Assessing the Spread of the Novel Coronavirus In The Absence of Mass Testing

Preliminary draft May 5 2020

Oscar Dimdore-Miles[1] and David Miles[2]

[1]*Department of Physics, Oxford University, Email: oscar.dimdore-miles@physics.ox.ac.uk*
[2] *The Business School, Imperial College London, Email: d.miles@imperial.ac.uk*

Abstract

This note outlines a simple method for estimating the spread of the COVID 19 virus in the absence of data on test results for a large, random sample of the population. It applies the method to the UK, and other countries, and finds that to match data on daily new cases of the virus, the estimated model favours high values for the number of people infected but asymptomatic. That result is very sensitive to whether the transmission rate of the virus is different for symptomatic and asymptomatic cases, something about which there is significant uncertainty. This illustrates how difficult it is to estimate the spread of the virus until very large samples of the population can be tested.

1 introduction

There is significant uncertainty about the degree to which the novel coronavirus (COVID19) has spread and infected people who show no obvious symptoms. This has very significant policy implications - Stock 2020 shows that different policies aimed at controlling the virus can have very different effects on the numbers who become infected and show symptoms depending on the proportion of those who are asymptomatic. It is those with symptoms who are at risk of death from the virus and so the relative size of the populations of symptomatic to asymptomatic amongst the infected is of enormous significance to welfare, including mortality rates, and to policy (Fauci, Lane, and Redfield 2020).

The degree of uncertainty about that asymptomatic rate is large enough to mean that neither 0.3 or 0.9 is outside the range of plausible values, though the implications of those two numbers are very different. Li *et al.* 2020 estimate that 86 per cent of all infections were undocumented prior to the Wuhan travel shutdown (on January 23, 2020). In contrast, estimates based on infections amongst passengers on the cruise ship Diamond Princess put the proportion of asymptomatic (or near asymptomatic) cases at around 50 per cent. Manski and Molinari 2020 report enormous ranges for the possible values of the infection rates in Illinois, New York and Italy. As of April 6th 2020 these ranges are estimated as [0.001, 0.517], [0.008, 0.645], and [0.003, 0.510] respectively.

While large scale testing of a random sample of the population would greatly narrow the range of plausible values for the asymptomatic proportion of the infected (Stock 2020), such testing seems some way off in most countries. In most countries, including the the USA and the UK, testing up to the end of April 2020 was concentrated on those who display symptoms or are at high risk; it was certainly not random.

In this note we implement a simple version of the SIR (susceptible-infected-recovered) model to estimate the asymptomatic rate from data on the non-random sample of those tested. We use data from the UK, where relatively few of those showing no symptoms had been tested up to the end of April 2020, to provide highly provisional estimates of the asymptomatic proportion of the infected. We find quite striking differences between what the simple model suggests is the asymptomatic rate and the much lower estimates based on the limited data of (near) random sampling.

We also apply the model to the US, Italy, Spain, France and Sweden. The results for those countries are similar to the UK - the value of the asymptomatic rate that seems to best fit the data is very high; far higher than is estimated based on the limited amount of results from more widespread testing which went beyond those who showed symptoms.

We consider what might account for such differences and find that the degree to which the asymptomatic spread the virus - and whether it is substantially lower than for the symptomatic - is of great significance. There is limited evidence on this which makes it hard to assess the spread of the virus, creating great challenges for policies on easing "lockdown" measures. If the spread of the virus is wide, and had been a factor in the observed decline of new cases up to the end of April, an easing of restrictions poses fewer risk of a sharp upwards spike in infections. However, the confidence interval for the estimated ratio of the numbers of infected with mild (or no) symptoms to the numbers of symptomatic is quite wide.

2 The Model

We use a version of the SIR model which closely follows Stock 2020. At each point in time the population is made up of three distinct groups: those who are currently infected (I_t); those who are susceptible (S_t) and those who have recovered (R_t). We assume a constant population and that the death rate is low enough to mean that this is reasonable. We also assume that some constant proportion (π_a) of those infected do not develop symptoms – or that they are so mild as to count as asymptomatic. There is some evidence that the degree to which the asymptomatic are infectious may be different from those who have symptoms (Ferguson *et al.* 2020), but we will initially assume that the transmission rates are the same for all those infected. We denote the population of the symptomatic infected at time t by I_{st} and the asymptomatic as I_{at} such that $I_t = I_{st} + I_{at}$. The evolution of S_t, I_t and R_t in discrete time is given by the dynamic system:

$$\Delta S_t = -\beta_t I_{t-1} \frac{S_{t-1}}{N} \tag{1}$$

$$\Delta R_t = \gamma I_{t-1} \tag{2}$$

$$\Delta I_t = \beta_t I_{t-1} \frac{S_{t-1}}{N} - \gamma I_{t-1}, \tag{3}$$

ΔS is the change in the population of the susceptible; N is the total population, β_t is the transmission rate of the virus at a time t (the mean number of people an infectious person will infect per unit time) and γ is the rate of recovery. The initial infection rate over the infectious period, the reproduction number, is defined as $R_0 = \frac{\beta_t}{\gamma}$. We assume that the infectious group I_t is made up of symptomatic and asymptomatic groups in fixed proportions such that

$$I_{st} = (1 - \pi_a) I_t \tag{4}$$

and

$$I_{at} = \pi_a I_t. \tag{5}$$

The number of new cases at time t (y_t) can be calculated as

$$y_t = \Delta I_t + \gamma I_{t-1}. \tag{6}$$

New cases are the sum of the change in the number of outstanding cases plus the numbers recovered. The number of new symptomatic cases (y_{st}) is

$$y_{st} = (1 - \pi_a)(\Delta I_t + \gamma I_{t-1}) = (1 - \pi_a)\left(\beta_t I_{t-1} \frac{S_{t-1}}{N}\right). \tag{7}$$

We make the key assumptions that: a large fraction of the symptomatic are tested; that a small proportion of the asymptomatic are tested; and that the test is reliable. This would mean that the observable number of those who test positive would closely track the quantity (y_{st}). The strategy that we pursue is to use the data on the numbers of new cases who test positive for the virus and to assume that this closely follows the true

number of newly infectious symptomatic people. We then seek the values of the parameters of the model - and in particular π_a - that give a predicted y_{st} that matches the data. The strategy is similar to that adopted in the study by Lourenço *et al.* 2020 who estimated that a high proportion of the UK population may have been infected even by early March 2020. But there are two important differences with the procedure we follow. First, we use data on the numbers of those who test positive (in the UK and in other countries) as the variable we are trying to match; the Gupta study used the number of deaths. There would seem to be significant ambiguity over assignment of the cause of death to the virus, perhaps more than over whether a positive test is reliable or not. Second, the study based on deaths looked at a short period when deaths were low and rising fast by early March. We use data on tests up the end of April by which time nearly 500,000 had been tested in the UK and around 175,000 had tested positive (according to data from the Office of National Statistics).

To implement the estimation of the model we need to make assumptions about the transmission rate of the virus β_t and the recovery rate γ. The transmission rate will not have been constant because of policy measures introduced to slow the spread of the infection. In the UK "lockdown", which began on March 23rd, has been strict and social distancing will likely have brought it down significantly. Similar policies were adopted at various times in March 2020 in other countries. We assume a constant value of β_t before the lockdown date (of β_0), followed by a gradual reduction in the β_t value after this date to simulate the effect the measures have on transmission. The initial value of β_0 is derived from assumed values of the initial reproduction rate R_0 and the recovery rate γ, using the relation $\beta_0 = \gamma R_0$. We try all values for an initial transmission rate ranging from 2.2 up to 3.9 at intervals of 0.005. We try three values of the recovery rate implied by half lives of the period of infectiousness - that is the number of days it takes for half an initial number of infected people to recover - of 4 days, 6 days (as used by Stock 2020) and 8 days. The corresponding three values of γ are 0.159, 0.109 and 0.0833.

We assume that after the lockdown date there is a lag until the value of β_t starts to change from β_0. The lag is between the lockdown measures starting and the impact on the numbers testing positive for the virus. That lag reflects several distinct factors: it must include the lag in the impact on new infections, the lag before symptoms show, the lag before testing the symptomatic and finally the lag before results are known and recorded in the daily measure. We set the overall lag at 14 days, but also assess sensitivity of results to shorter lags. After this lag, β decays exponentially towards a value of β_L, the post lockdown asymptotic β. The time path for β_t can be expressed as

$$\beta_t = \begin{cases} \beta_0, & \text{if } t \leq t^* \\ \beta_0 - (\beta_0 - \beta_L)(1 - e^{-(t-t^*)\lambda}) & \text{if } t > t^*, \end{cases} \tag{8}$$

where t^* is the lockdown time plus the 14 day lag period and λ is the speed of adjustment in β after lockdown measures begin to take effect. We assume that once the lockdown does begin to affect numbers testing positive it quite quickly reaches its full effectiveness, bringing the transmission rate down so that half of its long run impact on β comes through in 3 days, implying that $\lambda = 0.231$.

For given values of γ, β_0 and λ we search for the values of the two free parameters - β_L and π_a - so as to maximise the fit of the model. We chose those two free parameters to minimise the sum of squared deviations between the daily data on the numbers of new positive tests for the virus and the model prediction of that number (y_{st}). The parameters we fit are a measure of how effective the lockdown is in bringing down the infection rate (measured by how much lower β_L is relative to β_0) and the ratio of those infected with no symptoms to the total population of the infected (π_a).

3 Sensitivity to key assumptions and calibration

Before showing results we stress that our model relies on a number of key assumptions.

We assume that those who have been tested up to the end of April 2020 are overwhelmingly those with symptoms and that a very high proportion of those who have significant symptoms are tested. Neither assumption is an exact approximation to reality in the UK or elsewhere for a number of reasons. In the UK there has not been an entirely consistent policy on testing in hospitals - some test those who present symptoms, others

with sufficient resources have done more widespread testing of patients regardless of symptoms. But, overall, relatively few in the UK with no symptoms have been tested up to the end of April 2020, and a high proportion of those with symptoms serious enough to be admitted to hospitals have been tested.

For the UK the model is initialised on data from the 31st of January, the date on which the first non zero value of positive test cases is recorded. At this time, testing was only applied to those who had travelled to certain regions of China and presented with symptoms and therefore data in the first week or so may not be fully representative of all symptomatic cases. Nevertheless, the criteria for testing was quickly widened. The data we have on recorded positive tests is clearly imperfect but it is the closest available to the model variable y_{st}.

We rely on estimates of R_0 and of γ to generate a value for β_0. There is considerable uncertainty about both. At the lower end of the ranges of values used in simulations are those chosen by Ferguson *et al.* 2020, who assume a value of 2.4, and Lourenço *et al.* 2020 who take figures centred around 2.25 or 2.75. Stock 2020, who draws on estimates using data from Wuhan, uses a much higher figure for simulations with a pre-shutdown value for R_0 of 3.8. The range of estimates of R_0 from several studies is between 2.2 and as high as 3.9. A team at the London School of Hygiene and Tropical Medicine found 11 published estimates of R0 for Covid-19, which averaged 2.68 with a standard deviation of 0.57 (see Paul Taylor, London Review of Books, May 2020, vol 42, no 9). The range we use for simulations is 2.2 to 3.9 - values outside this range gave a poor fit to the data for all countries we analysed for any values of the other parameters. For our estimate of γ we assume the half life of the infection as x days and therefore that γ satisfies the equation $(1-\gamma)^x = 0.5$. We take x as 4, 6 or 8 days - a range which encompasses those used in several studies.

As noted above we assume that once lockdown begins beta is reduced so that it declines asymptotically towards a value that would then be maintained as long as the lockdown remains in place (β_L in our equations). Our choice of the speed with which beta declines towards its steady-state value, after the initial lag, is such that the transition is fairly rapid, corresponding to a half life of 3 days ($\lambda = 0.231$).

A final key assumption is that new symptomatic infections are generated from the total current population of the infected (symptomatic and asymptomatic) and that the degree of infectiousness is the same across the infected. The number of new cases of those with symptoms will therefore be higher, for a given number of existing infected people with symptoms, the higher is the asymptomatic rate (π_a). It is also higher the larger β is. It is this dependence which allows us to use data on the numbers of newly tested infected with symptoms to infer something about π_a and also to learn about the impact of policy (the lockdown) from the change in the trajectory of new cases after it came into effect.

The data we try to fit is the number of new infections recorded where we assume that all such new cases have some symptoms. Testing of people with no symptoms has (up to late April 2020) been relatively small scale in the countries we analyse and to a large extent limited to those at high risk. There has been no very large scale testing of a reliably random sample of the population. We use a grid search to find the values of the two unknown and free parameters (β_L and π_a) to minimise the root mean squared deviation between the observations and (y_{st}), given the choice of other parameters.

4 Results

Figure 1 shows the data on new cases of those testing positive for the virus in the UK. The data start on January 31st. The data is from the Office for National Statistics. (The spike in reported new cases on 11/04/2020 coincides with an expansion in testing capacity).

Figure 2 shows the best fit of the model when we set the half life of the infection to 6 days (γ=0.109). The best fit for this value of γ was when $R_0 = 2.5$ and β_L and π_a are 0.1928 and 0.996 respectively. These values imply that the transmission rate started to turn down sharply by the end of first week of April, some 2 weeks after the lockdown began. The value for π_a is very high - implying that there are around 250 people who have had the infection with no symptoms (or very mild symptoms) for every person infected with symptoms. If that were true then by April 20th - by which time around 120,000 had tested positive for the virus (and the great majority of whom had shown symptoms) close to 45% of the UK population might have had the virus.

Figure 3 shows the root mean square error of the model for all combinations of parameters π_a and β_L. The

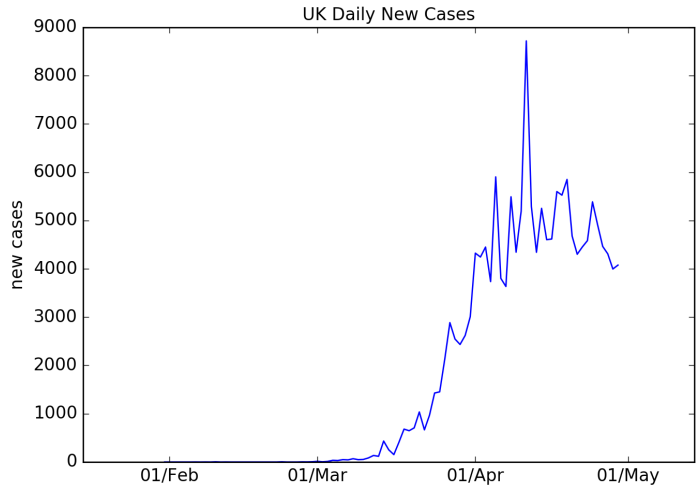

Figure 1. Time series of daily new positive COVID19 cases recorded in the UK between January 31st and April 30th 2020.

shape of this measure of fit illustrates several things. The dark shaded area on the far right of the figure (highest RMS errors) suggest the lockdown had an impact. The steeply downward sloping bands of different shades suggest that in terms of fitting the data if one increases (reduces) the assessed effectiveness of the lockdown the assumed asymptomatic rate would be lower (higher). In other words, to fit the data reducing the estimate of the share of the asymptomatic (π_a) can be partially offset by raising the assumed effectiveness of the lockdown (reducing β_L).

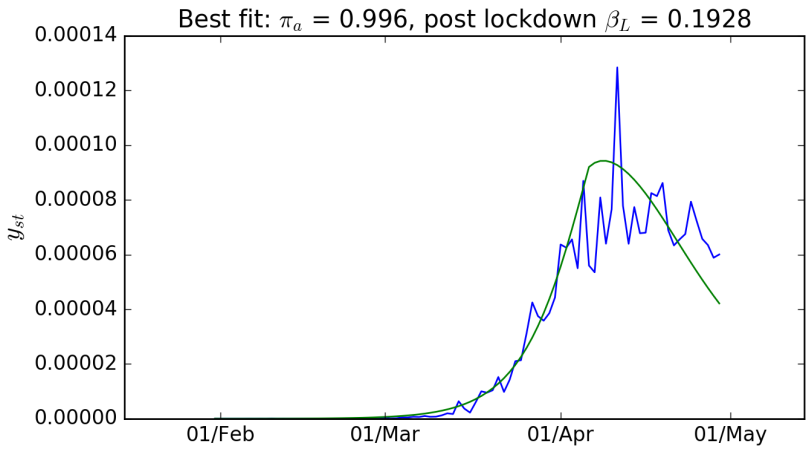

Figure 2. Results from the free parameter optimisation of the model with γ = 0.109. **a**: UK new case data from figure 1 given as a proportion of total population (blue) and model simulation for y_{st} that gives the best fit to the UK data (green).

Figure 4 shows the model fit to the UK data when we set the half life of the virus at 8 days (γ =0.0833). The best fit here was with a value of R0 of 2.95 and β_L and π_a of 0.1598 and 0.996 respectively. Once again the best fit value of π_a is very high and once again it implies that approximately 45% of the UK population may have been infected by late April. Figure 5 shows the parameter combinations that have a goodness of fit within 10% of the best pair of values; once again these are bunched fairly closely around the best fit values with all such pairs generating a value of π_a close to 0.996. The fit of the model deteriorates so sharply when we set the half life of the virus to be only 4 days that we do not show those results.

The parameter space in figure 5 shows the best fit parameters (red dot) and also the parameter combinations that generate a root mean squared error within 10% of the best value. This is illustrative of the degree of uncer-

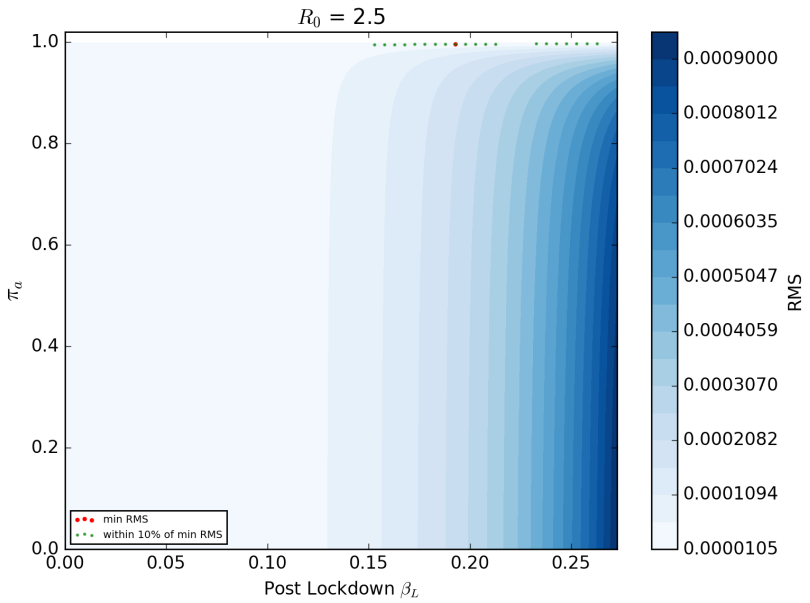

Figure 3. RMS difference between UK new case data and $y_s t$ for all combinations of π_a and β_L for simulations with $\gamma = 0.109$. The combination that gives the lowest RMS value (the best fit) is shown as a red dot. The combinations that give and RMS value within 10% of the minimum RMS are shown in green.

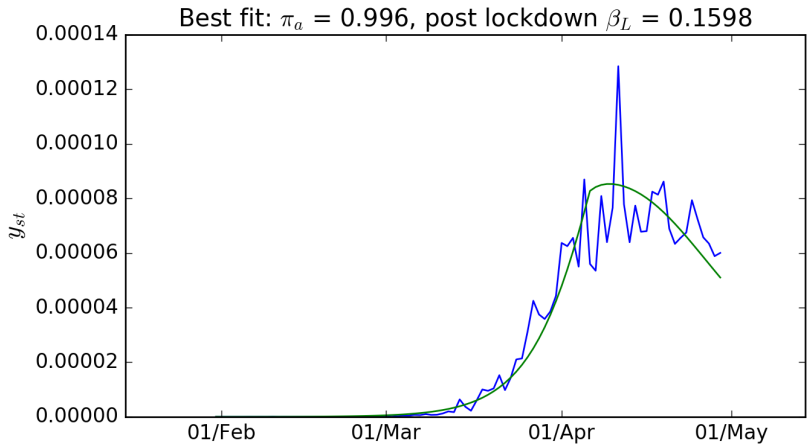

Figure 4. Like figure 2 with $\gamma = 0.0833$

tainty around the best fit values of the two parameters. However, it is difficult to construct precise confidence regions around the best-fit parameter estimates. Under restrictive assumptions, the parameter space within which the standard error of the model is within 10% of the best fit would very likely contain the true parameter values. Standard tests based on the assumption of independent and normally distributed residuals between data and the fit of the model would imply a small chance of parameters lying outside this area. The statistic $T(In(\frac{RSS_r}{RSS^*}))$, where RSS^* is the unrestricted minimum residual sum of squares, RSS_r is the sum of squared residuals at some other restricted value of the parameters and T is the sample size (in this case number of days we run the simulation over) would follow a χ^2 distribution if all the ideal assumptions for OLS estimation were satisfied. At a T value of 90 the 1% confidence region for that statistic with two estimated parameters would include only values where the standard error of the model were within around 5.2% of the best value. An F version of this test, based on the statistic $[\frac{(RSSr-RSS*)}{k}]/[\frac{RSS*}{T-k}]$ and where $T = 90$, $k = 2$ would imply a

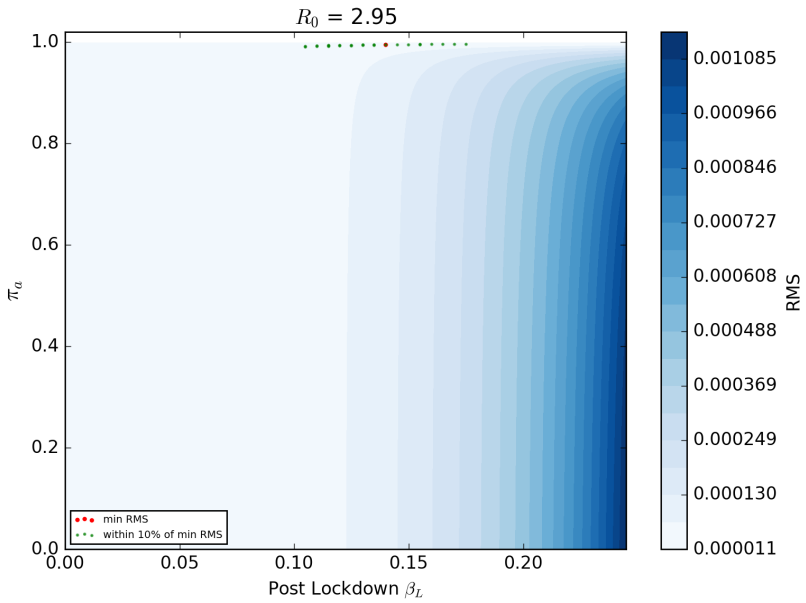

Figure 5. Like figure 3 with $\gamma = 0.0833$

1% confidence region including parameters generating a standard error no more than around 5.6% above the best fit value. However, the conditions for these parametric methods to be a reliable guide to the uncertainty over parameter estimates do not hold: the model is highly non-linear and the values of state variable used to generate predictions of y_{st} - that is I_{t-1} and S_{t-1} - are themselves generated using the estimated parameters. To overcome this, we use a simple bootstrap technique to judge confidence intervals for the parameters. We take the set of T squared residuals between data and the model using the best fit parameter values and also construct T squared residuals at some other point in the parameter space we wish to compare to. We construct a pooled square residual dataset by combining the two sets (giving $2T$ values), from which we randomly draw 2 samples (without replacement) each of size T. For each pair of samples we calculate the mean difference between them. We repeat this 10,000 times and construct the frequency distribution of outcomes. We then calculate where the actual mean difference in squared residuals between the two parameter estimates is in this sample distribution. An example of such a distribution is shown in figure 6. It is produced by taking the point in figure 5 which gives the best fit and comparing the residuals to another point in parameter space defined by ($pi_a = 0.9$ and $\beta_L = 0.007$). This value of β_L is chosen as it minimises the RMS for $\pi_a = 0.9$. The mean of the distribution of constructed differences in sums of square residuals is very close to zero (its expected value) and the actual difference in squared residuals based on the two sets of parameter estimates lies at around the 91st percentile of the distribution. We find this to be the case when the best fit parameters are compared to all combinations of values when β_L is lower than approximately 0.12 and pi_a is less than 0.9. This suggests that these regions of parameter space can be rejected but only with moderate (90%) confidence.

While the parameter space that generates a standard error within 10% of the best fit value (green dots) suggests that parameters significantly far from the red dot do significantly less well in accounting for the UK data (conditional on the wide range of assumptions we have made), the interval at a less than 10% significance level defined by our bootstrapping method is relatively wide and encompasses low values of π_a. Using these intervals, one could reject a value of π_a below 0.9 at the 10% level, but not at higher levels. In short, one cannot be sure that the main reason why test cases of those newly infected turned down was because a large fraction of the population had already been infected (very high π_a) rather than a low value of β_L (a very effective lockdown).

Nonetheless, as we describe in more detail below, we consistently find the best fit for the data (for both the UK and other countries) is for a very high value of π_a.

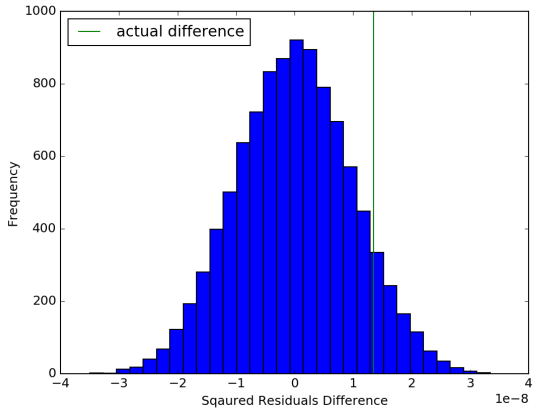

Figure 6. Frequency distribution (produced using the bootstrap method) of squared residual differences between points in parameter space defined by the best fit from figure 5 where $pi_a = 0.996$ and $\beta_L = 0.1598$ and the point at $\pi_a = 0.9$ and $\beta_L = 0.07$. The Green line represents the actual difference between mean squared residuals of the fits and lies at the 91st percentile of the distribution.

5 Results for France, US, Italy, Spain and Sweden

We estimated the same model for other countries where up to the end of April 2020 testing had largely confined to those with symptoms or those at high risk. Data for those testing positive comes from the Johns Hopkins data bank. Dates at which measures to reduce the spread of the virus became severe (the lockdown date) were taken from the Blavatnik Centre at Oxford university which has constructed an index of the severity of measures. We choose the date at which that index rises most sharply to be our starting date for lockdown measures. The dates used are outlined in table 1. For the US, the date is problematic because actions vary substantially across states.

Country	Lockdown Date
France	16 March
Spain	10 March
Italy	23 February
Sweden	19 March (partial lockdown)
USA	16 Mar (localised lockdown)

Table 1. Lockdown dates for various countries used for simulations.

The estimated impact of the measures (along with the asymptomatic) rate is freely estimated for each country. Since lockdown measures differ significantly across countries (mild in Sweden; severe in France) we expect estimates of the difference between β_L and β_0 could be substantial across countries. We would expect smaller differences in the estimated asymptomatic rate, π_a

Figures 7-11 show the fit of the model for each country. The most striking result is that the values of π_a that best fit the national data on positive tests for the virus are consistently at very high levels - generally around 0.995 (though lower for the USA). As with the UK results, taken at face value this would mean that there are 200 or so people who have had the virus with few symptoms for every infected person who has had symptoms.

But what is equally striking, and much less reassuring, is that these best-fit estimates for π_a are much higher than those based on the rather limited test results from countries that went beyond testing only those with symptoms and which are therefore closer to being based on a random sample of the population. Ultimately tests based on a large and random sample of the population is the only way to be confident about how far the virus has spread. Evidence based on test results from what is closer to a random sample (even if the sample

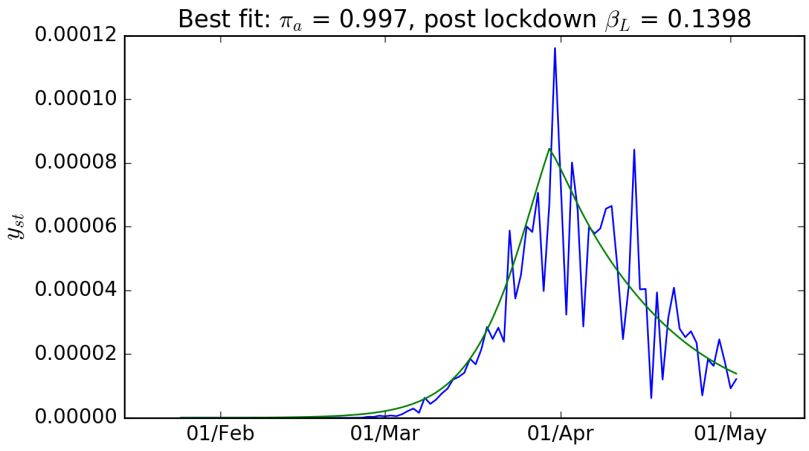

Figure 7. Model simulation for $y_s t$ that gives the best fit to data from France. R_0 = 2.95 and γ = 0.083

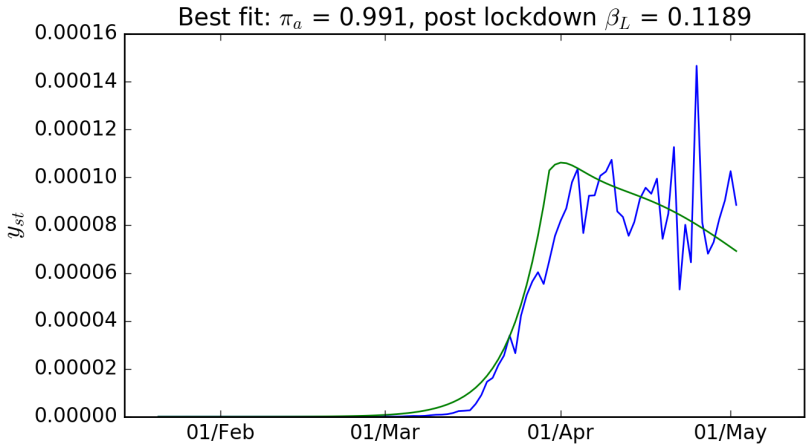

Figure 8. Model simulation for $y_s t$ that gives the best fit to data from USA. R_0 = 3.3 and γ = 0.083

size is not large and the sample not truly representative of the whole population) should be given great weight. Those results suggest a value of π_a very much lower than the values we find to fit the data on tests when most of those tested had symptoms. Some cross country studies based on deaths associated with the virus (for example Flaxman *et al.* 2020) also suggest a significantly lower value of π_a than we find best fits the data on test results.

6 Interpretation, caveats and implications

If it is really the case that those who have been infected but are asymptomatic may be 200 times as numerous as those who develop symptoms (and who are therefore more at risk) then based on the numbers who have tested positive with symptoms it would seem likely that a high proportion of the UK population (60% or so by end April 2020) had already been infected and that a substantial proportion then had some sort of immunity. This would be very good news. It would mean that the rate of new infections would be likely to die down, even if there was some rise in β as severe lockdown conditions were to be eased. Figure 12 illustrates by showing how the number of UK new daily symptomatic infections ($y_s t$) would evolve based on the model parameters (using R0=2.95; γ =0.0833 which generates best fit values for π_a and β_L of 0.996 and 0.1598 respectively) and assuming that the infection rate moves half way back to β_0 from early May to simulate some partial relaxation of lockdown measures.

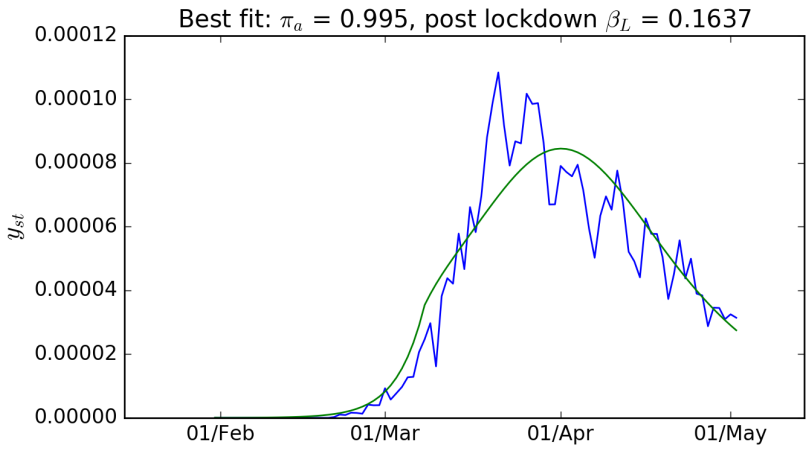

Figure 9. Model simulation for $y_s t$ that gives the best fit to data from Italy. R_0 = 3.9 and γ = 0.083

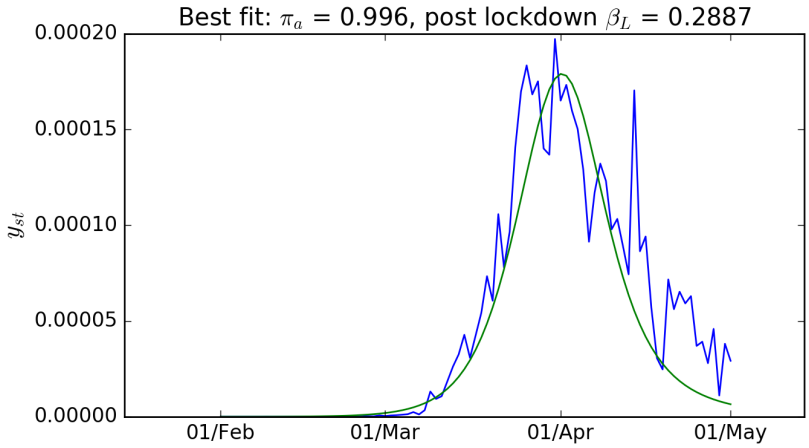

Figure 10. Model simulation for $y_s t$ that gives the best fit to data from Spain. R_0 = 3.9 and γ = 0.083

The reason that this projection shows such a decline in new cases is that the R number implied by the model would be comfortably under 1 by the end of April 2020 if the proportion of the population that is susceptible had already dipped down by as much as is implied by as high a value of π_a of 0.996. The final figure illustrates the trajectory of R implied by the model parameters π_a and β_L that best fit the UK data.

But why should results of the SIR model designed to fit the UK test data (and which also seem to best fit data from Italy, Spain, France, Sweden and the US) suggest a much higher rate of the spread of the virus than test data from countries that have done more widespread (closer to random) testing? One answer is purely mechanical: if one wants to fit a model that tracks the data on positive tests it must be one where the number of infections rises very fast early on (a relatively high R0 and β_0). But the number of new infections amongst those tested in the UK and other countries (a very high proportion of whom had symptoms) did turn around quite sharply in April. There are two things in the model that between them can account for this turn: a big reduction in β as a result of the lockdown and a large and rising population of people who had already been infected which brings the susceptible population down fast as we moved through April. The only way the latter effect could be significant is if the population of those who have had the virus but had never been tested was very substantial.

Is it possible that we have made assumptions which force the model to explain much more of the slowdown in new positive test cases by a fast rise in the immune population (which implies a very large group have had the

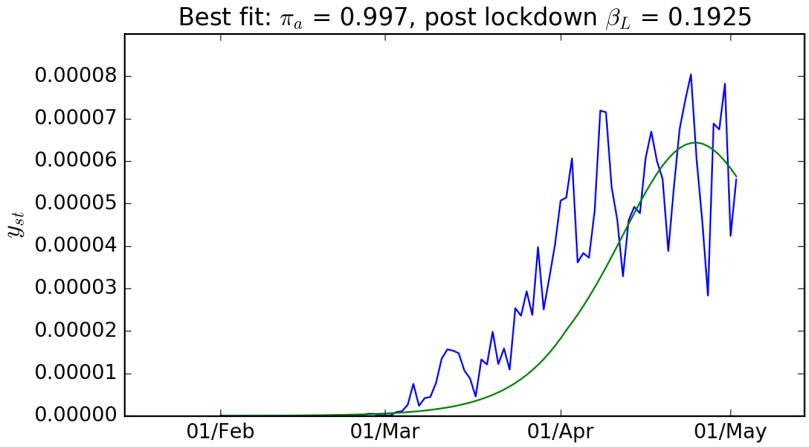

Figure 11. Model simulation for $y_s t$ that gives the best fit to data from Sweden. R_0 = 2.5 and γ = 0.083

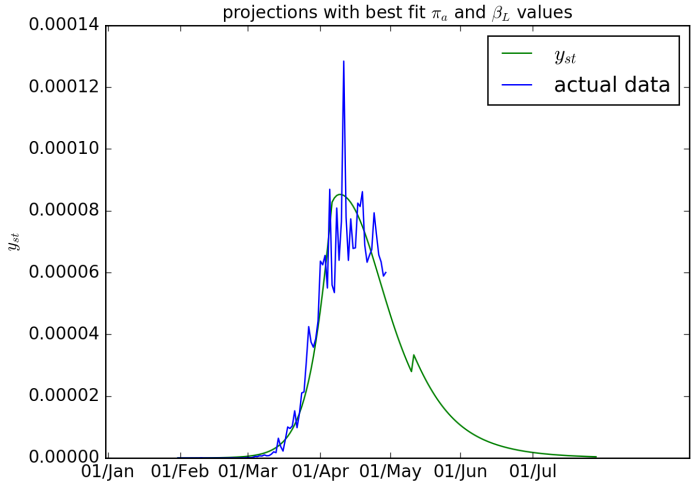

Figure 12. Actual and predicted (y_{st}) new cases. The model is run with R_0 = 2.95, γ = 0.0833 and an increase in beta on May 10th half way back to β_0

virus with few symptoms) rather than attribute it to a very effective lockdown? One factor may be significant: we have assumed a 14 day delay between the start of the lockdown and its beginning to affect the rate of new positive tests for the virus. If that lag were much smaller, more of the turn around in new cases might be attributed to the lockdown and correspondingly less to a rise in mass immunity. But in fact, when we halve the lag between the start of the lockdown and its effect on β we still find that the value of π_a that best fits the data remains very close to 1.

There is, however, one assumption that does have a significant impact on the estimated asymptomatic rate. This is the assumption that β is the same for both symptomatic and asymptomatic groups. If the rate at which the asymptomatic infect people is significantly lower than for the symptomatic, the best way for our SIR model to explain the UK data is to have a much lower number of asymptomatic (π_a). If the transmission rate of the asymptomatic is one half that of the symptomatic, but the weighted average of the two keeps the overall β as it was, π_a falls to approximately 0.5. But, the fit of the model deteriorates and the RMS error is around 16% higher than the lowest value obtained in simulations with identical transmission rates.

There is limited evidence that the transmission of the virus is weaker for those with few symptoms (Li *et al.* 2020). But, it is clear that it matters for modelling the spread of the virus (Park *et al.* 2020). The influential

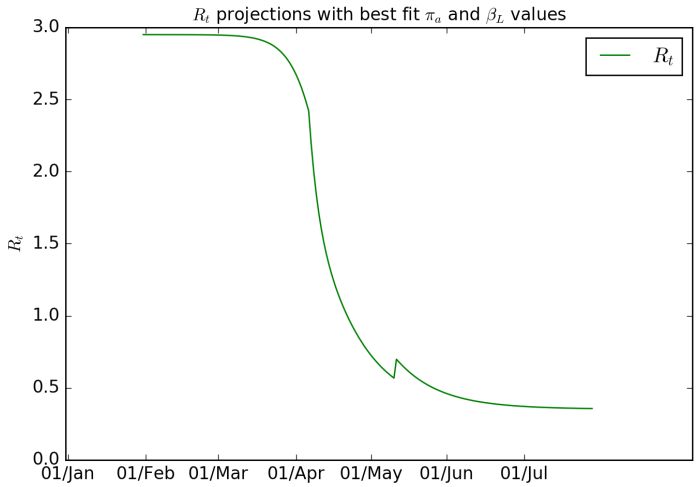

Figure 13. Forward projection of R_t which is defined as $\frac{\beta_t S_t}{\gamma N}$. The model is run with $R_0 = 2.95$, $\gamma = 0.0833$ and an increase in beta on May 10th half way back to β_0

Imperial College study (Ferguson *et al.* 2020) does assume a lower asymptomatic transmission rate (by 50%). The analysis of Gupta and her team (Lourenço *et al.* 2020), designed to explain the early spread of the virus in the UK, appears to assume a common transmission rate amongst the infected. That study suggested that the asymptomatic were a very high proportion of the infected and that the virus had spread very widely by early March. Our study suggests that estimates of the spread of the virus that best account for the data are sensitive to whether the transmission rate is assumed to be the same for asymptomatic and symptomatic groups.

We have found that when trying to match data on the recorded cases of the virus our model appears to favour high values of π_a (the asymptomatic proportion of the total infected people). This is a consistent finding across a number of scenarios where we vary the mean transmission rate, the recovery rate and lockdown measures. It is only when the transmission rate for the asymptomatic is much lower than for the symptomatic that the best fitting estimate of π_a is reduced. These two facts lead to two conclusions: First, that previous estimates of π_a near 0.9 (Li *et al.* 2020), or even higher, are consistent with versions of a simple SIR model designed to track results of tests for the virus in the UK and other countries; but we do not make the stronger claim that the evidence clearly proves such a high value. Second, that reliable modelling of the evolution of the spread of the virus requires accurate measurement of transmission rates for symptomatic and asymptomatic groups and is sensitive to whether these are different.

Finally our results indicate that it is hard to be very confident about which of two quite different factors is the primary reason why a corner has been turned in the trajectory of new cases of positive tests for the virus: i. that the lockdown is very effective; ii. that the infection has spread so far that new infections naturally slow down. In all of the countries whose data we analysed, the best fit to that data favours the second explanation and that there have been a very large number of asymptomatic infected for each infected person with symptoms. But in no cases can the alternative hypothesis be rejected with very high (0.05 or 0.01) confidence. The data by no means overwhelmingly reject the hypothesis of a value of π_a lower by enough to mean that the main cause of the slowdown (and then reversal) in the arrival of new positively tested cases of the virus were the measures taken to curb it. But there is another way of looking at the same results. This is that there is evidence that the infection may have spread far enough to mean that the trajectory of falling new cases could be maintained with some easing of restrictions.

Policy on how far to ease restrictions will inevitably have to be made in a fog of considerable uncertainty.

References

Fauci, A. S., H. C. Lane, and R. R. Redfield. 2020. "Covid-19 — Navigating the Uncharted." PMID: 32109011, *New England Journal of Medicine* 382 (13): 1268–1269. doi:10.1056/NEJMe2002387. https://doi.org/10.1056/NEJMe2002387.

Ferguson, N., D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg, *et al.* 2020. "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand."

Flaxman, S., S. Mishra, A. Gandy, H. Unwin, H. Coupland, T. Mellan, H. Zhu, T. Berah, J. Eaton, P. Perez Guzman, *et al.* 2020. "Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries."

Li, R., S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. 2020. "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2)." *Science.*

Lourenço, J., R. Paton, M. Ghafari, M. Kraemer, C. Thompson, P. Simmonds, P. Klenerman, and S. Gupta. 2020. "Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic." *medRxiv.*

Manski, C. F., and F. Molinari. 2020. *Estimating the COVID-19 Infection Rate: Anatomy of an Inference Problem.* Technical report. National Bureau of Economic Research.

Park, S. W., D. M. Cornforth, J. Dushoff, and J. S. Weitz. 2020. "The time scale of asymptomatic transmission affects estimates of epidemic potential in the COVID-19 outbreak." *medRxiv.* doi:10.1101/2020.03.09.20033514. eprint: https://www.medrxiv.org/content/early/2020/04/14/2020.03.09.20033514.full.pdf. https://www.medrxiv.org/content/early/2020/04/14/2020.03.09.20033514.

Stock, J. H. 2020. *Data gaps and the policy response to the novel coronavirus.* Technical report. National Bureau of Economic Research.