Supplementary Materials for: Model-based analysis of tuberculosis genotype clusters in the United States reveals high degree of heterogeneity in transmission, and state-level differences across California, Florida, New York and Texas.

Sourya Shrestha[1,*], Kathryn Winglee[2], Andrew Hill[2], Tambi Shaw[3], Jonathan Smith[4],
J. Steve Kammerer[2], Benjamin J. Silk[2], Suzanne Marks[2], David Dowdy[1]

[1]Department of Epidemiology, Johns Hopkins School of Public Health, Baltimore, USA

[2]Division of Tuberculosis Elimination, Centers for Disease Control and Prevention, Atlanta, USA

[3]California Department of Public Health, California, USA

[4]Yale University, New Haven, CT, USA

[*]To whom correspondence should be addressed; E-mail: sourya@jhu.edu.

S-1 Branching Process with Poisson-Lognormal Distribution

We use a classical Galton-Watson branching process, where the offspring distribution, Z is taken to be the number of secondary infections resulting from an individual in the next generation. We modeled this distribution as a Poisson process conditional on an observed value ν of the individual reproduction number: $Z|\nu \sim$ Poisson(ν). The probability mass function of this distribution is given by following expression: $P(Z = z|\nu) = \frac{\nu^z\, e^{-\nu}}{z!}$. Using convention described previously [1], we take ν to be the observed value of a random variable X describing the individual reproductive number. It follows that $E[Z] = E[E[Z|X]] = E[X] = R_0$, which is the reproductive number of the transmission described by the branching process Z.

For Poisson lognormal model, we take X to follow a lognormal distribution, i.e., $X \sim$ lognormal(μ, σ), such that the resulting mixture distribution Z is a Poisson-lognormal distribution. Consequently, the probability mass function of Z is given by the following expression:

$$P(Z = z) = \frac{(2\,\pi\,\sigma^2)^{-1/2}}{z!}\int_0^\infty \nu^{z-1}\, e^{-\nu}\, e^{-\frac{(\ln\nu-\mu)^2}{2\,\sigma^2}}\, d\nu$$

The mean of this distribution is reproductive number R_0, and is given by $E[Z] = R_0 = e^{\mu+\frac{\sigma^2}{2}}$ [2].

The expression for the variance of this distribution can be derived by using Law of Total Variance, which states that $Var(Z) = E(Var(Z|X)) + Var(E(Z|X))$. So, if $Z|X = \nu \sim$ Poisson(ν), and $X \sim$ lognormal(μ, σ^2), then, $Var(Z) = E(Var(Z|X)) + Var(E(Z|X)) = E[X] + Var[X] = \exp(\mu + \sigma^2/2) + (\exp(\sigma^2) - 1)\, \exp(2\mu + \sigma^2) = R_0(1 + R_0(\exp(\sigma^2) - 1))$.

While there is no closed form expression for the probability mass function, prior works have contributed to several numerical approximations of this function [2–5]. We use a mixture of these approximations in our model.

For $P(Z = 0)$, we approximate the following integral transformed on the interval $(-\infty, \infty)$ as proposed by Grundy [4].

$$P(Z = 0) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left[-e^{\nu} - \frac{(\nu-\mu)^2}{2\sigma^2}\right] d\nu$$

For $1 \leq z < 60$ we approximate the following integral transformed on the interval $(0, 1)$ as proposed by Izsak [5].

$$P(Z = z) = \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{z!}\left[\int_0^1 F_1(\nu) d\nu + \int_0^1 F_2(\nu) d\nu\right] \qquad \text{, where,}$$

$$F_1 = \nu^{z-1}\, e^{-\nu}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

$$F_2 = \nu^{-z-1}\, e^{-\frac{1}{\nu}}\, e^{\frac{-(\ln\frac{1}{\nu}-\mu)^2}{2\sigma^2}}$$

For $60 \leq z \leq 100$ we use the following approximation proposed by Bulmer [2].

$$P(Z = z) = \frac{1}{\sigma\, z\sqrt{2\pi}}\, e^{-\frac{(\ln z-\mu)^2}{2\sigma^2}}\left[1 + \frac{1}{2\, z\sigma^2}\left[\frac{(\ln z - \mu)^2}{\sigma^2} + \ln z - \mu - 1\right]\right]$$

For $z > 100$ we approximate the following integral transformed back on the interval $(-\infty, \infty)$, and use the saddle point method to approximate the integral [5].

$$P(Z=z) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-f(x)} dx \qquad \text{, where,}$$
$$f(x) = e^x + \ln z! + \frac{(x-\mu)^2}{2\sigma^2} - z\,x$$

Let $I = \int_{-\infty}^{\infty} e^{-f(x)} dx$. Taylor expansion of $f(x)$ around x_{min} is:

$$f(x) = f(x_{min}) + (x - x_{min})\, f'(x_{min}) + \frac{1}{2}(x - x_{min})^2\, f''(x_{min}) + \dots$$

$f'(x_{min}) = 0$, since $f(x)$ has minimum at x_{min}. Hence,

$$I \approx e^{-f(x_{min})} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\, f''(x_{min})\,(x-x_{min})^2} dx$$

Since, the above is Gaussian integral,

$$I \approx e^{-f(x_{min})} \sqrt{\frac{2\pi}{f''(x_{min})}}.$$

Hence,

$$P(Z=z) \approx \frac{e^{-f(x_{\min})}}{1+\sigma^2 e^{x_{\min}}}.$$

S-2 Recursion for outbreak size

Consider a branching process, where Z is the number of secondary infections resulting from an individual in the next generation. We assume that the reproductive number associated with this branching process is less than 1, such that the probability of extinction is 1. In such a case, we define S to be random variable describing the final size of an outbreak resulting from a single case. Consider transmission chains starting from n separate index cases. The cumulative final size of the outbreak that includes all n separate chains starting from n index cases will be the sum of n i.i.d. random variables $S_1 + S_2 + \cdots + S_n$. We define this sum to be C_n, where $C_n = S_1 + S_2 + \cdots + S_n$, where S_i's are sizes of each individual outbreaks.

Then the probability mass function describing the size of the outbreak, S, can be written as follows:

$$\begin{aligned} P(S = n) &= \sum_{i=1}^{n} P(Z = i|C_i = n - i)P(C_i = n - i) \\ &= \sum_{i=1}^{n} P(Z = i)P(C_{n-i} = n - i) \\ &= \sum_{i=1}^{n} p_Z(i)\, p_{C_{n-i}}(n - i). \end{aligned}$$

Here, $p_Z(z)$ is the probability mass function of Z, and $p_{C_k}(z)$ is the probability mass function of C_k. In other words, the probability of getting an outbreak of size n can be partitioned in to (i) probability of getting i secondary cases in the first generation, where $1 \leq i \leq n$, and (ii) probability that sum of outbreak sizes of all chains starting with i cases results in the remaining $n - i$ cases.

Furthermore, we can write p_{C_k} as a function of p_S and $p_{C_{k-1}}$:

$$\begin{aligned} P(C_k = j) &= \sum_{i=1}^{k} P(S = i|C_{k-1} = j - i)P(C_{k-1} = j - i) \\ &= \sum_{i=1}^{k} P(S = i)P(C_{k-1} = j - i) \\ &= \sum_{i=1}^{k} p_S(i)\, p_{C_{k-1}}(j - i) \end{aligned}$$

Combining both, we get the following recursive relationship for S:

$$\begin{aligned} P(S = n) &= \sum_{i=1}^{n} p_Z(i) \sum_{j=1}^{n-i} p_S(j)\, p_{C_{n-i-1}}(n - i - j) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n-i} p_Z(i)\, p_S(j)\, p_{C_{n-i-1}}(n - i - j) \end{aligned}$$

$P(S = 1) = p_S(1) = p_Z(0)$, i.e., the probability of outbreak of size 1 is extinction in the first generation, thus only counting the index case.

S-3 Branching Process with Negative Binomial Distribution

If the individual reproductive number follows a gamma distribution, i.e., $\nu \sim$ Gamma(R_0, k), where R_0 is the mean of the distribution, and k is the dispersion parameter, then the number of secondary infections resulting from an individual follows a negative binomial distribution: $Z \sim$ negative binomial(R_0, k). This particular distribution has been used quite extensively to characterize individual-level heterogeneity for a range of infectious processes including tuberculosis. [1, 6–10]

A particular quantity of interest is the final size of the outbreak, and a closed form expression has been derived independently by Nishiura, et. al. [6], and Blumberg, et. al. [7]. We use the following expression for final size distribution:

$$P(S = x) = \frac{\prod_{j=0}^{x-2}(j/k + x)}{x!} \left(\frac{k}{R_0 + k}\right)^{kx} \left(\frac{R_0\, k}{R_0 + k}\right)^{x-1}$$

S-4 Likelihood function and parameter estimation

We used a likelihood-based framework to ascertain model fits. The likelihood of observing a cluster distribution in which there are n_j clusters of size j, under a model with parameters x, is taken as the product of probabilities of observing the number of clusters of each size j as seen in the empirical cluster distribution. Hence, if $P(S = s|x)$ is the probability of a cluster of size j under a given model with parameters x, the likelihood function is given by the following equation:

$$\mathcal{L}(x) = \prod_{s=1}^{\infty} P(S = s|x)^{n_j}$$

Maximum likelihood estimates (MLEs), the parameters that yield the highest likelihood, and corresponding 95% confidence regions/intervals were estimated by conducting grid searches across parameter space. The areas for the grid search were adjusted to ensure that the entire estimated 95% confidence regions were included. The 95% confidence regions were taken to be $\chi_2^2(0.95)/2 \approx 3$ log-likelihood units below the maximum — bivariate confidence limits using the χ^2 distribution.

S-5 Simulation study for method validation

We conducted a simulation study to assess the ability of our likelihood based approach to infer model parameters using cluster distribution data. We generated synthetic cluster distributions by simulating the branching process models with either a negative binomial distribution or a Poisson lognormal distribution. Synthetic clusters were generated with model parameters sampled using Latin Hypercube Sampling. For the negative binomial model, R_0 was sampled from a uniform distribution between 0.1 and 0.4, and k from a uniform distribution between 0.01 and 0.25. For the Poisson lognormal model, R_0 was sampled from a uniform distribution between 0.15 and 0.45, and σ from a uniform distribution between 1.5 and 2.5. For each of the models, we generated 10,000 synthetic cluster distributions, each with 10,000 clusters.

Using these synthetic cluster distributions, we attempted to estimate the model parameters using the likelihood-based framework. For each synthetic cluster distribution, we maximized the likelihood function using `optim` function in `R`. We compared the estimated parameter values against the true values in Fig. S-1 for the negative binomial models, and in Fig S-2 for the Poisson lognormal model. We find that about 95% of the estimates were within 15% of the true values.

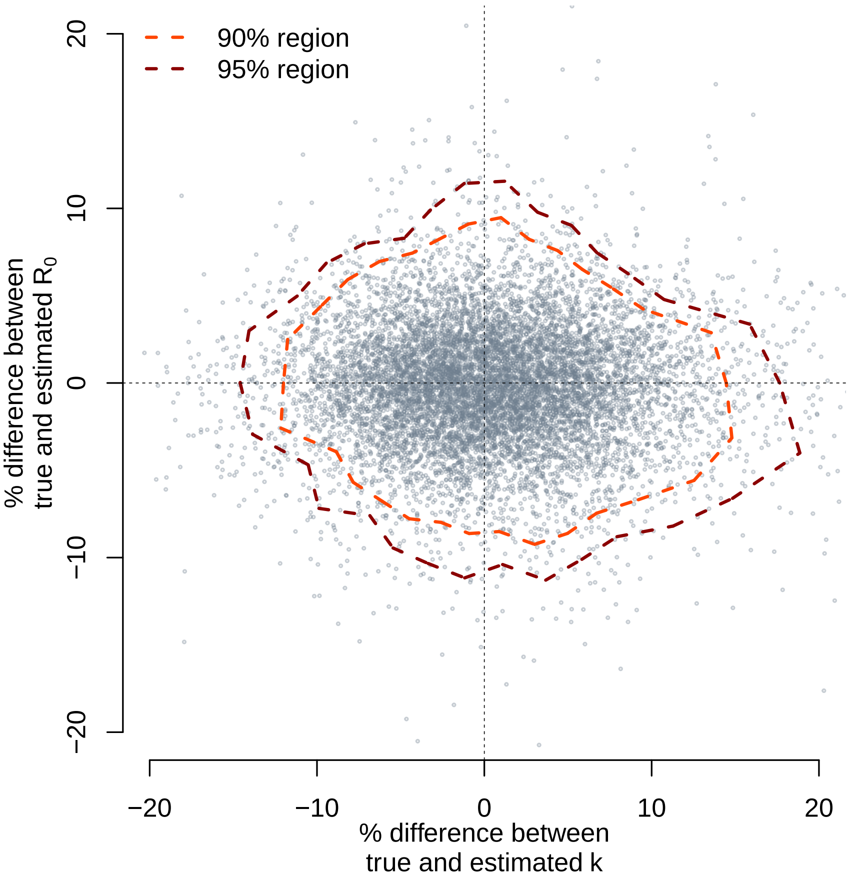

Figure S-1: Simulation Study: inferred parameter estimates for the negative binomial model. Shown are percentage difference between the true and the estimated parameters – R_0 on the y-axis and k on the x-axis. Each dot indicates an inference based on a synthetic distribution; the regions enclosed by light and dark red lines represent 90% and 95% regions, respectively.

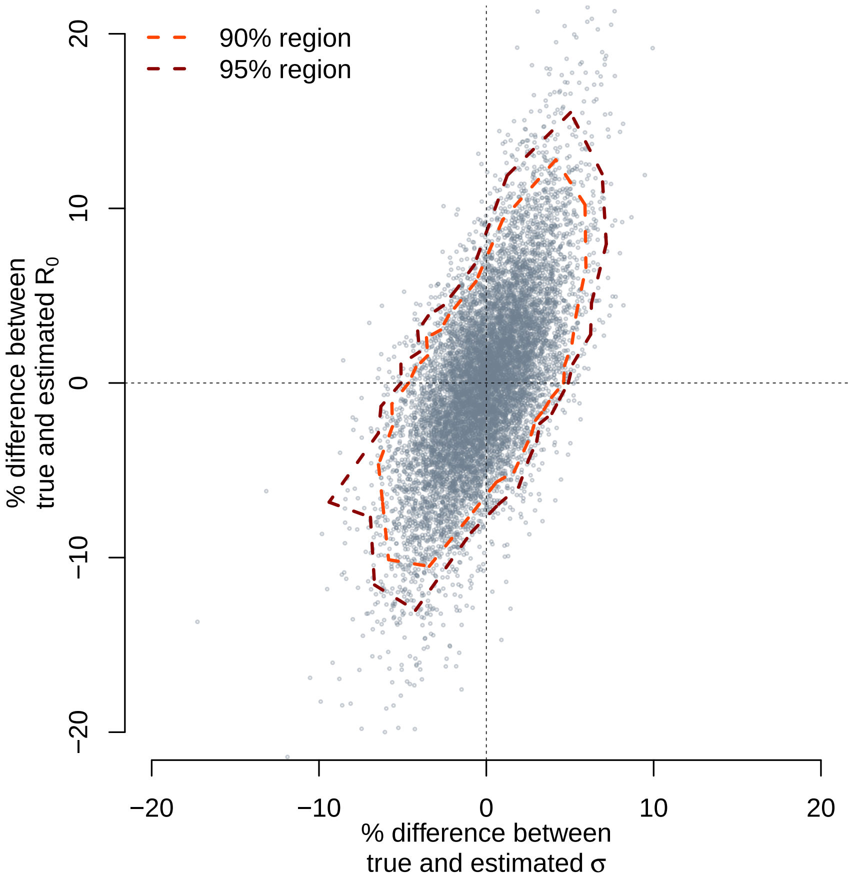

Figure S-2: Simulation Study: inferred parameter estimates for the Poisson lognormal model. Shown are percentage difference between the true and the estimated parameters – R_0 on the y-axis and σ on the x-axis. Each dot indicates an inference based on a synthetic distribution; the regions enclosed by light and dark red lines represent 90% and 95% regions, respectively.

S-6 Under- and over- ascertainment

We assess how potential under- and over- ascertainment of clusters are likely to impact model-based inferences, we conducted a simulation study in which we incorporated potential underreporting and over-ascertainment of clusters. We assumed both under- and over- ascertainment to be binomially distributed. In particular, if the distribution of true secondary cases, Z, is given by $Z \sim \text{Poisson}(\nu)$, we assumed that observed secondary cases $\tilde{Z}$ followed a binomial distribution, $\tilde{Z} \sim \text{Bin}(Z, p)$, if cases were under-reported, and $0 < p < 1$ is the reporting ratio. For over-ascertainment, we assumed that the observed secondary cases $\tilde{Z} = Z + \text{Bin}(Z, p - 1)$, where $p - 1 > 0$ is the rate of over-ascertainment.

Similar to the simulation study conducted for method validation, we generated synthetic cluster distributions by simulating the branching process models with either a negative binomial distribution or a Poisson lognormal distribution. Synthetic clusters were generated with model parameters sampled through Latin Hypercube Sampling. For the negative binomial model, R_0 was sampled from a uniform distribution between 0.1 and 0.4, and k from a uniform distribution between 0.01 and 0.25. For the Poisson lognormal model, R_0 was sampled from a uniform distribution between 0.15 and 0.45, and σ from a uniform distribution between 1.5 and 2.5. For both models, we sampled p between 0.5 and 1.5, i.e., underreporting rates and over- ascertainment rates of up to 50%, and generated 10,000 cluster distributions, each with 10,000 clusters.

We then estimated the model parameters using the likelihood-based framework. For each synthetic cluster distribution, we maximized the likelihood function using `optim` function in `R`. The results for the Poisson lognormal model are shown in Fig. S-3, and the results for the negative binomial model are shown in Fig. S-4.

Under- and overascertainment of clusters had a predictable effect on the inference of R_0, for both models. When the observed clusters were affected by underreporting of cases, R_0 was underestimated (bottom left quadrant in Fig. S-3A or Fig. S-4A), and when the observed clusters included over-ascertainment, R_0 was overestimated (top right quadrant in Fig. S-3A, oFig. S-4A). In both instances, the degree of under- or overestimation was linearly associated with the level of underreporting or overascertainment. Estimates of individual-level heterogeneity in transmission in both models (σ and k) were unaffected by underreporting. When the observed clusters included overascertainment, σ and level of heterogeneity were slightly underestimated in Poisson lognormal model(Fig S-3B). In negative binomial model, inclusion of overascertainment resulted in overestimate of k, which is underestimate of individual-level heterogeneity (since heterogeneity increases with decreasing k).

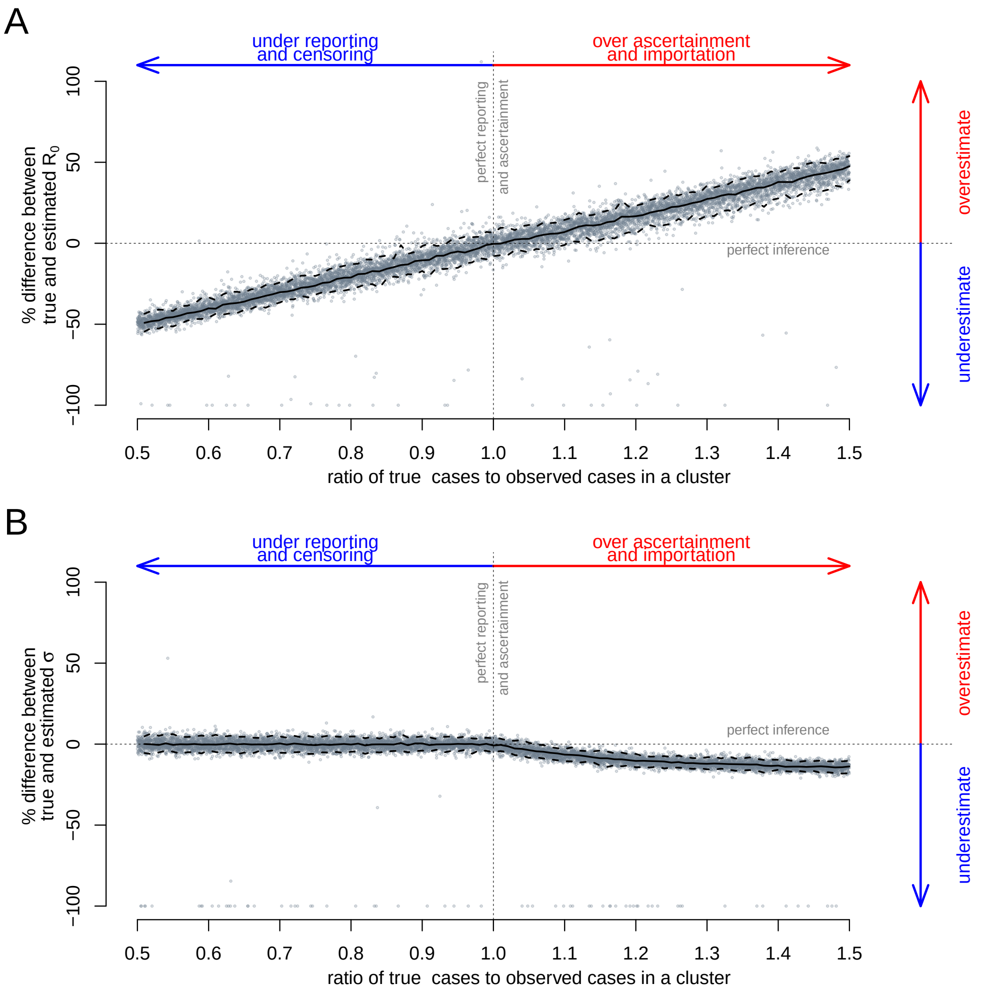

Figure S-3: Sensitivity of model inference to possible imperfections in reporting and cluster ascertainment for Poisson lognormal model. Shown are comparisons between true and estimated values of (A) R_0 and (B) σ (standard deviation of the underlying normal model) as a percentage difference between the true and estimated values (y-axis), under different levels of the ratio between true cases and observed cases (plotted on the x-axis). A ratio of 1 (dashed vertical) indicates perfect reporting of cases and cluster ascertainment (or an equal balance of underreporting and overreporting), ratios < 1 (to the left of the vertical line) indicate net underreporting, and ratios > 1 (to the right of the vertical line) indicate net overascertainment. In both panels, each dot indicates inference result based on an individual cluster, the solid line indicates the median at each y-value, and the dashed lines above and below enclose 95% of the simulation results for the y-value.

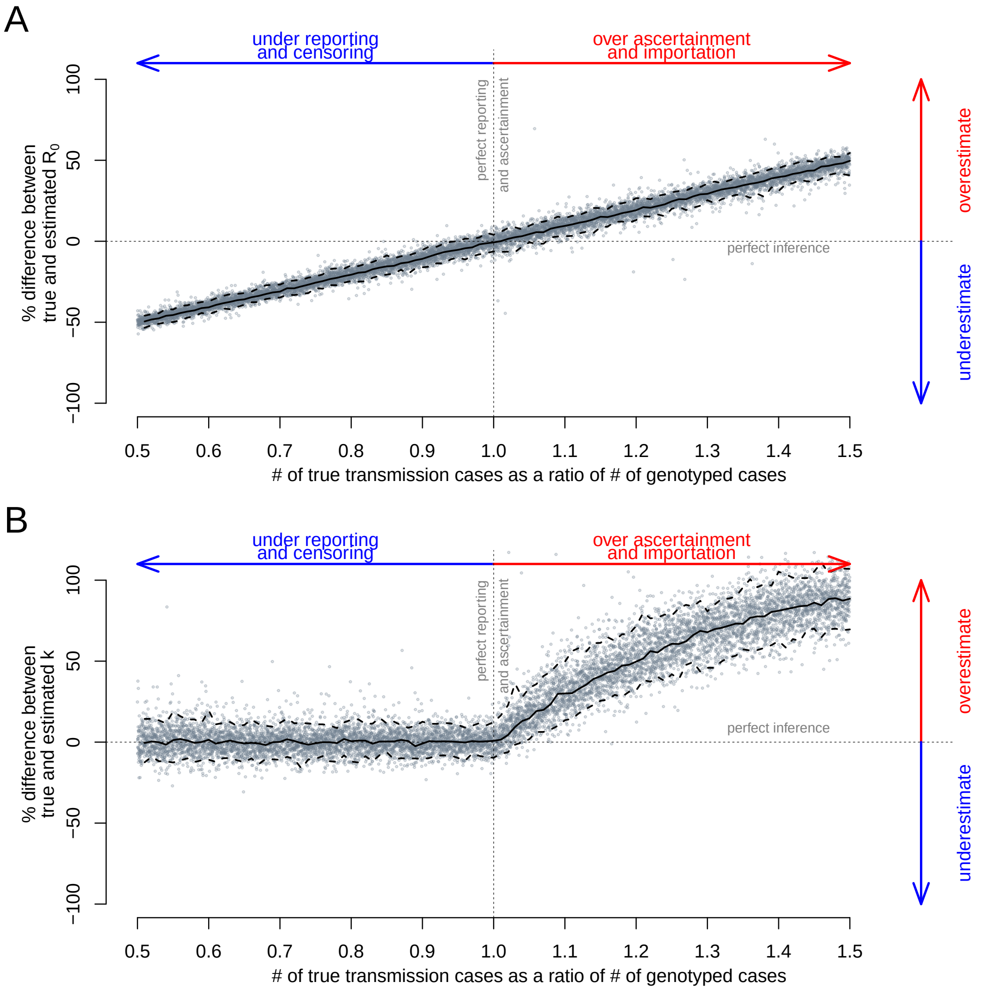

Figure S-4: Sensitivity of model inference to possible imperfections in reporting and cluster ascertainment for negative binomial model. Shown are comparisons between true and estimated values of (A) R_0 and (B) k (dispersion) as a percentage difference between the true and estimated values (y-axis), when the number of true transmission cases varied as a ratio of the number of observed genotyped cases (x-axis). Ratio of 1 (dashed vertical) indicates perfect reporting of cases and cluster ascertainment, ratios < 1 (to the left of the vertical line) indicate scenarios of underreporting, and ratios > 1 (to the right of the vertical line) indicate scenarios of over ascertainment. In both panels, each dot indicates inference result based on an individual cluster, the solid line indicates the median at each y-value, and the dashed lines above and below enclose 95% of the simulation results at the y-value.

S-7 Estimates using negative binomial model

We fit negative binomial model to four separate cluster distributions of TB in the US, where the clusters were defined to be cases reported within: (i) state boundaries and occurring within 5-year time window (Fig. S-5 green); (ii) state boundaries and occurring with 3-year time window (Fig. S-5 purple); (iii) county boundaries and occurring with 5-year time window (Fig. S-5 blue); and (iv) county boundaries and occurring with 3-year time window (Fig. S-5 red). In Fig. S-5, colored circles show the data and colored lines show the predicted distribution using best-fit negative binomial models.

Likelihood surfaces corresponding to negative binomial model fits, including MLE estimates are shown in Fig. S-6.

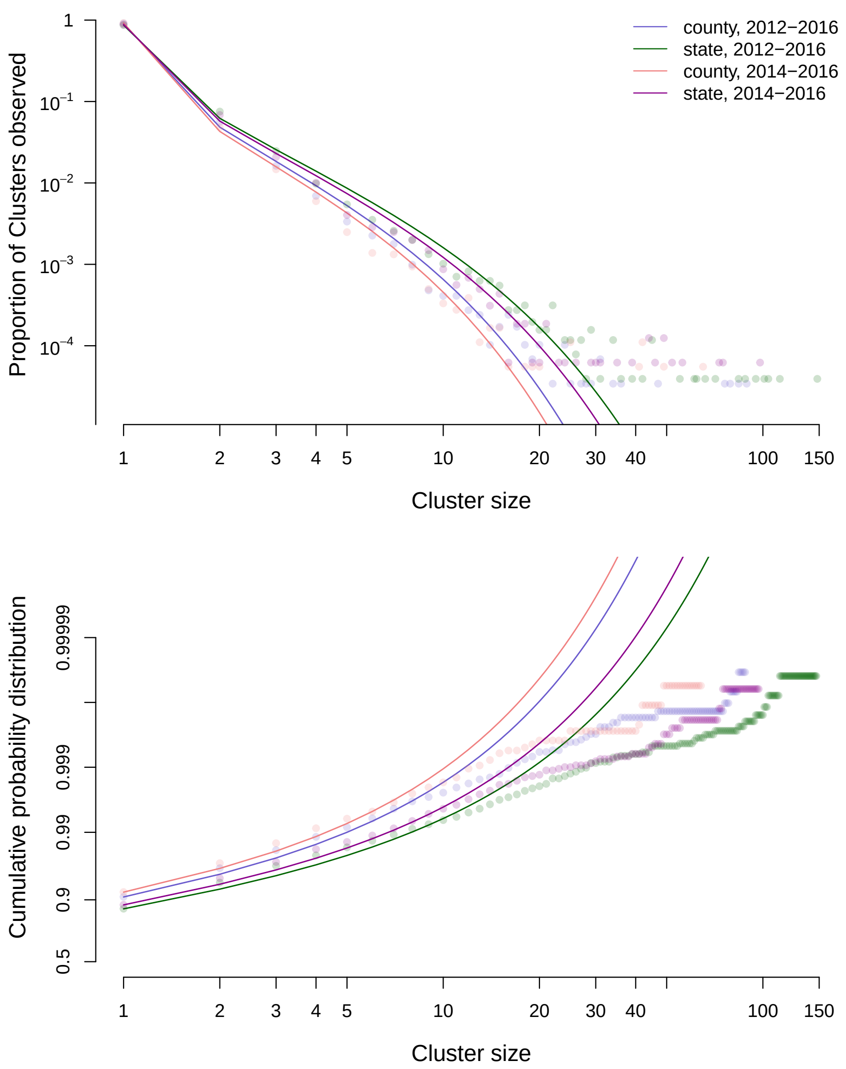

Figure S-5: Comparison of model fits with negative binomial model to US cluster data. We fit negative binomial model to four separate cluster distributions, each using different geographic boundary and time window for cluster ascertainment. This includes clusters defined to be cases reported within (i) state boundaries and occurring within 5-year time window (green); (ii) state boundaries and occurring with 3-year time window (purple); (iii) county boundaries and occurring with 5-year time window (blue); and (iv) county boundaries and occurring with 3-year time window (red). Shown are (top panel) frequency distributions and (bottom panel) cumulative probability distributions corresponding to the best-fit models (shown by colored lines) against the data (shown in colored dots). Cluster size and frequency distributions are plotted on a log-scale, and the cumulative probability distribution on a logit scale.

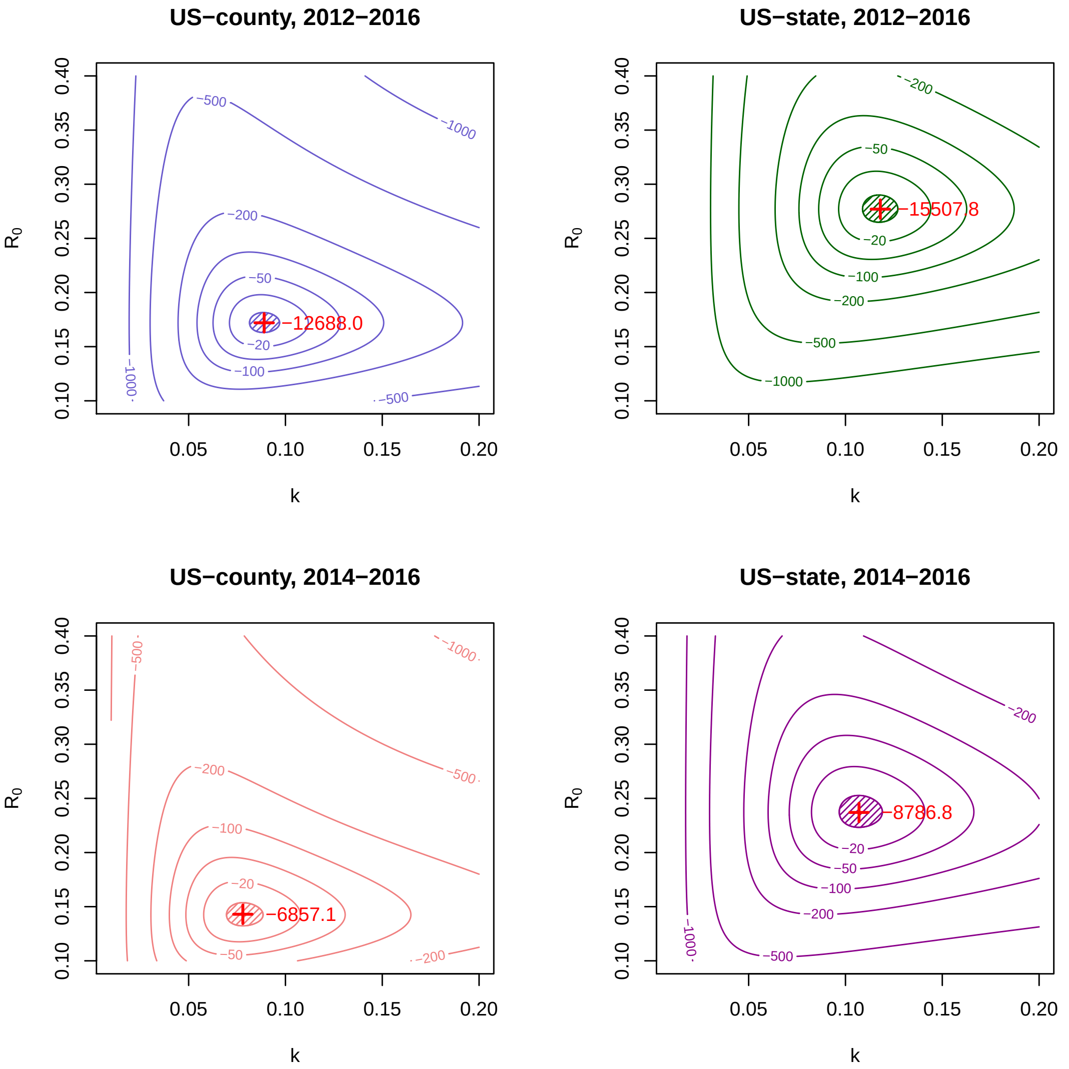

Figure S-6: Likelihood surfaces corresponding to negative binomial model fitting. Each panel shows the likelihood surface (in log scale) corresponding to likelihood estimates using the negative binomial model, when fit to clusters consisting of cases reported within (i) state boundaries and occurring within 5-year time window (top-right); (ii) state boundaries and occurring with 3-year time window (bottom-right); (iii) county boundaries and occurring with 5-year time window (top-left); and (iv) county boundaries and occurring with 3-year time window (bottom-left). Two model parameters, the mean of the distribution, R_0, and the dispersion parameter, k, are plotted on the y- and the x-axis, respectively. In each panel, the red cross indicates the MLE estimate (along with the log likelihood estimate in red), the hatched region around it shows the 95% confidence region, and the countour lines show log likelihood surfaces at various levels of log likelihood values (and the numbers indicate the difference in log likelihood values from the MLE).

Table S-1: Maximum likelihood estimates and 95% confidence intervals for R_0 and variance in the United States based on various cluster definitions.

	Cluster definitions			
	State, 2012-2016	State, 2014-2016	County, 2012-2016	County, 2014-2016
R_0: MLE (95% CI)	0.29 (0.28-0.31)	0.25 (0.24-0.27)	0.19 (0.18-0.2)	0.16 (0.15-0.17)
Variance: MLE (95% CI)	3.8 (2.7-5.3)	3.2 (2.1-5.4)	2.3 (1.6-3.4)	1.9 (1.2-3.6)

S-8 Additional results with the Poisson lognormal model

Figures S-7 and S-8 show detailed results regarding Poisson lognormal model fits to US cluster data. In Fig. S-7, colored circles show the data and colored lines show the predicted distribution using best-fit Poisson lognormal model. Corresponding likelihood surfaces, including MLE estimates are shown in Fig. S-8, and in Table S-1.

The distributions of the individual reproductive numbers for US cluster distributions using the four different cluster definitions, and for the four states are shown in Fig. S-9 and Fig. S-10, respectively.

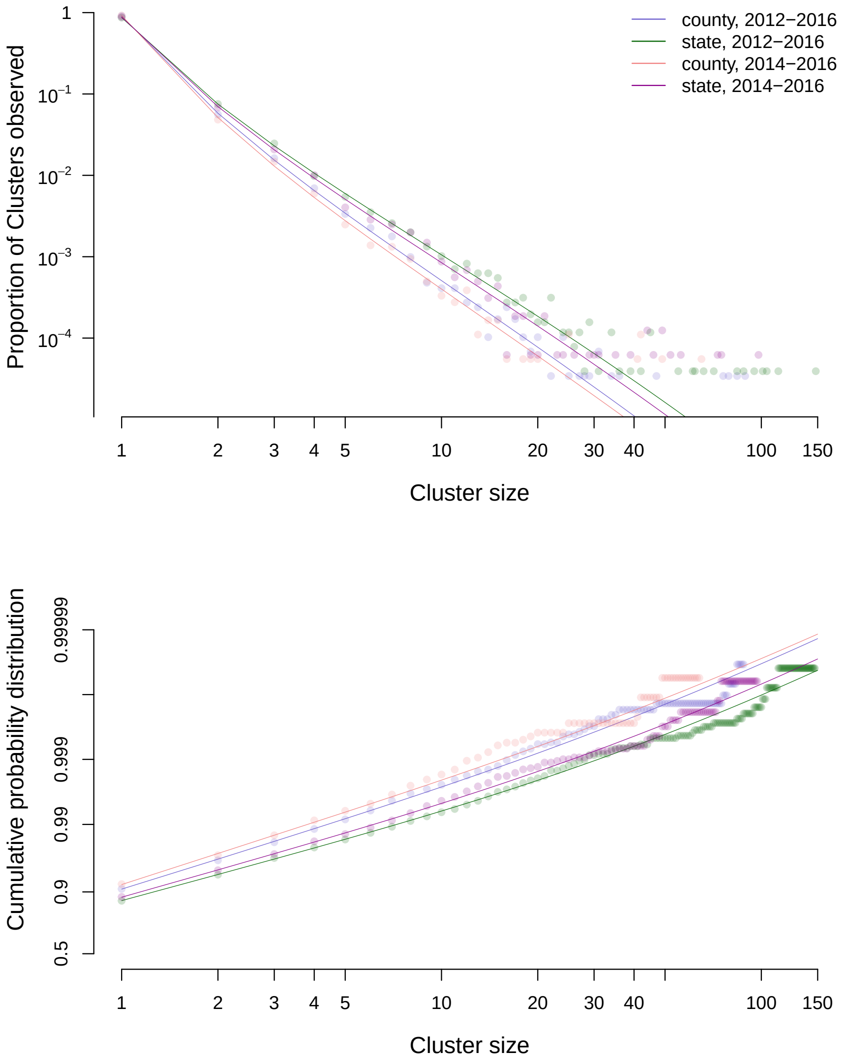

Figure S-7: Comparison of model fits with Poisson lognormal model to US cluster data. We fit negative binomial model to four separate cluster distributions, each using different geographic boundary and time window for cluster ascertainment. This includes clusters defined to be cases reported within (i) state boundaries and occurring within 5-year time window (green); (ii) state boundaries and occurring with 3-year time window (purple); (iii) county boundaries and occurring with 5-year time window (blue); and (iv) county boundaries and occurring with 3-year time window (red). Shown are (top panel) frequency distributions and (bottom panel) cumulative probability distributions corresponding to the best-fit models (shown by colored lines) against the data (shown in colored dots). Cluster size and frequency distributions are plotted on a log-scale, and the cumulative probability distribution on a logit scale.

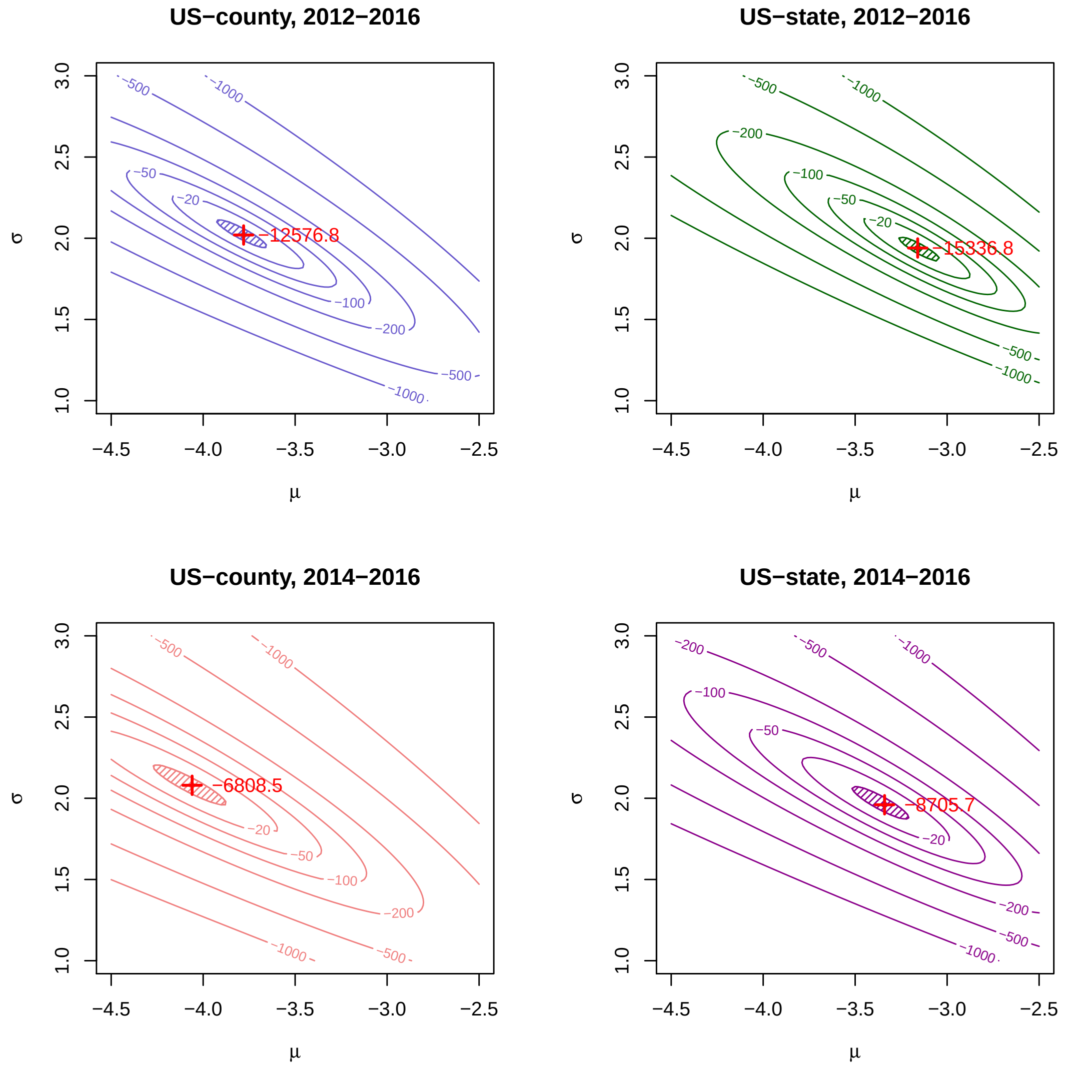

Figure S-8: Likelihood surfaces corresponding to Poisson lognormal model fitting. Each panel shows the likelihood surface (in log scale) corresponding to likelihood estimates using the Poisson lognormal model, when fit to clusters consisting of cases reported within (i) state boundaries and occurring within 5-year time window (top-right); (ii) state boundaries and occurring with 3-year time window (bottom-right); (iii) county boundaries and occurring with 5-year time window (top-left); and (iv) county boundaries and occurring with 3-year time window (bottom-left). Two model parameters, the mean of the underlying normal distribution, μ, and the standard deviation of the underlying normal distribution, σ, are plotted on the x- and the y-axis, respectively. In each panel, the red cross indicates the MLE estimate (along with the log likelihood estimate in red), the hatched region around it shows the 95% confidence region, and the countour lines show log likelihood surfaces at various levels of log likelihood values (and the numbers indicate the difference in log likelihood values from the MLE).

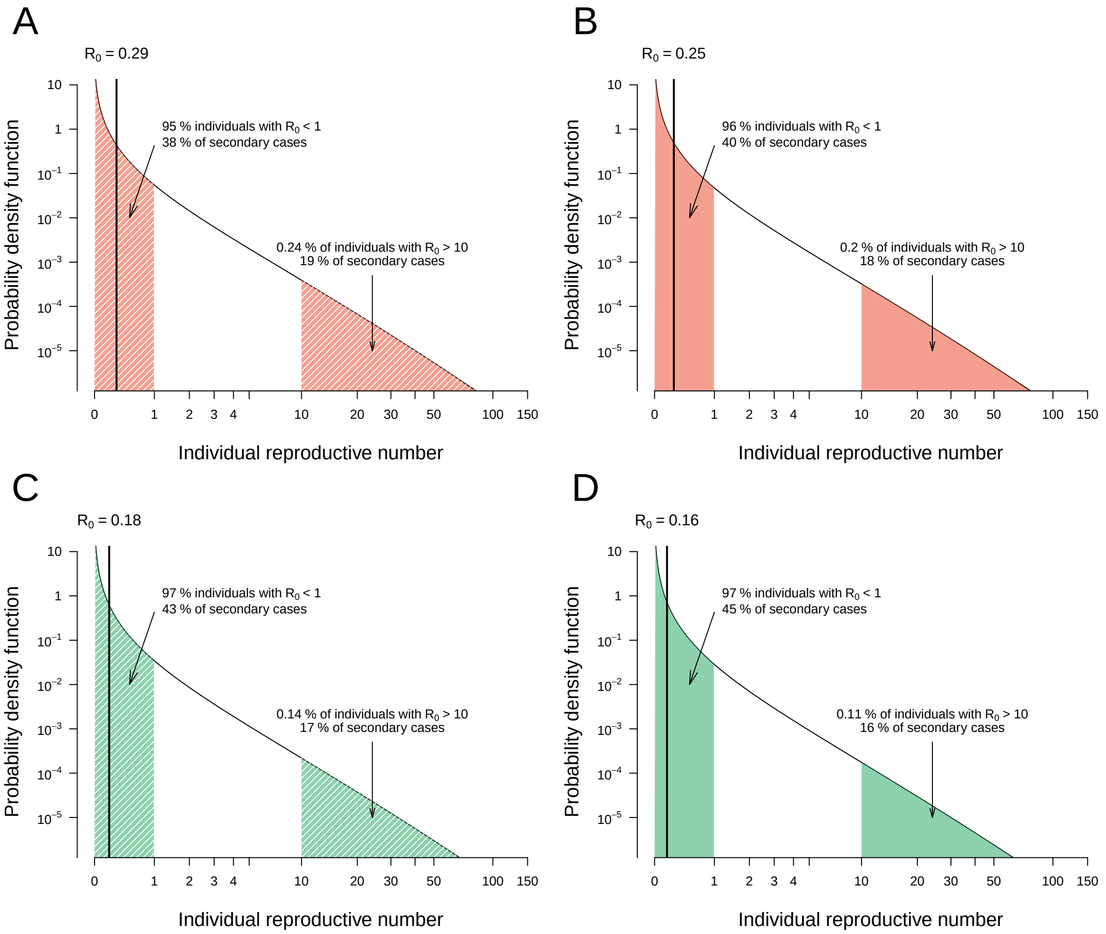

Figure S-9: Comparison of estimated heterogeneity with different various cluster ascertainments. (A) State, 2012-2016; (B) State, 2014-2016; (C) County, 2012-2016; and (D) County, 2014-2016.

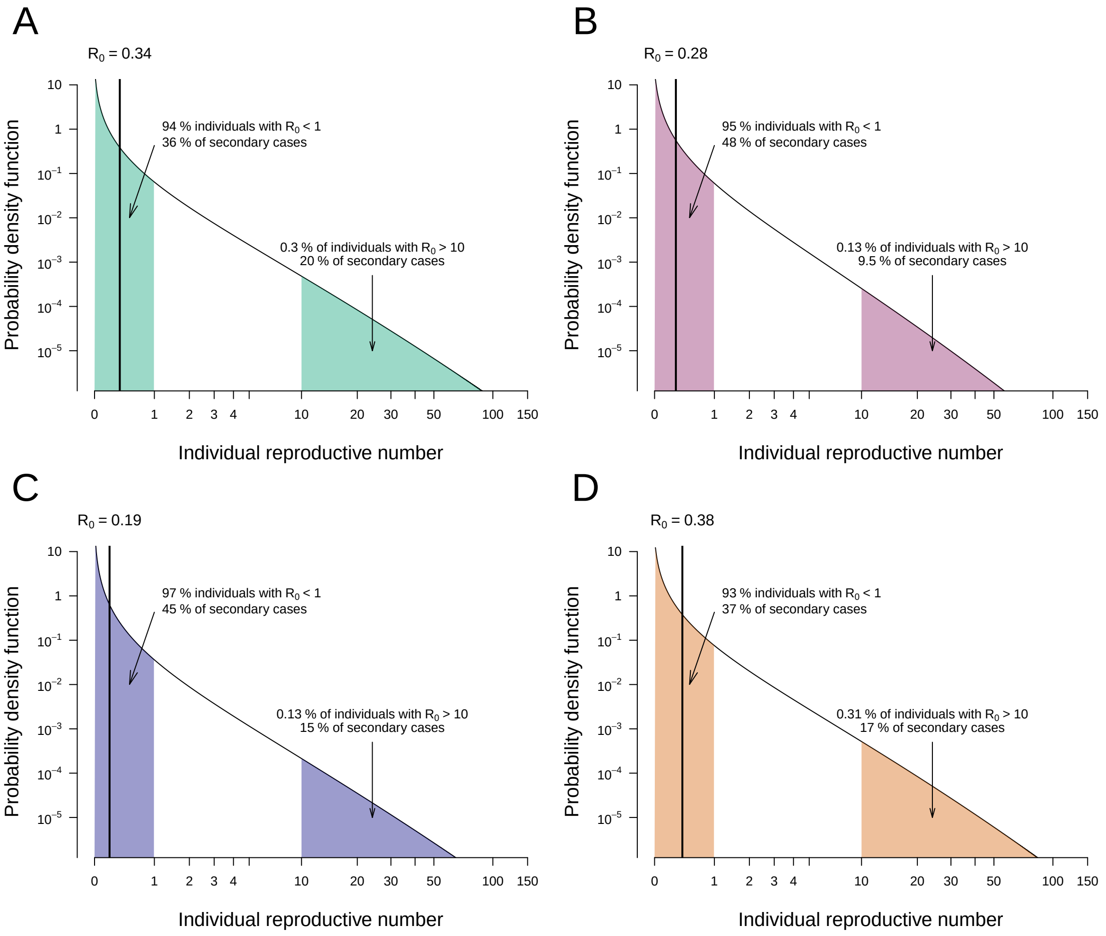

Figure S-10: Estimated heterogeneity in California, Florida, New York, and Texas. (A) California; (B) Florida; (C) New York; and (D) Texas.

Table S-2: Maximum likelihood estimates and 95% confidence intervals for R_0 and variance in California, Florida, New York, and Texas.

	States			
	California	Florida	New York	Texas
R_0: MLE (95% CI)	0.34 (0.3-0.4)	0.28 (0.24-0.36)	0.19 (0.15-0.27)	0.38 (0.33-0.46)
Variance: MLE (95% CI)	4.6 (2.2-12.8)	1.6 (0.8-8.4)	2 (0.6-22)	4 (2-14)

S-9 Comparing model fits for California, Florida, New York, and Texas.

Figures S-11 and S-12 show likelihood surfaces corresponding to model fits in four states: California, Florida, New York, and Texas using either the Poisson lognormal model (Fig. S-11 and Table S-2), or the negative binomial model (Fig. S-12). The maximum likelihood values are indicated on the likelihood surface. For California, Florida, and Texas, the maximum likelihood values corresponding to the Poisson lognormal model are larger than those corresponding to the negative binomial model by more than 3 log units, indicating that Poisson lognormal model is significantly better fit to the data that negative binomial model. For New York, the maximum likelihood values corresponding to the two models were within 3 log units, indicating that neither model was significantly better fit to NY data over the other.

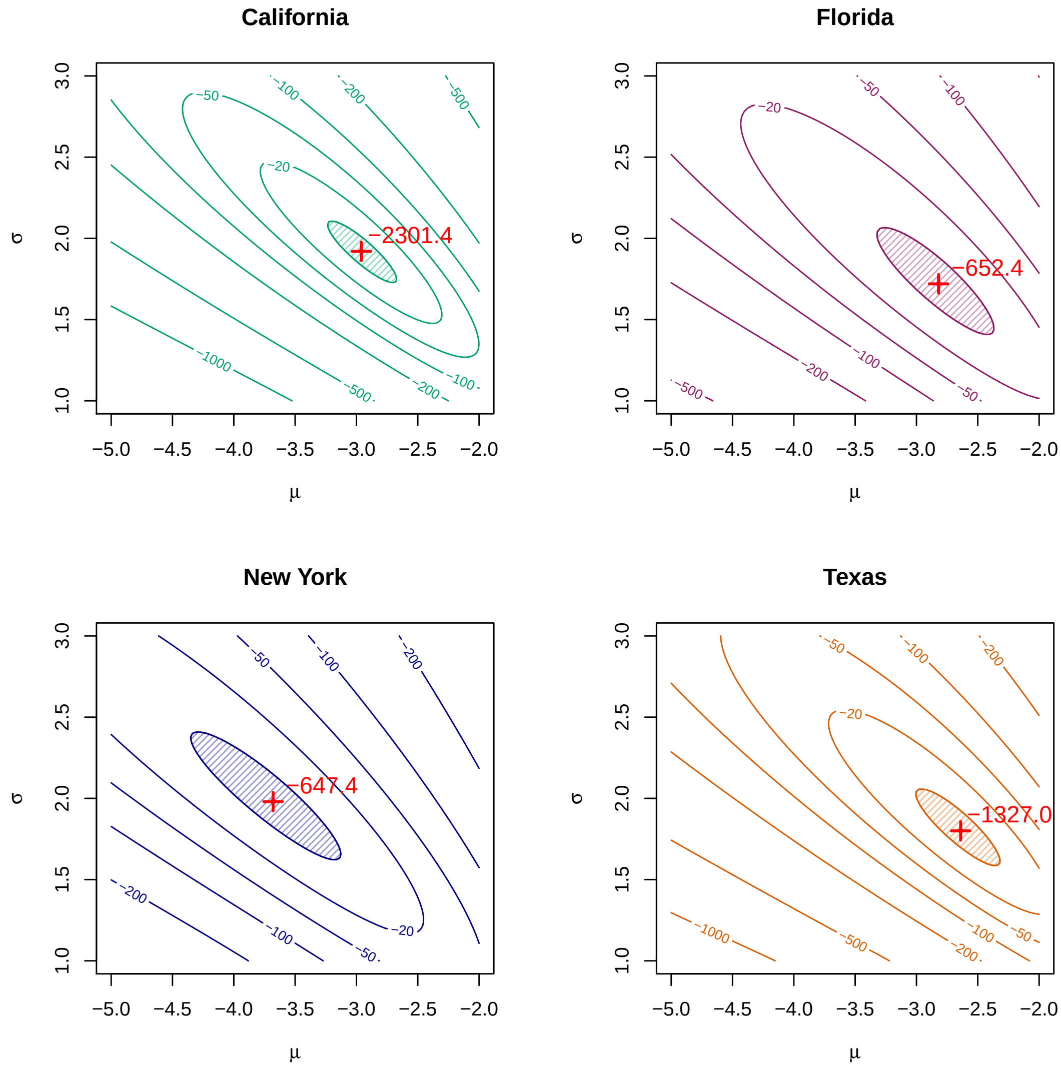

Figure S-11: Likelihood surfaces corresponding to Poisson lognormal model fitting for California, Florida, New York, and Texas. Each panel shows the likelihood surface (in log scale) corresponding to likelihood estimates using the Poisson lognormal model, when fit to clusters consisting of cases reported within state boundaries and occurring within 3-year time window in: (i) California (top-right); (ii) Florida (bottom-right); (iii) New York (top-left); and (iv) Texas (bottom-left). Two model parameters, the mean of the underlying normal distribution, μ, and the standard deviation of the underlying normal distribution, σ, are plotted on the x- and the y-axis, respectively. In each panel, the red cross indicates the MLE (along with the log likelihood estimate in red), the hatched region around it shows the 95% confidence region, and the countour lines show log likelihood surfaces at various levels of log likelihood values (and the numbers indicate the difference in log likelihood values from the MLE).

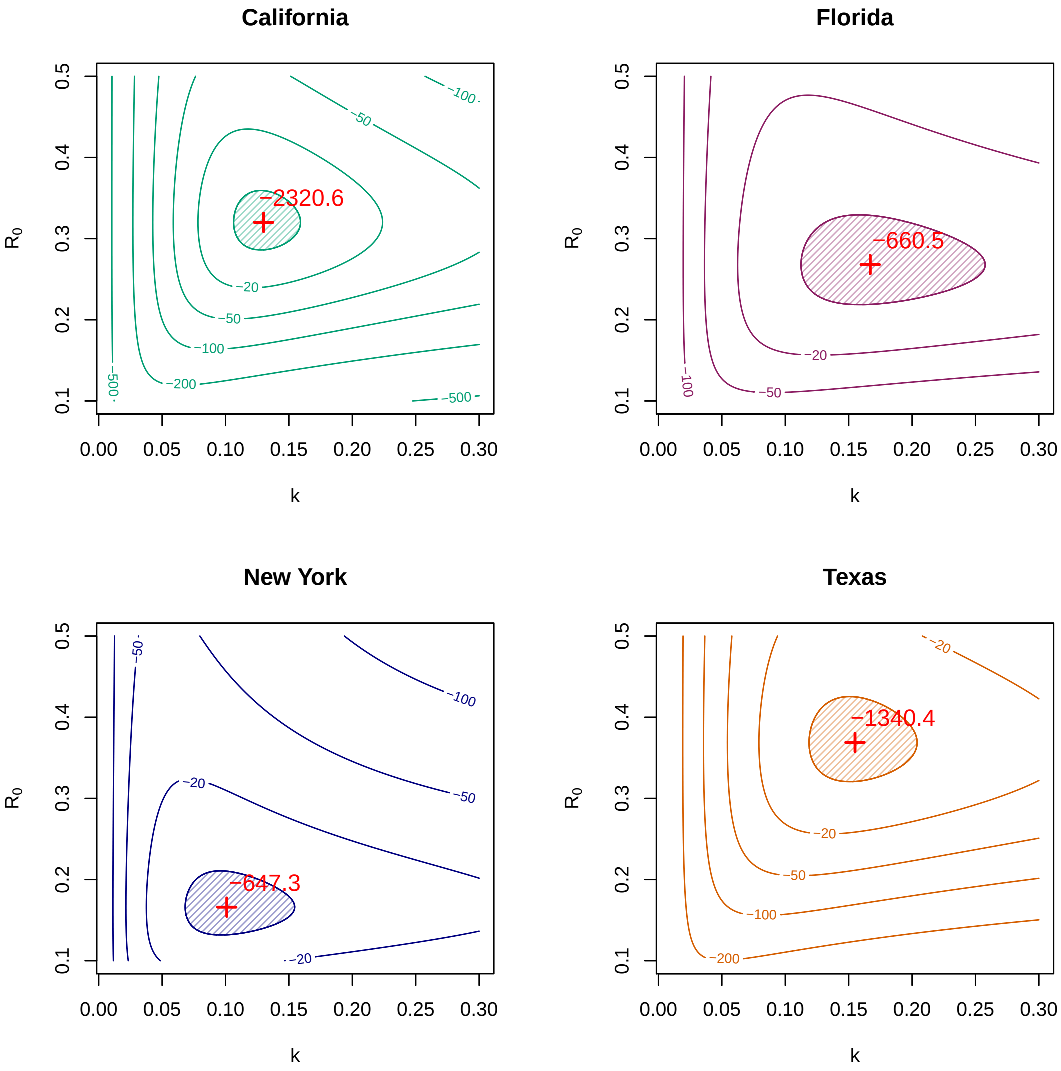

Figure S-12: Likelihood surfaces corresponding to negative binomial model fitting for California, Florida, New York, and Texas. Each panel shows the likelihood surface (in log scale) corresponding to likelihood estimates using the negative binomial model, when fit to clusters consisting of cases reported within state boundaries and occurring within 3-year time window in: (i) California (top-right); (ii) Florida (bottom-right); (iii) New York (top-left); and (iv) Texas (bottom-left). Two model parameters, the reproductive number R_0, and the dispersion parameter, k, are plotted on the y- and the x-axis, respectively. In each panel, the red cross indicates the MLE (along with the log likelihood estimate in red), the hatched region around it shows the 95% confidence region, and the countour lines show log likelihood surfaces at various levels of log likelihood values (and the numbers indicate the difference in log likelihood values from the MLE).

S-10 Comparing model fits in the Netherlands and the United Kingdom.

Using similar data on cluster size distribution of TB cases from the Netherlands and the United Kingdom, made available by Brooks-Pollock et al [3], we compared the models fits using Poisson lognormal model and the negative binomial model. Data and model fits using the Poisson lognormal model are shown in Figure S-13, and model fits using the negative binomial model are shown in Figure S-14. Our results confirm the authors' finding that Poisson lognormal models better fit to these cluster size distributions: likelihood estimates are larger for Poisson lognormal models for dataset, and the models fits better capture the tail of the cluster size distributions.

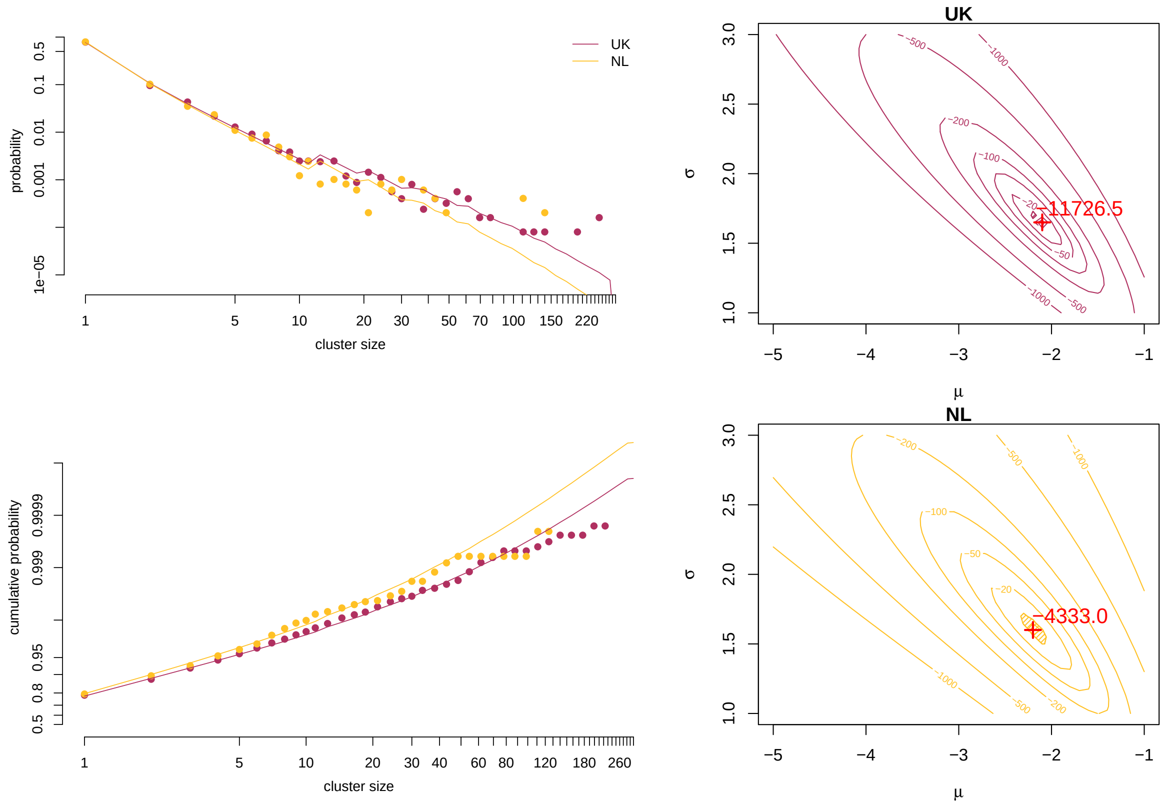

Figure S-13: Poisson lognormal model fits for the data from the Netherlands and the United Kingdom. [Top-left] Shown in colored dots are cluster size distribution for the United Kingdom (purple) and the Netherlands (yellow); and the colored lines indicate distributions corresponding to the best model fits using the Poisson lognormal model. [Bottom-left] Shown are the same data and model fits with the vertical axis presented as a cumulative probability distribution. [Top-right] Shown in contour lines are likelihood surfaces (in log scale) for models fits on the UK data: two model parameters, the mean of the underlying normal distribution, μ, and the standard deviation of the underlying normal distribution, σ, are plotted on the x- and the y-axis, respectively. [Bottom-right] Shown in contour lines are likelihood surfaces (in log scale) for models fits on the Netherlands data: two model parameters, the mean of the underlying normal distribution, μ, and the standard deviation of the underlying normal distribution, σ, are plotted on the x- and the y-axis, respectively. In each panel, the red cross indicates the MLE (along with the log likelihood estimate in red), the hatched region around it shows the 95% confidence region, and the countour lines show log likelihood surfaces at various levels of log likelihood values (and the numbers indicate the difference in log likelihood values from the MLE). Data presented from the Netherlands and the United Kingdom are taken from Brooks-Pollock et al [3].

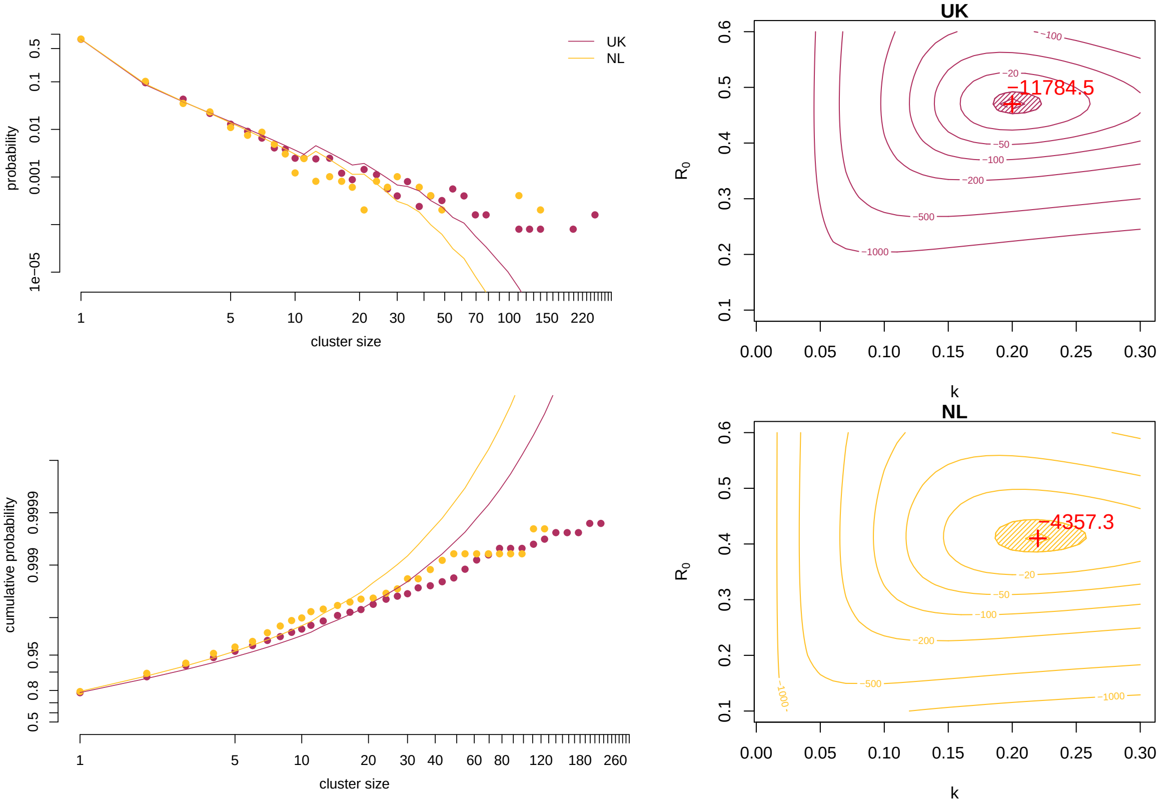

Figure S-14: Negative binomial model fits for the data from the Netherlands and the United Kingdom. [Top-left] Shown in colored dots are cluster size distribution for the United Kingdom (purple) and the Netherlands (yellow); and the colored lines indicate distributions corresponding to the best model fits using the negative binomial model. [Bottom-left] Shown are the same data and model fits with the vertical axis presented as a cumulative probability distribution. [Top-right] Shown in contour lines are likelihood surfaces (in log scale) for models fits on the UK data: two model parameters, the reproductive number R_0, and the dispersion parameter, k, are plotted on the y- and the x-axis, respectively. [Bottom-right] Shown in contour lines are likelihood surfaces (in log scale) for models fits on the Netherlands data: two model parameters, the reproductive number R_0, and the dispersion parameter, k, are plotted on the y- and the x-axis, respectively. In each panel, the red cross indicates the MLE (along with the log likelihood estimate in red), the hatched region around it shows the 95% confidence region, and the countour lines show log likelihood surfaces at various levels of log likelihood values (and the numbers indicate the difference in log likelihood values from the MLE). Data presented from the Netherlands and the United Kingdom are taken from Brooks-Pollock et al [3].

References

[1] Lloyd-Smith JO, Schreiber SJ, Kopp PE, et al. Superspreading and the effect of individual variation on disease emergence. *Nature* 2005;438(7066):355–359. doi:10.1038/nature04153.

[2] Bulmer MG. On Fitting the Poisson Lognormal Distribution to Species-Abundance Data. *Biometrics* 1974;30(1):101–110.

[3] Brooks-Pollock E, Danon L, Korthals Altes H, et al. A model of tuberculosis clustering in low incidence countries reveals more transmission in the United Kingdom than the Netherlands between 2010 and 2015. *PLOS Computational Biology* 2020;16(3):1–14. doi:10.1371/journal.pcbi.1007687.

[4] Grundy PM. The Expected Frequencies in a Sample of an Animal Population in Which the Abundances of Species are Log-Normally Distributed. Part I. *Biometrika* 1951;38(3/4):427–434.

[5] Izsák R. Maximum likelihood fitting of the Poisson lognormal distribution. *Environmental and Ecological Statistics* 2008;15(2):143–156. doi:10.1007/s10651-007-0044-x.

[6] Nishiura H, Yan P, Sleeman CK, et al. Estimating the transmission potential of supercritical processes based on the final size distribution of minor outbreaks. *Journal of Theoretical Biology* 2012;294:48–55. doi:https://doi.org/10.1016/j.jtbi.2011.10.039.

[7] Blumberg S, Lloyd-Smith JO. Inference of R0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. *PLOS Computational Biology* 2013;9(5):1–17. doi: 10.1371/journal.pcbi.1002993.

[8] Blumberg S, Lloyd-Smith JO. Comparing methods for estimating R0 from the size distribution of subcritical transmission chains. *Epidemics* 2013;5(3):131–145. doi: https://doi.org/10.1016/j.epidem.2013.05.002.

[9] Farrington CP, Kanaan MN, Gay NJ. Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics* 2003;4(2):279–295. doi:10.1093/biostatistics/4.2.279.

[10] Ypma RJF, Altes HK, van Soolingen D, et al. A Sign of Superspreading in Tuberculosis: Highly Skewed Distribution of Genotypic Cluster Sizes. *Epidemiology* 2013;24(3):395–400.