Estimating SARS-CoV-2-positive Americans using deaths-only data

James Johndrow, Kristian Lum, and Patrick Ball
University of Pennsylvania and Human Rights Data Analysis Group

April 7, 2020

Abstract

We fit a Bayesian model to data on the number of deaths attributable to COVID-19 with the goal of estimating the number of infected individuals. Our model links an underlying Susceptible Infectious Removed (SIR) model of disease dynamics to observed deaths via a time-to-death distribution informed by previous studies. This allows us to actually fit a statistical model to the data, unlike many epidemiological studies in which the SIR model parameters are simply "calibrated" to obtain outputs that look similar to the real data. The main outputs of our model are estimates of the number of infections currently, as well as forecasts of the number of infections and deaths under various scenarios for the effectiveness of social distancing measures. All of our outputs have attached Bayesian credible intervals. An important conclusion is that the confirmed case counts greatly underestimate the total number of infected individuals.

1 Intro

The total number of people infected with the SARS-CoV-2 virus is unknown. In the United States, there is a severe shortage of SARS-CoV-2 tests, and so even people in hospital with obvious symptoms of Covid-19 are rarely tested. Furthermore, it is by now well-understood that many, perhaps most people infected with SARS-CoV-2 have mild or no symptoms. The combination of these factors means that only a small fraction of the SARS-CoV-2 cases are ever confirmed by a positive test. Because the widely-reported case counts are based only on the number of positive tests, official case counts cannot possibly reflect the true number of people infected with the SARS-CoV-2 virus. Indeed, the confirmed case counts may more accurately represent the distribution of tests than the prevalence of infections.

The true number of SARS-CoV-2 positive individuals is necessarily much higher than the reported counts. But how much higher? Understanding the total size of the infected population is important for planning, policy, and public communication because it determines how many people will need critical care, and how many are at risk of death. Knowledge of the true number of affected individuals could also

help with public adoption of containment or mitigation measures, which may seem an overreaction when the only publicly available numbers make it appear that the disease affects relatively few. It is therefore important for healthcare planning, economic policy, and transparency to estimate the population prevalence of SARS-CoV-2 infection.

This analysis models the spread of SARS-CoV-2 infection in the United States based on data on the number of Covid-19 attributable deaths, the infection fatality rate reported by previous studies, and the time-to-death reported by previous studies. We chose this approach because, while we believe the case count data from which epidemiological models would normally be built to be highly unreliable, we expect the data on the number of deaths from Covid-19 to be reasonably close to the truth. Our approach is founded on the belief that vast majority of Covid-19-related deaths are recorded as such. If this is not true, the model would likely underestimate the true prevalence of the disease.

The Bayesian model we present here is a version of the standard Susceptible-Infectious-Removed (SIR) model. However, rather than "calibrating" the parameters of the SIR model to obtain a reasonably close match to the observed deaths, we develop an explicit probabilistic model linking the dynamics of new infections in the SIR model to observed deaths. This allows us to give estimates of uncertainty in parameters such as the number of infected individuals. The sampling model for the daily number of deaths uses a model of disease progression conditional on the past time series of new infections. The parameters defining the disease progression model are fixed and based on existing clinical data. The underlying process generating the time series of new infections is the SIR model, the parameters of which are estimated by Markov Chain Monte Carlo (MCMC). We place informative priors on these parameters using existing estimates of the R_0 and the infectious period of SARS-CoV-2.

Using this model, we are able to "work backwards" from the time series of observed deaths to estimates of the standard epidemic curves in a principled way. This fully model-based approach allows us to incorporate detailed information on disease progression, such as a distribution over the time to death as opposed to assuming every person dies exactly x days after their infection, as has been done in similar more informal analyses. Our approach allows us to provide uncertainty intervals on parameters estimated by our model, and using our fitted model, we make rough preliminary estimates of the actual number of SARS-CoV-2 positive Americans.

The goal of this work is not to provide specific forecasts of the infected population and likely deaths, although our initial estimates here fit the observed patterns of deaths. Accurate forecasts will require location-specific measures of containment and treatment efficacy, as well as age- and comorbidity-specific infected fatality rates. We don't have these data at present, but our model could incorporate them as better information becomes available. Our paper offers a modeling approach using minimal but probably-good data, describes a likelihood and priors, is fitted to data, and is underpinned by a widely-used epidemiological model that is designed to approximate the real dynamics of disease spread.

2 Related Work

Several previous studies have offered methods for estimating the total number of infections, although estimating the size of the infected population has been largely a side-effect of other analytic goals. For example, Li et al. [5] propose that in China in the first month of the epidemic 82-90 percent of infections were undocumented. Riou et al. [8] use a SEIR model—an analogue of the SIR model in which there is a lag between exposure and the individual becoming infectious—and calibrate their model to the time series of reported deaths and reported infections. By modeling the under-reporting of symptomatic cases, and by assuming that approximately half of infections lead to symptomatic cases, they estimate the infected population in Hubei, finding that approximately 30% of infections were documented.

Ferguson et al. [1] model the effect of transmission between susceptible and infectious individuals using a microsimulation model built on synthetic populations designed to mimic the populations of the United Kingdom and United States. They assume a fixed time-to-onset and a range of R_0 values from 2.0-2.6, and they assume symptomatic cases to be 50% more infectious than asymptomatic cases. They calibrate their model to the cumulative number of deaths seen by March 14, 2020. This is conceptually similar to our approach in that the model ignores case counts and is fitted only to death information. However, the calibration is based on only one observation—the cumulative number of deaths at a fixed time point—rather than fitting to the full time series of deaths, and there is no likelihood or sampling model from which to estimate parameter values.

Having calibrated their model to the cumulative number of deaths, Ferguson et al. estimate deaths and hospital loads under different non-pharmaceutical interventions (NPI) involving social distancing and isolation; the estimated infected population is a side-effect of their model. By contrast, Perkins et al. [7] estimate directly that in the US, more than 90% of infections have been undocumented by tests using Chinese data and initial reports in the US.

The CHIME app[1] is an online tool created by researchers at the University of Pennsylvania to help hospitals anticipate the number of incoming Covid-19 patients and their needs. It takes a conceptually similar approach to the work we present here. The CHIME model uses the current number of Covid-19 hospitalizations to back out the total number of cases based on external estimates of the hospitalization rate of the disease. Similar to our work, they do not use data on the case counts. They make forward projections for the number of hospital admissions, ICU admissions, and ventilators needed over the coming weeks. They allow the user to specify the parameters of their underlying epidemiological model as inputs in terms of the doubling times.

Two other models are worth noting. The *New York Times* online tool allows the user to specify inputs to understand how those inputs affect likely infections, hospital loads, and deaths; infections are a side-effect of the rest of the model. The model given in Murray [6] has a goal similar to CHIME (hospital use planning). However, Murray [6] does not use an SIR model, or any mathematical model of disease spread. Instead his projection uses only the pattern of observed deaths, which he fits to an arbitrary

[1] chime

curve chosen because it fits the data well (p.4). He uses location-specific parameters and the time between infection and death as predictors; this model avoids predicting the infected population at all.

The size of the infected population and the extent to which the reported case count underestimates it are important questions. The size of the infected population could inform health care providers about the number of cases likely to require hospitalization in the coming 1-2 weeks. Furthermore, knowing the probable size of the infected population would alert policymakers and the public about the likely number of deaths in the next 2-4 weeks.

3 Model

We model the infection spread in the entire U.S. population with a single SIR model. With the unfortunate recent growth in the number of reported deaths and hospitalizations, there will soon be enough data to disaggregate the modeling to individual states or even cities. The use of a single model for the entire U.S. population is clearly a limitation of our approach and we anticipate disaggregating geographically in the near future.

Let ν_t be the number of *new* infections on day t of the epidemic, and let p denote the infection fatality rate, i.e. the probability of death given infection. We denote the day of the first infection by T_0. Let $\theta = \{\theta_s : s = 0, 1, ..., m\}$ be the set of probabilities defining the discrete time-to-death distribution, where θ_s denote the probability that, for those who die, death from Covid-19 occurs s days after the initial infection. Let $X(t, t')$ denote the number of individuals newly infected on day t who die on day t'. Our death model is

$$X(t, t') \mid p, \theta \sim \text{Poisson}(p\nu_t\theta_{(t'-t)}).$$

The observed deaths on day r are thus given by

$$D(r) = \sum_{t=1}^{r} \sum_{t'=t}^{r} X(t, t'),$$

the sum over all previous days of the number of individuals infected on that day who went on to die on day r. This has marginal distribution

$$D(r) \sim \text{Poisson}\left(p \sum_{t=1}^{r} \sum_{t'=t}^{r} \nu_t\theta_{(t'-t)}\right).$$

The use of a Poisson distribution in specifying our model may seem unnatural compared to the specification $X(t, t') \mid p, \theta \sim \text{Binomial}(\nu_t, p\theta_{(t-t')})$. The Poisson specification simplifies computation (as we will see in Section 4), and fortunately in cases where $p\theta_s$ is small and ν is large (precisely the situation in which we find ourselves after the very early days of the epidemic), the Poisson$(p\theta_s\nu)$ distribution is a good approximation to Binomial$(\nu, p\theta_s)$. Furthermore, allowing for ν_t to take real values (as opposed to integer values, as would be required for a Binomial distribution), allows us

to use simpler, deterministic models for the underlying epidemiological curves defining the ν_ts. This also simplifies computation.

The observed number of deaths D are linked to the compartmental model via the total number of newly infected individuals on day t, ν_t, which appears in the above equations. In our preliminary analysis, we have used a discrete-time version of the Susceptible-Infectious-Removed (SIR) model with a modification of the state evolution to make the model consistent with the observed number of deaths on each day that enter into our likelihood. The state evolution of this model is deterministic and given by

$$\begin{aligned}
\nu_t &= S_{t-1} I_{t-1} \beta N^{-1} \\
R_t &= R_{t-1} + \gamma_r (I_{t-1} - D_{t-1}) + D_{t-1} \\
S_t &= S_{t-1} - \nu_t \\
I_t &= I_{t-1} + \nu_t - \gamma_r (I_{t-1} - D_{t-1}) - D_{t-1}
\end{aligned}$$

where S_t is the number of susceptible individuals at time t, I_t is the number of infected individuals at time t, N is the total population size (set to 330 million in the United States) and R_t is the number of removed (recovered or deceased) individuals at time t. In this model, infected individuals are considered contagious. This model is nearly identical to the classic SIR model with the exception that γI_{t-1} in the canonical model has been replaced by $\gamma_r(I_{t-1} - D_{t-1}) + D_{t-1}$ in our specification. In our evolution model, the expected number of individuals who move from state I to state R is explicitly decomposed into the number who died (which is an integer count that comes directly from our observed data) and the expected number who recover each day, $\gamma_r(I_{t-1} - D_{t-1})$. The logic behind the latter term is that $(I_{t-1} - D_{t-1})$ is the number of people who were infected as of time $t-1$ who did not die during that time period. Of those, we expect γ_r to recover per unit time, so in our model γ_r is the rate of recovery for those who will eventually recover. Thus, the mean time between infection and "recovery" (the end of the contagious period) for those who do recover is given by γ_r^{-1}, and R_0 is approximately given by $\beta\gamma_r^{-1}$. These equivalences are approximate in our case, since our compartmental model is discrete-time rather than continuous-time, and we explicitly split the removed population between deceased and recovered.

Figure 1 shows one realization of our modified SIR model. The top panel shows one draw of the daily number of deaths. Notice that this is not a smooth curve. When fitting our model, these Ds will be our data, and will be the only thing we observe directly. The other panels are what we will infer from the Ds. The middle panel of the figure shows the νs, the number of daily new infections. The series of Ds has roughly the same shape as the νs, though it lags it by about 25 days. This lag is due to how we have set θ, the parameter governing the time-to-death distribution which we describe below. The bottom panel gives the standard view of the SIR model, showing the total number of susceptible, infected, and removed at any given time point.

This model has several parameters. Due to identifiability constraints, we do not estimate all of them. Rather, we fix those for which there exists high quality information on reasonable values that is applicable in the context studied here. By and large, this includes those parameters pertaining more to the biology of the disease than to the

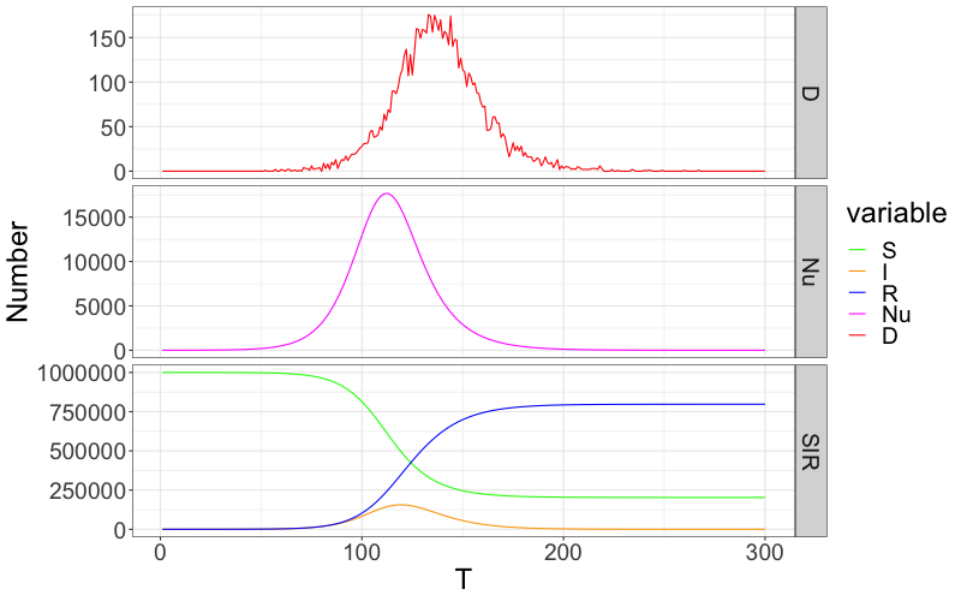

Figure 1: One realization of our modified SIR models

social dynamics of its spread.[2] These include θ and p. For θ, we draw primarily from two studies. Zhou et al. [10] reports that in Wuhan, China, the time from symptom onset to death had a median time of 18.5 days with an interquartile range of 15 to 22 days. A Poisson-Gamma distribution with parameters $\{\alpha, \beta\} = \{27.75, 1.5\}$ matches the reported quantiles well. Lauer et al. [3] estimate the incubation period of the disease, also using data from China. They report a median incubation time of 5.1 days with an estimated 97.5th percentile of 11.5 days and a 99th percentile of 14 days. The quantiles of a Poisson-Gamma distribution with parameters $\{\alpha, \beta\} = \{5.5, 1.1\}$ match these reported quantiles well. We calculate the distribution of the total time from infection to death by generating 100,000 samples from the described distribution of the incubation period and the described distribution of the time from symptom onset to death. The time to death is the sum of these two numbers. We truncate the maximum time to death from infection to be the 99th percentile of the generated samples. This results in a time to death distribution shown in Figure 2.

To set p we rely on several external data points. Russell et al. [9] use data from individuals on the Diamond Princess cruise ship to estimate an infection fatality rate of 1.2% (95% CI: 0.38%-2.7%) after adjusting for delays between infection confirmation and death. The ship was a closed population: we know who was on the ship and therefore who to test, so we have confidence in the denominator (all individuals infected

[2]Clearly, it is impossible to make a clear distinction between these two types of parameters. For example, if the rate of spread allows for the number of people requiring care to overwhelm existing healthcare infrastructures, the case fatality rate will increase, as those who require medical support to survive but cannot get it will die.

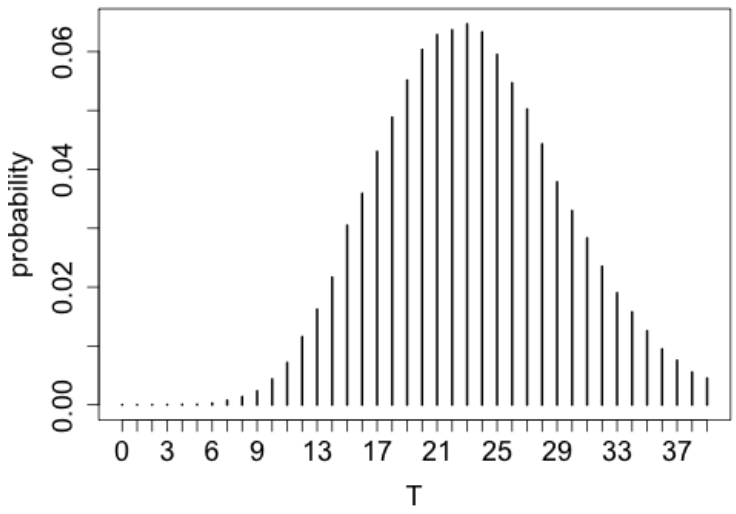

Figure 2: Conditional on death occurring, the probability of death on each day following infection.

with COVID-19 were identified because all people on the ship could be tested).

As of 25 March 2020 there were 9,137 confirmed positive cases and 126 deaths in South Korea.[3]. This gives a crude case fatality rate of 1.4%. Although this statistic focuses on confirmed cases, this number may be reasonably close to the true infection fatality rate because South Korea has aggressively tested and traced contacts, including testing of asymptomatic individuals. Moreover, South Korea's epidemic is relatively old, giving sufficient time that many of the infected individuals who will go on to die have already done so. Keep in mind that computing an infection fatality rate in this way—by dividing total deaths at time t by total cases that have been identified up to time t—would give an underestimate of the infection fatality rate because the denominator includes some people who are currently infected who will eventually die. To give some estimate of how much of an underestimate this may be, South Korea as of 25 March reports 59 patients in serious or critical condition. If 1/4 of these patients die, then the empirical case fatality rate would rise to 1.6%. On the other hand, despite the very high level of testing in South Korea (as of 25 March, they had performed about 7 tests per 1000 population), the number of reported cases is still an underestimate of the total number of cases.

The other country that was very aggressive about testing and surveillance from the early days of the epidemic is Germany. As of 27 March, Germany reports 281 deaths and 47,278 total cases, giving an IFR estimate of 0.59 percent. Germany's case count

[3] `https://www.cdc.go.kr/board/board.es?mid=a30402000000&bid=0030`

still appears to be rising, whereas South Korea appears to have largely controlled the spread. So, many of the fatalities in Germany may be yet to come. If this is true, then the raw IFR from Germany's data could be an under-estimate if this effect is enough to offset the infections that have gone undetected despite the aggressive testing taking place. In both the comparison to Germany and South Korea, differences in underlying demographics, comorbities, and other risk factors could limit the applicability of these estimates to the United States. However, such factors are difficult to adjust for, since data on risk factors is preliminary and limited. We believe that the comparison is still useful in providing a rough ballpark estimate.

Finally, a recent paper in *Science* attempts to estimate the total size of the infected population during the epidemic in China [5]. Their estimates of the total number of infected implies an adjusted infection fatality rate of about 0.4 percent (see the note between Li et al. (2020) co-author J. Shaman here) . Thus, consistent with all of these estimates, and because the infection fatality rate is arguably the most important parameter in our model that cannot be learned from the data, we consider two different cases: 0.5 percent and 1.0 percent.

We estimate the parameters of the model that govern the spread of the disease rather than its progression conditional on infection. The parameters we estimate in our model are γ_r, β, and T_0. Previous estimates of the length of the infectious period γ_r^{-1} vary widely, so we choose a fairly diffuse prior $\gamma_r^{-1} \sim \text{Uniform}(3, 25)$, corresponding to an infectious period that ranges from 3 to 25 days. To set the prior on $\beta \mid \gamma_r$, we use the reported 95 percent confidence interval in [4] to establish upper and lower bounds on the R_0, and put $\beta \mid \gamma_r \sim \text{Uniform}(1.4\gamma_r, 3.9\gamma_r)$, encompassing possible R_0 values of between 1.4 and 3.9 for all of the allowed values of γ_r. We place a uniform prior on T_0 between January 1 and January 30.

4 Computation

We do computation by MCMC using the following algorithm. This algorithm is an adaptive Metropolis-within-Gibbs algorithm (see e.g. [2]). The algorithm produces samples from the Bayesian posterior distribution of the parameters of our model.

1. Sample X given D, ν, p, θ from

$$X(\cdot, s) \sim \text{Multinomial}(D_s, \pi)$$

where

$$\pi_t \propto p\nu_t\theta_{(s-t)}.$$

Each multinomial sample of $X(\cdot, s)$ imputes the number of people infected on each day that went on to die on day s, conditional on having observed exactly D_s deaths at day s. This follows because a vector of independent Poisson random variables constrained to sum to some positive integer has a Multinomial distribution with probabilities proportional to the Poisson rate parameters (see [4]).

[4] `https://ecommons.cornell.edu/bitstream/handle/1813/32480/BU-39-M.pdf?sequence=1` for a proof

2. We update β, γ_r, and T_0 using the adaptive Metropolis algorithm. To do this, we propose a new set of parameters, $\{\beta^*, \gamma_r^*, T_0^*\}$ from $N((\beta, \gamma_r, T_0), \Sigma)$, with the time-inhomogeneous covariance Σ computed using the method of [2]. We then calculate the corresponding set of ν_t^*s. We accept or reject the proposed $\{\beta^*, \gamma_r^*, T_0^*\}$ using Metropolis-Hastings, with target density propportional to

$$\ell(\nu, \beta, \gamma_r) = \sum_{t=1}^{T} \rho_t \log(p\nu_t \eta_{(T-t)}) + \log(\pi(\beta, \gamma_r, T_0)),$$

with $\rho_t = \sum_{t'=1}^{T} X(t, t')$ and $\pi()$ representing the prior.

We fit our model to the daily number of deaths in the United States as recorded on `https://worldometer.com`. We use data up until 28 March 2020 to fit our model.

We run for 50,000 iterations, begin adaptation after 5,000 iterations, and use 10,000 iterations of burn-in. Trace plots for the parameters β, γ_r, T_0 are shown in the appendix. R code for all of the analysis here is available at [XXX].

5 Results

Figure 8 shows the observed deaths in the U.S. (red), as well as 1,000 posterior samples of the number of deaths (gray) and the pointwise posterior mean (blue). The model provides a reasonably good fit to the observed data.

Figure 4 shows posterior samples of the R_0. The mean value is about 2.75, with a 95 percent posterior credible interval of approximately [2.2, 3.75]. This is in line with previous estimates of R_0 for this disease.

The left panel of Figure 9 shows the posterior distribution of the number of newly infected individuals on March 18, and the right panel shows the total number of infected plus recovered individuals on 18 March. We focus on an estimate of the number of infections on March 18, because this is roughly the time at which social distancing measures began in earnest in the United States. For example, California put into place statewide social distancing measures on 19 March. Because we fit our model to only deaths through March 28, and very few individuals die within 10 days of infection, we do not expect that social distancing has had much effect yet on our estimates. However, it does affect the interpretation of our estimates. The deaths being observed today are of individuals infected several weeks ago, at a time when few social distancing measures were in place and the virus was spreading essentially unchecked in the United States. Thus, our estimates of R_0 reflect the transmissibility of the virus *without social distancing in place*. Once social distancing measures took effect, we expect that the R_0 would decline. We study several cases for how social distancing has changed R_0 in the next section.

Our mean estimate of the total number of infected plus recovered on March 18 is about 475,000. Comparing this to the number of reported cases on 18 March (9,187), we find that the reported case numbers underestimate the total number of infected by a factor of about 50. Of course, infected individuals take some amount of time to develop symptoms, and thus we would not expect everyone who had ever been infected as of March 18 to have sought or received testing. One study [4] reports the

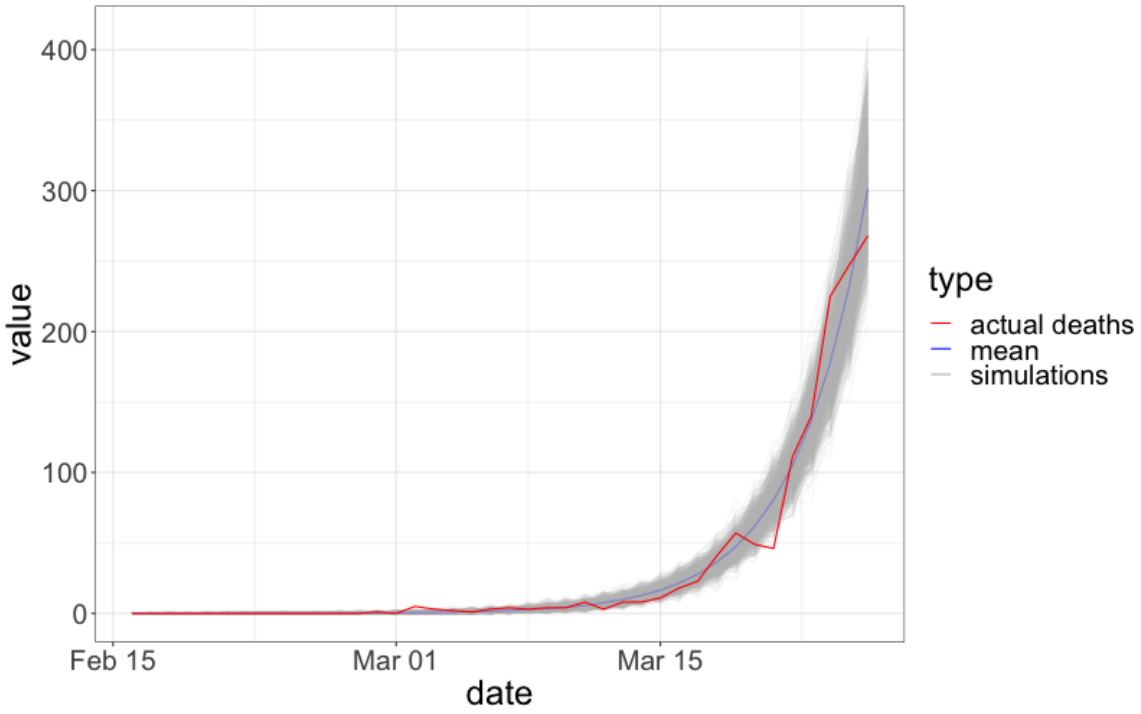

Figure 3: The observed deaths in the United States (red), posterior samples of the number of deaths (gray), and pointwise mean of the posterior samples of the number of deaths (blue).

incubation period is about 5 days, and the 90th percentile is about 12.5 days, while another study Lauer et al. [3] reports the median incubation period is about 5.1 days, the 2.5th percentile about 2.2 days, and the 97.5th percentile about 11.5 days. Thus a better comparison may be between our estimate of the total number of infected on March 18,2020 and the total number of reported cases on 23 March 2020 (43,781) or on 27 March 2020 (85,435), which suggests that the case count underestimates the total number of infected who are "visible" to testing by a factor of between 5 and 10. The latter is in line with the recent paper on the undercount in Hubei [5]. Although the extent of under-testing could be very different in the United States and in Hubei, we find it interesting that these estimates correspond and suspect this is because the numbers are linked by the fact that in both places, asymptomatic people were rarely tested.

Next we look at model predictions under several scenarios. Since significant mitigation strategies began to be put in place about ten days ago, in each scenario we assume that the R_0 is as estimated by our model up until March 18, at which point the R_0 drops. Several options for after-mitigation R_0s are considered in each of the figures, each new R_0 corresponding to a different colored trajectory. Plots of the number of

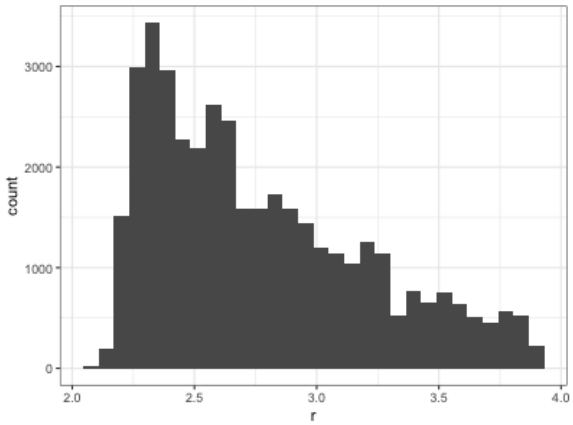

Figure 4: Estimated approximate R_0 under our model.

new infections by day along with a 95 percent posterior credible bands, are shown in the left panel of Figure 10.

The right panel of Figure 10 shows the estimated number of deaths by day in each of the scenarios. In the absence of social distancing measures, the number of daily deaths continues to rise. If this curve were extended out further, we'd find that the number of daily deaths peaks at around 10^5 in early to mid May under this scenario. On the other hand, if social distancing measures have the effect of reducing the R_0 by about 36 percent to 1.75, then the number of deaths would be less by an order of magnitude at this time, though will not yet have peaked. Similarly, under the "no effect of social distancing" scenario, new infections peak in early to mid-April at about $10^{7.4}$ at more than 10 million new infections. If social distancing measures reduce R_0 to 1.75, the number of new infections at this time is less by more than an order of magnitude. In a more hopeful scenario where the R_0 drops to 1.25—almost to the point of suppression—the number of daily deaths in mid-May is only about 4,000 and the number of daily new infections in mid-April is only about 400,000.

All of these results have been conditional on the assumption that the IFR is 1%. This is the most important assumption in the model, and the value of the IFR cannot be learned directly from the data used in this analysis. It must be included as an input or assumption to the model. Because of this, we also give results for an alternative, lower IFR of 0.5% in the appendix.

6 Discussion and Conclusion

Our model strives to use only information which has a reasonable chance of being measured correctly. We condition our findings on the observed number of deaths, and we use the infected fatality rate (p) and the time from infection to death (θ_s) measured in specific clinical settings to estimate the number of people infected and the number likely to die. Our model predictions of the number of deaths in the US are consistent

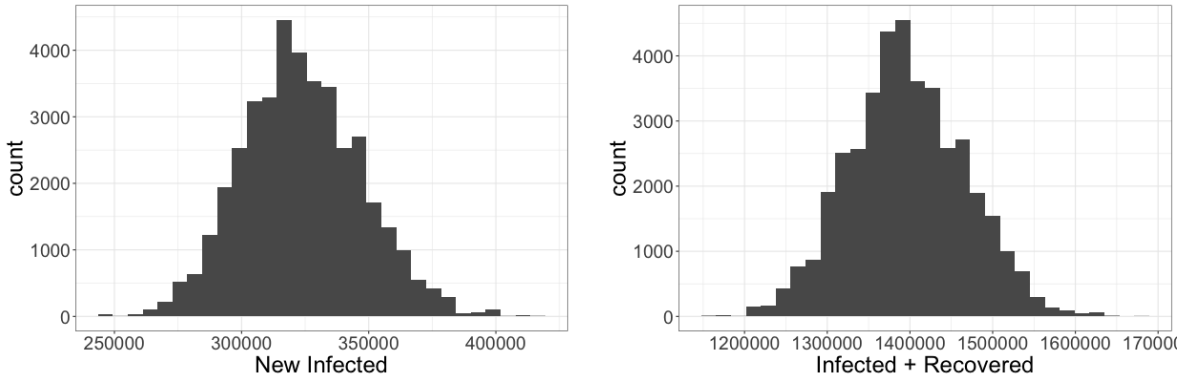

Figure 5: Histograms of the number of newly infected people on March 18 (left) and the cumulative number of infections on March 18 (right).

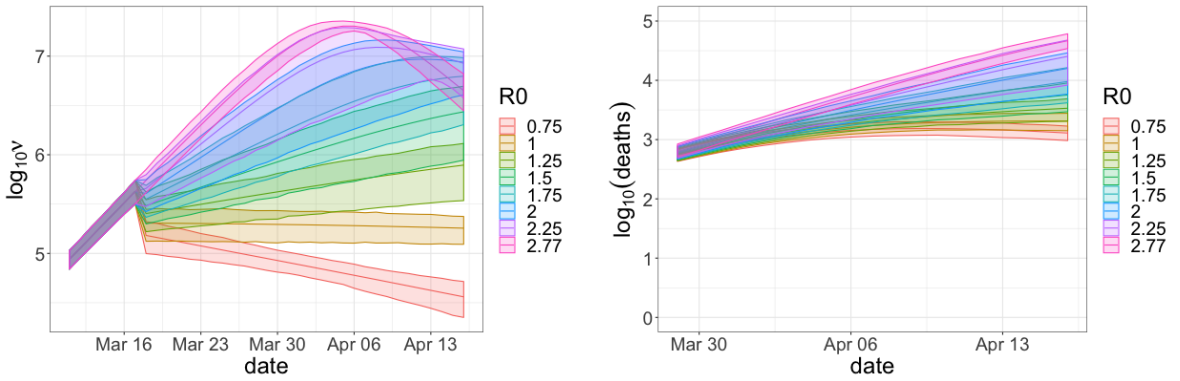

Figure 6: $\log_{10}$ of new infections (left) and deaths (right) under different scenarios for the effect of social distancing on R_0.

with the time series of deaths observed from the beginning of the pandemic to the present.

In contrast to models that use a wider range of information that may be less precisely measured, our approach minimizes information use and prioritizes parameter identifiability. The uncertainty from the clinical measures is propagated to the estimates of new infections ν_t and likely deaths D_t by means of explicit priors that reflect the ranges of the reported measures.

The model has several underlying assumptions: perhaps most importantly, we assume that after an individual has recovered, they cannot be reinfected. The parameters we draw from prior studies, p and θ, may be better measured as more data is collected and clinicians improve reporting. It is easy to update our model with improved values of these parameters.

Finally, our model uses the SIR approach to link observed deaths to the underlying

unobserved infections. It would not be difficult with our approach to change the linkage for some other model.

7 Acknowledgements

The authors thank Dr Megan Price and Tarak Shah for their comments and suggestions. This work was supported by grants to the Human Rights Data Analysis Group by the John D. and Catherine T. MacArthur Foundation and the Oak Foundation.

A Trace plots

Figure 7 shows traceplots of β, γ_r, and the implied R_0 for $p = 0.01$. Once adaptation begins, the algorithm evidently mixes quite well.

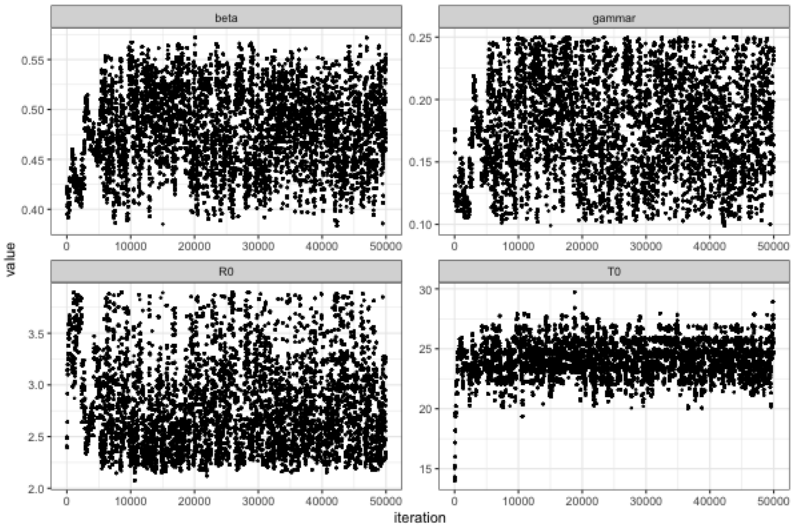

Figure 7: MCMC samples of β, γ_r, the implied approximate R_0, and T_0 in the case where $p = 0.01$.

B Alternative IFR

All figures in this section pertain to the scenario where we assume the IFR is $p = 0.005$.

References

[1] Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., et al. (2020).

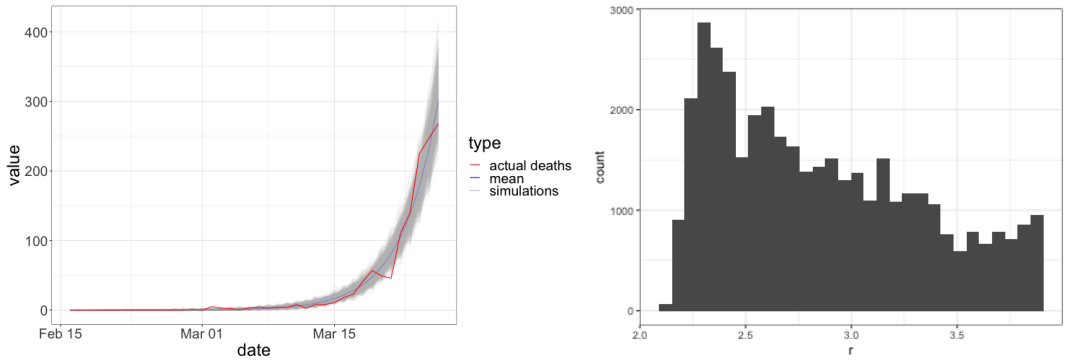

Figure 8: (left) The observed deaths in the United States (red), posterior samples of the number of deaths (gray), and pointwise mean of the posterior samples of the number of deaths (blue); (right) estimates of R_0. Both figures pertain to the scenario where $p = 0.005$.

Impact of non-pharmaceutical interventions (npis) to reduce covid-19 mortality and healthcare demand. *London: Imperial College COVID-19 Response Team, March*, 16.

[2] Haario, H., Saksman, E., Tamminen, J., et al. (2001). An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242.

[3] Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., and Lessler, J. (2020). The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*.

[4] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., et al. (2020a). Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*.

[5] Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., and Shaman, J. (2020b). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov2). *Science*.

[6] Murray, C. (2020). Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator days and deaths by us state in the next 4 months. *MedRxiv preprint*.

[7] Perkins, A., Cavany, S. M., Moore, S. M., Oidtman, R. J., Lerch, A., and Poterek, M. (2020). Estimating unobserved sars-cov-2 infections in the united states. *medRxiv*.

[8] Riou, J., Hauser, A., Counotte, M. J., and Althaus, C. L. (2020). Adjusted age-specific case fatality ratio during the covid-19 epidemic in hubei, china, january and february 2020. *medRxiv*.

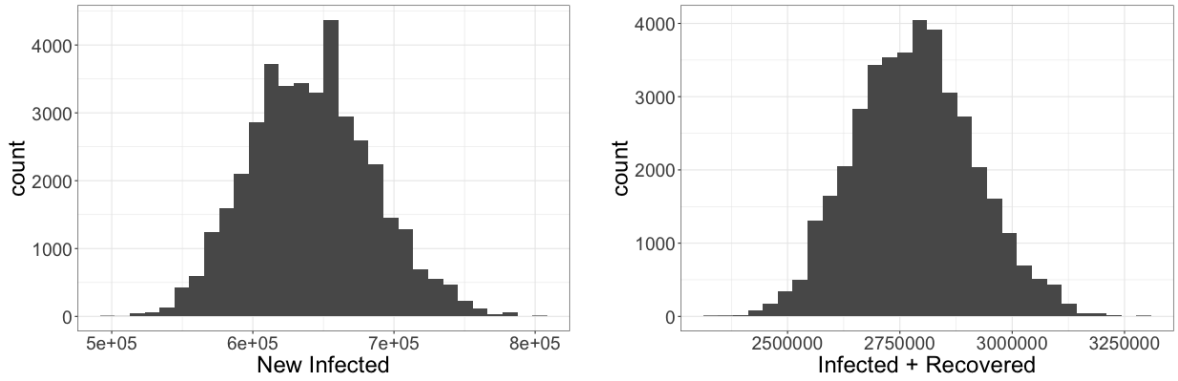

Figure 9: Histograms of the number of newly infected people on March 18 (left) and the cumulative number of infections as of March 18 (right).

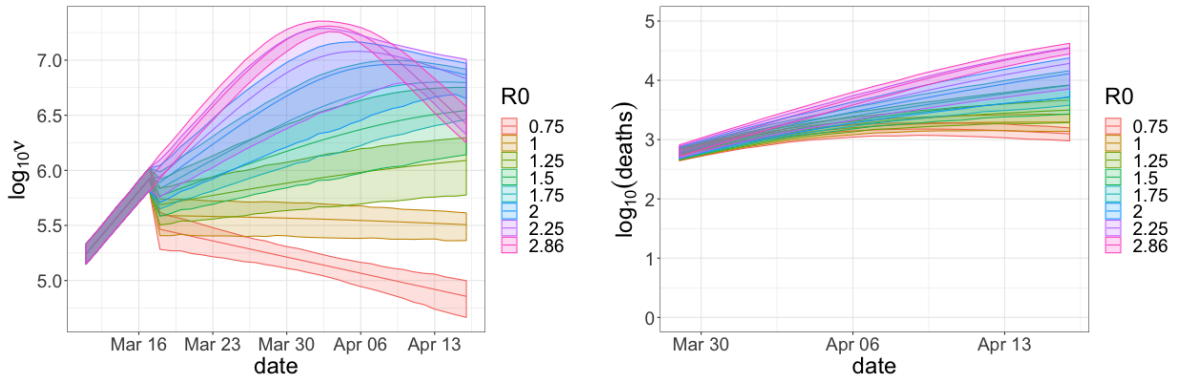

Figure 10: $\log_{10}$ of new infections (left) and deaths (right) under different scenarios for the effect of social distancing on R_0.

[9] Russell, T. W., Hellewell, J., Jarvis, C. I., van Zandvoort, K., Abbott, S., Ratnayake, R., Flasche, S., Eggo, R. M., Kucharski, A. J., and nCov working group, C. (2020). Estimating the infection and case fatality ratio for covid-19 using age-adjusted data from the outbreak on the diamond princess cruise ship. *medRxiv*.

[10] Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., et al. (2020). Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. *The Lancet*.