Epidemics 41 (2022) 100637

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

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

H. Juliette T. Unwin[a,b,*], Anne Cori[a,b], Natsuko Imai[a,b], Katy A.M. Gaythorpe[a,b], Sangeeta Bhatia[a,b], Lorenzo Cattarino[a,b], Christl A. Donnelly[a,b,c], Neil M. Ferguson[a,b], Marc Baguelin[a,b,d]

[a] *MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, UK*
[b] *The Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), School of Public Health, Imperial College London, UK*
[c] *Department of Statistics, University of Oxford, UK*
[d] *Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK*

ARTICLE INFO

Keywords:
Mathematical modelling
Disease outbreaks
Epidemiological methods

ABSTRACT

Contact tracing, where exposed individuals are followed up to break ongoing transmission chains, is a key pillar of outbreak response for infectious disease outbreaks. Unfortunately, these systems are not fully effective, and infections can still go undetected as people may not remember all their contacts or contacts may not be traced successfully. A large proportion of undetected infections 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 method for estimating the proportion of infections that are not detected during an outbreak. Our method uses next generation matrices that are parameterized by linked contact tracing data and case line-lists. We validate the method using simulated data from an individual-based model and then investigate two case studies: the proportion of undetected infections in the SARS-CoV-2 outbreak in New Zealand during 2020 and the Ebola epidemic in Guinea during 2014. We estimate that only 5.26% of SARS-CoV-2 infections were not detected in New Zealand during 2020 (95% credible interval: 0.243 – 16.0%) if 80% of contacts were under active surveillance but depending on assumptions about the ratio of contacts not under active surveillance versus contacts under active surveillance 39.0% or 37.7% of Ebola infections were not detected in Guinea (95% credible intervals: 1.69 – 87.0% or 1.70 – 80.9%).

1. Introduction

There are many non-pharmaceutical interventions for controlling infectious disease epidemics. Some control measures, such as case isolation and safe and dignified burials avoid secondary infections but others, such as contact tracing, avoid tertiary infections. Measures, which avoid secondary infections, are most effective when tertiary infections are also avoided and all (or nearly all) infections are identified so that interventions can be targeted (Salathé et al., 2020). 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 (Saurabh and Prateek, 2017; Lópaz et al., 2015; Smith et al., 2015). However, if many infections are not detected, outbreaks can grow rapidly as undetected infections usually infect more people than detected cases (Li et al., 2020).

Infections or deaths may not be reported for a variety of reasons (Gamado et al., 2014). 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 (Adalja et al., 2020). Asymptomatic individuals may also not know they are infected unless tested for other reasons, such as through contact tracing (Lavezzo et al., 2020). Undetected infections are not unique to SARS-CoV-2 and under-reporting is common in Ebola outbreaks due to barriers to accessing health care and limited hospital capacity (Dalziel et al., 2018). 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 (Enserink, 2014).

Infectious disease analysis and modelling are important tools for managing epidemics and can help provide quantitative evidence and situational awareness to public health responses (Rivers et al., 2019).

* Corresponding author at: MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, UK.
E-mail address: h.unwin@imperial.ac.uk (H.J.T. Unwin).

https://doi.org/10.1016/j.epidem.2022.100637
Received 1 March 2022; Received in revised form 17 September 2022; Accepted 3 October 2022
Available online 6 October 2022

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. (Enserink and Kupferschmidt, 2020; Flaxman et al., 2020; Ferguson N.M., Laydon D., Nedjati-Gilani, 2020). 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 (Xia et al., 2015; Heesterbeek and Dietz, 1996), however, this can lead to an underestimation of the infection rate (Gamado et al., 2014). Other methods assume a constant under-reporting rate (Meltzer et al., 2014), use data augmentation techniques (Gamado et al., 2014) or rely on more complex models to merge multiple data streams through evidence synthesis (Knock et al., 2021). 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 (Flaxman et al., 2020). 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 infections that are not detected in an outbreak. This method is not disease specific, is simple to implement from contact tracing and surveillance data and can be repeated throughout the outbreak to provide time varying estimates. We investigate the suitability of our method using simulated data and present two applications of our method: the SARS-CoV-2 outbreak in New Zealand (NZ) in 2020 and the Ebola epidemic in Guinea in 2014.

2. 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) e.g. (Baguelin et al., 2013). 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 eigenvalue of this matrix (Diekmann et al., 1990; Diekmann et al., 2010). They have also been used by Grantz et al. (Grantz et al., 2021) to evaluate contact tracing systems. Similarly, 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 infections: i) infections that are not detected (ND), ii) infections (or cases) that are detected but not under active surveillance (NAS), and (iii) infections (or cases) that are detected and under active surveillance (AS).

Contact follow-up or surveillance might take 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 (WHO, 2015), whereas for SARS-CoV-2 in some settings, a contact under active surveillance may be notified by contact tracers, or through a mobile phone application, and asked to self-isolate for up to 10 days (NHS, 2020; Verrall, 2020).

2.1. 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) infections 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) *AS* 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 infected contacts recalled and reported, γ as the proportion of contacts actively under surveillance, and π as the proportion of cases detected or "re-captured" by community surveillance for example by routine testing.

We identify 12 pathways through which individuals can become infected (Fig. 1). These pathways are described as follows:

1. A case that was detected (with probability π), who was infected by an infection that was not detected and was therefore not under active surveillance.
2. An infection that was not detected (with probability 1-π), who was infected by an infection 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. An infection that was not detected (with probability 1-π), 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. An infection 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. An infection 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. An infection 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 infections resulting from the pathways 2, 5, 7, 10 and 12 are not detected by the surveillance system. We use the notation F_X to denote the expected number of infections stemming from pathway X, for example F_1 equals $R\pi$..

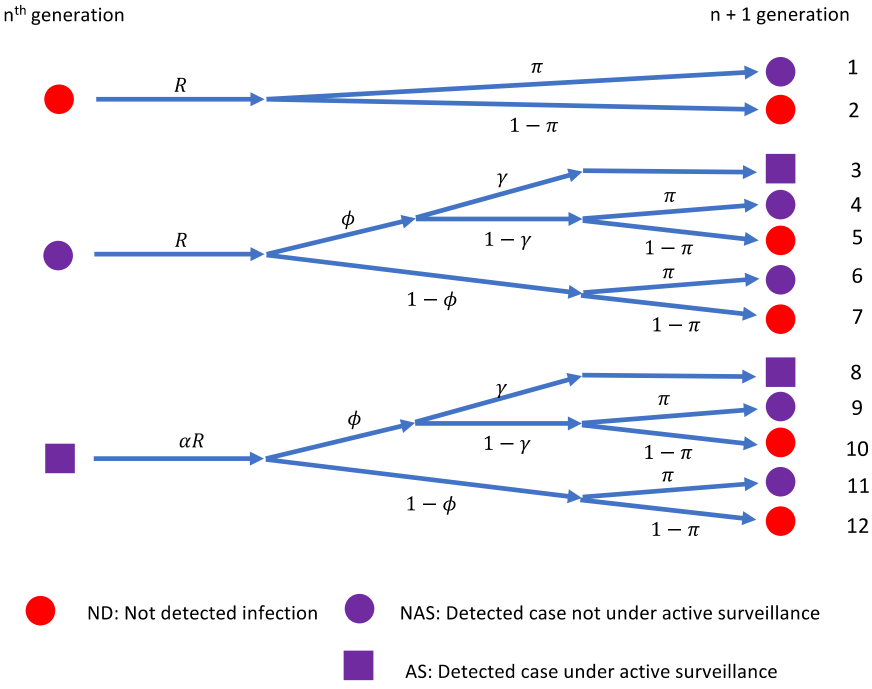
n^th generation
n + 1 generation
R
π
1 − π
R
φ
γ
1 − γ
π
1 − π
1 − φ
π
1 − π
αR
φ
γ
1 − γ
π
1 − π
1 − φ
π
1 − π
1
2
3
4
5
6
7
8
9
10
11
12
ND: Not detected infection
NAS: Detected case not under active surveillance
AS: Detected case under active surveillance

Fig. 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 infected 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 coloring 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).

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:

$$Z_{n+1} = AZ_n \tag{1}$$

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

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

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

$$\mu_{ND} = \lim_{n\to\infty}\frac{ND_n}{ND_n+NAS_n+AS_n} = \frac{(-1+\pi)\left(1+\alpha(-2+\gamma\phi)-\pi\gamma\phi+\sqrt{-2\pi(1+\alpha(-2+\gamma\phi))\gamma\phi+\pi^2\gamma^2\phi^2+(-1+\alpha\gamma\phi)^2}\right)}{2(\alpha-1)} \tag{3}$$

As shown in the calculation in the SIA and illustrated in Fig. S1, convergence to this equilibrium value is fast. A different equivalent formulation where four different types of cases are considered is shown in SIB. Here the *NAS* cases are broken down into not under active surveillance cases that are contacts, NAS^C, and those that are not, NAS^{NC}.

2.2. 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 Fig. 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 $\left[\frac{F_4+F_9}{F_3+F_8}\right]$. 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,

$$r_1 = \frac{R\phi\pi(1-\gamma)\mu_{NAS}C_n + \alpha R\phi\pi(1-\gamma)\mu_{AS}C_n}{R\phi\gamma\mu_{NAS}C_n + \alpha R\phi\gamma\mu_{AS}C_n} = \frac{(1-\gamma)\pi}{\gamma} \tag{4}$$

We re-write this as

$$\gamma = \frac{\pi}{r_1+\pi} \tag{5}$$

We also expand r_2 (the ratio of de novo cases versus detected cases that were contacts and under surveillance) as $\left[\frac{F_1+F_6+F_{11}}{F_3+F_8}\right]$. Therefore,

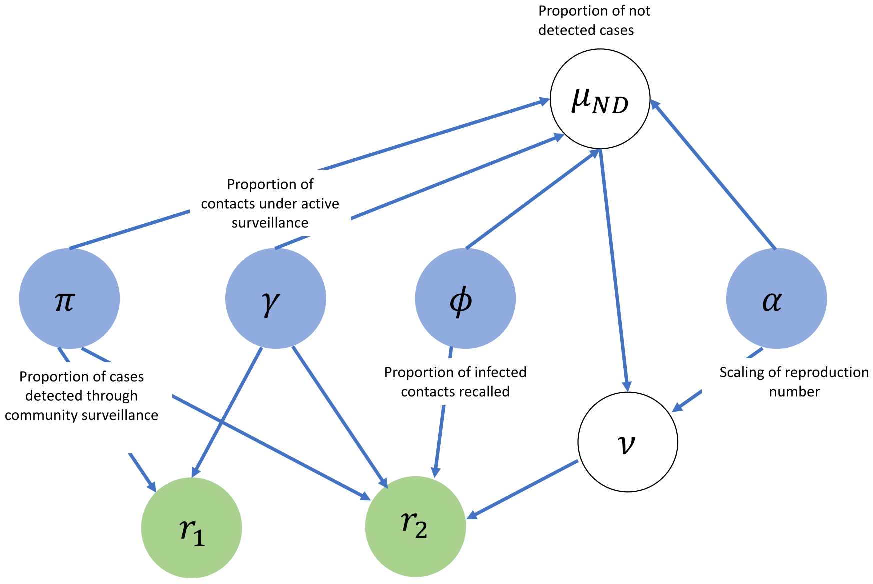
Proportion of not detected cases
μ_{ND}
Proportion of contacts under active surveillance
π
γ
ϕ
α
Proportion of cases detected through community surveillance
Proportion of infected contacts recalled
Scaling of reproduction number
ν
r_1
r_2

Fig. 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 infected contacts recalled and reported 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.

$$r_2 = \frac{R\pi(\mu_{ND}C_n + (1-\phi)\mu_{NAS}C_n \ + \ \alpha(1-\phi)\mu_{AS}C_n)}{R(\phi\gamma\mu_{NAS}C_n + \alpha\phi\gamma\mu_{AS}C_n)} = \frac{\pi(\beta + (1-\phi)\,)}{\phi\gamma}, \tag{6}$$

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

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

Fig. 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 Eqs. (5) and (7), we also have three more relationships that we can use: 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 parameters, we can then estimate the other parameters. We choose here to fix α since this could be estimated from additional data such as genetic data and π since that can be informed by information about the contract tracing system through Eq. (5).

2.3. *Application to the estimation of the proportion of infections that were not detected.*

We estimated the proportion of infections that are not detected using a quasi-Bayesian framework for each scenario. For each run of each scenario, we sampled 10,000 values from $[0,1]^2$ uniformly for (π, α), which is comparable to assuming a uniform prior distribution, and computed the other parameters (γ, ϕ, μ_{ND}) if a solution was viable. We note that there is no solution for some values of (π, α), (see SIC, Fig. S5). Our credible intervals (CrI) reflect the values between which 95% of our viable samples lie.

Simulated data. We investigate the suitability of our method using an individual-based model developed using NetLogo (Center for connected learning and computer-based modeling. NetLogo, 1999. http//ccl.northwestern.edu/NetLogo/) (see SID) for 3 scenarios:

1) Contact tracing similar to SARS-CoV-2 example in New Zealand (NZ);
2) Contract tracing similar to Ebola in Guinea;
3) Contact tracing similar to Ebola in Guinea and then improves to match the SARS-CoV-2 example in NZ after 500 days.

For each scenario, we simulated 1000 runs and sampled each run 10,000 times. Here we assumed prior knowledge about the values of π and α so uniformly sampled between 0.2 above and below the true values of π and α (see SID for parameter value). We compared the probability that the true parameters in each of our scenarios lie within the 95% CrI estimates. We consider two time periods for scenario 3, before and after the parameter change.

We also undertook a sensitivity analysis to investigate relaxing our assumption on α, where we compared the estimated values of missing cases when we varied the reduction in the scaling for a NAS case. This enabled us to evaluate the performance of the model if the NAS group did for example isolate. We compared the probability that the true value of the proportion of infections that were not detected lies within our 95% CrI for scenario one with values of α for NAS cases of 0.6 and 0.8 and 1.0 (initial scenario one). We again ran 1000 simulations of each and assumed the parameter were equal to the SARS-CoV-2 scenario.

SARS-CoV-2 in New Zealand 2020. Well performing contact tracing systems have been partially credited for the success of NZ's response to the SARS-CoV-2 epidemic in 2020 (Baker et al., 2020; Jefferies et al., 2020; James, 2020). NZ's Ministry of Health reported 570 locally acquired cases up until 14th December 2020 that had an epidemiological link to a previous case and 90 cases without an epidemiological link (New Zealand Ministry of Health, 2020). We assume that 80% of contacts were under active surveillance before diagnosis, since 80% was determined as the minimum requirement for the NZ system (Verrall, 2020). 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. (Dixon et al., 2014), 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, 27,936 contacts were not under active surveillance, whereas 167,419 were (Swanson et al., 2018). 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 worse 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 again estimated the proportion of infections that are not detected using our quasi-Bayesian framework for both case studies and took 100,000 samples for each case study, sampling π and α between 0 and 1. For the *SARS-CoV-2* example we also sampled 100,000 values for (γ,α) to investigate the impact of placing a prior distribution over different parameters. Here we sampled α uniformly between 0 and 1, but between 0 and $1/(r_1 + 1)$ for γ due Eq. (5). All code necessary to implement the analysis is included open source in the "*MissingCases*" R package on GitHub (Unwin, H.J.T., Baguelin M. MissingCases, 2020. doi: https://github.com/mrc-ide/MissingCases. Accessed 15 Dec 2020).

3. Results

3.1. Simulated data

We find that in our three scenarios, the true proportion of infections that are not detected always lie within the uncertainty intervals of the NGM estimates even in scenario 3 where our parameters are not constant (Fig. 3, Table S2). However, not all parameters perform consistently well as shown in Table S2, where γ only lies within the interval 75.4% of the time in scenario 1 and ϕ only 24.6% of the time in scenario 2. We found that performance remained similar if we reduced alpha for NAS cases (Table S3).

3.2. SARS-CoV-2 in New Zealand 2020

We estimate that only 5.26% (95% CrI: 0.245–16.0%) of cases were not detected during this wave of the SARS-CoV-2 pandemic in NZ (see Table 1 for all parameter estimates) assuming surveillance targets were met, which would correspond to a well-functioning and rigorous contact tracing and surveillance system in NZ. In Fig. 4, we find that this estimate comes from a feasible parameter space that is focused along the right-hand side of the parameter space, where the proportion of cases detected in the community (π) is high. However, we do not learn anything about the scaling in transmission for traced cases so the uncertainty intervals in the proportion of not detected infections account for this. Similar results were obtained when a prior distribution was applied over γ instead of π (Table S4).

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

Parameter	Description	Median estimates (95% CrI)
π	Proportion of cases detected in the community	84.8% (61.9, 99.2)
α	Scaling of the reproduction number for traced cases	50.2% (4.28, 98.3)
ϕ	Proportion of infected contacts recalled	91.9% (86.6, 99.5)
γ	Proportion of contacts under active surveillance	77.2% (71.2, 79.9)
μ_{ND}	Proportion of infections not detected	5.26% (0.243, 16.0)

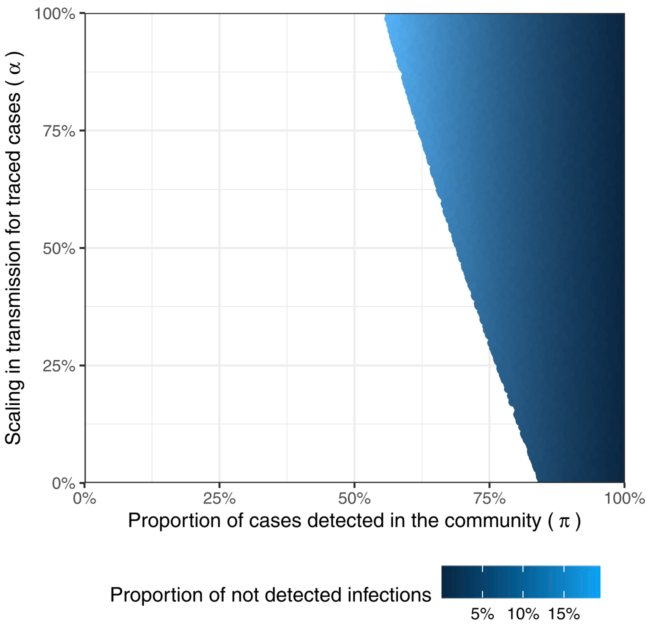

Fig. 4. Region of the parameter space compatible with the observed data from New Zealand. Values of π and α are sampled uniformly from $[0, 1]^2$. The coloured area show our feasible samples with the colour indicating the proportion of not detected infections.

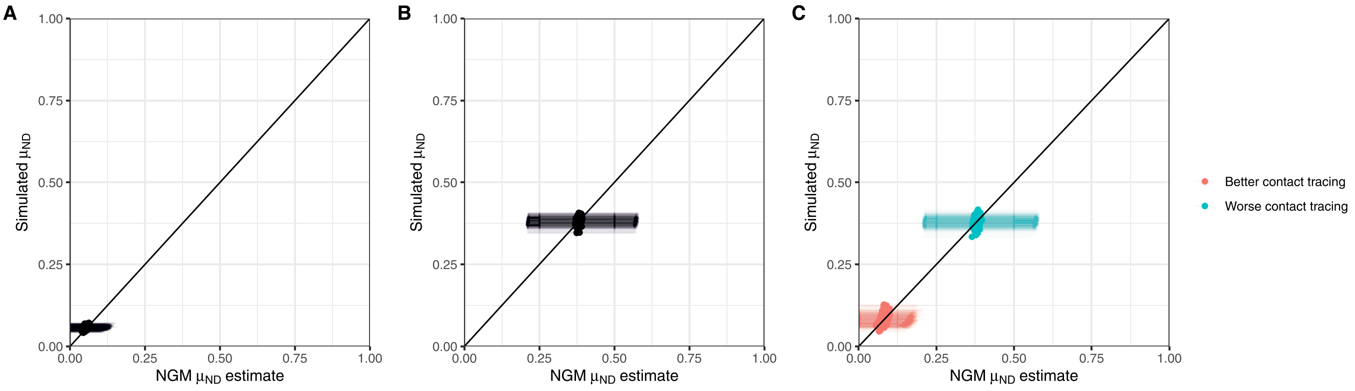

Fig. 3. Comparison of NGM estimate of proportion of infections not detected against simulated proportion for 3 scenarios. The error bars parallel to the x-axis depict the 95% CrIs from the NGM estimates. Fig. 3A shows a scenario with contact tracing like SARS-CoV-2 in NZ, Fig. 3B shows a scenario with contact tracing like Ebola from Guinea and Fig. 3C shows a scenario in which contact tracing starts like the Ebola scenario and improves to be like the SARS-CoV-2 scenario. The colors in Fig. 3C refer to the two different time periods considered (worse contact tracing: days 100–500, better contact tracing days 500–900) in our scenarios.

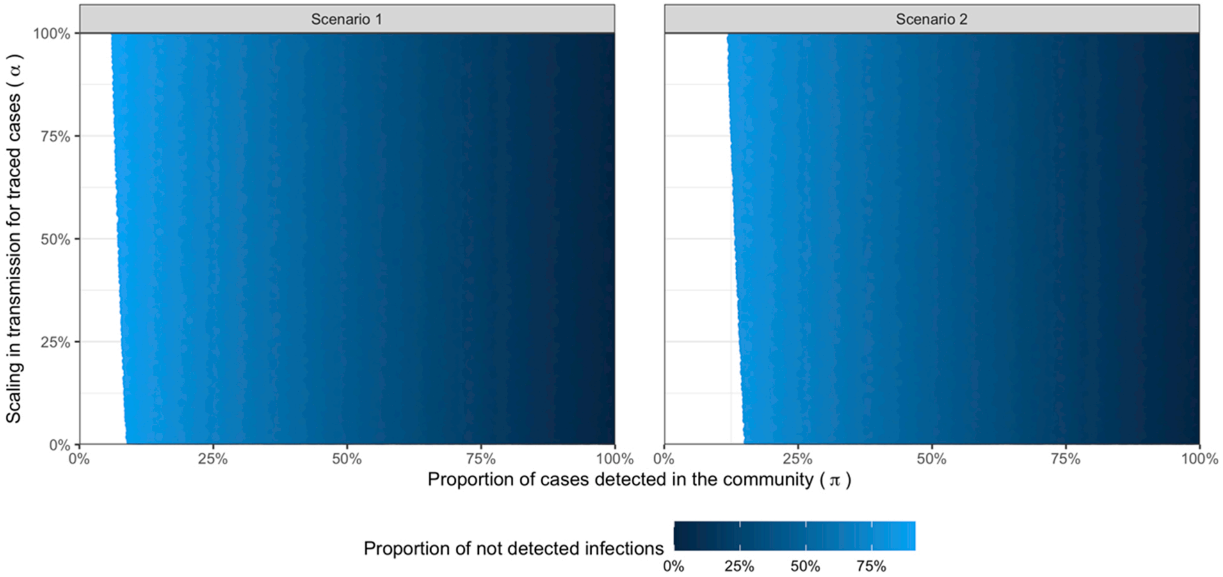

Fig. 5. Region of the parameter space compatible with the observed data for the two scenarios in Guinea. Values of π and α are sampled uniformly from $[0,1]^2$. The coloured area show our feasible samples with the colour indicating the proportion of not detected infections.

Table 2
Estimates of the parameters for Ebola in Guinea.

		Median 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	54.0% (10.1, 97.8)	57.03% (16.0, 97.9)
α	Scaling of the reproduction number for traced cases	49.8% (2.51, 97.4)	49.7% (2.54, 97.6)
ϕ	Proportion of infected contacts recalled	35.7% (29.8, 83.1)	38.6% (29.9, 89.1)
γ	Proportion of contacts under active surveillance	73.0% (33.6, 83.0)	53.3% (24.2, 66.2)
μ_{ND}	Proportion of infections not detected	39.0% (1.69, 87.0)	37.7% (1.70, 80.9)

3.3. *Ebola in Guinea 2014*

We estimate that the proportion of Ebola cases that were not detected in Guinea was 39.0% (95% CrI: 1.69–87.0%) or 37.7% (95% CrI 1.70 – 80.9%) for our two scenarios where $r_1 = 0.2$ and $r_1 = 0.5$ respectively (Fig. 5). The corresponding model parameter estimates for both scenarios are given in Table 2 and we note wide uncertainty due to our uninformative prior distributions on π and α. The only parameter that differs substantially between our scenario is the proportion of contacts under active surveillance, which is directly impacted by the ratio of contacts not under active surveillance versus contacts under active surveillance. We find that we do not learn much about the feasible values of α and π for these scenarios but as proportions of cases detected in the community fall, the proportion of not detected infections increases.

4. 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 in simulated examples and two different disease outbreaks. Our method requires much less data to parameterize our model that other methods, such as capture re-capture (Enserink, 2014), 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. As highlighted by our example, the data required for this analysis may not be available publicly and we suggest that people involved in contract systems recognise the benefit of this analysis and routinely link contact tracing data with line lists to parameterise r_1 and r_2.

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 (WHO, 2014). We find that our estimates for the proportion of cases not detected in Guinea (39.0% (95% CrI: 1.69–87.0%) or 37.7% (95% CrI 1.70 – 80.9%) 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. Dalziel et al. (Dalziel et al., 2018) suggested reporting rates in Sierra Leone of 68% (32% under reporting) in the Western Area Urban on 20 October 2014 using burial data. However, higher under reporting has also been estimated: the US Centers for Disease Control and Prevention (Centers for Disease Control and Prevention (CDC), 2020) estimated a 40% reporting rate (60% under-reporting) from Ebola treatment unit bed data and Gignoux et al. (Gignoux et al., 2015) estimated a 33% (67% under-reporting) from a capture and recapture study in Liberia between June and August 2014. We acknowledge that parameters are likely to change during an outbreak, so repeated analysis may give a better understanding of performance over a given time.

Our estimates of the proportion of cases that were not detected during the SARS-Cov-2 outbreak in NZ of 5.26% (95% CrI 0.243–16.0%) is in-line with the good health care facilities and the low community transmission of SARS-CoV-2 in NZ (New Zealand Ministry of Health, 2020), 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 infected contacts recalled and under surveillance. The wide CrI, especially in second and third simulated data scenarios and the Ebola case study, come from the uniform sample of (π, α). This is a limitation of the method but could be improved with better understanding of the performance of the routine surveillance (π) and changes in transmissibility due to contact tracing status (α), which would narrow the region in the parameter space. A second limitation is our assumption on α that only detected cases under active surveillance have a reduced transmissibility. In this simple framework, it is not possible to relax this assumption; however if additional information such as genetic or behavioral data was available, we believe this could be used to form a prior distribution on this parameter and potentially allow users to further vary the number of people NAS and ND individuals infect or improve the accuracy of some of the other parameter estimates. As we see in our sensitivity analysis, this does not impact our estimation of the proportion of infections that were not detected but potentially other parameters. A third limitation is that we do not account for differing times to locate contacts within each group, which would further vary the number of cases each case goes on to infect. A fourth limitation is that this method may not be suitable for every outbreak due to delays in classifying cases and the large numbers of individuals involved.

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 (Unwin, 2020) to assess under-reporting throughout an epidemic. This would help outbreak responses, especially during the early and late phases, target resources and quantify how effective their surveillance systems were. As Figs. 4 and 5 suggest, decreasing the scaling in reproduction number for traced cases (α) results in a lower proportion of unknown cases, which have higher transmissibility. This can be obtained by using non-pharmaceutical interventions such as isolating pre-symptomatic contacts. Additionally, improving surveillance systems through higher resource allocation and the aid of digital solutions so that higher proportions of cases are detected in the community (π), more contacts are recalled (ϕ) and more are placed under active surveillance (γ) will also reduce the proportion of not detected cases and thus lower transmission. Finally, 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.

CRediT authorship contribution statement

Marc Baguelin, Anne Cori, Neil M. Ferguson, H. Juliette T. Unwin: Designed the study. **H. Juliette T. Unwin**: Implemented the initial analysis and **Marc Baguelin** oversaw the study. **Natsuko Imai, Katy A. M. Gaythorpe, Sangeeta Bhatia, Lorenzo Cattarino, Christl A. Donnelly**: Contributed to discussion of the results and conclusions.

Competing interests

None.

Acknowledgments

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.

Data availability

All code necessary to implement the analysis is included open source in the "*MissingCases*" R package on GitHub https://github.com/mrc-ide/MissingCases.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.epidem.2022.100637.

References

Adalja, A.A., Toner, E., Inglesby, T.V., 2020. Priorities for the US Health Community Responding to COVID-19. JAMA 323, 1343. https://doi.org/10.1001/jama.2020.3413.

Baguelin, M., Flasche, S., Camacho, A., Demiris, N., Miller, E., Edmunds, W.J., 2013. Assessing optimal target populations for influenza vaccination programmes: an evidence synthesis and modelling study. PLoS Med 10, e1001527. https://doi.org/10.1371/journal.pmed.1001527.

Baker, M.G., Kvalsvig, A., Verrall, A.J., Telfar-Barnard, L., Wilson, N., 2020. New Zealand's elimination strategy for the COVID-19 pandemic and what is required to make it work. N. Z. Med. J. 133, 10–14. ⟨https://www.nzma.org.nz/journal-articles/new-zealands-elimination-strategy-for-the-covid-19-pandemic-and-what-is-required-to-make-it-work⟩. Accessed 14 Dec 2020.

Center for connected learning and computer-based modeling. NetLogo, 1999. http//ccl.northwestern.edu/NetLogo/.

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, 2020. ⟨https://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/estimating-future-cases/december-2014.html⟩. Accessed 15 Dec 2020.

Dalziel, B.D., Lau, M.S.Y., Tiffany, A., McClelland, A., Zelner, J., Bliss, J.R., et al., 2018. Unreported cases in the 2014-2016 Ebola epidemic: Spatiotemporal variation, and implications for estimating transmission. PLoS Negl. Trop. Dis. 12, e0006161.

Diekmann, O., Heesterbeek, J.A.P., Metz, J.A.J., 1990. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. J. Math. Biol. 28, 365–382.

Diekmann, O., Heesterbeek, J.A.P., Roberts, M.G., 2010. The construction of next-generation matrices for compartmental epidemic models. J. R. Soc. Interface 7, 873–885.

Dixon M.G., Taylor M.M., 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. doi:10.3201/EID2111.150684.

Enserink, M., Kupferschmidt, K., 2020. Mathematics of life and death: How disease models shape national shutdowns and other pandemic policies. Science (80-). https://doi.org/10.1126/science.abb8814.

Enserink, M., How many Ebola cases are there really? Sci, 2014; 4. ⟨http://search.ebscohost.com/login.aspx?direct=true&db=a2h&AN=99172119&site=ehost-live⟩.

Ferguson N.M., 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. doi:10.25561/77482.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H.J.T., Mellan, T.A., Coupland, H., et al., 2020. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature 584, 257–261. https://doi.org/10.1038/s41586-020-2405-7.

Gamado, K.M., Streftaris, G., Zachary, S., 2014. Modelling under-reporting in epidemics. J. Math. Biol. 69, 737–765.

Gignoux, E., Idowu, R., Bawo, L., Hurum, L., Sprecher, A., Bastard, M., et al., 2015. Use of capture–recapture to estimate underreporting of Ebola Virus disease, Montserrado County, Liberia. Emerg. Infect. Dis. 21, 2265–2267. https://doi.org/10.3201/eid2112.150756.

Grantz, K.H., Lee, E.C., D'Agostino McGowan, L., Lee, K.H., Metcalf, C.J.E., Gurley, E.S., et al., 2021. Maximizing and evaluating the impact of test-trace-isolate programs: A modeling study. PLOS Med 18, e1003585. https://doi.org/10.1371/journal.pmed.1003585.

Heesterbeek, J.A.P., Dietz, K., 1996. The concept of R $_0$ in epidemic theory. Stat. Neerl. 50, 89–110. https://doi.org/10.1111/j.1467-9574.1996.tb01482.x.

James, A., Plank, M.J., 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. doi:10.1101/2020.06.10.20125013.

Jefferies, S., French, N., Gilkison, C., Graham, G., Hope, V., Marshall, J., et al., 2020. COVID-19 in New Zealand and the impact of the national response: a descriptive epidemiological study. Lancet Public Heal 5, e612–e623. https://doi.org/10.1016/S2468-2667(20)30225-5.

Knock, E.S., Whittles, L.K., Lees, J.A., Perez-Guzman, P.N., Verity, R., FitzJohn, R.G., et al., 2021. Key epidemiological drivers and impact of interventions in the 2020 SARS-CoV-2 epidemic in England. Sci. Transl. Med. https://doi.org/10.1126/SCITRANSLMED.ABG4262.

Lavezzo, E., Franchin, E., Ciavarella, C., Cuomo-Dannenburg, G., Barzon, L., Del Vecchio, C., et al., 2020. Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'. Nature 584, 425–429. https://doi.org/10.1038/s41586-020-2488-1.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., et al., 2020. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science (80-) 368, 489–493. https://doi.org/10.1126/SCIENCE.ABB3221.

Lópaz, M.A., Amela, C., Ordobas, M., Domínguez-Berjón, M.F., Álvarez, C., Martínez, M., et al., 2015. First secondary case of Ebola outside Africa: epidemiological characteristics and contact monitoring, Spain, September to November 2014.

Eurosurveillance 20, 21003. https://doi.org/10.2807/1560-7917.ES2015.20.1.21003.

Meltzer, M.I., Atkins, C.Y., Santibanez, S., Knust, B., Petersen, B.W., Ervin, E.D., et al., 2014. Estimating the Future Number of Cases in the Ebola Epidemic — Liberia and Sierra Leone, 2014–2015. Mmwr. Morb. Mortal. Wkly. Rep. https://doi.org/10.15620/cdc.24900.

New Zealand Ministry of Health, COVID-19: Souce of cases, 2020. doi https://www.health.govt.nz/covid-19-novel-coronavirus/covid-19-data-and-statistics/covid-19-source-cases. Accessed 15 Dec 2020.

NHS, If you're told to self-isolate by NHS Test and Trace - NHS, 2020. ⟨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/⟩. Accessed 10 Dec 2020.

Rivers, C., Chretien, J.-P., Riley, S., Pavlin, J.A., Woodward, A., Brett-Major, D., et al., 2019. Using "outbreak science" to strengthen the use of models during epidemics. Nat. Commun. 10, 3102. https://doi.org/10.1038/s41467-019-11067-2.

Salathé, M., Althaus, C.L., Neher, R., Stringhini, S., Hodcroft, E., Fellay, J., et al., 2020. COVID-19 epidemic in Switzerland: On the importance of testing, contact tracing and isolation. Swiss Med. Wkly. 150. https://doi.org/10.4414/smw.2020.20225.

Saurabh, S., Prateek, S., 2017. Role of contact tracing in containing the 2014 Ebola outbreak: a review. Afr. Health Sci. 17, 225–236. https://doi.org/10.4314/ahs.v17i1.28.

Smith, C.L., Hughes, S.M., Karwowski, M.P., Chevalier, M.S., Hall, E., Joyner, S.N., et al., 2015. Addressing needs of contacts of Ebola patients during an investigation of an Ebola cluster in the United States - Dallas, Texas, 2014. Mmwr. Morb. Mortal. Wkly. Rep. 64, 121–123. ⟨http://www.ncbi.nlm.nih.gov/pubmed/25674993%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4584687⟩. Accessed 10 Dec 2020.

Swanson, K.C., Altare, C., Wesseh, C.S., Nyenswah, T., Ahmed, T., Eyal, N., et al., 2018. Contact tracing performance during the Ebola epidemic in Liberia, 2014-2015. PLoS Negl. Trop. Dis. 12, e0006762 https://doi.org/10.1371/journal.pntd.0006762.

Unwin, H.J.T., Baguelin M. MissingCases, 2020. doi: https://github.com/mrc-ide/MissingCases. Accessed 15 Dec 2020.

Verrall, A., Rapid Audit of Contact Tracing for Covid-19 in New Zealand, 2020. ⟨https://apo.org.au/sites/default/files/resource-files/2020–04/apo-nid303350.pdf⟩. Accessed 14 Dec 2020.

WHO. No early end to the Ebola outbreak, 2014. doi: ⟨https://apps.who.int/mediacentre/news/ebola/overview-20140814/en/index.html⟩. Accessed 15 Dec 2020.

WHO, EMERGENCY GUIDELINE Implementation and management of contact tracing for Ebola virus disease, 2015. ⟨https://apps.who.int/iris/bitstream/handle/10665/185258/WHO_EVD_Guidance_Contact_15.1_eng.pdf;jsessionid=1BA73A77042B8EA4BE60F9A971E37D46?sequence=1⟩. Accessed 10 Dec 2020.

Xia, Z.-Q., Wang, S.-F., Li, S.-L., Huang, L.-Y., Zhang, W.-Y., Sun, G.-Q., et al., 2015. Modeling the transmission dynamics of Ebola virus disease in Liberia. Sci. Rep. 5, 13857. https://doi.org/10.1038/srep13857.