Web Appendix

Details of the Statistical Model

Occurrence of symptoms among confirmed MERS-CoV infections

Our approach to estimating the true number of cases of MERS-CoV is based on the assumption that those who are tested through active surveillance develop symptoms suggestive of MERS-CoV and die at the same rate as all infections (regardless of detection status). That is:

$$\Pr(y = 1 \mid a, s = 0) = \Pr(y = 1) = \xi_a$$

where y is an indicator variable for the development of symptoms suggestive of MERS-CoV infections, a is the age class of the case; $s = 1$ if a case was detected through passive surveillance (i.e., was tested for MERS-CoV due to having symptoms) and $s = 0$ if they were detected through active surveillance; and ξ_a is the symptomatic infection ratio (SymIR) for age group a. Likewise:

$$\Pr(z = 1 \mid a, s = 0) = IFR_a$$

where $z = 1$ if a case eventually dies, and IFR_a is the age specific probability of death among age group a.

We assume that the vast majority of people detected through the passive surveillance system (i.e., reported as tested because of their symptoms) will have had severe symptoms, or die. We further assume that all cases who die without having previously been detected, will be detected by the passive surveillance system upon death. However, some people may be detected through this system despite having less severe symptoms, not considered suggestive of MERS-CoV infection. Hence, the probability of symptoms or death in those detected via passive surveillance is (noting that $z = 1 \Rightarrow s = 1$):

$$\Pr(\, y = 1 \mid a, s = 1\,) =$$

$$\frac{\left(\Pr\left((y = 1 \wedge s = 1) \vee (z = 1)\middle|a\right)\right)}{\Pr(s = 1)} = \frac{1 - \Pr(z = 0|a)\Pr(y = 0 \vee s = 0|a)}{\Pr(s = 1)}$$

$$\frac{[1 - (1 - IFR_a)(1 - \psi_2\xi_a)]}{1 - (1 - IFR_a)(1 - (\psi_1(1 - \xi_a) + \psi_2\xi_a))}$$

where ψ_1 is the probability that a case with symptoms not suggestive of MERS-CoV infection is tested for MERS-CoV through passive surveillance, and ψ_2 is the probability that a case with symptoms suggestive of MERS-CoV infection is tested. Note that in our current data we expect ψ_1 to be near 0, hence the above probability will be near one.

Under the assumption that 100% of deaths due to MERS-CoV infections are identified by passive surveillance upon dying, the proportion of those detected by passive surveillance who die is:

$$\Pr(z = 1 \mid a, s = 1) = \frac{IFR_a}{1 - (1 - IFR_a)(1 - (\psi_1(1 - \xi_a) + \psi_2\xi_a))}$$

Observed incidence of MERS-CoV

Given the above parameters, it is possible to calculate the expected number of cases observed by active and passive surveillance in each time step (in our model taken to be weeks). Let $I_{a,s,t}$ be the observed incidence of MERS-CoV at time t:

$$I_{a,s,t} \sim \begin{cases} Poisson(\phi\lambda_{a,t}P_a) & \text{if } s = 0 \\ Poisson((1-\phi)(\psi_1(1-\xi_a) + \psi_2\xi_a)\lambda_{a,t}P_a) & \text{if } s = 1 \end{cases}$$

where ϕ is the probability of being detected through active surveillance, $\lambda_{a,t}$is the force of infection at time t and P_a is the population in age class a.

Model specification and priors

The above model was specified in the RStan probabilistic modeling language and fit using MCMC methods. Uninformative priors were used for all parameters except the probability of detection of symptomatic MERS-CoV cases, where we assumed we were 95% confident that the true value was at least 90%. Four chains of 1,000 iterations were run, and final parameter estimates come from pooling the last 500 runs of all chains. Convergence was assessed using Gelman and Rubin's $\hat{R}$ statistic[11].

Parameter Estimates

Below are descriptions and estimates of the parameters of the statistical model. Age specific symptomatic rates (ξ_a) infection fatality rates (IFR_a), and total attack rates ($\sum_t \lambda_{a,t} P_a$) are shown in Table 1 of the main text.

parameter	estimate	description
ψ_1	0.018 (0.008, 0.311)	mildly/asymptomatic passive detection rate
ψ_2	0.956 (0.769, 1.00)	severe symptoms passive detection rate
ϕ	0.122 (0.096, 0.150)	active detection rate

Performance of Method on Simulated Data

In order to evaluate the performance of our method given that our assumptions hold, we simulated an epidemic and attempted to estimate the true population and underlying parameters.

We considered three possibilities for how the IFR and probability of developing symptoms varied by age: (1) the IFR and probability of developing symptoms increase linearly by age

category on the logit-scale, (2) slight deviations from linearity, and (3) complete non-linearity. Under each scenario we estimated the true total number of infections when nearly all severe cases are detected through passive surveillance ($\psi_2 = 0.95$), a moderate number of severe cases are so detected ($\psi_2 = 0.50$) and few severe cases are so detected ($\psi_2 = 0.30$). In all scenarios 5% of cases are identified through active surveillance ($\phi = 0.05$) and 1% of cases without severe symptoms are still detected through passive surveillance ($\psi_1 = 0.01$).

In each case the true number of total infections is 4,096, with 2,603 of those cases being in age class 1, 1,120 being in age class 2, and 373 in age class 3. Of the nine scenarios examined, in only one scenario was the true value not covered by the 95% credible interval: when only 30% of symptomatic cases were passively detected and the probability of symptoms and death did not increase linearly, or nearly linearly, with age.

Web Table 1

Scenario	Total Cases	Age Class 1	Age Class 2	Age Class 3
Logit-linear				
high detection	4087 (3180, 5622)	2738 (1982, 3821)	1016 (813, 1439)	354 (281, 497)
mid detection	3460 (2480, 5764)	2165 (1513, 3649)	986 (706, 1671)	302 (208, 528)
low detection	4182 (2680, 6857)	2678 (1687, 4508)	1092 (691, 1798)	392 (241, 667)
Near logit-linear				
high detection	4523 (3812, 5538)	2776 (2256, 3482)	1132 (1102, 1673)	415 (335, 538)
mid detection	3508 (2117, 5859)	2307 (1363, 3880)	877 (519, 1480)	322 (186, 564)
low detection	5453 (3688, 8210)	3380 (2256, 5154)	1574 (1056, 2413)	484 (311, 771)
Non logit-linear				
high detection	3808 (3590, 4260)	2577 (2410, 2851)	894 (767, 1123)	323 (272, 396)
mid detection	3741 (2986, 5106)	2412 (2015, 3477)	1012 (702 1316)	311 (228,447)
low detection	6614 (5767, 8014)	4571 (3832, 5436)	1512 (1138, 2086)	568 (434, 795)

Sensitivity to Reporting Periods

To assess the possible sensitivity of our results to reporting delays we re-ran our main analysis allowing the rate of active and passive case detection to change at two key events: the change in health ministers (April 21), and the release of an updated case definition (May 13). This analysis showed no significant difference in active or passive surveillance case detection rates over the course of the outbreak. However, it did estimate a non-statistically significant decrease in passive case detection after the revised case definition was released on May 13th. This analysis predicts a marginal, but not statistically significant, increase in the total number of cases versus our main analysis.

Web Table 2

age group	observed cases	varying surveillance estimated infections	main manuscript estimated infections
0-9	15	53 (28, 97)	50 (26, 89)
10-19	28	103 (65, 156)	97 (61, 150)
20-29	97	366 (272, 502)	347 (257, 469)
30-39	132	411 (308, 560)	384 (296, 504)
40-49	115	248 (190, 337)	235 (182, 307)
50-59	138	214 (169, 278)	204 (163, 263)
60-69	78	109 (82, 148)	101 (78, 133)
70+	113	134 (106, 175)	127 (101, 166)
total	721	1,639 (1,377, 2,051)	1,548 (1,327, 1,883)

Sensitivity to relative probability of symptoms among the actively detected.

To test the sensitivity of the model to differences in the probability of developing symptoms among the actively surveilled population compared to the general population, we considered models where those infections identified through active surveillance had twice the odds, and half the odds of developing severe symptoms or dying compared to the general population across age groups. In the former case, the estimated total number of cases goes up by 31% (2,035 vs 1,548) while in the latter it decreases by 20% (1,234 vs. 1,548).

Web Table 3

age group	observed cases	2x the odds of dying or developing symptoms	main manuscript estimated infections	0.5x the odds of dying or developing symptoms
0-9	15	65 (34, 120)	50 (26, 89)	40 (22, 71)
10-19	28	127 (80, 195)	97 (61, 150)	76 (49, 114)
20-29	97	472 (348, 638)	347 (257, 469)	263 (197, 356)
30-39	132	528 (406, 687)	384 (296, 504)	285 (222, 374)
40-49	115	322 (248, 423)	235 (182, 307)	181 (144, 238)
50-59	138	259 (202, 336)	204 (163, 263)	171 (138, 217)
60-69	78	117 (88, 155)	101 (78, 133)	91 (72, 119)
70+	113	135 (103, 185)	127 (101, 166)	120 (97, 154)
Total	721	2,035 (1,705, 2,475)	1,548 (1,327, 1,883)	1,234 (1,069, 1,500)