As at 15 June 2015, a large transmission cluster of Middle East respiratory syndrome coronavirus (MERS-CoV) was ongoing in South Korea. To examine the potential for such events, we estimated the level of heterogeneity in MERS-CoV transmission by analysing data on cluster size distributions. We found substantial potential for superspreading; even though it is likely that $R_0 < 1$ overall, our analysis indicates that cluster sizes of over 150 cases are not unexpected for MERS-CoV infection.

**MERS-CoV transmission**

There have been 1,288 cases of Middle East respiratory syndrome (MERS) reported worldwide as at 10 June 2015 [1]. Many of these have been index cases, likely to have been infected from an animal reservoir, but there have also been several clusters of human-to-human transmission. An imported MERS case with a travel history to the Arabian Peninsula resulted in a new cluster in South Korea, with 150 cases reported as at 15 June 2015 [2]. This raises two important questions about the transmission dynamics of MERS coronavirus (MERS-CoV). First, how much heterogeneity is there in MERS-CoV transmission in the absence of animal–human infection? Second, given such heterogeneity, what are the chances of observing an outbreak as large as the one in South Korea?

The dynamics of an outbreak depend on both $R_0$ – the average number of secondary cases generated by a typical infectious individual – and individual heterogeneity in transmission. Such heterogeneity can be estimated by describing the distribution of secondary cases as a negative binomial distribution with dispersion parameter $k$, where $k < 1$ suggests that transmission is overdispersed, and hence outbreaks can include superspreading events [3,4]. However, there is currently no measure of transmission heterogeneity for MERS-CoV. Using reported outbreak data, we examined the extent of individual variation in MERS-CoV transmission, and estimated the probability of observing clusters as large as the one in South Korea.

### Analysing cluster data

We analysed data on MERS cluster sizes for cases reported up to 31 August 2013 [5]. For comparison, we also considered data from two other reports, up to 21 June 2013 [6] and 8 August 2013 [7]. Cases with known epidemiological links were classified as a cluster. Single index cases were considered as independent clusters of size one. Although more cases have since been reported [1], it is not entirely clear how many clusters there have been. We therefore chose to focus on published cluster data (Table), which also made it possible to compare our results with previous analyses.

To estimate $R_0$ and $k$ from the distribution of cluster sizes, we used a likelihood-based inference method based on branching processes with the offspring distribution following a negative binomial distribution with mean $R_0$ and dispersion parameter $k$. This distribution is widely used to describe overdispersed count data in biology and epidemiology [4], and has the useful property that Poisson ($k = ∞$) and geometric offspring distributions ($k = 1$) are special cases of it. The probability that an index case generates a cluster of size $j$ is [8,9]:

$$r_j = \frac{\Gamma(kj + j - 1)}{\Gamma(kj) \Gamma(j + 1)} \left(\frac{R_0}{k}\right)^{j-1} \left(1 + \frac{R_0}{k}\right)^{kj+j-1}.$$  

Therefore the likelihood of observing $n_j$ clusters of size $j$ is:

$$L = \prod_{j=1}^{\infty} r_j^{n_j}.$$
For given values of $R_0$ and $k$, the probability that an index case generates a transmission cluster of size $j$ or greater is:

$$p_j = 1 - \sum_{i=0}^{j-1} \tau_i.$$  

Assuming $N$ introductions of infections into the human population, the probability that at least one cluster of size $j$ or greater occurs is $1 - (1 - p)^N$. All analyses were done in the R software environment for statistical computing [10].

**Findings**

Using available cluster data, we jointly estimated $R_0$ and the dispersion parameter $k$ for MERS-CoV (Figure 1).

Analysis of severe acute respiratory syndrome (SARS) coronavirus transmission during the early stages of the outbreak in Singapore suggested $k = 0.16$ (90% confidence interval (CI): 0.11–0.64) [3] (the study cited 90% CI owing to the paucity of available data). Our estimate for MERS-CoV is similar, with $k = 0.26$ (90% CI: 0.11–0.87, 95% CI: 0.09–1.24). As it is not always clear from case reports which cases are epidemiologically linked, we also estimated $k$ using data from two other studies of clusters [6,7]. These data included fewer clusters and were less conclusive regarding the amount of over-dispersion, with $k = 0.61$ (95% CI: 0.16–∞) [7] and $k = 2.94$ (95% CI: 0.23–∞) [6].

Our estimate for $R_0$ was 0.47 (95% CI: 0.29–0.80). The maximum likelihood estimate (0.47), which is independent of $k$ [8], agrees with previous work [5-7]. However unlike earlier studies, which assumed the distribution of secondary cases to be either geometric (i.e. $k = 1$) [5,7] or Poisson ($k = \infty$) [6], our upper 95% CI is larger. This is because allowing for potential over-dispersion increases the uncertainty surrounding the estimate of $R_0$ (Figure 1).

There is an intricate relationship between the basic reproduction number, $R_0$, the dispersion parameter, $k$, and the probability of observing a large transmission cluster (Figure 2A). For a given value of $k$, increasing $R_0$ also increases the probability of observing large clusters. If $R_0$ is low, a higher variation in the number of secondary cases (i.e. smaller $k$) increases the probability of observing large transmission clusters owing to the potential for superspreading. The effect of $k$ is reversed for values of $R_0$ near one, where a smaller $k$ reduces the probability of observing large clusters. This is because a higher variation in the number of secondary cases increases the probability that an infected index case does not generate further cases [3]. Interestingly, the
area where the effect of overdispersion for a given value of $R_0$ switches from increasing to decreasing the probability to observe large cluster sizes lies near the maximum likelihood estimate for MERS-CoV (Figure 2B).

Finally, we calculated the expected probability of observing a MERS-CoV transmission cluster of a given size or greater, by integrating across the full parameter distribution in Figure 1. Using the estimated distribution of $k$ substantially increases the probability that index cases generate large clusters (Figure 3A), compared with the situation in which the number of secondary cases are assumed to be geometrically distributed ($k = 1$). The probability that a single index case infected with MERS-CoV results in a cluster of 150 cases or more – as observed in South Korea – is 0.04%. Assuming different numbers of MERS-CoV introductions into human populations, the probabilities that at least one such outbreak occurs are 2.5% (100 introductions), 5.6% (500 introductions), 7.4% (1,000 introductions) and 9.3% (2,000 introductions).

**Discussion**

Our results suggest that MERS-CoV transmission is highly overdispersed, and hence there is substantial potential for superspreading events. This finding is corroborated by a similar analysis of MERS-CoV outbreak size distributions [11]. Given that hundreds of MERS-CoV index cases have been reported to date, our analysis indicates that occasional cluster sizes of over 150 cases – such as the one in South Korea – should not be unexpected. We also found a non-linear relationship between the basic reproduction number, $R_0$, dispersion parameter, $k$, and outbreak size: when $R_0 < 0.9$, the probability of obtaining a large cluster increases as the process becomes more overdispersed; as $R_0$ approaches one, the effect is reversed and a higher level of overdispersion reduces the chances of a large cluster for a given value of $R_0$.

There are some limitations to our study. Case data may be subject to bias or under-reporting. However, such factors will generally drive up estimates of overdispersion [4] and hence are unlikely to alter our overall conclusions. It can also be difficult to conclusively identify outbreak clusters from case data. We therefore considered three different data sources, and found evidence of overdispersion in the two largest and most recent data sets.

Other infections, including SARS [3] and Ebola virus disease [12], also exhibit overdispersed transmission patterns. However, it can be difficult to establish precisely which factors drive superspreading events. For MERS-CoV, the observed overdispersion may result
from a combination of factors, including individual viral shedding and contact rates, hospital procedures and location, as well as population structure and density [13]. Even if such factors cannot be disentangled, measuring the overall extent of overdispersion – as we have done here – can help with the interpretation of surveillance data, and enable more realistic analysis of disease transmission and control [14].

* Authors’ correction

A typo in the code that was used for the analysis resulted in erroneous estimates for the dispersion parameter k and the confidence intervals surrounding the basic reproduction number $R_0$. All numbers in the text and the figures have been updated using the corrected estimates of k. Furthermore, figures have been updated to include parameter estimates derived from the largest data set of cluster sizes as reported by Poletto et al. [5]. The study now refers to a more recent analysis of MERS-CoV outbreak size distributions that showed very similar results [11]. These changes were made on 10 August 2015, at the request of the authors.

The authors have made the code available on GitHub (https://github.com/calthaus/MERS) to ensure reproducibility of the analysis.

---

**Figure 3**

Probability that a MERS-CoV transmission cluster of a given size or greater occurs*

Panel A: Probability that a single index case generates a transmission cluster of a given size or greater. Panel B: Probability that at least one transmission cluster of a given size or greater occurs assuming different numbers of MERS-CoV introductions into human populations. We assumed $R_0 = 0.47$ (0.29–0.80). Note the logarithmic and linear scale of the horizontal axis in panels A and B.

**Acknowledgements**

AK is supported by the Medical Research Council (MR/K021524/1) and the RAPIDD program of the Science & Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health. CLA received funding through an Ambizione grant from the Swiss National Science Foundation (grant 136737).

**Conflict of interest**

None declared.

**Authors’ contributions**

AJK and CLA designed the study, carried out the analysis, and wrote the manuscript.

**References**


