INDIVIDUAL BASED MODELLING OF THE COVID-19 PANDEMIC

MURAD BANAJI[*], 2ND MAY 2020

Abstract. This draft document presents some insights from individual based stochastic simulations of the COVID-19 pandemic. Results from the modelling are used to make claims about several data sets including from India, the UK, and Sweden, whose pandemics are at different stages and taking different courses. It is work in progress, and supercedes earlier documents on results of modelling the pandemic. Associated code is at https://github.com/muradbanaji/COVIDSIM with change logs tracking how the code itself has evolved. The latest version of this document is available at http://maths.mdx.ac.uk/research/modelling-the-covid-19-pandemic/.

1. Introduction and aims. The basic desire which motivates this work is to gain some insight into the underlying dynamics of the COVID-19 pandemic as it unfolds in different contexts. The reality is that in any given country we only ever have access to incomplete data about infections and fatalities, and the goal is to see if we can come to any robust conclusions by combining modelling and this incomplete data. One assumption is that fatality data, while often incomplete, is more consistent than data on the levels of infection; the latter are, after perhaps the very early days of an outbreak, likely to consistently underestimate the true levels of infection, possibly by orders of magnitude. This said, fatality data too must be approached with caution: fatality figures are known to be underestimates in many contexts. Bearing this in mind, even when we use the data as it is, it is always a good idea to explore the effects of underreporting of fatalities on model predictions.

Some basic prior knowledge of the behaviour of the pandemic is used in building the model. Knowledge about COVID-19 seems to be evolving rapidly and all modelling-based claims must be treated with caution. The construction aims to be sufficiently flexible so that as more information comes to light, the model can be updated accordingly, whether this be by altering the values of parameters, putting distributions on parameters which are currently fixed, or introducing new phenomena and events into the model.

Built into the modelling are attempts to model the effects of various interventions, including quarantining, physical distancing and hygiene measures, and restrictions on freedom of movement ("lockdown"). All such measures will be referred to collectively as **mitigation**. Any model which ignores mitigation can at best be useful in the very early stages of an outbreak. It is hard to estimate the mitigation parameters and often our knowledge of these is limited to the dates on which they begin, and some intuition about how severe they were. However, we'll see that playing a little with mitigation parameters allow us to reproduce several data sets reasonably well.

2. Methodology and model assumptions. The model, termed `COVIDSIM`, is stochastic and agent based. The agents are infected individuals and it is assumed that there is a pool of individuals "available" to be infected (this may or may not be infinite). Time steps are days, and each day an infected individual can either (i) infect new individuals; (ii) do nothing; (iii) recover, or (iv) die. The decision on whether a given individual infects others or not at a given time step is a probabilistic one, determined by overlaying all the different effects in play at that moment, including any mitigation measures in force. Roughly speaking, predetermined infection events are associated with individuals but can be "cancelled" by mitigation. We list the key features of the model.

1. **Time course.** People who are infected either recover after a certain number of days or die after a certain number of days. The day of infection is termed day 1, so day n means $n-1$ days after the infection event. The mean day of death or recovery are parameters to be controlled (their current default values are 17 and 20 respectively). Individual values may deviate from the mean values, with deviations chosen from a binomial distribution.

[*] Middlesex University, London, Department of Design Engineering and Mathematics. m.banaji@mdx.ac.uk.

2. The **fatality rate**, namely the probability with which an infected person dies, is a model parameter which we vary. Several studies, many not as yet peer-reviewed, suggest fatality rates in the range of 0.3-0.7%, for example [1, 2, 3]. Lower estimates exist, but are controversial (e.g., [4]). In our model, the fatality rate hardly affects the dynamics of the epidemic, except inasmuch as events (such as lockdowns) can be triggered by fatalities reaching a certain level. Rather, varying the fatality rate can be used to achieve a range of estimates on underlying levels of infection consistent with the available data. Thus we end up with conclusions such as "*If the fatality rate is k%, then there are m unreported infections for every reported infection.*"

3. **Infectivity.** Each individual would, in the absence of quarantining, lockdown, or other mitigation measures, infect a certain number of new individuals. This number, termed the "infectivity" of the individual, is drawn from some distribution. Currently we work with truncated approximations to the Poisson and geometric distributions with expected values which are parameters to be varied. The default choice is the geometric distribution. The expected value of infectivity can be regarded as R_0, the **basic reproduction number** of the epidemic, and it will be referred to as such. It varies from outbreak to outbreak and can be estimated from the initial dynamics of the outbreak in the absence of mitigation. The choice of distribution has most effect in the early days of an outbreak when there are relatively few infected individuals and consequently stochastic effects can be more dramatic. Currently there seems to be considerable ignorance about the extent to which asymptomatic or presymptomatic individuals spread the disease [5]. We make no assumptions about whether the infected individuals in the model are symptomatic or not. There will, in any case, be infected individuals who are not infective simply as a consequence of the distribution of infectivities.

   Remark 2.1 (Truncation of distributions and R_0). The truncation of the distributions slightly alters the true R_0 values from the prescribed values. In earlier versions of the model, the truncation was done too crudely and consequently the R_0 values mentioned in documents before April 24th are too high.

4. **Infective window.** All infected people are infective during a certain number of days post-infection. We'll call this the *infective window*. The start and end day of the infective window are model parameters which can be changed. The assumption of a fixed infective window for all individuals is to simplify matters and is **open to question and exploration**. A median incubation period of 5.1 days is given in the summary [6], and a typical time course of disease lasting approximately 14 days after the onset of symptoms is described at [7]. The window default values are days 3 to 14 (inclusive) post-infection. This represents a rough guess about the period of infectivity based on these sources and other such as [8]. With greater knowledge, it would be possible to put distributions on the start point and duration of the infective window.

5. **Time distribution of infection events.** The times at which an individual infects others are, for simplicity, sampled from a uniform distribution over the infective window. Again, **this assumption is open to question** – it is likely that individuals are more or less infective during certain phases of the disease [8].

6. **Quarantining.** Some percentage of infected people may be quarantined on some day post-infection. The details are obviously different in different settings. The proportion quarantined and the mean day of quarantining are parameters to be varied. Deviations from the mean day of quarantining can be chosen, if so desired, from a binomial distribution. If an individual is quarantined, then they do not infect any more people. Quarantining can be effective as a means of controlling the disease provided it occurs early enough in the disease cycle and reaches a sufficient number of infected individuals.

7. **Testing.** Some percentage of those quarantined are also tested. This is how we obtain simulated testing data to compare with official numbers of infections. Note that the simulated infections reported today are not simply a proportion of today's true infection levels – there is an assumed delay. Thus testing data from simulations is, roughly speaking, a scaled snapshot of the situation some time in the past. The assumption of a fixed percentage of infections being picked up by testing over the whole time-course of an epidemic is unlikely to be accurate – testing strategies and protocols can change over time. Some simulations when matched with actual data suggest precisely this.

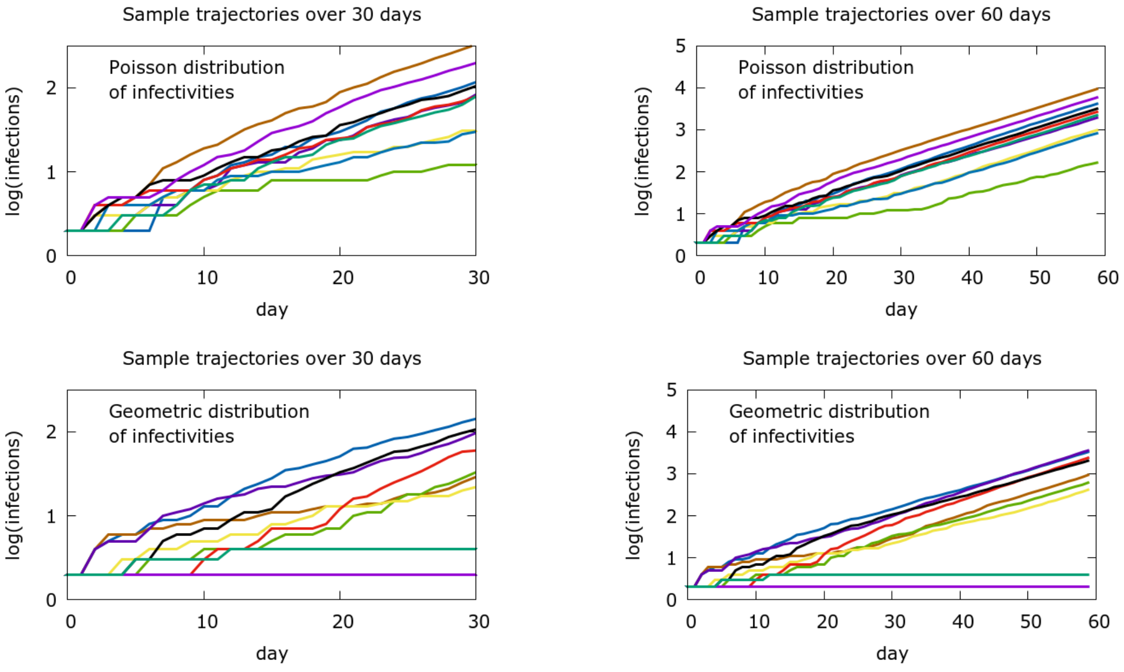
8. **Physical distancing** is used as a catch-all term for all measures which decrease contact. If there is physical distancing, then with some probability infection events which would have taken place do not take place. This probability is a parameter to be varied. If we say "30% effective physical distancing, then this means that with probability 3/10 each infection event which would have taken place is stopped by this measure. Physical distancing measures can be triggered by crossing a certain number of fatalities or a certain number of confirmed infections.
9. **Infectible population and herd immunity.** The total "infectible" population is a notional quantity representing the number of individuals available to be infected. We choose not to use the word "susceptible" in order not to give the impression that it truly represents all those who are susceptible to infection. Even without mitigation measures, the infectible population need not be the whole population: a situation with little movement of indviduals could be modelled via a smaller infectible population. As the model is not a spatial model, varying the infectible population can be used as a crude mechanism to model diffusion effects and localisation of the disease. The infectible population affects the probability with which an infection event takes place. For example if the total number of infected individuals are 75% of the total infectible population, then there is a 75% probability that a predetermined infection event fails to take place. This allows modelling of **herd immunity** without explicitly introducing non-infected individuals into the model.
10. **Lockdown.** Lockdown is likely to result in some level of physical distancing; but in addition lockdown means restrictions on freedom of movement. This is modelled by reducing the total infectible population. The extent to which this population is reduced would, in general, depend on the stage at which lockdown occurs and the strictness of the lockdown. It is possible to model imperfect lockdown as "leakage" into the infectible population at some prescribed rate and beginning at some prescribed moment after lockdown. Lockdown can be triggered by crossing a certain number of fatalities or a certain number of confirmed infections.
11. **Initiating an outbreak.** An outbreak may be initiated by introducing a certain number of infected individuals – this number is a parameter which can be controlled. These in turn may "spawn" further infected individuals as described above. The dynamics of any infected individual, namely the moments at which they will spawn further infected individuals are predetermined the moment the individual is "created" by sampling from distributions as described above. Of course, if they are quarantined or there is physical distancing or lockdown, then some or all of these infection events do not take place. We see in Figure 1 that, because of the stochastic nature of the model, introducing a small number of infected individuals means that initial dynamics of the model can vary significantly from simulation to simulation. It is more or less likely, given certain model parameters, that an introduction of infected individuals fails to cause an outbreak.

3. Basic behaviour of the model.

The stochastic nature of the model leads to variations in the initial dynamics. The greater the variance in infectivities, the greater the variation in initial dynamics in different simulations. In particular, the choice of geometric distribution for infectivities leads to greater initial variation in model trajectories than the choice of Poisson distribution as illustrated in Figure 1. Moreover, if we introduce a small number of infected individuals there is a greater probability that there will be no outbreak if infectivities follow the geometric distribution rather than the Poisson distribution with its lower variance. Provided there is an outbreak then, in the absence of herd immunity and with high probability, the slope of a trajectory in logarithmic space eventually becomes approximately constant, namely we have exponential growth in the number of infections with some fixed doubling time.

Physical distancing. If we introduce physical distancing into a model, then this reduces the rate of infection, and the effective value of R_0 (more properly speaking, of R_t). This results in slowing the spread of infection, and can even end the outbreak if this value is reduced to less than 1 (Figure 2).

Herd immunity can also lead to the end of an outbreak, more of less quickly depending on the size of the infectible population. Note that we do not explicitly model the infectible individuals, but if we allow herd immunity then whenever a scheduled infection event is due to occur, with some probability it fails to occur as the event actually involves a non-susceptible individual.

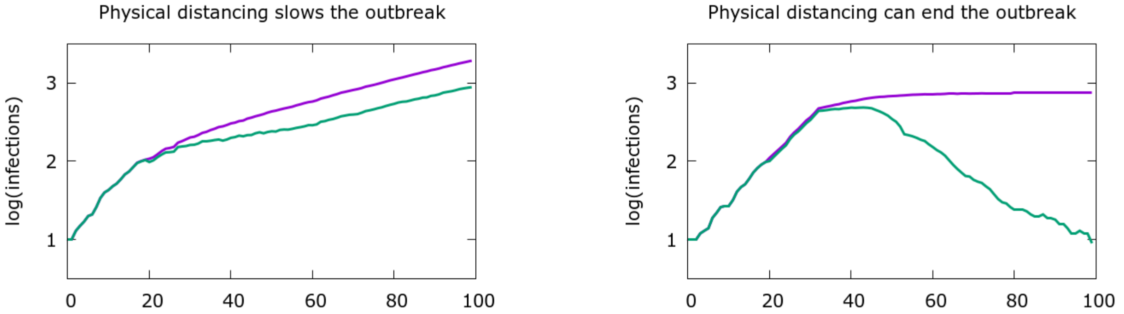

Sample trajectories over 30 days

Sample trajectories over 60 days

Sample trajectories over 30 days

Sample trajectories over 60 days

Fig. 1. *The outcomes of 10 runs of the model are shown, first with a Poisson distribution of infectivities (top) and then with a geometric distribution of infectivities (bottom). The time course of total infections is shown. In each case* $R_0 = 2.5$*, and simulation is initiated with 2 infected individuals. There are no mitigation measures, and other model parameters are set at their default values discussed above.*


Physical distancing slows the outbreak

Physical distancing can end the outbreak

Fig. 2. *The outcomes of one run of the model where physical distancing is introduced at the first death. Total infections and active infections are shown.* R_0 *is set at* 2.5 *and the simulations are initiated with 10 infected individuals with a Poisson distribution of infectivities. Left:* 50% *effective physical distancing introduced at the moment of the first fatality leads to a drop in apparent* R_0 *observed as a drop in the slope of total infections. But the effective* R_0 *remains above* 1 *and the active infections still increase. Right:* 80% *effective physical distancing at the first fatality leads to a drop in apparent* R_0 *to less than* 1 *and active infections decrease. Note, nevertheless, the slow time-course of decrease.*

Quarantining is another mitigation strategy that can slow or even reverse an outbreak. In Figure 4 we see the effect of a highly effective quarantining regime where 1 in 2 infected individuals are quarantined before becoming infective. As we would expect this reduces the effective R_0 to half of its value. As $R_0 = 2.5$ initially in the simulations shown in Figure 4, quarantining reduces it effectively to 1.25; however, stochastic effects mean that even with $R_0 > 1$, the outbreak may be avoided altogether. In simulations we see the reversal of the outbreak much more frequently when infectivity is chosen from a geometric distribution, rather than a Poisson distribution.

So far, we have considered the behaviour of infections in the model without considering fatalities. The time course of fatalities in the model lags that of infections, with the extent of the delay dependent on the

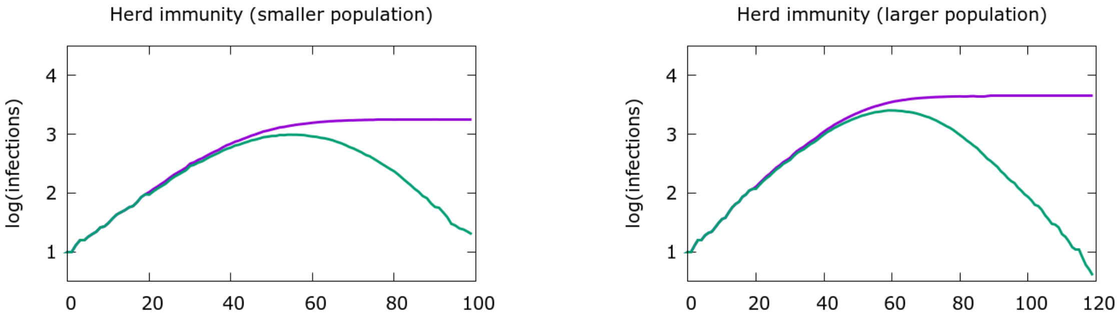


Fig. 3. *The outcomes of running of the model with herd immunity.* R_0 *is set at* 2.5 *and the simulations are initiated with 10 infected individuals with a Poisson distribution of infectivities. Total infections and active infections are shown. Left: a smaller infectible population of* 2000 *individuals. Right: a larger infectible population of* 5000 *individuals. In both cases the outbreak dies away, but this happens faster and with a lower peak for a smaller population. In both simulations, a little over* 85% *of the population have been infected at the end of the outbreak.*

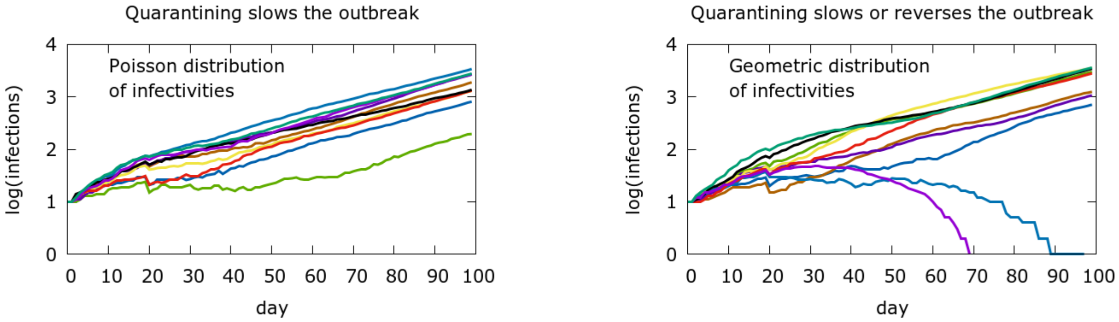


Fig. 4. *The outcomes of running the model ten times with each individual having a* 50% *probability of being quarantined on day 3 of the infection. Quarantining starts occurring from the very beginning of the outbreak.* R_0 *is initially set at* 2.5, *and simulations are initiated with 10 infected individuals. Only active infections are shown. Left: a Poisson distribution of infectivities leads to considerable slowing of the outbreak. Right: a geometric distribution of infectivities leads to slowing of the oubreak and in some cases reversal of the outbreak.*

number of days to death, the value of R_0, and the fatality rate. This is discussed further in Section 4.5. We see in Figure 5 that the trajectory of fatalities is similar to that for infections, but shifted down (according to the fatality rate) and to the right (by the typical delay to death). As we would expect, increasing either R_0 or the fatality rate decreases the delay between initial infections and the first deaths (see Section 4.5).

4. Applications of the model. The model has so far been applied to data from India as a whole, from the UK, from Sweden, and from two Indian states, namely Maharashtra and Gujarat. Note that the different places may be at different stages in this first wave of the pandemic.[1] In each case, the initial goal has been to see if it is possible to reproduce the dynamics of both infection and fatality data, bearing in mind the difficulties in interpreting these data sets discussed briefly earlier.

4.1. Modelling the Indian data. We are able to approximately match the time-course of the Indian outbreak using a variety of parameter choices. Infection and fatality data is drawn from the excellent site https://www.covid19india.org/. The sequence of advisories and mitigation measures which may have affected the disease transmission cannot easily be reduced to a couple of events. However we focus on the following. Around March 12th, when there was the first COVID-19 fatality in India,

[1] To refer to "waves" may turn out to be quite problematic. It is quite possible that the slow dynamics means that it is hard to pinpoint a moment when a "wave" ends. In simulations, even with parameters such that the epidemic dies out, we often see a trickle of deaths several months after the peak of infection.

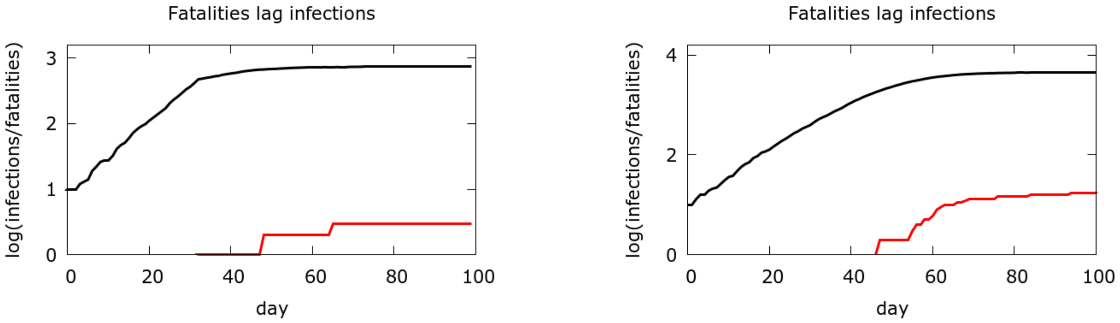


FIG. 5. *The model is run with $R_0 = 2.5$ and the fatality rate set to 0.5%. Left: a plot of total infections and fatalities in the simulation illustrating the effect of 80% physical distancing introduced at the first death as shown in Figure 2 above. Right: a plot of total infections and fatalities in the second simulation illustrating herd immunity in Figure 3 above. We see that there is typically a considerable delay between the first infections and the first deaths. This reflects both the time course of the disease in an individual, and the length of time taken for a sufficient number of infections to lead to a significant probability of seeing fatalities. In both cases, we see that, with stochastic fluctuations, the fatalities roughly follow the pattern of the infections.*

we assume that people started to be cautious, and some degree of physical distancing came into play. India went into national lockdown on March 25th when there were 12 fatalities, and we assume that this reduced the total infectible population (in addition to maintaing the physical distancing). Some state level lockdowns had started a few days earlier.

The outcome of one set of simulations is shown in Figure 6. R_0 is set at 4.0 to match the initial dynamics prior to mitigation. With a fatality rate of 0.5%, and the assumed mitigations above, the model predicts that testing is picking up about 5.5% of infections. Further details are in the Figure legend. The time-course of mitigation measures was set to match roughly the moments at which they occurred.

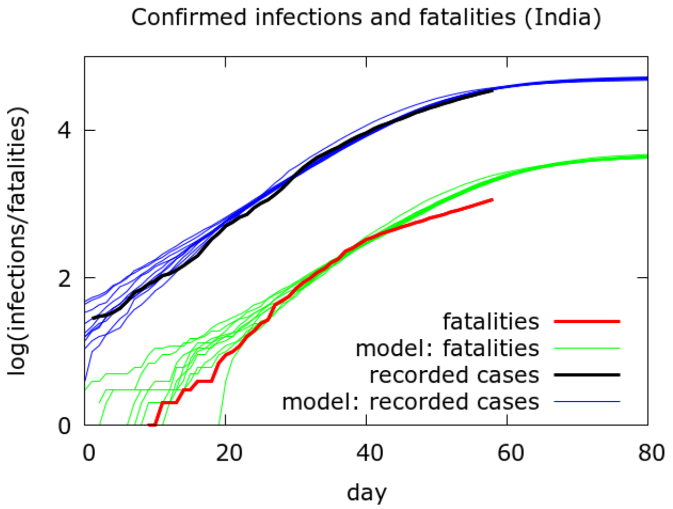


FIG. 6. *Simulations of the Indian outbreak. Day 0 is March 4th 2020. Ten simulations were carried out with $R_0 = 4.0$ and the fatality rate set to 0.5%. Simulations are initiated with 10 infected individuals and infectivities follow a geometric distribution. Initially, the infectible population is set to $12,000,000$, about 1% of the population of the country (this is essentially equivalent to setting the population to infinity in the early stages). From the start, 5.5% of all infected individuals are tested and quarantined, and this always occurs on day 10 of the infection cycle. At the first death, physical distancing starts, and this is 30% effective. At the 10th death, lockdown starts, physical distancing remains 30% effective, and the infectible population is now set to $720,000$. There is however a leak of 7000 individuals into this infectible population every day. The black curve is measured infection data. The blue curves are the model simulations of infection data. The red curve is measured fatality data. The green curves are model simulations of fatality data.*

With the parameters used in Figure 6, the model predicts that by mid-May there will have been about 3,700 fatalities and 900,000 infections, of which about 250,000 will be active, and 47,000 will have been reported. All such estimates must, however, be treated with extreme caution. If there is underreporting of fatalities, then numbers would need to be scaled up accordingly. Similarly, corrections to the infection numbers would need to be made if the fatality rate is found to differ from the "guesstimate" of 0.5%. Projections are made based on maintaining a strict lockdown over a considerable period, which is non-trivial – it is plausible that measures will be relaxed and/or the "leak" will increase with time, especially given the economic devastation that lockdown is causing.

It is worth commenting on a discrepancy which develops between modelled and measured fatalities in Figure 6, and which translates, by mid-may, into a couple of thousand "missing" deaths. If we trust the fatalities, and modify parameters to fit these fatalities better, then we find that it is hard to fit the measured infection data (see Figure 7). The discrepancy can be explained in various ways.

1. An increase in testing is capturing a higher percentage of total infections as time goes on (Figure 7 is then the "correct" view of things).
2. The epidemic is increasingly spreading in a younger or healthier demographic, and consequently the fatality rate is genuinely falling. The model, with its fixed fatality rate cannot capture this divergence between infection data and fatality data.
3. Fatalities are occurring but are not being recorded (Figure 6 is then the "correct" one). Several media reports such as [9, 10, 11] have suggested that there may be unrecorded COVID-19 deaths – the question is the extent of these. Given the fact that the disease is increasingly spreading in more deprived areas with less access to healthcare, this explanation needs to be considered seriously.

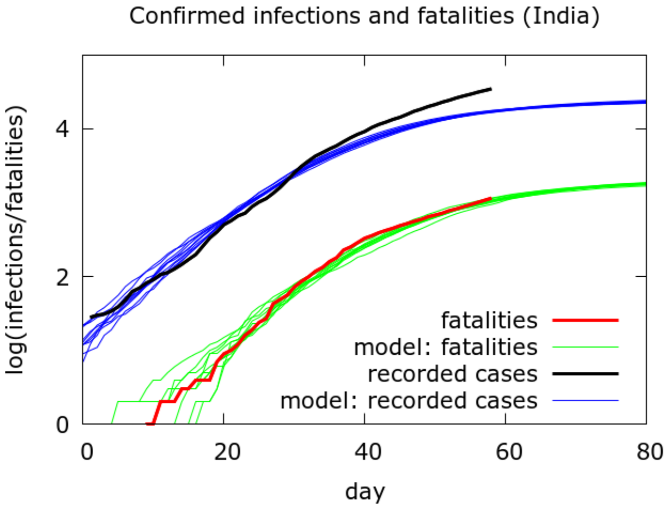


Fig. 7. *Simulations of the Indian outbreak. Ten simulations were carried out with* $R_0 = 3.8$ *and the infectible population post-lockdown set to* $276,000$*, corresponding to infection being much more strongly confined than in the simulations shown in Figure 6. Also physical distancing becomes* 35% *effective post-lockdown. All other parameters are as in Figure 6. This time we are better able to match the fatalities data, but at the cost of increasing divergence between the measured infections and the model simulations of these measurements.*

With the parameters used in Figure 7, the model predicts that by mid-May there will have been about $1,600$ fatalities and $375,000$ infections, of which about $90,000$ will be active.

Overall, the results of simulating the Indian data suggest that although the spread of COVID is much wider than the testing data suggests, it is still a relatively small outbreak by the standards of several European outbreaks, unless either deaths are being considerably under-reported, or the true fatality rate is much lower than estimated. Containment measures appear to have diminished the spread considerably, although at a huge social and economic cost.

4.2. Modelling the UK data. As with the Indian data, we are able to approximately match the time-course of the UK outbreak using a variety of parameter choices. Infection and fatality data is drawn from the Wikipedia site https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_the_United_Kingdom. We remark at the outset that official UK COVID-19 fatalities upto April 28th are hospital deaths and considerably underestimate the true COVID death toll – more on this below. One set of simulations is shown in Figure 8. With a fatality rate of 0.8%, the model predicts that testing is picking up about 4.5% of infections. It also predicts that active cases have started to decline, albeit slowly.

To get some feeling for how to model mitigation measures in the case of the UK, the time-course of advice of different restictions was briefly examined. WHO declared COVID-19 a pandemic on March 11th when the UK had 456 reported COVID-19 cases and 8 reported fatalities. The government started issuing self-isolation and physical distancing advice over the coming week. I modelled this as the introduction of a moderate level of physical distancing (45%) into the model around the first death (this occurred a few days earlier). Then something like a "lockdown" started to come into place around March 20th (when there were 177 reported fatalities). I modelled this via two effects which set in after the 200th fatality.

(i) Lockdown was assumed to bring in an increased level of physical distancing, namely 60%.

(ii) Post-lockdown, there was an infectible population of $3,300,000$, namely about 5% of the population of the country with a further leak of 30000 individuals into this infectible population every day.

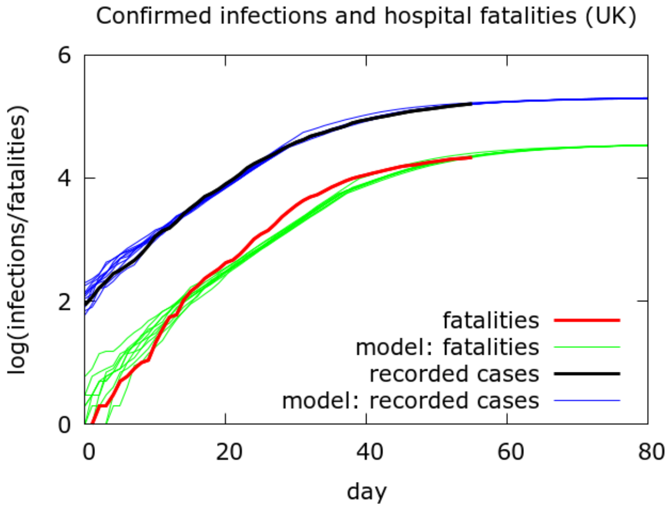


FIG. 8. *Simulations of the UK outbreak. Day* 0 *is March 4th 2020. Ten simulations were carried out with* $R_0 = 6.5$ *and the fatality rate set to* 0.8%*. Simulations are initiated with 10 infected individuals and infectivities follow a geometric distribution. Initially, the infectible population is set to* $66,000,000$ *(this is effectively infinite). Death occurs on day* 18 *post-infection. From the start,* 4.5% *of all infected individuals are tested and quarantined, and this always occurs on day 11 of the infection cycle. Thus the total number of infections is about* 25 *times the measured number. At the first death, physical distancing starts, and this is* 45% *effective. At the 200th death, lockdown starts, physical distancing becomes* 65% *effective, and the infectible population is now set to* $3,960,000$*. There is, moreover, a leak of* 40000 *individuals into this infectible population every day. The black curve is measured infection data. The blue curves are the model simulations of infection data. The red curve is measured fatality data. The green curves are model simulations of fatality data. With these parameters, the model predicts that 5 months after day* 0 *(early August 2020), as the wave subsides and daily deaths are in the tens, there will have been* $37,000$ *fatalities and* 4.7 *million infections.*

REMARK 4.1 (Remark about computation). With millions of infected individuals the model slows down considerably. In order to speed up the running of the model, a scaling can be applied – all populations are reduced to $1/n$ of their values, while the death rate is increased to n times its value. This scaling was done when simulating the UK outbreak. Other than heightening certain stochastic fluctuations in the early days, this scaling does not affect gross model predictions.

As is well known, UK recorded COVID-19 deaths are hospital deaths and considerably underestimate the true COVID death toll, which could easily be 50% higher (see, e.g., [12]). If the true COVID fatalities are higher than reported fatalities by some fixed percentage, we can still simultaneously reproduce infection and fatality data. We would need only to appropriately scale up populations, scale down testing and quarantining levels, and adjust the triggers for mitigation measures coming into force. For example, if we assume 50% more deaths than reported hospital deaths then, with these parameters and as this wave dies out, the model predicts that there will have been about $55,000$ deaths (i.e., both 50% higher than the numbers in Figure 8). If the death rate is lower than the assumed 0.8%, then infections again get revised proportionately upwards. Figure 9 shows the results of simulations with fatality data revised upwards by 50%, and a lower death rate of 0.5%. Based on such simulations I consider it very plausible that by late May 15% of the UK population will have been infected. In the simulations, about 90% of these infections have already occurred in early May 1st.

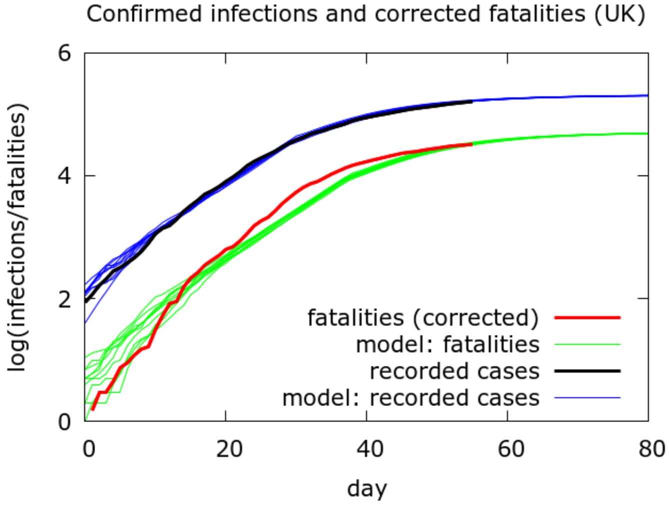


FIG. 9. *Simulations of the UK outbreak. Ten simulations were carried out with the fatality rate set to* 0.5% *and true fatalities set to be* 50% *higher than the measured (hospital) data. Now* 2% *of all infected individuals are tested and quarantined. The lockdown starts at the 300th death and the infectible population is set, post-lockdown, to* $9,504,000$*. There is, moreover, a leak of* $72,000$ *individuals into this infectible population every day. The black curve is measured infection data. The blue curves are the model simulations of the measured infection data. The red curve is corrected fatality data. The green curves are model predictions of the corrected fatality data. Other parameters are as in Figure 8. With these changes, the model predicts about* 10.5 *million infections as the current wave dies out.*

4.3. Modelling the Swedish data. Sweden was chosen as an interesting case, because it has no "lockdown" as such, although there have been restrictions on freedom of assembly, and voluntary social distancing [13]. At face value, these measures appear to have had a significant impact on the progress of the disease. The recorded infections data has the interesting feature that there appears to be a very clear signature of mitigation setting in around March 14th, with a significant change of slope occurring at this point; but this feature does not appear in the fatality data which is much "smoother". Simulations for one set of parameter choices are given in Figure 10. Computation was speeded up with a scaling (see Remark 4.1 in Section 4.2 above). In this simulation, active cases peak around day 55 (late April 2020), at which point there are in the region of $300,000$ active cases (these numbers depend, of course on the death rate).

To illustrate that parameter choices are by no means unique we present results of another simulation in Figure 11. With these parameters, active infections peak around day 40 (11th April), at which point there are in the region of $260,000$ active cases.

In both simulations, as the wave dies out there will have been in the region of $800,000$ to $950,000$ infections (about $8-10\%$ of the total population). If the death rate is higher (resp., lower) than 0.8%,

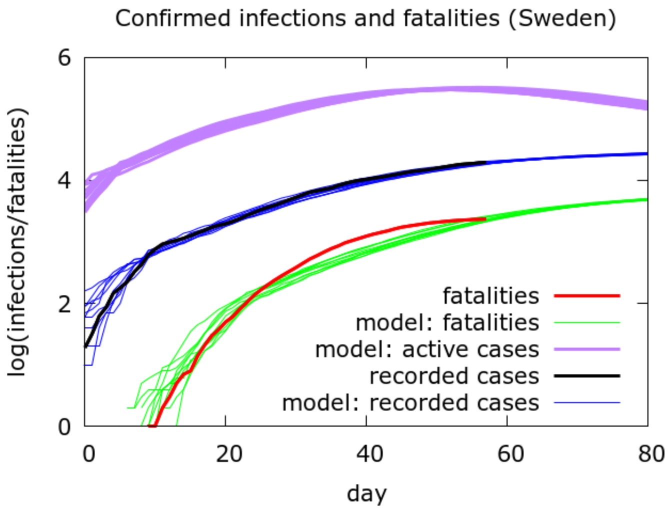


FIG. 10. *Simulations of the Swedish outbreak. Day* 0 *is March* 2*nd 2020. Ten simulations were carried out with* $R_0 = 8.5$ *and the fatality rate set to* 0.8%*. Simulations are initiated with 10 infected individuals and infectivities follow a geometric distribution. The average day of death is day* 18*. Initially, the infectible population is set to* 10,000,000 *(effectively infinite). From the start,* 3.8% *of all infected individuals are tested and quarantined, and this occurs on day 4. At the* 550*th positive case (about* 11*th March), mitigation begins, physical distancing becomes* 74% *effective, and the infectible population is now set to* 400,000*. There is, moreover, a leak of* 10000 *individuals into this infectible population every day. Remaining parameters are at their default values. The black curve is measured infection data. The blue curves are the model simulations of infection data. The red curve is measured fatality data. The green curves are model simulations of fatality data.*

then this number needs to be scaled down (resp., up).

4.4. Modelling an Indian state: Maharashtra. Maharashtra accounts for about 30% of all recorded positive cases in India, and about 40% of all deaths nationwide (9282 out of 31,324, and 400 out of 1008, respectively, on 28th April 2020). This means that there is sufficient data – in particular fatality data – to attempt modelling of the dynamics within Maharashtra on its own. In Figure 12 we see an attempt which matches the time course of recorded infections reasonably well. Active infections peak around day 55 (May 8th) at which point there are about 600,000 active infections. Computation was speeded up with a scaling (see Remark 4.1 in Section 4.2 above).

Figure 12 shows the same divergence between predicted and obtained fatality data that was remarked on for India as a whole, and the possibilities for why this occurs outlined in Section 4.1 for India as a whole hold in this case too. In Figure 13 we see a second set of simulations which match the time course of recorded fatalities better. In this case, as expected, the model somewhat underestimates the extent of recorded infections. Active infections peak around day 70 (May 23th) at which point there are about 350,000 active infections. Note that assuming stronger mitigation both lowers the peak number of active infections, and delays it. The more optimistic scenario in Figure 13 corresponds to a later peak.

Note that the first, more pessimistic, simulation makes sense if there are COVID-19 deaths going unrecorded, while the second, more optimistic, simulation makes sense if testing is capturing an increasing proportion of actual infections after about day 20, namely April 3rd. According to data on https://www.covid19india.org/, between April 5th and April 28th, total tests increased 7.5 times (16,008 to 120,620), recorded fatalities increased about 9 times (45 to 399), and recorded infections increased about 12.5 times (747 to 9282). Although at face value testing is not keeping pace with other measures of spread, this could be a consequence of time-delays – the model simulations resulting in Figures 12 and 13 suggest that total infections increased about 3.5 to 5 times during this period, and it is plausible that the increased pace of testing means that a higher proportion of infections are being picked up. On the other hand, media reports cited above [9, 10, 11] focussed on Maharashtra (Mumbai in particular) suggest that

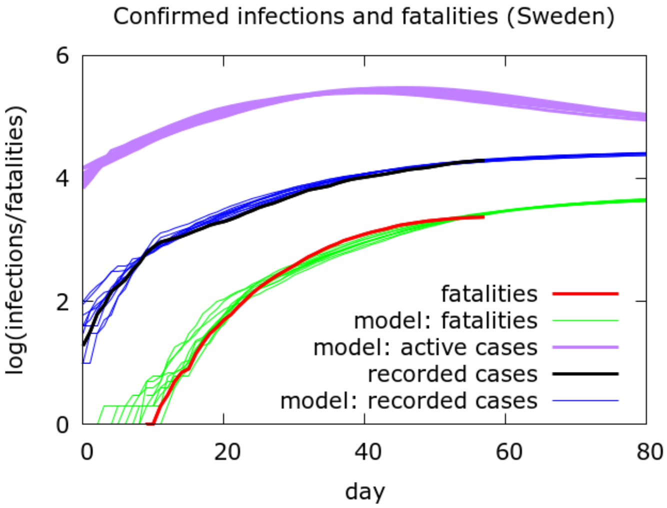


Fig. 11. *Simulations of the Swedish outbreak with some altered parameters. The mean time to death is now set at 19 days. Death occurs on average on Day 20. From the start, 4% of all infected individuals are tested and quarantined, and this occurs on day 7 of the infection cycle. At the* 600*th positive case (about* 11*th March), mitigation begins, physical distancing becomes* 70% *effective, and the infectible population is now set to* 230, 000*, namely disease is much more confined than in the simulations of Figure* 10*. Remaining parameters are as in Figure* 10*. The blue curves are the model simulations of infection data. The red curve is measured fatality data. The green curves are model simulations of fatality data.*

there are unrecorded COVID-19 fatalities occurring. It is possible that the divergence between recorded infections and recorded fatalities is a consequence of some combination of the two effects: there is some effect from increased testing and some from “missing” deaths, in which case, the “true” behaviour may be intermediate between that shown in Figures 12 and 13.

4.5. Estimating the beginning of an outbreak: Gujarat. Can we use stochastic individual-based modelling of the kind described here to tell us when infection began to spread? This would be relevant to a situation where the origins of an outbreak are not clear. This is the case for Gujarat where the origins of early cases were unclear – already on April 2nd 2020 four positive cases were found in Ahmedabad through “public surveillance” (i.e., without known contact history) [14].

The answer to this question is, roughly, yes, if we trust the model. We choose parameters which fit the data set of interest, initiate outbreaks in the model, and let it run. We then check when we arrive at some marker event in the model simulations, e.g., the first death, or the 10th reported case, or whatever, and compare this with when it occurred in real life. Because the model is stochastic we should do this multiple times, to get an estimate for when the outbreak was initiated.

There are issues with this method. For example,

1. They assume a single introduction of the disease, when in reality there may be multiple introductions at the start of an outbreak.
2. They depend on assumptions about typical transmission time in the model. (The default window of transmission in the model day 3 to day 14 of infection. The upper limit of day 14 can be argued to be too low.)
3. They depend on the R_0 value used, generally chosen to fit the early dynamics in the data. However, stochastic effects are greater both in data and simulation in the early period, making this fitting process less reliable.

Consider a key event, the first COVID-19 death in an outbreak. We expect to reach this event quicker if we have a more rapid time-course of the disease, either at the level of the individual because an individual progresses from infection to death quicker, or at the level of the population because, for example, early

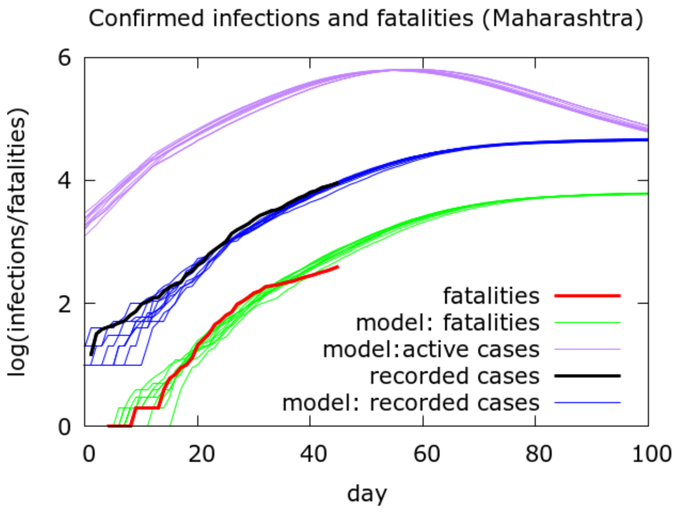


FIG. 12. *Simulations of the Maharashtra outbreak. Day* 0 *is March* 14*th* 2020. *Ten simulations were carried out with* $R_0 = 4.0$ *and the fatality rate set to* 0.5%. *Simulations are initiated with 10 infected individuals and infectivities follow a geometric distribution. Initially, the infectible population is set to* 10, 000, 000. *From the start,* 3.7% *of all infected individuals are tested and quarantined, and this occurs, on average, on day 12 of the infection cycle. At the* 70*th positive case (about* 23*rd March), mitigation begins, physical distancing becomes* 42% *effective, and the infectible population is now set to* 900, 000. *There is, moreover, a leak of* 10000 *individuals into this infectible population every day. The black curve is measured infection data. The blue curves are the model simulations of infection data. The red curve is measured fatality data. The green curves are model simulations of fatality data.*

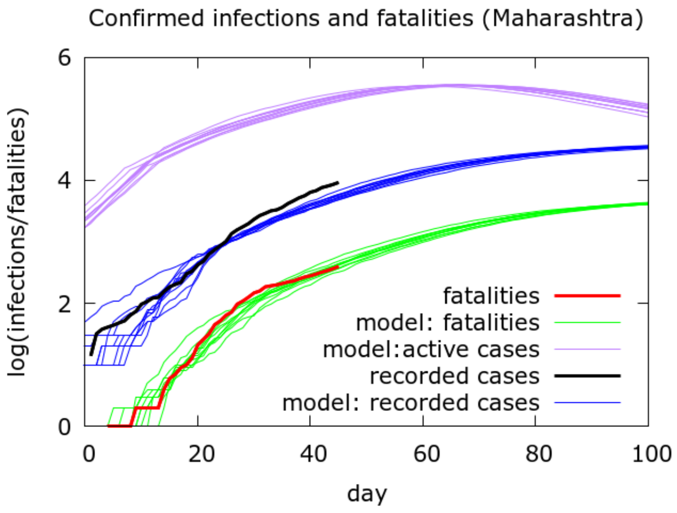


FIG. 13. *Simulations of the Maharashtra outbreak which match the fatality data quite well. Death now occurs on day* 18 *on average, and* 3.9% *of all infected individuals are tested and quarantined. Physical distancing is* 55% *effective after mitigation starts, and the infectible population is set to* 700, 000. *All other details and parameters are as in Figure 12.*

infectivity or a high R_0-value means that infection spreads quicker.

To test this hypothesis, we first ran the model 100 times with $R_0 = 2.5$, a fatality rate of 0.5%, and all other parameters at their defaults. A single infection was used to seed the outbreak. For those simulations where an outbreak occurred (59 out of 100 simulations), the mean (± SD) day of the first death was day 56.1(±15.3). Note how surprisingly long it takes for the first death to occur on average – almost two months! Increasing R_0 to 5 resulted in an outbreak in 81 out of 100 cases with mean day of first death

42.5(±10.9). Maintaining R_0 at 2.5 but increasing the fatality rate to 2% resulted in an outbreak in 61 out of 100 cases with mean day of first death 45.2(±12.9). Initiating an outbreak with a two infections instead of one, increased the likelihood of the outbreak occurring but did not greatly decrease the average time to the first death: with $R_0 = 2.5$, a fatality rate of 0.5%, and two initial infections, there were 86 outbreaks with a mean day of first death being 51.1(±13.3) days.

This illustrates how either a more rapid early dynamics or an increased death rate decreases the time to the first death. A greater number of initial infections seems to have a more modest effect on the time to the first death.

Running the model with parameters aimed at reproducing the Gujarat data gave the plots shown in Figure 14. An R_0 value of 8 was needed to approximately match the initial rapid increase in cases, although it may be that this was a function of when the decision was made to start testing, rather than reflecting a genuine rapid increase in cases. It was assumed that mitigation started at the first reported infection (this was on March 19th, a few days before national lockdown). Outbreaks were initiated with 2 infected individuals, and testing was assumed to occur, on average, 10 days after infection. Some parameters were altered from their defaults as described in the legend to Figure 14. In 3 out of 100 model runs, the outbreak died out. In the remaining simulations death occurred on day 31.2(±8.7). Note that the average time between between first infection and first death is about a month despite the large initial R_0. A sample of simulations are plotted in Figure 14.

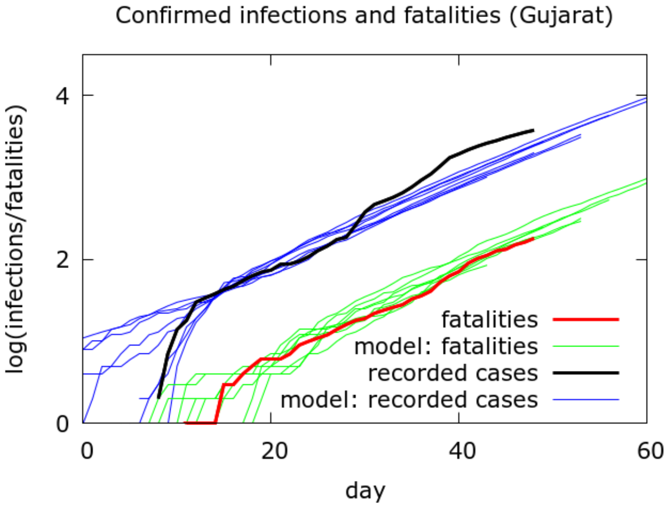


Fig. 14. *Simulations of the Gujarat outbreak. Day* 0 *is March* 12*th* 2020*. Ten simulations were carried out with* $R_0 = 8.0$ *and the fatality rate set to* 0.5%*. Simulations are initiated with* 2 *infected individuals and infectivities follow a geometric distribution. Initially, the infectible population is set to* 10,000,000 *(this is effectively infinite given the small number of cases). The average day of death is day* 14*, and the infective window is from day* 3 *to day* 13*. From the start,* 3.2% *of all infected individuals are tested and quarantined, and this occurs, on average, on day 10 of the infection cycle. At the first tested positive case, mitigation begins (this occurred on March* 19*th). At this point physical distancing becomes* 70% *effective, and the infectible population is now set to* 900,000*. There is, moreover, a leak of* 10,000 *individuals into this infectible population every day. Remaining parameters are at default values. The black curve is measured infection data. The blue curves are the model simulations of infection data. The red curve is measured fatality data. The green curves are model simulations of fatality data.*

The first recorded COVID-19 death in Gujarat occurred on March 22nd and involved a patient with no travel history abroad – although they had travel history within India [15]. While it is unclear where the disease was contracted, it is possible that the disease was contracted in Gujarat itself. Based on this possibility, the model predicts (with all the notes of caution listed above) that the first infections occurred about a month earlier, namely around Feb. 21st (one standard deviation on each side gives Feb. 12th to March 1st). The *Namaste Trump* event [16] took place during this period on Feb. 24th, shortly after Donald Trump arrived in India with his entourage, and involved about 100,000 people. Thus it is plausible that it played a part in the spread of the outbreak in Gujarat.

It is also possible, though less likely, that *Namaste Trump* played a part in the initiation of the outbreak. Genomic studies showed that COVID-19 was likely quite widespread in the USA by mid-late February [17]. Although there had been only a dozen diagnosed cases and a couple of deaths by this point, it is noteworthy that the first two fatalities in the USA had no known travel history. Knowing what we know about the time-course of the disease, it is likely that the diagnosed cases were only the tip of the tip of the iceberg. Simulations show that by the time of the first "community" fatalities, there can be tens of thousands of cases, and although we haven't simulated the data from the US, the data suggests that the initial growth rate was rapid.

Thus while we cannot be certain without adequate contact tracing whether *Namaste Trump* played a part in the spread, or even initiation, of COVID-19 in Gujarat, based on model simulations it cannot be argued that the event occurred too early to contribute to the dynamics of COVID-19 in Gujarat. It is plausible that it was a major event in the initiation/early spread of the disease.

5. Conclusions. Individual-based modelling is able to reproduce qualitatively, and to some extent quantitatively, both infection and fatality data observed in several different contexts. Some of the more interesting insights come from questioning discrepancies between modelled and observed behaviour. Although such discrepancies may be artifacts arising from the stochastic nature of the model and the noisy data, they sometimes appear robust enough to warrant extra thought.

Another interesting set of insights arises from considering the effects of mitigation measures which can often be observed in infection and/or fatality data. Although any mitigation measure can, if sufficiently strong, reverse an outbreak, different measures have different mechanisms and leave different signatures in the data. Measures which restrict freedom of movement reduce the population available for infection, effectively localise a disease and, in the short term, introduce high levels of spatial heterogeneity into infection levels. More global measures which lead to social distancing can slow the spread by reducing the probability of infection events occurring and thus effectively reducing R_0; but if R_0 is high to begin with, such measures need to be very effective to bring it to below 1 and reverse an outbreak.

We see very clearly in the modelling how the slow time course of the disease complicates management and mitigation. Even when mitigation has proved successful in the sense that active infections have peaked, there is often a long delay before the pool of infection dwindles to manageable levels. To be effective, measures need to be maintained for a considerable period, making it imperative to consider carefully how to reduce their social and economic costs alongside considering their benefits in disease control.

REFERENCES

[1] Timothy W Russell, Joel Hellewell, Christopher I Jarvis, Kevin van Zandvoort, Sam Abbott, Ruwan Ratnayake, CMMID nCov working group, Stefan Flasche, Rosalind Eggo, W John Edmunds, and Adam J Kucharski. Estimating the infection and case fatality ratio for covid-19 using age-adjusted data from the outbreak on the diamond princess cruise ship. *CMMID repository*, 2020, available at https://cmmid.github.io/topics/covid19/severity/diamond_cruise_cfr_estimates.html.

[2] Robert Verity, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick G T Walker, Han Fu, Amy Dighe, Jamie T Griffin, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Anne Cori, Zulma Cucunubá, Rich FitzJohn, Katy Gaythorpe, Will Green, Arran Hamlet, Wes Hinsley, Daniel Laydon, Gemma Nedjati-Gilani, Steven Riley, Sabine van Elsland, Erik Volz, Haowei Wang, Yuanrong Wang, Xiaoyue Xi, Christl A Donnelly, Azra C Ghani, and Neil M Ferguson. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*, 2020, available at https://www.thelancet.com/action/showPdf?pii=S1473-3099%2820%2930243-7.

[3] Antonio Regalado. Blood tests show 14% of people are now immune to covid-19 in one town in germany. *MIT Technology Review*, 2020, available at https://www.technologyreview.com/2020/04/09/999015/blood-tests-show-15-of-people-are-now-immune-to-covid-19-in-one-town-in-germany/.

[4] Michael Schulson in *Undark*. On Covid-19, a Respected Science Watchdog Raises Eyebrows, April 24th 2020, available at https://undark.org/2020/04/24/john-ioannidis-covid-19-death-rate-critics/. Accessed on 25th April 2020.

[5] Apoorva Mandavilli. Infected but Feeling Fine: The Unwitting Coronavirus Spreaders. *New York Times*, 31-03-2020, available at https://www.nytimes.com/2020/03/31/health/coronavirus-asymptomatic-transmission.html.

[6] Kenneth McIntosh. Coronavirus disease 2019 (COVID-19): Epidemiology, virology, clinical features, diagnosis, and prevention. *UpToDate*, 2020, available at https://www.uptodate.com/contents/coronavirus-disease-2019-covid-19-epidemiology-virology-clinical-features-diagnosis-and-prevention.

[7] Dawei Wang, Bo Hu, Chang Hu, and more. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA Network*, 2020, available at https://jamanetwork.com/journals/jama/fullarticle/2761044.

[8] Roman Wölfel, Victor M. Corman, Wolfgang Guggemos, Michael Seilmaier, Sabine Zange, Marcel A. Müller, Daniela Niemeyer, Terry C. Jones, Patrick Vollmar, Camilla Rothe, Michael Hoelscher, Tobias Bleicker, Sebastian Brünink, Julia Schneider, Rosina Ehmann, Katrin Zwirglmaier, Christian Drosten, and Clemens Wendtner. Virological assessment of hospitalized patients with COVID-2019. *Nature*, 2020, available at https://www.nature.com/articles/s41586-020-2196-x.

[9] BBC News India. Video available at https://twitter.com/BBCIndia/status/1250030616801763328, April 14th 2020. Accessed on 25th April 2020.

[10] Gaurav Sarkar on *mid-day.com*. COVID-19: 17-year-old boy tests positive at JJ Hospital after death; family not quarantined, April 18th 2020, available at https://www.mid-day.com/articles/covid-19-17-year-old-boy-tests-positive-at-jj-hospital-after-death-family-not-quarantined/22737202. Accessed on 25th April 2020.

[11] Laxman Singh and Tabassum Barnagarwala in *The Indian Express*. Mumbai: Kin say man died as eight hospitals turned him away, April 19th 2020, available at https://indianexpress.com/article/india/mumbai-kin-say-man-died-as-eight-hospitals-turned-him-away-6368832/. Accessed on 25th April 2020.

[12] Dominic Gilbert, Ashley Kirk, and Henry Bodkin in *The Telegraph*. How accurate are UK coronavirus death toll numbers?, April 19th 2020, available at https://www.telegraph.co.uk/news/2020/04/19/how-accurate-coronavirus-uk-death-numbers/. Accessed on 25th April 2020.

[13] Jenny Anderson in *Quartz*. Swedens very different approach to Covid-19, April 28th 2020, available at https://qz.com/1842183/sweden-is-taking-a-very-different-approach-to-covid-19/. Accessed on 29th April 2020.

[14] *Times of India*. Ahmedabad: In a 1st, 4 positive cases found in public surveillance, April 2nd 2020, available at https://timesofindia.indiatimes.com/city/ahmedabad/in-a-1st-4-ve-cases-found-in-public-surveillance/articleshow/74939488.cms. Accessed on 28th April 2020.

[15] *The Economic Times*. Surat man first victim of coronavirus in Gujarat, March 22nd 2020, available at https://economictimes.indiatimes.com/news/politics-and-nation/surat-man-first-victim-of-coronavirus-in-gujarat/articleshow/74760199.cms?from=mdr. Accessed on 28th April 2020.

[16] Wikipedia. Namaste Trump, Available at https://en.wikipedia.org/wiki/Namaste_Trump. Accessed on 28th April 2020.

[17] Carl Zimmer in *The Economic Times*. Most New York Coronavirus Cases Came From Europe, Genomes Show, April 8th 2020, available at https://www.nytimes.com/2020/04/08/science/new-york-coronavirus-cases-europe-genomes.html. Accessed on 28th April 2020.