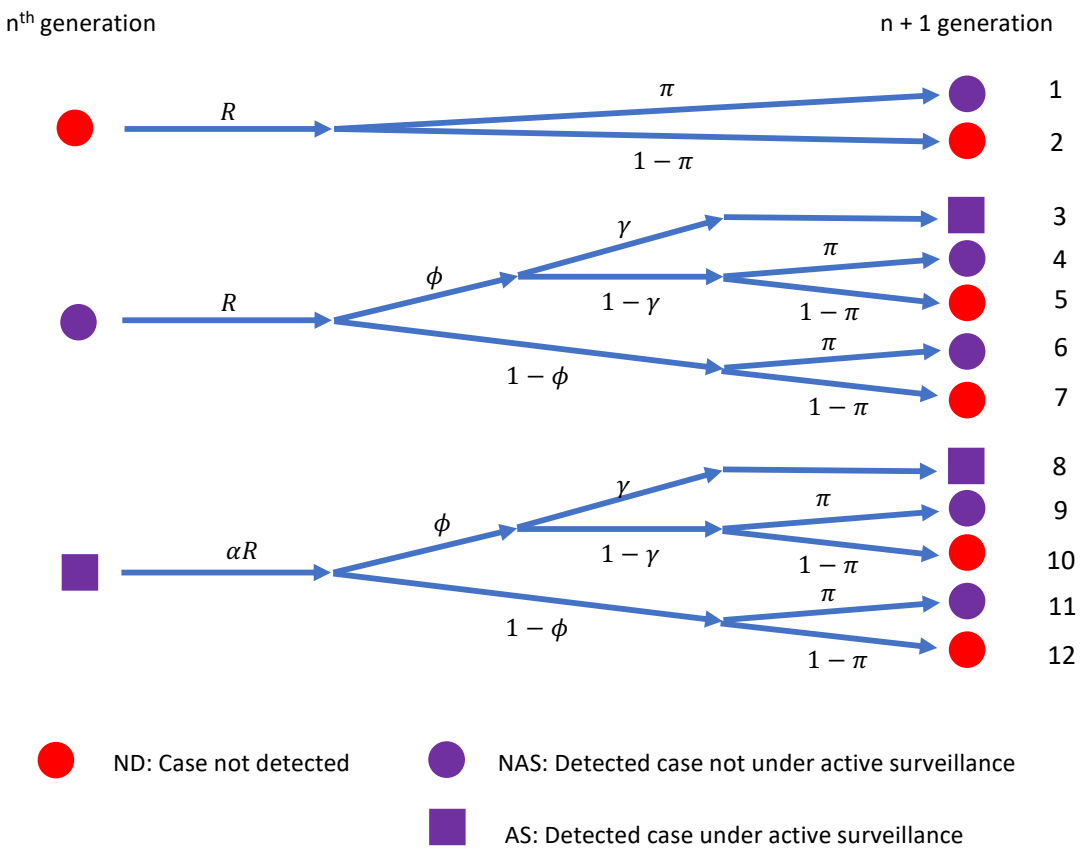

Figure 1: Potential pathways for a three-state model of Ebola surveillance (ND, AS, NAS). R is the effective reproduction number, α is the scaling of the reproduction number due to active surveillance (rapid isolation upon symptom onset), ϕ is the proportion of contacts recalled and reported by a case, γ is the proportion of contacts actively under surveillance, and π is the proportion of cases detected or "re-captured" by community surveillance. We assume that all cases under active surveillance are detected. The colouring and shape of the end points of the paths are described as follows: red circle - any case that was not detected (so cannot be under active surveillance), purple circle - an eventually detected case that was not under active surveillance at the time of symptom onset (e.g. a contact of an earlier case lost to follow-up or who refused follow-up), purple square: a detected case that was under active surveillance at the time of symptom onset (e.g. a contact of a previously detected case, correctly recalled and reported, and under surveillance).

where $\beta = \frac{\mu_{ND}}{\mu_{NAS}+\alpha\mu_{AS}}$. This can be rewritten as

$$\mu_{ND} = \nu(\mu_{NAS} + \alpha\mu_{AS}),\ \nu = \frac{r_2\phi\gamma}{\pi} - 1 + \phi. \qquad (7)$$

Figure 2 illustrates the dependencies between these two ratios and the parameters in our model in a directed acyclic graph where the green nodes are our data, blue nodes are model parameters and white nodes are calculated parameters.

In addition to equations (5) and (7), we also have three more relationships that we can use to parameterize our NGM: the proportions of each type of case (μ_{ND}, μ_{NAS} and μ_{AS}) that are found using the leading eigenvector of the NGM (see SIA). We therefore have five equations and seven unknown parameters (π, α, ϕ, γ, μ_{ND}, μ_{NAS}, μ_{AS}). If we fix two, we can derive estimates of the five remaining parameters. For example, if we fix π and ϕ, we can then estimate γ directly from equation 5. As shown by the derivation in SIA, the parameters μ_{ND}, μ_{NAS} and μ_{AS} are defined as a function of α, the scaling on the level of transmission when under surveillance ($\alpha = 0$ perfect control and $\alpha = 1$ no impact of surveillance on transmission). Equation 7 also expresses μ_{ND} as a function of α, μ_{NAS} and μ_{AS} (the proportions of the two contact traced groups). By replacing μ_{NAS} and μ_{AS} by their eigenvector relationships, we can express μ_{ND} as a function of α only. Therefore, the system can be solved finding the intersection of the (α, μ_1) curves defined by the two equations derived from the normalized eigenvector solution for μ_1 and equation 7, see SIB for a graphical example.

Application to the estimation of the proportion of cases that were not detected to two case studies

SARS-CoV-2 in New Zealand 2020. Well performing contact tracing systems have been partially credited for the success of the New Zealand's (NZ) response to the SARS-CoV-2 epidemic in 2020 (23–25). NZ's Ministry of Health reported 570 locally acquired cases up

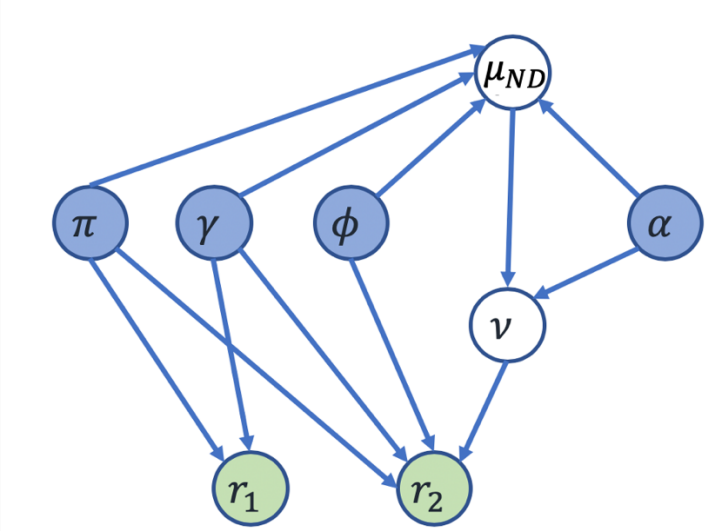

Figure 2: Directed acyclic graph showing the functional relationships of the surveillance model and the ratios observed in the surveillance. The blue nodes represent the parameters of the model that we want to infer (π is the proportion of cases detected or "re-captured" by community surveillance; γ is the proportion of contacts actively under surveillance; ϕ is the proportion of contacts recalled by a case and α is the scaling of reproduction number due to active surveillance (rapid isolation upon symptom onset)). The green terminal nodes are the potentially observable data (r_1 is the ratio of cases who were contacts but not under surveillance versus the cases who were contacts and under surveillance; and r_2 as the ratio of de novo cases versus detected cases that were contacts and under surveillance. The white nodes are our calculated terms (μ_{ND} is the proportion of cases that are not detected; and ν relates the proportion of not detected cases to the other two types of cases). The arrows show the direction of the dependence.

until 14th December 2020 that had an epidemiological link to a previous case and 90 cases without an epidemiological link (26). We assume that 80% of contacts were under active surveillance, since this was determined as the minimum requirement for the NZ system (22). Therefore, we estimate 456 cases were under active surveillance and 114 cases were not. This makes $r_1 = 0.25$ and $r_2 = 0.20$.

Ebola in Guinea 2014 We use data from Dixon et al. (27), which present contact tracing outcomes from two prefectures in Guinea between the 20th September and 31st December 2014. The authors found that only 45 cases out of 152 were registered as contacts of known cases across Kindia and Faranah prefectures.

Since there is little published data, we consider two scenarios based on different assumptions about r_1 (ratio of contacts not under active surveillance versus contacts under active surveillance).

1) We assume r_1 is equal to 0.2 (five times as many contacts under active surveillance than not under active surveillance, or 5 out of 6 contacts are under active surveillance). This is based on data from Liberia in 2014 and 2015 where, during the same epidemic as Guinea, 27936 contacts were not under active surveillance, whereas 167419 were (28). Since we know the total number of cases on the contact tracing list, 45, and assume $r_1 = 0.2$, we estimate the number of contacts under active surveillance to be 38 (denominator of r_2). The number of people not on the contact list for the two regions was 107 (numerator of r_2). Therefore, r_2 is equal to 2.85.

2) We assume r_1 is equal to 0.5 (twice as many contacts under active surveillance than not under active surveillance or two thirds of contacts are under active surveillance) to illustrate the impact of a slightly better surveillance system. Since we know the total number of cases on the contact tracing list, 45, and assume $r_1 = 0.5$, we estimate the number of contacts under active surveillance to be 30 (denominator of r_2). Therefore, r_2 is equal to 3.57.

We estimated the proportion of cases that are not detected using a quasi-Bayesian framework for both case studies. We sampled 100,000 values from $[0,1]^2$ uniformly for (π, ϕ), which is comparable to assuming a uniform prior distribution, and computed the other parameters $(\gamma, \alpha, \mu_{ND})$ if a solution was viable. We note that there is no solution for some values of (π, ϕ), (see SIB). Since we found our viable parameter space to be convex,

the mean parameter values calculated from our sampling may be a poor central estimate. Therefore, we define our central estimate as the solution for the point in our viable parameter space (π, ϕ) that is furthest from the boundary of our central region. We calculate this using the *polylabelr* R package (29). Our credible intervals reflect the values between which 95% of our viable samples lie. All code necessary to implement the analysis is included open source in the "*MissingCases*" R package on GitHub (30).

RESULTS

SARS-CoV-2 in New Zealand 2020. In Figure 3, we find that the region of feasible parameter space for SARS-CoV-2 was only 2.4% of the total space, which suggests high certainty in our parameter estimates. It is also located in the top right corner of the parameter space, where both the proportion of cases detected in the community(π) and proportion of contacts (ϕ) are high. This suggests a well-functioning and rigorous contact tracing and surveillance system in NZ, which our estimate that only 6.14% (95% CrI: 1.31 - 16.7%) of cases are not detected also suggests. All parameter estimates for this model are given in Table 1.

Ebola in Guinea 2014. We find that our two assumptions for r_1 result in different feasible parameter spaces, which overlap at high values (>75%) of the proportion of cases detected in the community (π), see Figure 4. The feasible region was again small for Scenario 1 when $r_1 = 0.2$ was 5.4% of the region $[0,1]^2$ and 4.3% for Scenario 2 when $r_1 = 0.5$. Our estimates of the proportion of Ebola cases that were not detected in Guinea was 64.5% (95% CrI: 15.5-89.3%) or 63.3% (95% CrI 15.9 – 83.0%) for our two scenarios where $r_1 = 0.2$ and $r_1 = 0.5$ respectively. These estimates are not significantly different despite our two assumptions. The corresponding model parameter estimates for both scenarios are given in Table 2. The only parameter that differs significantly between our scenario is the

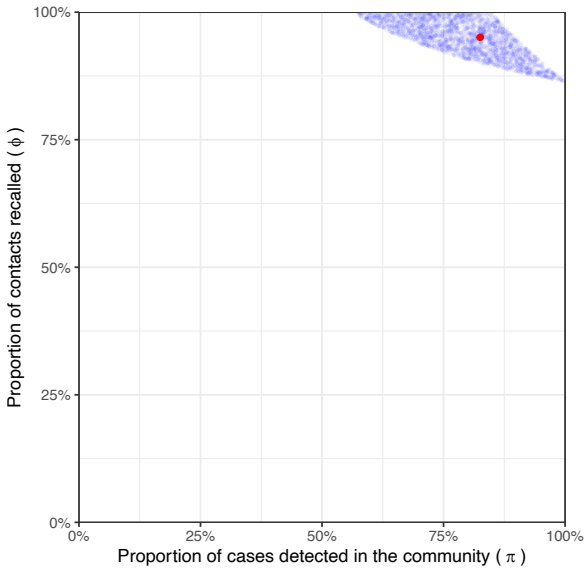

Figure 3: Region of the parameter space compatible with the observed data New Zealand. Values of π and ϕ are sampled uniformly from $[0,1]^2$ in all cases. The blue dots show our feasible samples and the red dots indicate the point inside each polygon that is most distant from the outline.

Table 1: Estimates of the parameters for SARS-CoV-2 in New Zealand

Parameter	Description	Central estimates (95% CrI)
π	Proportion of cases detected in the community	82.5% (60.6, 95.9)
α	Scaling of the reproduction number for traced cases	30.8% (2.39, 97.5)
ϕ	Proportion of contacts recalled	95.1% (87.9, 99.8)
γ	Proportion of contact under active surveillance	76.8% (70.8, 79.3)
μ_{ND}	Proportion of cases not detected	6.14% (1.31, 16.7)

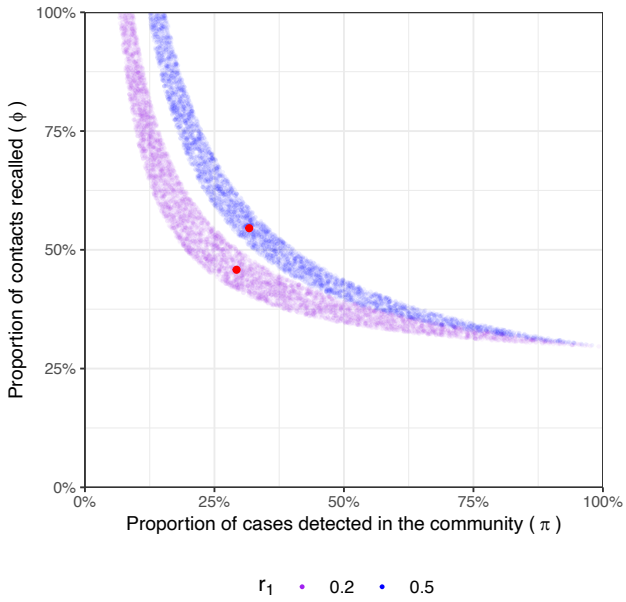

Figure 4: Region of the parameter space compatible with the observed data for the two scenarios in Guinea. Values of π and ϕ are sampled uniformly from $[0,1]^2$ in all cases. The blue dots show the feasible samples for $r_1 = 0.2$ and the purple dots for $r_1 = 0.5$. The red dots indicate the point inside each polygon that is most distant from the outline.

Table 2: Estimates of the parameters for Ebola in Guinea

		Central estimates (95% CrI)	
Parameter	Description	Scenario 1 ($r_1 = 0.2$)	Scenario 2 ($r_1 = 0.5$)
r_2	Ratio of de novo cases versus detected cases that were contacts and under surveillance	2.85	3.57
π	Proportion of cases detected in the community	29.3% (8.26, 80.4)	31.7% (14.1, 81.0)
α	Scaling of the reproduction number for traced cases	44.3% (2.06, 96.9)	47.9% (2.17, 97.3)
ϕ	Proportion of contacts recalled	45.8% (31.4, 95.3)	54.5% (32.4, 97.1)
γ	Proportion of contact under active surveillance	59.4% (29.2, 80.1)	38.8% (22.1, 61.8)
μ_{ND}	Proportion of cases not detected	64.5% (15.5, 89.3)	63.3% (15.9, 83.0)

proportion of contact under active surveillance, which is directly impacted by the ratio of contacts not under active surveillance versus contacts under active surveillance.

DISCUSSION

Contact tracing is an important control mechanism for infectious disease outbreaks. However, its efficiency depends on detecting as many cases as possible. We show in this paper that NGMs can be easily used to estimate the proportion of cases that were not detected for two different disease outbreaks. Our method requires much less data (only 5 parameters) that other methods, such as capture re-capture (10), which is an alternative method suggested for estimating under-reporting and is highly data intensive. This means that it is feasible to repeat this analysis in near real time as the epidemic unfolds.

During the West African Ebola epidemic, the WHO acknowledged that their reported case and death figures "vastly underestimate(d)" the true magnitude of the epidemic (31). We find that our estimates for the proportion of cases not detected in Guinea (64.5% (95% CrI: 15.5-89.3%) or 63.3% (95% CrI 15.9 – 83.0%) for our two scenarios where $r_1 = 0.2$ and $r_1 = 0.5$ respectively) are in line with values in the literature for neighbouring countries. The US Centers for Disease Control and Prevention (32) estimated a 40% reporting rate (60% under-reporting) from Ebola treatment unit bed data and Gignoux et al. (33) estimated a 33% (67% under-reporting) from a capture and recapture study in Liberia between June and August 2014. However, Dalziel et al. (9) suggested higher reporting rates in Sierra Leone, for example 68% (32% under reporting) in the Western Area Urban on 20 October 2014 using burial data.

Our estimates of the proportion of cases that were not detected during the SARS-Cov-2 outbreak in NZ of 6.14% (95% CrI 1.31 - 16.7) is in-line with the good health care facilities and the low community transmission of COVID-19 in NZ (26), but we did not find any estimates in literature to compare our estimates to.

A benefit of this method is that we do not just estimate the proportion of cases that were not detected but also other useful quantities that are important for managing a response such as the proportion of cases that are detected in the community. Our central estimates for routine surveillance of 82.5% (95% CrI: 60.6 – 95.9%) suggests that NZ was very effective at detecting cases in the community during the SARS-Cov-2 outbreak. In contrast, our estimates of 29.3% (95% CrI: 8.26-80.4%) and 31.7% (95% CrI 14.1-81.0%) suggests that routine surveillance could have been an area for improvement during the West African Ebola outbreak in Guinea. We also find that a higher proportion of contacts were recalled, or membered, in the NZ SARS-Cov-2 outbreak 95.1% (95% CrI: 87.9-99.8%) compared to the Ebola epidemic in Guinea (45.8% (95% CrI: 31.4-95.3%) and 54.5% (95% CrI: 32.4-97.1%)). This could reflect more trust in the NZ health service. The wide credible intervals come from the uniform sample of (π, ϕ). This is a limitation of the method but could be improved with more information about the numbers of contacts people had (ϕ) and a better understanding of the performance of the routine surveillance (π) to narrow the region.

We believe this method highlights important lessons for responding to the ongoing SARS-CoV-2 pandemic and the unfortunate inevitability of future infectious disease outbreaks. By simply linking the case line-lists and contact tracing lists, we can use the very general method from our "MissingCases" package (30) to assess under-reporting throughout an epidemic. This would help outbreak responses, especially during the early and late phases, target resources and quantify how effect their surveillance systems were. In addition, these estimates can be used to improve the accuracy of other models, such as for the time varying reproduction number, which are key tools for the outbreak response themselves.

REFERENCES

1. Salathé M, Althaus CL, Neher R, Stringhini S, Hodcroft E, Fellay J, et al. COVID-19 epidemic in Switzerland: On the importance of testing, contact tracing and isolation. Swiss Med Wkly [Internet]. 2020 Mar 19 [cited 2020 Dec 10];150(11–12).

Available from: https://doi.emh.ch/smw.2020.20225

2. Saurabh S, Prateek S. Role of contact tracing in containing the 2014 Ebola outbreak: a review. Afr Health Sci [Internet]. 2017 Mar [cited 2020 Dec 10];17(1):225–36. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29026397
3. López MA, Amela C, Ordobas M, Domínguez-Berjón MF, Álvarez C, Martínez M, et al. First secondary case of Ebola outside Africa: epidemiological characteristics and contact monitoring, Spain, September to November 2014. Eurosurveillance [Internet]. 2015 Jan 8 [cited 2020 Dec 10];20(1):21003. Available from: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES2015.20.1.21003
4. Smith CL, Hughes SM, Karwowski MP, Chevalier MS, Hall E, Joyner SN, et al. Addressing needs of contacts of Ebola patients during an investigation of an Ebola cluster in the United States - Dallas, Texas, 2014. [Internet]. Vol. 64, MMWR. Morbidity and mortality weekly report. 2015 [cited 2020 Dec 10]. p. 121–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25674993%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4584687
5. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science (80-) [Internet]. 2020 May 1 [cited 2020 Dec 10];368(6490):489–93. Available from: https://science.sciencemag.org/content/368/6490/489
6. Gamado KM, Streftaris G, Zachary S. Modelling under-reporting in epidemics. J Math Biol. 2014;69(3):737–65.
7. Adalja AA, Toner E, Inglesby T V. Priorities for the US Health Community Responding to COVID-19. JAMA [Internet]. 2020 Apr 14 [cited 2020 Dec 10];323(14):1343. Available from: https://jamanetwork.com/journals/jama/fullarticle/2762690
8. Lavezzo E, Franchin E, Ciavarella C, Cuomo-Dannenburg G, Barzon L, Del Vecchio C, et al. Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'. Nature [Internet]. 2020 Aug 20 [cited 2020 Dec 10];584(7821):425–9. Available

from: http://www.nature.com/articles/s41586-020-2488-1

9. Dalziel BD, Lau MSY, Tiffany A, McClelland A, Zelner J, Bliss JR, et al. Unreported cases in the 2014-2016 Ebola epidemic: Spatiotemporal variation, and implications for estimating transmission. PLoS Negl Trop Dis. 2018;12(1):e0006161.
10. Enserink M. How many Ebola cases are there really? Sci Now [Internet]. 2014;4. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=a2h&AN=99172119&site=ehost-live
11. Rivers C, Chretien J-P, Riley S, Pavlin JA, Woodward A, Brett-Major D, et al. Using "outbreak science" to strengthen the use of models during epidemics. Nat Commun [Internet]. 2019 Dec 15 [cited 2020 Dec 10];10(1):3102. Available from: http://www.nature.com/articles/s41467-019-11067-2
12. Enserink M, Kupferschmidt K. Mathematics of life and death: How disease models shape national shutdowns and other pandemic policies. Science (80-) [Internet]. 2020 Mar 25 [cited 2020 Dec 10]; Available from: https://www.sciencemag.org/news/2020/03/mathematics-life-and-death-how-disease-models-shape-national-shutdowns-and-other
13. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature [Internet]. 2020 Aug 13 [cited 2020 Dec 10];584(7820):257–61. Available from: http://www.nature.com/articles/s41586-020-2405-7
14. Ferguson NM, Laydon D, Nedjati-Gilani G, Imai N, Ainslie K, Baguelin M, et al. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. 2020 [cited 2020 Dec 10]; Available from: https://doi.org/10.25561/77482.
15. Xia Z-Q, Wang S-F, Li S-L, Huang L-Y, Zhang W-Y, Sun G-Q, et al. Modeling the transmission dynamics of Ebola virus disease in Liberia. Sci Rep [Internet]. 2015 Nov 8 [cited 2020 Dec 11];5(1):13857. Available from:

http://www.nature.com/articles/srep13857

16. Heesterbeek JAP, Dietz K. The concept of R $_o$ in epidemic theory. Stat Neerl [Internet]. 1996 Mar 1 [cited 2020 Dec 11];50(1):89–110. Available from: http://doi.wiley.com/10.1111/j.1467-9574.1996.tb01482.x
17. Meltzer MI, Atkins CY, Santibanez S, Knust B, Petersen BW, Ervin ED, et al. Estimating the Future Number of Cases in the Ebola Epidemic — Liberia and Sierra Leone, 2014–2015 [Internet]. MMWR. Morbidity and mortality weekly report. 2014 [cited 2020 Dec 11]. Available from: http://stacks.cdc.gov/view/cdc/24900
18. Diekmann O, Heesterbeek JAP, Metz JAJ. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. J Math Biol. 1990;28(4):365–82.
19. Diekmann O, Heesterbeek JAP, Roberts MG. The construction of next-generation matrices for compartmental epidemic models. J R Soc Interface. 2010;7(47):873–85.
20. WHO. EMERGENCY GUIDELINE Implementation and management of contact tracing for Ebola virus disease [Internet]. 2015 [cited 2020 Dec 10]. Available from: https://apps.who.int/iris/bitstream/handle/10665/185258/WHO_EVD_Guidance_Contact_15.1_eng.pdf;jsessionid=1BA73A77042B8EA4BE60F9A971E37D46?sequence=1
21. NHS. If you're told to self-isolate by NHS Test and Trace - NHS [Internet]. 2020 [cited 2020 Dec 10]. Available from: https://www.nhs.uk/conditions/coronavirus-covid-19/testing-and-tracing/nhs-test-and-trace-if-youve-been-in-contact-with-a-person-who-has-coronavirus/
22. Verrall A. Rapid Audit of Contact Tracing for Covid-19 in New Zealand [Internet]. 2020 [cited 2020 Dec 14]. Available from: https://apo.org.au/sites/default/files/resource-files/2020-04/apo-nid303350.pdf
23. Baker MG, Kvalsvig A, Verrall AJ, Telfar-Barnard L, Wilson N. New Zealand's elimination strategy for the COVID-19 pandemic and what is required to make it work [Internet]. Vol. 133, New Zealand Medical Journal. 2020 [cited 2020 Dec 14]. p. 10–4. Available from: https://www.nzma.org.nz/journal-articles/new-zealands-

elimination-strategy-for-the-covid-19-pandemic-and-what-is-required-to-make-it-work

24. Jefferies S, French N, Gilkison C, Graham G, Hope V, Marshall J, et al. COVID-19 in New Zealand and the impact of the national response: a descriptive epidemiological study. Lancet Public Heal [Internet]. 2020 [cited 2020 Dec 14];5(11):e612–23. Available from: www.thelancet.com/
25. James A, Plank MJ, Hendy S, Binny R, Lustig A, Steyn N, et al. Successful contact tracing systems for COVID-19 rely on effective quarantine and isolation 4 August 2020. [cited 2020 Dec 14]; Available from: https://doi.org/10.1101/2020.06.10.20125013
26. New Zealand Ministry of Health. COVID-19: Souce of cases. 2020.
27. Dixon MG, Taylor MM, Dee J, Hakim A, Cantey P, Lim T, et al. Contact Tracing Activities during the Ebola Virus Disease Epidemic in Kindia and Faranah, Guinea, 2014 - Volume 21, Number 11—November 2015 - Emerging Infectious Diseases journal - CDC. [cited 2020 Dec 14]; Available from: https://wwwnc.cdc.gov/eid/article/21/11/15-0684_article
28. Swanson KC, Altare C, Wesseh CS, Nyenswah T, Ahmed T, Eyal N, et al. Contact tracing performance during the Ebola epidemic in Liberia, 2014-2015. Althouse B, editor. PLoS Negl Trop Dis [Internet]. 2018 Sep 12 [cited 2020 Dec 14];12(9):e0006762. Available from: https://dx.plos.org/10.1371/journal.pntd.0006762
29. Larsson J. polylabelr: Find the pole of inaccessibility (visual center) of a polygon [Internet]. 2020. Available from: https://github.com/jolars/polylabelr
30. Unwin HJT, Baguelin M. MissingCases. 2020;
31. WHO. No early end to the Ebola outbreak. 2014.
32. Centers for Disease Control and Prevention (CDC). Updating the Estimates of the Future Number of Cases in the Ebola Epidemic—Liberia, Sierra Leone, and Guinea, 2014–2015 | Ebola (Ebola Virus Disease) | CDC [Internet]. 2020 [cited 2020 Dec 15]. Available from: https://www.cdc.gov/vhf/ebola/outbreaks/2014-west-

africa/estimating-future-cases/december-2014.html

33. Gignoux E, Idowu R, Bawo L, Hurum L, Sprecher A, Bastard M, et al. Use of Capture–Recapture to Estimate underreporting of Ebola Virus disease, Montserrado County, Liberia [Internet]. Vol. 21, Emerging Infectious Diseases. 2015 [cited 2020 Dec 15]. p. 2265–7. Available from: http://wwwnc.cdc.gov/eid/article/21/12/15-0756_article.htm

ACKNOWLEDGEMENTS

All authors acknowledge funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union; and acknowledge funding by Community Jameel. HJTU also acknowledges funding from Imperial College, London for her fellowship. CAD also acknowledge the NIHR Health Protection Research Unit in Emerging and Zoonotic Infections.

Using next generation matrices to estimate the proportion of cases that are not detected in an outbreak Supplementary Information

Supplementary Information A: Convergence of the surveillance system

Here we derive the proportion cases that are undetected using a framework where the proportion of cases that are not detected does not directly depend on the values of the reproduction number, R. Following on from equation 1, we re-write

$$Z_n = \prod_{i=0}^{n} R\,B Z_0, \qquad (S1)$$

where if we assume $\xi = \gamma\phi$,

$$B = \begin{bmatrix} 1-\pi & (1-\pi)(1-\xi) & \alpha(1-\pi)(1-\xi) \\ \pi & \pi(1-\xi) & \alpha\pi(1-\xi) \\ 0 & \xi & \alpha\xi \end{bmatrix}. \qquad (S2)$$

Assuming B is diagonalisable, we can rewrite equation S2 as:

$$B = \lambda_1 P \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{\lambda_2}{\lambda_1} & 0 \\ 0 & 0 & \frac{\lambda_3}{\lambda_1} \end{bmatrix} P^{-1} \qquad (S3)$$

where λ denote eigenvalues of which λ_1 is the biggest, and

$$P = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix}, \qquad (S4)$$

is the matrix composed of the eigenvectors $[a, b, c]^\intercal$, $[d, e, f]^\intercal$ and $[g, h, i]^\intercal$ in order of descending eigenvalues. Following this decomposition, the evolution of the system can be rewritten as:

$$Z_n = \lambda_1^n \prod_{i=0}^{n} R\,P \begin{bmatrix} 1 & 0 & 0 \\ 0 & \left(\frac{\lambda_2}{\lambda_1}\right)^n & 0 \\ 0 & 0 & \left(\frac{\lambda_3}{\lambda_1}\right)^n \end{bmatrix} P^{-1} Z_0. \qquad (S5)$$

We define

$$\begin{bmatrix} j \\ k \\ l \end{bmatrix} = P^{-1} Z_0 \qquad (S6)$$

as the vector of the initial conditions in the coordinate system defined by the P^{-1} matrix.

Therefore, we can re-write equation S5 as

$$\begin{aligned} Z_n &= \lambda_1^n \prod_{i=0}^{n} R\,P \begin{bmatrix} 1 & 0 & 0 \\ 0 & \left(\frac{\lambda_2}{\lambda_1}\right)^n & 0 \\ 0 & 0 & \left(\frac{\lambda_3}{\lambda_1}\right)^n \end{bmatrix} \begin{bmatrix} j \\ k \\ l \end{bmatrix}, \\ &= \lambda_1^n \prod_{i=0}^{n} R\,P \begin{bmatrix} j \\ \left(\frac{\lambda_2}{\lambda_1}\right)^n k \\ \left(\frac{\lambda_3}{\lambda_1}\right)^n l \end{bmatrix}. \end{aligned} \qquad (S7)$$

This means we can define the number of each type of case (*ND, NAS* and *AS*) in generation n as

$$
\begin{aligned}
ND_n &= \left\{aj + dk\left(\frac{\lambda_2}{\lambda_1}\right)^n + gl\left(\frac{\lambda_3}{\lambda_1}\right)^n\right\}\lambda_1^n\prod_{i=0}^{n}R\,,\\
NAS_n &= \left\{bj + ek\left(\frac{\lambda_2}{\lambda_1}\right)^n + hl\left(\frac{\lambda_3}{\lambda_1}\right)^n\right\}\lambda_1^n\prod_{i=0}^{n}R\,,\\
AS_n &= \left\{cj + fk\left(\frac{\lambda_2}{\lambda_1}\right)^n + il\left(\frac{\lambda_3}{\lambda_1}\right)^n\right\}\lambda_1^n\prod_{i=0}^{n}R\,.
\end{aligned}
$$

and the proportion of missing cases

$$
\begin{aligned}
&\frac{ND_n}{ND_n + NAS_n + AS_n}\\
&= \frac{aj + dk\left(\frac{\lambda_2}{\lambda_1}\right)^n + gl\left(\frac{\lambda_3}{\lambda_1}\right)^n}{j(a+b+c) + k(d+e+f)\left(\frac{\lambda_2}{\lambda_1}\right)^n + l(g+h+i)\left(\frac{\lambda_3}{\lambda_1}\right)^n}. \qquad (S8)
\end{aligned}
$$

As $\left(\frac{\lambda_2}{\lambda_1}\right)^n < 1$ and $\left(\frac{\lambda_3}{\lambda_1}\right)^n < 1$ by construction, we have:

$$\lim_{n\to\infty}\frac{ND_n}{ND_n + NAS_n + AS_n} = \frac{a}{a+b+c} = \mu_{ND}$$

where $[a, b, c]^\intercal$ is the eigenvector associated with the biggest eigenvalue of B. This eigenvector can be found by calculating the determinant of $B - \lambda I$, were I is the 3x3 identity matrix.

The largest eigenvalue of B is equal to

$$\frac{1}{2}\left(1 - \pi\xi + \alpha\xi + \sqrt{(-1 + \pi\xi - \alpha\xi)^2 - 4(\alpha\xi - \pi\alpha\xi)}\right)$$

and corresponds to the eigenvector

$$\begin{bmatrix} (-1+\pi)\dfrac{-1+\pi\xi+\alpha\xi-\sqrt{-2\pi(1+\alpha(-2+\xi))\xi+\pi^2\xi^2+(-1+\alpha\xi)^2)}}{2\pi\xi} \\ \dfrac{1-\pi\xi-\alpha\xi+\sqrt{-2\pi(1+\alpha(-2+\xi))\xi+\pi^2\xi^2+(-1+\alpha\xi)^2}}{2\xi} \\ 1 \end{bmatrix}.$$

Therefore,

$$\mu_{ND} = \frac{(-1+\pi)(1+\alpha(-2+\xi)-\pi\xi+\sqrt{-2\pi(1+\alpha(-2+\xi))\xi+\pi^2\xi]^2+(-1+\alpha\xi)^2})}{2(\alpha-1)}. \quad (S9)$$

We now show an example of the convergence of the system. We assume $\pi = 0.83$, $\alpha = 0.30$, $\gamma = 0.77$ and $\phi = 0.95$. We chose a distribution for R to get some variability while keeping a relatively contained epidemic.

$$R \sim \mathcal{U}(0.2,2)$$

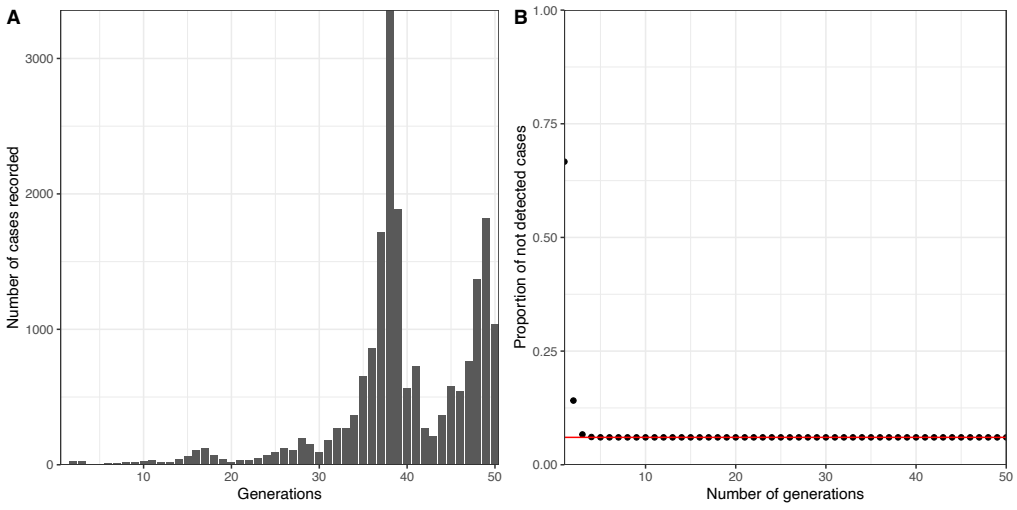

Figure S1: Evolution of simulated outbreak. A) Number of cases for each generation of the example model where $\pi = 0.83$, $\alpha = 0.30$, $\gamma = 0.77$, $\phi = 0.95$ *and* $R \sim \mathcal{U}(0.2,2)$. *B) Proportion of not detected cases over*

a number of generations. The black dots are values derived from simulation while the red line shows the value derived analytically.

We see that, as predicted by the analytic derivation, despite the unpredictable course of the epidemics (Figure S1A), the proportion of missing cases quickly reaches an equilibrium (Figure S1B). This proportion at equilibrium can be calculated using the formula derived above. We note that the convergence is exponential and thus fast proving that $\lambda_1 \gg \lambda_2$ and $\lambda_1 \gg \lambda_3$.

Supplementary Information B: Solution of the system of equations assuming parameters

We assume that $\pi = 0.83$, $\phi = 0.95$ and get $r_1 = 0.25$ and $r_2 = 0.20$ from the NZ data. Using Equation 5, we estimate $\gamma = 0.77$. Solving the system gives the following solution: $\alpha = 0.30$ and $\mu_1 = 0.06$ meaning that in this configuration, 6% of the cases are not detected and missing from the records. A graphical representation is given in Figure S2, where the solution of the system can be seen at the intersection of the two curves. No solution is found if the two curves do not intersect.

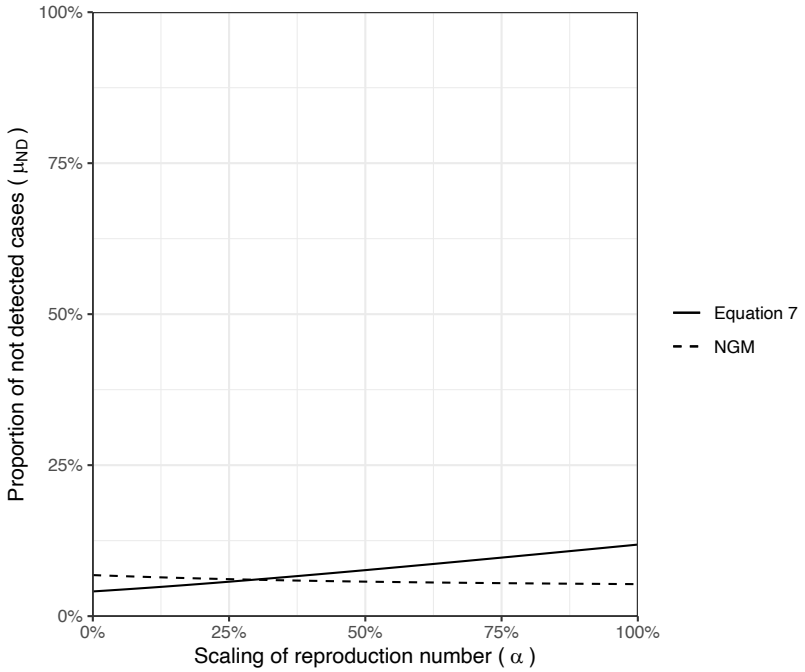

Figure S2: Derivation of α *the scaling of the reproduction number following contactt tracing and* μ_1 *the proportion of not detected cases, assuming* $\pi = 0.82$ *and* $\phi = 0.95$*. The solid line links* μ *and* α *by using the proportion of de novo cases observed vs cases who are contact and followed up. The dashed line indicates the relationship given by the dynamics of the next generation model.*