Received: 8 March 2022 | Accepted: 9 February 2023

DOI: 10.1111/biom.13842

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Modeling COVID-19 contact-tracing using the ratio regression capture–recapture approach

Dankmar Böhning[1] | **Rattana Lerdsuwansri**[2] | **Patarawan Sangnawakij**[2]

[1]Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK

[2]Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani, Thailand

Correspondence
Dankmar Böhning, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton SO17 1BJ, UK.
Email: d.a.bohning@soton.ac.uk

Funding information
Bualuang ASEAN Chair Professor Research Fund, Thammasat University, Grant/Award Number: TUBC 02/2022

Abstract

Contact-tracing is one of the most effective tools in infectious disease outbreak control. A capture–recapture approach based upon ratio regression is suggested to estimate the completeness of case detection. Ratio regression has been recently developed as flexible tool for count data modeling and has proved to be successful in the capture–recapture setting. The methodology is applied here to Covid-19 contact tracing data from Thailand. A simple weighted straight line approach is used which includes the Poisson and geometric distribution as special cases. For the case study data of contact tracing for Thailand, a completeness of 83% could be found with a 95% confidence interval of 74%–93%.

KEYWORDS

contact tracing, count distribution modeling, Covid-19 transmission in Thailand, ratio regression, zero-truncation

1 | INTRODUCTION

This note is motivated by the effort in estimating the completeness of contact tracing (CT) in the Covid-19 epidemic. Completeness is defined as the proportion of identified cases out of the total of identified cases and cases that have been missed. Completeness of CT is of high importance in any disease outbreak and it is valuable to have methods at hand that help to determine how successful CT has been (Doyle et al., 2002). We consider as case study the first wave of the Covid-19 outbreak in Thailand. However, the methodology developed here is in principal applicable to any infectious disease outbreak involving CT.

Evaluating completeness of surveillance systems using capture–recapture methods is nowadays an established and accepted method. A recent example on HIV surveillance is given in Wesson et al. (2018). Also, capture–recapture methods are now widely applied in the social and medical sciences, beyond their origin in ecology and wildlife (Böhning et al., 2018; McCrea and Morgan, 2015), although applications in medicine have a long tradition (McKendrick, 1926). However, completeness assessment by means of capture–recapture is typically done using a multiple systems approach. Different registers or sources provide evidence on the registration of individuals with a specific condition. Using the overlap of the different lists an estimate of the number of missing individuals with the condition can be constructed. In the case of CT, there is only one list available, the list of cases with the count of contacts they had. The number of contacts of an identified case can be viewed as how often the case has been identified, and, in the capture–recapture terminology, how often the individual has been recaptured. Cases with no contacts could be those with truly no contacts or those cases where contacts could not be traced and the frequency of the latter is the target of the inference to establish the magnitude

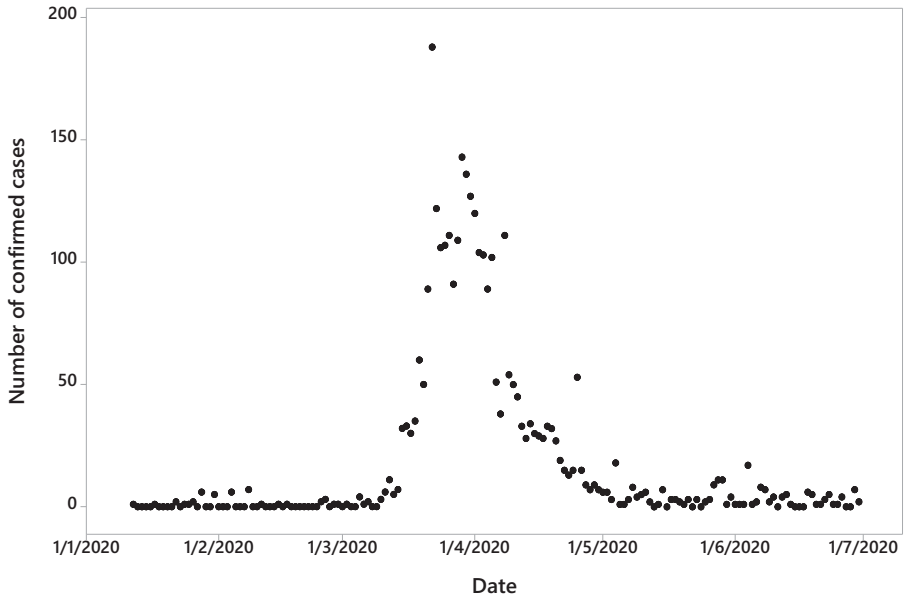

FIGURE 1 Number of Covid-19 cases by day for Thailand in the period from January 2020 to June 2020.

of the completeness of CT. A review of uni-list capture–recapture approaches is given in Wilson and Collins (1992) and an application to Scrapie surveillance in Böhning and Del Rio Vilas (2008). In the following, we describe in a case study of the Covid-19 outbreak in Thailand the uni-list CT data source.

Covid-19 contact-tracing in Thailand. Covid-19 is an infectious disease caused by a novel virus of the corona family. It was first detected in Wuhan, China, in November 2019. A huge number of Covid-19 cases were observed in various countries from December 2019 onward and the World Health Organization (WHO) announced Covid-19 a global pandemic in March 2020 (WHO, 2020). Thailand became the second country in Asia to register cases of Covid-19. The first confirmed case, who had traveled from Hubei (China), was reported on 12 January 2020. As of 30 June 2020, Thailand had 3,171 cases with 58 deaths. Figure 1 shows confirmed cases reported per day since 12 January 2020 until the end of June 2020.

The infection spread rapidly and increased in the middle of March 2020 with the highest number of 188 patients per day in the last week of March 2020. As of late May 2020, the cases were less than 10 on average. A main reason that Thailand could efficiently control the spreading of Covid-19 in a short time was that the government announced a lockdown of the entire country by the emergency decree outright on March 26, 2020. The government also imposed a nationwide curfew between 10 pm and 4 am from 3 to 30 April 2020. Extremely important approaches to prevent transmission of the infection included social distancing, quarantine, and use of face masks.

Not only the infection control measure, but CT is also a crucial tool for effectively breaking chains of transmission. CT for Covid-19 in Thailand during 2020 was an operation that involved multiple institutions including the Department of Disease Control (DDC), the Ministry of Public Health (MOPH), the Rapid Response Teams (RRT), and Village Health Volunteers who were trained during earlier major infectious disease outbreaks such as H1N1, SARS, and Avian Influenza (Kaweenuttayanon et al., 2021). Once a person is confirmed by the polymerase chain reaction (PCR) test as a Covid-19 confirmed case, teams interviewed the confirmed case to collect information about clinical history and close contacts. In general, CT methods include a mixture of disease investigation form, patient interviews, and contact verifications to map the social and work encounters of an infected individual (Ferretti et al., 2020). The identified contacts are classified as either high-risk contacts (HRC) or low-risk contacts (LRC) following investigation guidelines (MoPH, 2020). HRC is defined as a contact who is more likely to contact the virus through exposure to respiratory secretions of the confirmed case while not wearing personal protective equipment (PPE) according to standard precautionary guidelines. LRC is defined as a contact who is less likely to contract the virus from the confirmed case. This includes contacts who have not met the definition for HRC. Only high-risk contacts were quarantined in designated places. CT of a confirmed case was closed 14 days after the last successful tracing attempt and the completion of CT was marked as the date when the last contact was successfully trace. During the first wave of the Covid-19 pandemic in Thailand, from early January 2020 to June 30, 2020, 3,171 cases were confirmed. According to Thailand's regular Covid-19 CT operations, for a total of 352 (11.1%) confirmed cases CT could be completed. These are the fully traced index cases. It should

TABLE 1 Frequency f_x of fully traced index cases with exactly x contact-counts; there is also a frequency $f_0 = 11$ of fully-traced index cases with no contacts.

x	1	2	3	4	5	6	7	8	9	10
f_x	44	22	24	16	15	10	11	10	9	9
x	11	12	13	14	15	16	17	18	19	20
f_x	9	10	6	7	8	16	3	2	3	8

be noted that only HRCs were included in analysis. To be clear, in total there were 3,171 cases, but only 352 cases could be fully traced for their contacts. The reason for this was that the authorities could not cope with the amount of efforts required for tracing contacts. The 352 index cases stem from the early phase of the epidemic where follow up of contacts were still manageable. The $3,171 - 352 = 2,819$ cases (from later phases of the first wave) are not considered for this application. It is emphasized here that interest is not in estimating the total number of index cases but the completeness of CT for the population of index cases it has been applied to.

It is appropriate to explain the use of the term index case in this setting. During the epidemic period, we consider here for Thailand Covid-19 cases occur and are brought to the attention of the public health institutions (reported cases of notifications). For each such reported case, a process of investigation is started to determine to whom this case had contacts. We call this case the index case but this is not important, another name could be initial case. However, we need to separate this from the number of contacts this case had (and some very few contacts were in fact also cases, see also Section 9 on this point). This is simply the reality that took place during the epidemic (or pandemic) in Thailand, which defines our target population (notified cases with their characteristics including the number of contacts) and that we mirror here. For each of these identified cases, the number of contacts is determined and is modeled. We then focus on how complete this process of CT has been by considering index cases with no contacts. These could be those with truly no contacts or those cases where contacts could not be traced and the frequency of the latter is the target of the inference. To accomplish this task, we consider the count distribution as zero-truncated and predict from the fitted model the number of index cases with contacts which have not been traced.

Among these 341 index cases with non-zero contacts, there were 44 index cases with one contact, 22 index cases with two contacts, 24 index cases with three contacts, and so on. Table 1 illustrates the frequency f_x of index cases with exactly x contacts for $x = 1, 2, \ldots, 20$ (the entire distribution is provided in Table S1).

The largest observed count was $x = 167$ and in total 6,359 HRCs were identified, indicating the enormous effort of CT. Note that the 6,359 contacts arise from the 341 fully traced index cases. Here, interest lies in determining the completeness of CT. Capture–recapture approaches will be applied to estimate the true number of index cases with contacts missed by CT. This will be done by estimating the frequency f_0 of index cases with contacts which remained unobserved. The observed number of 11 index cases with zero contacts is not relevant for this purpose and is ignored in the further analysis.

Estimating f_0 will be accomplished by modeling the count distribution of contacts by means of ratio regression (RR) in Section 2. In Section 3, it will be shown how this can be utilized for capture–recapture modeling. Section 4 discusses parameter estimation and upper truncation. Section 5 considers potential one-inflation, Section 6 discusses confidence interval (CI) estimation and Section 7 applies the modeling to the case study. Section 8 adds a simulation study to investigate model misspecification and variance and CI performance. The paper ends with a short discussion in Section 9.

2 | COUNT DISTRIBUTION AND RATIO REGRESSION MODELING

We consider a count random variable X taking values $x \in \{0, 1, \ldots, m\}$. Here, m is a positive integer or $m = \infty$, depending on the setting. Let p_x denote the associated probability mass function $P(X = x) = p_x$ for which we seek an appropriate model. A key idea is that it frequently be easier to develop an appropriate model for p_x if we consider ratios of neighboring probabilities

$$R_x = \frac{p_{x+1}}{p_x},$$

for $x = 0, \ldots, m-1$. If $p_x = \exp(-\theta)\theta^x/x!$ is the Poisson distribution ($\theta > 0$) then $R_x = \theta/(x+1)$. If $p_x = \theta(1-\theta)^x$, where $\theta \in (0,1)$, is the geometric distribution then $R_x = 1 - \theta$. Given a sample $X_1, \ldots, X_n$ of size n, we can estimate R_x by $r_x = f_{x+1}/f_x$ where f_y is the frequency of sample elements X_i equal to y. This allows consideration which models from a candidate list might be appropriate. The ideas of using the ratios of neighboring frequencies have some tradition. The Poissonness Plot of Hoaglin (1980) is well-known and this has been also developed in Hoaglin and Tukey (1985). There is also the work by Friendly (2001) who discusses the Poissonness plot and provides an SAS macro for it. A similar graphical idea is provided by the so-called Ord plot (Ord, 1967). These concepts can be developed for the binomial distribution or geometric distribution among others. In fact, for any member of the power series family log ratios of neighboring probabilities follow a straight line if considered as a

function of the count x. These concepts have been further graphically explored in Böhning et al. (2013). The difference of the RR approach as suggested in Böhning (2016) to these early more graphical ideas is that these are now taken forward and developed into a more rigorous modeling approach. In addition, the application to CT in disease outbreak situations appears to be a novel development. Furthermore, in contrast to Böhning (2016) f_0 is estimated on the basis of the entire model (hence using all the data used for model fitting). The key equations are Equations (6) and (7) in Section 4. In Böhning (2016), f_0 was estimated using $f_1/\hat{r}_0$ where f_1 is observed and $\hat{r}_0$ is arising from the model fit. We believe that it is advantageous to base the estimate of f_0 on the entire model fit rather than on f_1 and $\hat{r}_0$ alone. Note that the RR approach is particularly suitable for zero-truncated distributions as the zero-truncated and untruncated ratio are identical. It is also suitable for distributional families, where the normalizing constant is more difficult to compute as it cancels out in the ratio.

Here, we elaborate on the connection between the probability mass function p_x and R_x. Suppose that we consider a set of candidate probability mass functions $\Pi = \{p_x(\theta)|\theta \in \Theta\}$, where Θ is some real scalar- or vector-valued interval. We call this the *P-space*. Then, there is a unique associated space generated by $\Omega = \{R_x(\theta) = p_{x+1}(\theta)/p_x(\theta)|\theta \in \Theta\}$. We call this the *R-space*. To illustrate these spaces, we consider the two-parameter Conway–Maxwell–Poisson (COM) distribution as an example which is defined by

$$p_x = \frac{\mu^x/(x!)^\lambda}{c(\mu,\lambda)},$$

where $c(\theta)$ is the normalizing constant defined by $c(\theta) = c(\mu,\lambda) = \sum_{x=0}^{\infty} \mu^x/(x!)^\lambda$ for μ and λ both positive, or $\mu \in (0,1)$ for $\lambda = 0$. For $\lambda = 1$ the COM-distribution corresponds to the Poisson and for $\lambda = 0$ it is the geometric distribution. More details on the COM-distribution including an illustration of its flexibility is given in Sellers and Shmueli (2010). The corresponding R-space of the COM-distribution is generated by

$$R_x = \frac{\mu}{(x+1)^\lambda}. \tag{1}$$

Here, we see a *first* benefit of moving into the R-space as we reach a simplified model, where the normalizing constant has canceled out. Taking logarithms on both sides of Equation (1), we achieve

$$\log R_x = \log\mu - \lambda\log(x+1) = \beta_0 + \beta_1\log(x+1). \tag{2}$$

It is convenient to think of Equation (2) as regression of R_x on $\log(x+1)$ using a log-link function. Then, $\log\mu$ corresponds to the intercept and λ to the slope. As shown in Figure S1, the three distributions are illustrated. The geometric distribution is characterized by a slope of zero, whereas the Poisson distribution has a fixed negative slope of -1. The COM-distribution has an arbitrary intercept and arbitrary negative slope. From Equation (2), we have $\mu = \exp(\beta_0)$ and no restriction on β_0 as $\mu > 0$ implies $\beta_0 \in (-\infty,\infty)$. However, we must constrain $\beta_1 < 0$ due to $\lambda > 0$.

Now, as p_x is unknown so is R_x. However, we can replace R_x by its estimate r_x and then consider more general models

$$\log r_x = \beta_0 + \beta_1 g_1(\log(x+1)) + \cdots + \beta_p g_p(\log(x+1)) + \epsilon_x, \tag{3}$$

where $g_j(.)$ are known functions for $j = 1,\ldots,p$ and ϵ_x is a random error. An example would be the simple extension of the straight line model by a quadratic term such as

$$\log r_x = \beta_0 + \beta_1\log(x+1) + \beta_2\log(x+1)^2 + \epsilon_x.$$

If we allow arbitrary regression models such as in Equation (3), the question arises if such a model corresponds to a discrete probability mass function. This is answered by the following argument. Suppose we have the fitted model $\hat{r}_x = \exp\{\hat{\beta}_0 + \hat{\beta}_1 g_1(\log(x+1)) + \cdots + \hat{\beta}_p g_p(\log(x+1))\}$, then we can use the recursive relationship $\hat{p}_{x+1} = \hat{r}_x\hat{p}_x$ or

$$\hat{p}_{x+1} = \hat{p}_0\prod_{j=0}^{x}\hat{r}_j$$

for $x = 0,\ldots,m-1$. Finally, we need to find $\hat{p}_0$. This can be accomplished by noting that

$$1 = \sum_{x=0}^{m}\hat{p}_x = \hat{p}_0\left(1 + \hat{r}_0 + \hat{r}_0\hat{r}_1 + \hat{r}_0\hat{r}_1\hat{r}_2 + \cdots + \prod_{x=0}^{m-1}\hat{r}_x\right),$$

so that $\hat{p}_0$ can be found as the inverse of

$$1 + \sum_{j=0}^{m-1}\prod_{x=0}^{j}\hat{r}_x. \tag{4}$$

Hence, any regression model of the type given in Equation (3) can be related to a unique element in the P-space. Note the importance of the link function as it guarantees that all fitted ratios are positive. This argument could have been made also using R_x instead of $\hat{r}_x$, but we prefer here the latter as it illustrates the strategic concept.

3 | CAPTURE–RECAPTURE COUNT MODELING

We are interested in applying the ideas of the previous section to zero-truncated count modeling as it typically arises in capture–recapture studies. Here, X_i represents the number of identifications of the ith member of the target population within a given time period. It is assumed that we have a sample $X_1, \dots, X_N$ of these, where N is the size of the target population. However, as a count of zero corresponds to the situation that the associated unit has not been observed, only a reduced, zero-truncated sample $X_1, \dots, X_n$ of positive counts has been observed with $n \le N$. In the application, we have in mind, the target population consists out of index cases (those with confirmed infections) who had contacts to other individuals. The population of index cases with contacts of size N can be further partitioned into a group of size n for which we observe a positive count: X_i is the number of contacts for the ith member of this sub-population of index cases, $i = 1, \dots, n$. Furthermore, we assume that there is a second group of index cases with contacts but not identified as such by the CT system. So, there are cases (counts of zeros) with truly no contacts and there are cases (counts of zeros) which had in fact some positive number of contacts. In Figure S2, the setting is illustrated. It is assumed that we have $N - n$ index cases of the latter type. There might be a number of reasons why these could not be successfully traced. Due to the mixing of zeros of different types, we truncate these altogether. A different view would be to consider this setting as a special one-point mixture and it shown in Böhning and Ogden (2021) that these can be dealt with by truncated the relevant point, here counts of zero. Hence, we consider the distribution of positive contact counts as zero-truncated as we want to estimate the number of index cases which transmitted the infection but were not successfully traced. Consequently, the focus is on estimation of the completeness of CT.

RR is also very suitable for zero-truncated count data modeling as the additional re-normalizing constant $1 - p_0$ cancels out. We have that

$$R_x = \frac{p_{x+1}}{p_x} = \frac{p_{x+1}/(1-p_0)}{p_x/(1-p_0)},$$

so that the estimate r_x refers to the same estimand R_x, independent of whether f_x arises from the zero-truncated distribution or from the associated untruncated distribution. The only difference is that $\hat{r}_0$ will be a *predicted* ratio as we have no observation for it in the zero-truncated case. Having found $\hat{r}_0, \hat{r}_1, \dots, \hat{r}_{m-1}$, we can find $\hat{p}_0$ as the inverse of Equation (4). This allows population size estimation by means of a Horvitz–Thompson-type estimator

$$\hat{N} = n/(1 - \hat{p}_0), \tag{5}$$

where n is the observed sample size of members of the target population with positive counts.

In Figure 2, an application to the CT data of Thailand is given. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version. For each case i infected with Covid-19, the index case i, its associated frequency of contacts X_i is used to create Figure 2. The largest contact count was 167, 50% of all index cases had less than 10 contacts and 75% had less than 23 contacts. Only index cases with contact counts less than 20 were used to generate Figure 2. The graph clearly indicates that the Poisson distribution is not appropriate here and that a straight line model seems feasible to capture the structure as the embedded LOWESS smoother is not substantially different from the straight line. The LOWESS smoother uses the default values (a fraction of 50% for the inclusion of data points and the tricube weight function; for more details, see Cleveland (1979)). Note that the line has a positive slope which means that the corresponding distribution is not a COM-distribution, but the construction process of going back from the *R-space* to the *P-space* will guarantee that it is a probability distribution.

4 | PARAMETER ESTIMATION AND UPPER TRUNCATION

We apply conventional least-squares estimation to find estimates of the regression coefficients. As the variance of $\log r_x = \log f_{x+1} - \log f_x$ can be estimated by $1/f_{x+1} + 1/f_x$, assuming both frequencies are positive, we use weighted least-squares estimation with weights as the inverses of $1/f_{x+1} + 1/f_x$. The choice of the weights is motivated by the common assumption in frequency table modeling that the frequencies follow a Poisson distribution, so that estimates of the asymptotic variances of the log-frequencies are given by the inverse frequencies. Figure S3 illustrates the difference between the weighted and ordinary least squares in this case. Evidently, weighted least-squares gives more weight to lower values of x.

With increasing value of x the frequency f_x becomes small. This means that the variability of r_x becomes large, to the extent that it can no longer be estimated when $f_x = 0$. For this reason, we limit regression modeling to an upper truncation point $x = T$, where $T < m$. As a consequence, we need to incorporate this limitation into the inference which we do as follows. We have now that

$$1 - \sum_{x=T+1}^{m} p_x = \sum_{x=1}^{T} p_x = p_0\left(1 + R_0 + R_0R_1 + R_0R_1R_2 + \cdots + \prod_{x=0}^{T-1} R_x\right). \tag{6}$$

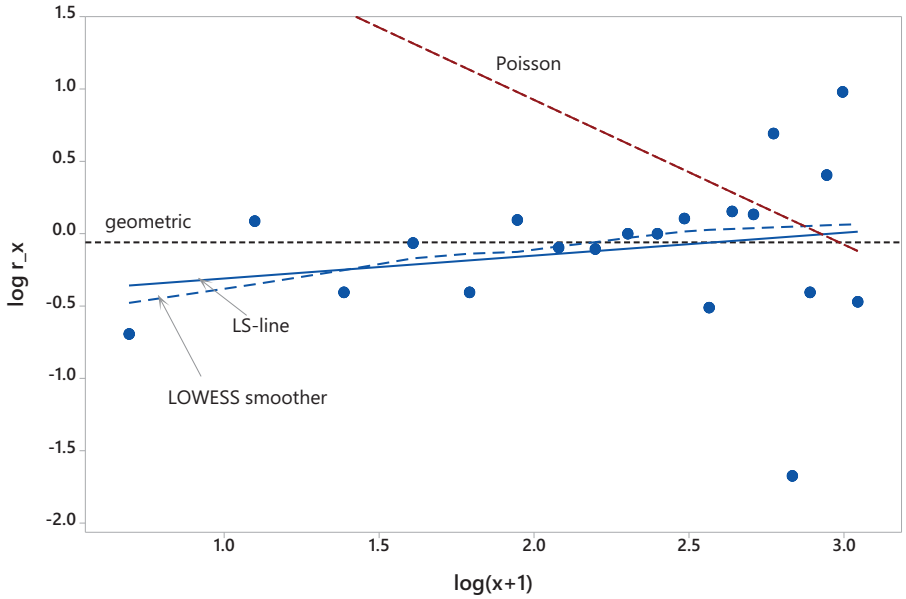

FIGURE 2 Illustration of the fitted geometric, Poisson and COM-distribution in the R-space using a log-link for the contact-tracing data from Thailand. The vertical axis represents $\log r_x$, where $r_x = f_{x+1}/f_x$ and the horizontal axis shows $\log(x+1)$ for positive integer values of x. The dashed curve represents the LOWESS smoother with the default values used in the software MINITAB which means that a fraction of 50% for the inclusion of data points is used as well as the tricube weight function. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.

We replace the left-hand side of Equation (6) by $\hat{q} = 1 - \sum_{x=T+1}^{m} f_x/n$ and R_x by the fitted value $\hat{r}_x$ of the respective regression model as before so that $\hat{p}_0$ can be found as

$$\hat{p}_0 = \frac{\hat{q}}{1 + \sum_{j=0}^{T-1} \prod_{x=0}^{j} \hat{r}_x}. \tag{7}$$

We will use then this value of $\hat{p}_0$ in Equation (7) to predict the population size N of the target population as $\hat{N} = n/(1-\hat{p}_0)$.

5 | ONE-INFLATION

Recently, one-inflation in capture–recapture modeling has attracted some attention. Here, one-inflation is defined as the occurrence of substantially more counts of ones (singletons) relative to what is predicted by the assumed model. In the dataset of CT counts from Thailand, we note that f_1 is by far the highest frequency. This alone does not speak for one-inflation. The question really is if $(1, \log r_1)$ is an influential point. Figure 3 shows an index plot of Cook's distance measure and there is no evidence of any influential point for the range of data considered for the RR.

We briefly outline the process if there were one-inflation. After truncation of the singletons, the model under consideration is fitted using $\log r_2$, ..., $\log r_{T-1}$. Removing the singletons results in no loss of generality, as it has been shown that one-inflation models can be fitted by truncating the counts of ones (Böhning and Ogden, 2021). This leads then to fitted values $\hat{r}_2, \ldots, \hat{r}_{T-1}$ and predicted values $\hat{r}_0$ and $\hat{r}_1$. From these estimates, $\hat{p}_0, \hat{p}_1, \ldots, \hat{p}_{T-1}$ can be constructed as previously. However, some modifications are required for estimating the population size. In the case that the observed sample will contain one-inflated singletons and non-inflated singletons, it is not known which singleton belongs to the inflated and which to the non-inflated part, so that the singletons are completely removed and estimation is based on the remaining counts of size $n - f_1$. This leads to a modified Horvitz–Thompson estimator

$$\hat{f}_0 = (n - f_1)\frac{\hat{p}_0}{1 - \hat{p}_1 - \hat{p}_0}, \tag{8}$$

from where the total population size estimator $\hat{N} = n + (n - f_1)\frac{\hat{p}_0}{1-\hat{p}_1-\hat{p}_0}$ follows.

6 | CONFIDENCE INTERVAL CONSTRUCTION

We are also interested in providing a $(1-\alpha)100\%$ CI for N. This is accomplished by using the following version of a semi-parametric bootstrap. Suppose we have a generic

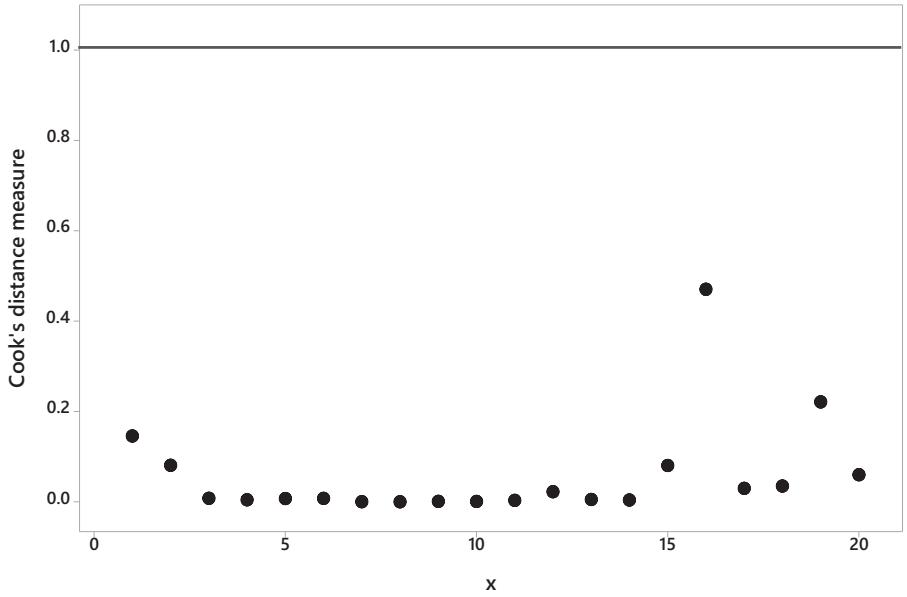

FIGURE 3 Index plot of Cook's distance measure for detecting influential points. The vertical axis represents Cook's distance measure and the horizontal line shows positive integer values x..

estimator $\hat{N}$ of the population size with an associated estimator $\hat{f}_0 = \hat{N} - n$ of the expected value $E(f_0)$ of f_0. Then, we can draw B samples of size $\hat{N}$ with replacement from the discrete distribution given mass $\hat{f}_0/\hat{N}$ to count 0 and mass $f_x/\hat{N}$ to x for $x = 1, \ldots, m$. For each of the B samples, we can find $\hat{N}_b$, leading to a bootstrap sample $\hat{N}_1, \ldots, \hat{N}_B$ from which we can derive a $(1-\alpha)100\%$ CI using, for example, the percentile method. In the application below we use a bootstrap replication size of $B = 10,000$. This form of semi-parametric bootstrap has been investigated in Anan et al. (2017) (called *imputed* bootstrap in the paper) and shown to perform well if the model is correctly specified.

As an alternative to the semi-parametric bootstrap, an analytic approximation might be considered. We look at the zero-truncated one-inflated setting which can be handled by means of truncating both, counts of zeros and ones. In this case, the Horvitz–Thompson estimator for f_0 is $\hat{f}_0 = n_1\hat{p}_0/(1-\hat{p}_1-\hat{p}_0)$ (see also Equation (8)), where $n_1 = n - f_1$, and we need to find its variance. This can be accomplished by using the technique of conditional moments (see Böhning (2008)) and builds on the result that $\text{Var}(\hat{f}_0)$ can be written as

$$\text{Var}(\hat{f}_0) = E(\text{Var}(\hat{f}_0|n_1)) + \text{Var}(E(\hat{f}_0|n_1)). \qquad (9)$$

The second term in Equation (9) can be estimated as

$$p_0^2/(1-p_0-p_1)^2\text{Var}(n_1) = p_0^2/(1-p_0-p_1)^2$$
$$N(1-p_0-p_1)(p_0+p_1)$$

which can be further estimated as $p_0^2/(1-p_0-p_1)^2 n_1(p_0+p_1)$. The first term in Equation (9) does not have such a general form but will rather depend more specifically on the parametric form of the distribution p_x. To give an illustration, we use the geometric distribution which is supported by the study data (as seen in the following section) and also is a special case of the straight line RR (slope is zero). Following the conditioning approach and the details in Böhning & Ogden (2021), we can find for the zero-one truncated geometric distribution that the variance of $\hat{f}_0$ can be estimated as

$$n_1^2\frac{(1+\hat{\theta})^2}{(1-\hat{\theta})^6}\widehat{Var}(\hat{\theta}) + n_1\frac{\hat{\theta}^3(2-\hat{\theta})}{\{1-\hat{\theta}-\hat{\theta}(1-\hat{\theta})\}^2}.$$

Here $\widehat{\text{Var}}(\hat{\theta}) = \{n_1/\hat{\theta}^2 + S/(1-\hat{\theta})^2\}^{-1}$ and $\hat{\theta} = n_1/(n_1+S)$ is the maximum likelihood estimator under the zero–one-truncated geometric model with $S = \sum_{x=0}^{m-2} xf_{x+2}$.

Again, using the technique of conditional moments we can estimate $\text{Var}(\hat{N}) = \text{Var}(\hat{f}_0 + n)$ as $\text{Var}(\hat{f}_0) + \text{Var}(n)$ with $\text{Var}(n) = Nq(1-q)$, where $1-q = P(X>0)$. Under the geometric distribution, the latter can be estimated as $n\hat{\theta}$. In total, we achieve

$$\widehat{Var}(\hat{N}) = n_1^2\frac{(1+\hat{\theta})^2}{(1-\hat{\theta})^6}\widehat{Var}(\hat{\theta}) + n_1\frac{\hat{\theta}^3(2-\hat{\theta})}{\{1-\hat{\theta}-\hat{\theta}(1-\hat{\theta})\}^2} + n\hat{\theta}.$$

For the non-inflated zero-truncated geometric distribution, again following Böhning & Ogden (2021), the variance of $\hat{f}_0 + n = n\hat{\theta}/(1-\hat{\theta}) + n = n/(1-\hat{\theta})$ is simply

TABLE 2 Estimated regression coefficients, estimated probability for observing a zero-count as well as the associated estimates for the number of index cases with contacts.

Model	$\hat{\beta}_0$ (SE)	$\hat{\beta}_1$ (SE, p-value)	$\hat{p}_0$	$\hat{N}$	$\hat{f}_0$
Unweighted					
All (with $m = 167$)	−0.3045 (0.4184)	0.0788 (0.1319, 0.553)	0.0024	341.81	0.81
With upper truncation $T = 20$	−0.5552 (0.4601)	0.2089 (0.1984, 0.307)	0.1753	413.47	72.47
With upper truncation $T = 40$	−0.4556 (0.4993)	0.1430 (0.1706, 0.407)	0.1717	411.69	70.69
Weighted					
All (with $m = 167$)	−0.474I1 (0.1967)	0.1555 (0.0808, 0.060)	0.0	341	0
With upper truncation $T = 20$	−0.6473 (0.2726)	0.2692 (0.1376, 0.067)	0.1684	410.07	69.07
With upper truncation $T = 40$	−0.5287 (0.2242)	0.1871 (0.0962, 0.059)	0.1077	382.17	41.17

estimated as

$$n^2\hat{\text{var}}(\hat{\theta})/(1-\hat{\theta})^4 + n\hat{\theta}^3/(1-\hat{\theta})^2,$$

where $\hat{\theta} = n/(n+S)$ with $S = \sum_{x=0}^{m-1} x f_{x+1}$. The variance of $\hat{\theta}$ is found from the negative inverse observed Fisher information as $\{n/\hat{\theta}^2 + S/(1-\hat{\theta})^2\}^{-1}$.

We find the bootstrap approach more appealing than using an approach based on an asymptotic normal approximation which depends on the derivation of an asymptotic variance. Although possible as seen above, it usually depends at least partly on the specific model and approximations such as the δ-method (Böhning, 2008). The bootstrap suggested here is more generically applicable.

7 | APPLICATION TO COVID-19 CONTACT-TRACING DATA FROM THAILAND

We now apply these RR concepts to the data of Table 1. We start by using weighted and unweighted linear regression with upper truncation points of $T = 20$ and $T = 40$. There are no strict rules for the choice of T. However, a guiding principle should be to obtain a stable estimate of the ratio, stable in the sense of a reasonable variance. We have been choosing $T = 20$ as frequencies start taking values smaller than 5 and also $T = 40$ as ratios become undefined. Table 2 presents estimated regression coefficients, estimated probability for observing a zero-count as well as the associated population size estimate. Both, weighted and unweighted, model estimates are included using the regression model $\log R_x = \beta_0 + \beta_1 \log(x+1)$. Columns 2 and 3 in Table 2 show that the weighted regression model produces the smaller standard errors for the estimated regression coefficients. We see that for the unweighted regression model the population size estimates remain similar when T changes from 20 to 40, whereas there is a slight decrease in $\hat{N}$ in the weighted case.

In Table 3, various bootstrap statistics are provided including the mean, median, standard error, and 95% CI for N. We also provide, for comparison, these statistics for the estimator of Chao (1989), also called Chao1-estimator, as

$$\hat{N}_C = n + f_1^2/(2f_2). \tag{10}$$

The estimator of Chao (Chao1) is developed under arbitrary heterogeneity for the parameter involved in the count distribution which counts the number of identifications X for a specific unit:

$$p_x = \int_\theta k(x|\theta)q(\theta)d\theta.$$

Here, $q(\theta)$ is an unspecified mixing distribution. However, it is crucial to make an appropriate assumption for the mixing kernel $k(x|\theta)$. The estimator Chao1 was originally developed for a Poisson kernel leading to Equation (10), but Chao estimation can be generalized for any kernel from the power series family (Böhning et al., 2019). If $p_x = \int_\theta k(x|\theta)q(\theta)d\theta$ holds, it can be shown that the Chao estimator provides a lower bound for N. However, the Chao estimator will depend on the form of the kernel assumed. Here, as we think a geometric kernel is more appropriate as contact counts show a long-tailed distribution and, in addition, the RR analysis has provided evidence for a geometric distribution. The Chao estimator for the geometric kernel is given by

$$\hat{N}_{\text{CG}} = n + f_1^2/f_2, \tag{11}$$

which is always larger than Equation (10) and also a sharper bound if the geometric kernel is appropriate. We call this estimator *ChaoG*. The estimator ChaoG was suggested in Böhning et al. (2019) and specifically used for CT data in Lerdsuwansri et al. (2022).

A further modified ChaoG estimator has been suggested in Böhning et al. (2019), $\hat{N}_{\text{MCG}} = n + f_2^3/f_3^2$. The

TABLE 3 Bootstrapped population size estimates of fully traced index cases with contacts; estimates are given with 95% confidence intervals (CIs) based upon weighted ratio regression (RR) with upper truncation ($T = 20$) including and excluding singletons, and for comparators Chao1, ChaoG, and modified ChaoG as well as the maximum likelihood estimates under the geometric distribution.

		Bootstrap statistics				
Estimator	$\hat{N}$	**Mean**	**Median**	**SE**	**95% percentile CI**	**Length of CI**
RR	410	408.55	406.85	16.04	(365.97, 461.56)	95.59
ChaoG $\hat{N}_{\mathrm{CG}}$	429	435.68	429.49	22.56	(379.48, 528.06)	148.58
Modified ChaoG $\hat{N}_{\mathrm{MCG}}$	359	365.43	359.76	22.42	(341.15, 423.39)	82.24
Chao1 $\hat{N}_C$	385	388.14	385.35	11.65	(358.0, 435.10)	77.10
Geometric	360	360.01	360.13	3.15	(350.58, 368.67)	18.09

modified ChaoG estimation was suggested in Böhning et al. (2019) to cope with one-inflation as this could lead to an overestimation of population size. The estimator $\hat{N}_{\mathrm{MCG}}$ avoids the use of f_1 and is, hence, less prone to overestimation. If $\hat{N}_{\mathrm{MCG}}$ and $\hat{N}_{\mathrm{CG}}$ are close, this indicates lack of evidence for one-inflation. Table 3 shows that there is little evidence for one-inflation as all three estimators, RR, maximum likelihood on the basis of the geometric and ChaoG, and their counterparts addressing potential one-inflation are quite close. Note that all lower bound estimators including Chao1 and ChaoG are close to the RR estimator. The latter has the benefit of providing a smaller standard error.

As the RR modeling provides evidence for a slope zero line, corresponding to a geometric distribution, we have also included in Table 3 population size estimates based upon the zero-truncated geometric distribution. Here, the benefit is that no issue of upper truncation exists and all data points can be included. It can be seen in Table 3 that this leads to smaller standard errors in comparison to all other approaches.

Using the proposed population size estimate (weighted RR) we can conclude that CT based on the available Covid-19 data reaches a completeness of 341/410 = 0.832, or 83.2%. In other words, only 16.8% of all cases with contacts can be assumed to have been missed by CT. For the zero-truncated geometric model estimate, we find a completeness of 341/360 = 0.947 or 94.7%. In other words, only 5.3% of all cases with contacts can be assumed to have been missed by CT. In Figure 4, all five estimators with their associated CIs are displayed. We see that they are fairly close together with all CIs overlapping. This motivates the question how these estimators compare and this will be investigated in the next section.

8 | SIMULATION STUDY

In this section, we take a closer look at the performance of the population size estimators introduced in Sections 3 and 4 by means of simulation. To be precise, we consider the RR-based population size estimators $\hat{N}_{\mathrm{RR}}$ for the unweighted, $w\hat{N}_{\mathrm{RR}}$ for the weighted case, and both with upper truncation denoted as $\hat{N}_{\mathrm{RR}_T}$ and $w\hat{N}_{\mathrm{RR}_T}$, respectively. For comparison, we also look at some other well-known estimators, namely Chao's lower bound estimators $\hat{N}_C = n + f_1^2/(2f_2)$ and $\hat{N}_{\mathrm{CG}} = n + f_1^2/f_2$, and the modified Chao estimator $\hat{N}_{\mathrm{MCG}} = n + f_2^3/f_3^2$ as well as the maximum likelihood based estimator $\hat{N}_{\mathrm{ztg}} = n/(1 - n/\sum x f_x)$ based on the zero-truncated geometric (ztg) model. For the design of this simulation, the data are generated using three settings. First, they are generated from a *Poisson distribution* with mean $\theta = 1.5$ and 3. Then, all zero-counts are discarded so that zero-truncated Poisson counts are obtained. Here, the population size parameter is taken as $N = 500, 1000$, and 5000. Second, the count data are sampled from a *geometric density* $\theta(1-\theta)^x$, where the probability parameter $\theta = 0.1$ and 0.25. Note that $\theta = 0.25$ corresponds to the mean $\mu = (1-\theta)/\theta = 3$ which is the same as in the second Poisson case. So, here both cases match the same mean, but they have quite different variances. Finally, we are interested in a setting where the data are not generated under a model which is covered by this simple form of RR $\beta_0 + \beta_1 \log(x+1)$. A distribution which meets this requirement is the *negative binomial* distribution which is also frequently used in the capture–recapture context. It can be viewed as a Poisson distribution mixed with a gamma distribution, hence adjusts already for some potential heterogeneity in the Poisson parameter. Also, the geometric and, as a limiting case, the Poisson are special cases of the negative binomial distribution. To be more precise, the data are generated from a negative binomial distribution with $\mu = 1.5$ and 3, where these settings are corresponding to mean of the previous considered distributions. The size parameter k is given by 2, 3, and 5, and the probability of success in each trial θ is computed by $\theta = k/(k+\mu)$. Zero-truncated counts are again obtained by discarding zeros.

We fit the RR using a straight line approach $\beta_0 + \beta_1 \log(x+1)$ after all zero counts are truncated. Both unweighted and weighted regression analyzes are applied.

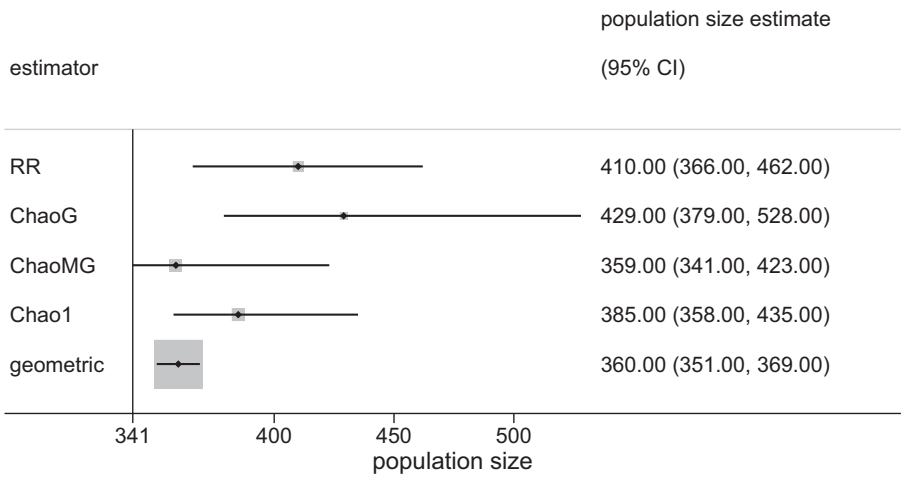

FIGURE 4 Illustration of all five estimators from Table 3 with their 95% confidence intervals.

From each generated dataset, the estimated p_0 and proposed N are obtained from Equation (5) with the inverse of Equation (4) or (7), depending whether the untruncated (all) or upper truncation approach is used, respectively. Under the truncation process, we truncate all counts which have frequency equal to or less than 5. We come back to this aspect again in Section 9. Furthermore, the well-known estimators noted above are also computed in all simulation settings. This means that a model misspecification situation is studied in this work. For example, Chao's lower bound estimator (Chao1) for the Poisson kernel is also provided although the data are simulated under a geometric distribution and vice versa. Each scenario is repeated for 10,000 times using the R programming language. To evaluate the performance of the various population size estimators, the relative bias (RB), relative standard deviation (RSd), and relative root mean squared error (RRmse) are computed, defined as

$$\mathrm{RB}(\hat{N}) = \frac{1}{N}E(\hat{N} - N) = \frac{1}{N}\mathrm{Bias}(\hat{N}),$$

$$\mathrm{RSd}(\hat{N}) = \sqrt{\frac{1}{N^2}E(\hat{N} - E(\hat{N}))^2},$$

$$\mathrm{RRmse}(\hat{N}) = \frac{1}{N}\sqrt{E(\hat{N} - N)^2} = \frac{1}{N}\sqrt{\mathrm{Var}(\hat{N}) + \{\mathrm{Bias}(\hat{N})\}^2}.$$

These quantities are estimated by replacing expected values with their corresponding simulation means.

The results under simulated data from the Poisson, the geometric, and the negative binomial distributions are presented in Tables S2–S4. Here, we first consider the scenarios of the first two models given in Tables S2 and S3. Under the Poisson distribution, $\hat{N}_C$ has RBs closest to zero, if compared to the other estimators. The RBs of $\hat{N}_{\mathrm{ztg}}$ and $\hat{N}_{\mathrm{CG}}$ are close to zero under the geometric distribution in all situations. This is not surprising as these estimators are considered under the true distribution. However, the estimators of Chao do less well if the model is misspecified, as $\hat{N}_C$ and $\hat{N}_{\mathrm{CG}}$ (also $\hat{N}_{\mathrm{MCG}}$ and $\hat{N}_{\mathrm{ztg}}$) have the large RBs in the case of a geometric and Poisson count distributions, respectively. $w\hat{N}_{\mathrm{RR}_T}$ behaves satisfactorily, as it provides small biases in estimating N, no matter if a Poisson or a geometric distribution is assumed. In general, the estimator using fitting RR with the weighted method performs better than the one obtained from the unweighted regression in terms of RB. However, their performances are less affected in terms of RSd and RRmse. Another interesting result refers to the estimator for N based on RR with upper truncation. As can be seen from these two cases, $w\hat{N}_{\mathrm{RR}_T}$ shows quickly decreasing bias with increasing population size and has small variance. Especially, its RB is much smaller and close to zero than that of $w\hat{N}_{\mathrm{RR}}$, while the variances do not differ much. Clearly, the population size estimator for N based on the RR with upper truncation is useful. These favorable results for the RR shown here clearly depend on the fact that the Poisson and the geometric are special cases of the RR model used. This implies that any RR based analysis should carefully check the validity of the model for the observed part of the data.

Let us return to the case under simulated data from the negative binomial distribution given in Table S4. $\hat{N}_{\mathrm{MCG}}$ and $\hat{N}_{\mathrm{ztg}}$ show large biases in estimating N. The RBs of $w\hat{N}_{\mathrm{RR}}$ and $w\hat{N}_{\mathrm{RR}_T}$ are slightly different and smaller than those of the comparators. However, the two estimators using unweighted RR are less satisfactory here as expected (in contrast to the Poisson and geometric distributional settings, where their performance was still acceptable). The reason is that $\hat{N}_{\mathrm{RR}}$ and $\hat{N}_{\mathrm{RR}_T}$ have large RBs, especially comparing to the well-known estimators $\hat{N}_C$ and

$\hat{N}_{CG}$ which are studied under model misspecification as well. Hence, we conclude that although the negative binomial distribution is not covered by this simple form of RR model, estimating population size using the *weighted RR* approach performs still rather well.

9 | DISCUSSION

We close with a few remarks. Completeness of outbreak detection is a crucial task in outbreak control. Doyle et al. (2002) had already pointed out the importance of surveillance for infectious diseases as a critical element in providing effective public health disease control and prevention. They point out that completeness of reporting is essential and emphasize the key role of capture–recapture methods. However, the traditional application of capture–recapture techniques in infectious disease surveillance lies in correcting the undercount of prevalence and incidence. In our case, we are not estimating the size of the outbreak (the number observed plus the hidden or dark number), we are targeting on the completeness of the tracing system for cases that have been notified. Capture–recapture methods provide a way of estimating the completeness of outbreak detection. RR provides a suitable and flexible class of count models and is particularly appropriate for ratios of neighboring probabilities as these are invariant with respect to zero-truncation.

CT of Covid-19 infections has been also considered in Lerdsuwansri et al. (2022). The difference of the current work to Lerdsuwansri et al. (2022) is that the latter is prepared for an applied medical audience and as such is brief on methodological issues. Modeling in Lerdsuwansri et al. (2022) was limited in the sense that only certain parametric families were considered such as the Poisson and geometric (and the negative binomial as the wider family), whereas here we take a more general approach by means of RR which allows modeling in a much wider class of distributions. As a result, the estimates for unobserved index cases with contacts are slightly revised here: the total number of cases with contacts is estimated as 410 using RR, whereas the best estimate in Lerdsuwansri et al. (2022) is 439 (based on the negative-binomial).

Even if RR estimators are not intended to be used, RR can be helpful in choosing an appropriate kernel in Chao-estimation as has been demonstrated in the previous section.

One disadvantage with RR is that it needs large frequencies to estimate ratios in a stable way. This may require larger sample sizes as they are occurring typically in routinely collected data such as CT distributions. In addition, for long-tailed distributions at some point T the frequencies f_x become small. As a rule of thumb, one can use $T = x$ for upper truncation if $f_x < 5$. A sensitivity analysis by choosing different values for T might be helpful as well. However, we believe that the prediction for f_0 depends much more on what is happening to the RR for x close to zero than what is happening in the right tail. Note that the procedure for upper truncation is not uncommon in capture–recapture; for example, the Chao–Bunge estimator uses this form of truncation (Chao and Bunge, 2002). In our case, we found that the geometric distribution was suggested by RR modeling which does not involve any truncation at all.

The incorporation of covariate information can be useful in providing more accurate population size estimates. One of the disadvantages of RR is that it builds on aggregated frequencies and not on individual case data which makes the inclusion of covariate information difficult. Of course, a stratified analysis could be done by running an RR separately for the strata and adding up size estimates over the strata, but this has its limitation as RR needs fairly stable frequency estimates. For the dataset at hand, we have age and gender information on the case basis. We investigated fitting a standard zero-truncated geometric model for these, but the two covariates were not significant.

Another issue is CT of infectious contacts as these are responsible for the spread of the epidemic in the population. Most of the contacts of an index case did not lead to a further infection. However, if an index case had a contact that turned out to be infected, we talk about an *infectious* contact. We note in passing that the infectious contacts are the basis from computing the reproduction number R_0, the average number of people a case infects in a period of time. For the Covid-19 CT data of Thailand from 341 index cases, only 30 has infectious contacts. We have the following frequencies: $f_1 = 16$, $f_2 = 9$, $f_3 = 4$, $f_4 = 1$. These need to be interpreted as follows: there are 16 index cases with 1 infectious contact, 9 index cases with 2 infectious contacts, 4 with 3 and 1 with 4 infectious contacts. The question arise how many unobserved index cases are there with infectious contacts. The answer can be reached by utilizing RR. In Figure S5, we see a scattergram of $\log r_x$ against $\log(x+1)$ with least-squares line. Here, the Poisson model (the line with slope $\beta_1 = -1$ in Figure S5) is quite close to the least-squares line. The associated zero-truncated Poisson likelihood is maximized for $\hat{\beta}_0 = 0.1189$ which gives a Horvitz–Thompson estimate of

$$\hat{N} = \frac{n}{1 - \exp\{-\exp(\beta_0)\}} = \frac{30}{1 - \exp\{-\exp(0.1189)\}} = 44.4,$$

which corresponds of a completeness of detection of cases with infectious contacts of 68%. The number is closely matched with Chao's estimate which is for the Poisson kernel $\hat{N}_C = n + f_1^2/(2f_2) = 30 + 16^2/18 = 44.2$. This shows

that the majority of cases with infectious contacts have been traced.

As a final point, we note the connection between RR and Bayesian inference. Suppose that $p_x = \int_\theta a_x \theta^x/\eta(\theta)q(\theta)d\theta$ is given as a mixture of a power series distribution with arbitrary mixing density $q(\theta)$. Here, a_x are the known coefficients defining the power series and $\eta(\theta)$ is the normalizing constant. Then, we have that

$$\frac{a_x}{a_{x+1}}R_x = \frac{a_x}{a_{x+1}}\frac{\int_\theta a_{x+1}\theta^{x+1}/\eta(\theta)q(\theta)d\theta}{\int_\theta a_x\theta^x/\eta(\theta)q(\theta)d\theta}$$
$$= \frac{\int_\theta \theta \times a_x\theta^x/\eta(\theta)q(\theta)d\theta}{\int_\theta a_x\theta^x/\eta(\theta)q(\theta)d\theta} = \int_\theta \theta p(\theta|x)d\theta,$$

the posterior mean with posterior density

$$p(\theta|x) = \frac{a_x\theta^x/\eta(\theta)q(\theta)}{\int_\theta a_x\theta^x/\eta(\theta)q(\theta)d\theta}.$$

Hence, $a_x r_x/a_{x+1}$ can be viewed as an estimate of the posterior mean of θ, given $X = x$. This result is closely related to empirical Bayesian inference (Carlin & Louis, 2011). The connection between RR and Bayesian inference is interesting as it offers another interpretation of the ratio as a posterior mean. However, it is not new and probably goes back to Robbins (1955) and the genesis of the nonparametric, empirical Bayes approach. Robbins (1955) showed that the posterior mean for a Poisson likelihood for X and arbitrary prior could be estimated as $(x+1)f_{x+1}/f_x$ with several advantageous properties. Carlin and Louis (2011) (see chapter on empirical Bayes approach) gave a nice review of the idea. Again, the difference to the approach considered here is that the ratio (or posterior mean) is viewed in its functional dependence on the count x. Finally, it is pointed out that the argument for this connection works more generally for the power series family.

ACKNOWLEDGMENTS

All authors are grateful to the Ministry of Public Health Thailand (MoPHT) for providing access to the Covid-19 contact tracing data. Thanks go to James Gallagher (Director of Statistical Services Centre Reading) for a critical reading of the manuscript. All authors are grateful to the Editor, an Associate Editor, and two Referees for their helpful comments. The first author is deeply grateful for receiving funding from Thammasat University (Thailand) to undertake this research. This study was supported by Bualuang ASEAN Chair Professor Fund, Agreement Number TUBC 02/2022.

DATA AVAILABILITY STATEMENT

The data that support the findings of this paper are available in the Supporting information section of this paper.

OPEN RESEARCH BADGES

This article has earned Open Data and Open Materials badges. Data and materials are available as supporting material.

ORCID

Dankmar Böhning https://orcid.org/0000-0003-0638-7106
Rattana Lerdsuwansri https://orcid.org/0000-0002-5904-7757
Patarawan Sangnawakij https://orcid.org/0000-0002-9614-6869

REFERENCES

Anan, O., Böhning, D. & Maruotti, A. (2017) Uncertainty estimation in heterogeneous capture–recapture count data. *Journal of Statistical Computation and Simulation*, 10, 2094–2114.
Böhning, D. (2008) A simple variance formula for population size estimators by conditioning. *Statistical Methodology*, 5, 410–423.
Böhning, D. (2016) Ratio plot and ratio regression with applications to social and medical sciences. *Statistical Science*, 31, 205–218.
Böhning, D. & Del Rio Vilas, V. (2008) Estimating the hidden number of Scrapie affected holdings in Great Britain using a simple, truncated count model allowing for heterogeneity. *Journal of Agricultural, Biological and Environmental Statistics*, 13, 1–22.
Böhning, D. & Ogden, H. (2021) General flation models for count data. *Metrika*, 84, 245–261.
Böhning, D., Baksh, M.F., Lerdsuwansri, R. & Gallagher, J. (2013) The use of the ratio-plot in capture–recapture estimation. *Journal of Computational and Graphical Statistics*, 22, 135–155.
Böhning, D., Bunge, J. & van der Heijden, P.G.M. (2018) *Capture–recapture methods for the social and medical sciences*. Boca Raton, FL: Chapman & Hall.
Böhning, D., Kaskasamkul, P. & van der Heijden, P.G.M. (2019) A modification of Chao's lower bound estimator in the case of one-inflation. *Metrika*, 82, 361–384.
Carlin, B.P. & Louis, T.A. (2011) *Bayesian methods for data analysis*. Boca Raton, FL: Chapman & Hall.
Chao, A. (1989) Estimating population size for sparse data in capture–recapture experiments. *Biometrics*, 45, 427–438.
Chao, A. & Bunge, J. (2002) Estimating the number of species in a stochastic abundance model. *Biometrics*, 58, 531–539.
Cleveland, W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836.
Doyle, T.J., Glynn, M.K. & Groseclose, S.L. (2002) Completeness of notifiable infectious disease reporting in the United States: an analytical literature review. *American Journal of Epidemiology*, 155, 866–874.
Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., et al. (2020) Quantifying SARS-CoV-2 transmission

suggests epidemic control with digital contact tracing. *Science*, 368, 1–9.

Friendly, M. (2001) *Visualizing categorical data*. Cary, NC: SAS Institute.

Hoaglin, D.C. (1980) A Poissonness plot. *American Statistical Association*, 34, 146–149.

Hoaglin, D.C. & Tukey, J.W. (1985) Checking the shape of discrete distributions. In: Hoaglin, D.C., Mosteller, F. & Tukey, J.W. (Eds.) *Exploring data tabkes, trends, and shapes*. New York: John Wiley & Sons, pp. 345–416.

Kaweenuttayanon, N., Pattanarattanamolee, R., Sornchaa, N. & Nakahara, S. (2021) Community surveillance of COVID-19 by village health volunteers, Thailand. *Bulletin of the World Health Organization*, 99, 393–397.

Lerdsuwansri, R., Sangnawakij, P., Böhning, D., Sansilapin, C., Chaifoo, W., Polonsky, J.A., et al. (2022) Sensitivity of contact-tracing for COVID-19 in Thailand: a capture–recapture application. *BMC Infectious Diseases*, 22, 1–10.

McCrea, R.S. & Morgan, B.J.T. (2015) *Analysis of capture–recapture data*. Boca Raton, FL: Chapman & Hall.

McKendrick, A.G. (1926) Application of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44, 98–130.

MoPH. (2020) Guidelines for surveillance and investigation of coronavirus disease 2019 (COVID-19). Available from: https://ddc.moph.go.th/viralpneumonia/eng/guidelines.php [Accessed 9th February 2020].

Ord, J.K. (1967) Graphical methods for a class of discrete distributions. *Journal of the Royal Statistical Society: Series A*, 130, 232–238.

Robbins, H. (1955) An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 157–164.

Sellers, K.F. & Shmueli, G. (2010) A flexible regression model for count data. *Annals of Applied Statistics*, 4, 943–961.

Wesson, P., Lechtenberg, R., Reingold, A., McFarland, W. & Murgai, N. (2018) Evaluating the completeness of HIV surveillance using capture-recapture models, Alameda County, California. *AIDS and Behavior*, 22, 2248–2257.

WHO. (2020) WHO announces COVID-19 outbreak a pandemic. Available from: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic [Accessed 23rd February 2020].

Wilson, R.M. & Collins, M.F. (1992) Capture–recapture estimation with samples of size one using frequency data. *Biometrika*, 79, 543–553.

SUPPORTING INFORMATION

Figures S1– S5 referenced in Sections 2–5 and 9, Tables S1– S4 referenced in Sections 1 and 8, and R code used in this study are available with this paper at the Biometrics website on Wiley Online Library.

How to cite this article: Böhning, D., Lerdsuwansri, R. & Sangnawakij, P. (2023) Modeling COVID-19 contact-tracing using the ratio regression capture–recapture approach. *Biometrics*, 1–13. https://doi.org/10.1111/biom.13842