How predictive are models of Covid-19?

Cliff Kerr, Institute for Disease Modeling (ckerr@idmod.org)
Daniel J Laydon, Imperial College (d.laydon@imperial.ac.uk)
Jean Helie, GitHub (jhelie@github.com)
Albert Ziegler, GitHub (wunderalbert@github.com)
Oege de Moor, GitHub (oegerikus@github.com)

Projections of Covid-19 models are used - among other factors - to support momentous policy decisions. Much remains unknown about the virus, and so the projections are bound to be wrong. Yet they are useful, because they help policy makers run “what if” scenarios, and understand the possible health consequences of proposed mitigation policies. Inga Holmdal and Caroline Buckee provide a very clear explanation of the limitations and merits of models.

In the first few months at the start of the pandemic, there was no easy way of testing how reliable the model projections were, and the data to help determine parameter values were extremely sparse. The projections could therefore only be judged by experts with a deep understanding of the models themselves, by checking their assumptions and implementations. Several months later, even non-experts can assess the accuracy of the projections, simply by running the models as black boxes and checking how the projections match up with reality. This note reports the results of one such experiment.

In writing this note, we wanted to answer the following questions:

- If we allow the models access to everything we knew up to the end of May, how well do they fit the data from the past?
- If we only allow the models access to data up until the end of April, how well would their predictions have held up until mid July?
- Are the models broadly in agreement over the future, or are the projections wildly different?

We should stress that most of us are not epidemiologists, and indeed the experiments reported here do not require epidemiological knowledge. All we did was to run the code created by the experts, and compare the projections to what actually happened. We are grateful to members of the modeling teams we worked with for their practical assistance with operating their models.

We are not advocating one model over another, nor do we offer new insights into Covid-19 and its containment. The code of the models themselves, and the scripts we used to compare their results, are all available in open source so others can repeat our experiment. In sharing these results, we hope to help other non-epidemiologists form an opinion on the utility of Covid-19 models.

Three sample models

The experiment was conducted with three sample models. There are many other models available, with different characteristics and advantages - here we choose just three that are representative of common model types. We do believe it would be valuable to extend the experiment to more models, and we shall return to that point at the end of this note.

Imperial CovidSim

The CovidSim microsimulation model was developed by Imperial College London. CovidSim models the transmission dynamics and severity of infections throughout a spatially and socially structured population over time. It enables modelling of intervention policies and healthcare provision, and how they affect the spread of Covid-19.

CovidSim takes detailed spatial and demographic inputs for each region where it is run, including data on schools, workplaces, age distribution, typical household sizes and so forth. To make CovidSim fit the outbreak in a particular location, it needs to calibrate to the cumulative number of deaths on a given date. Based on this one input, it fits the likely start of the outbreak and the original number of infections This limited calibration can be refined by specifying any intervention policies that have occurred previously, and is part of the CovidSim code base.

Further fitting of CovidSim to ground truth data is currently done externally (by scripts that are not currently part of the published code base), running the model systematically over an appropriate grid of parameters and recording how well the predictions match the reported daily number of deaths. In the experiment reported here, we used a 12x10 grid for the base transmission number R0 (ranging between 1.75 and 3.75) and another parameter named “relative spatial contact rates over time given social distancing” (ranging between 25% and 95%), which encapsulates the intervention effectiveness. We’ll further discuss the loss function used in fitting below.

CovidSim has many other parameters but those were set to fixed values for the experiment - better fits would be possible through judicious variation of more parameters. For now, we refrained from doing so for simplicity and to save computational resources.

IDM Covasim

Covasim is a stochastic agent-based simulator developed by the Institute for Disease Modeling. It provides projections of the numbers of infections, hospitalizations, and deaths. Covasim can also be used to explore the potential impact of different interventions, including social distancing, school closures, mask usage, testing, contact tracing, and quarantine.

Covasim takes in some demographic data for each region, but it does not require the same extensive data as Imperial CovidSim. Covasim can be automatically calibrated on the number of fatalities, the number of positive tests and the number of severe cases. For each of these quantities, it is fit both to the cumulative and daily numbers. In this example, the quantities automatically calibrated by Covasim were: number of "seed" infections, infectiousness (beta), the date when interventions to reduce transmission are assumed to have occurred, the magnitude of these interventions, and the likelihood that people with Covid-19 will be tested.

Stanford/Stripe Modeling Covid-19 (MC-19)

MC-19 is an SEIR model, which divides the population into four groups (Susceptible, Exposed, Infected and Recovered). It then models the transitions between these states through differential equations. One input is Google mobility data, used to estimate the level of social distancing.

MC-19 takes fairly limited demographic inputs - like Covasim, it only needs to know about the population size and age distribution. MC-19 fits its projections to the available data for fatalities, positive tests, hospitalizations and critical cases (that require treatment in intensive care).

The parameters that are automatically fit are the base reproduction number R0, the import time, reporting characteristics of a state, and a power of the distancing function. Full details of these can be found in the “Parameter Table” of MC-19 for each state (for example here for California).

Comparing to the ground truth

The models provide projections for the number of infected people, as well as the numbers of hospitalised patients and those requiring intensive care. However, these numbers are hard to measure in reality. In theory hospitalizations should be easy to collect, but this is not true for all geographies. We therefore settled on only comparing the number of reported daily fatalities. We use the number of fatalities as published by CovidTracking.

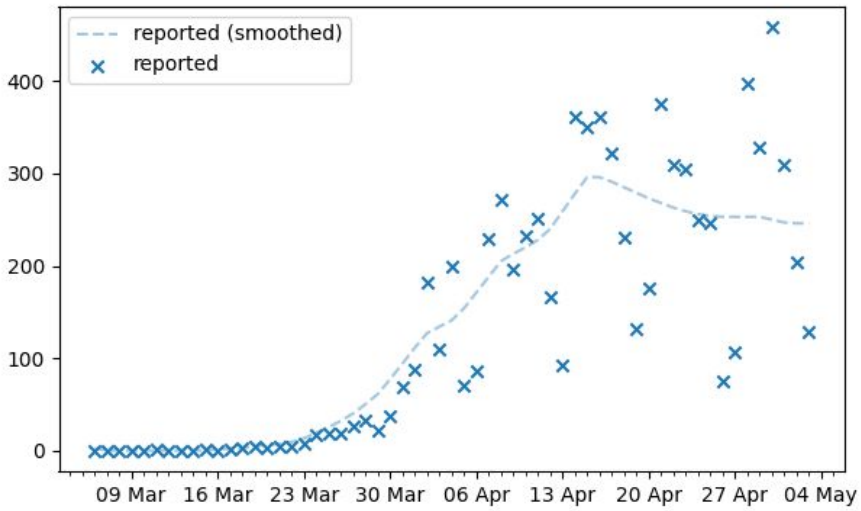
To have meaningful data to compare against, we decided to run the projections on six American states where the epidemic is fairly far advanced: California, Illinois, Massachusetts, Michigan, New Jersey, and New York.

Processing

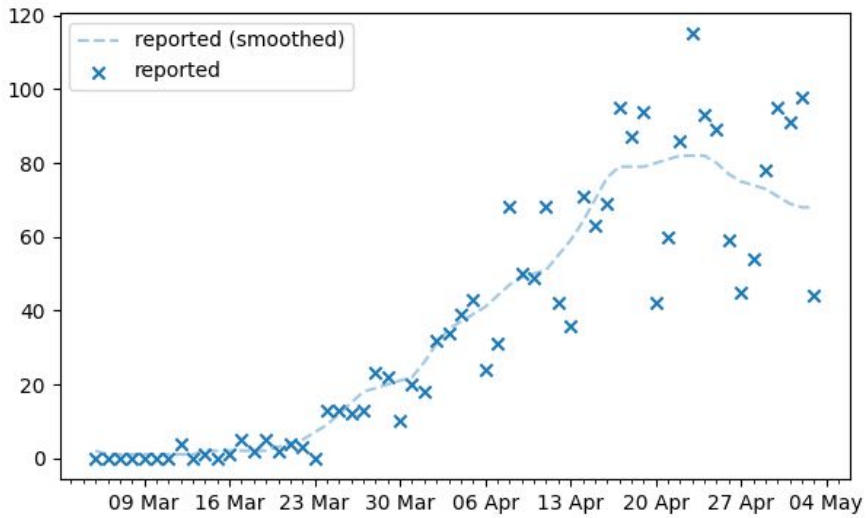
We now summarize how the data was processed:

- We interpret fatalities data as a noisy source where cases may be reported one or (less likely) more days late, resulting in the often observed zig-zag pattern. We’re applying a simple low-pass Gaussian filter to achieve the necessary smoothing.

As an example, take the data from New Jersey: it appears all over the place, but by applying the smoothing we can see it looks like the curve started bending in mid-April:



While in California the data also looks noisy at first glance, it is actually a pattern of alternating days with many and few fatalities, and smoothing removes that noise:



- We score a model prediction for a single day using a Poisson log-likelihood score, which is a standard way of comparing projections to ground truth data. The total score consists of the sum of the scores on single days.

- We do not want to “punish a model twice” for getting a single point in the outbreak wrong. For example, if in the beginning, the model is too pessimistic (predicting the initial growth x2), it may still model the effect of the coming intervention correctly. Hence,

instead of scoring the raw model predictions, for each day, we score the prediction that the model *would have given* knowing the current extent of the outbreak. So if the outbreak (due to bad modelling at the beginning) is only half the size that the model reported for the last week, then today it makes sense to scale its prediction downwards by ½.

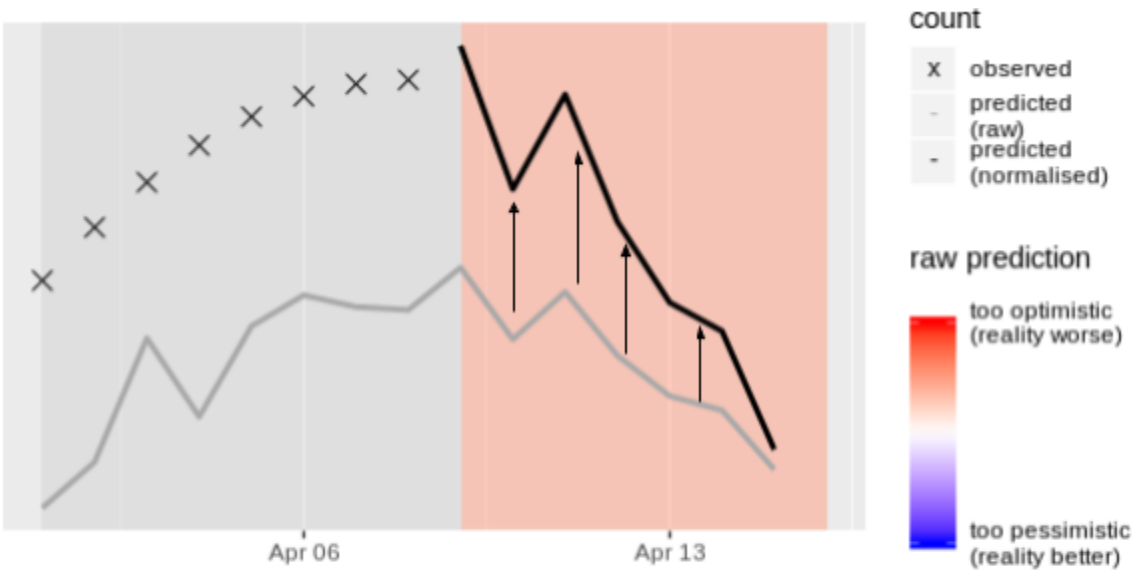
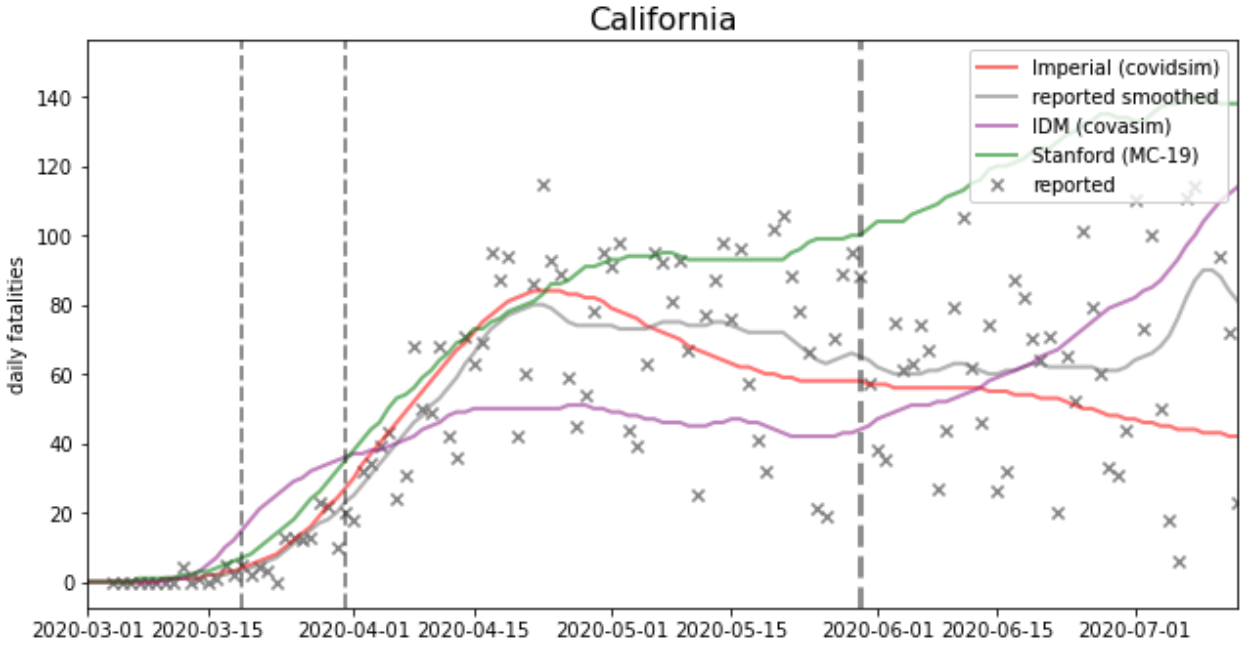


Fig 2: Dynamic scaling (schematic example). Having been too optimistic last week, the model's predictions are upscaled for the second week. The model's score for the first week of April suffers, but not for the second week.(In practice, the window is rolling and shorter.)

We call this approach "dynamic scaling" and implement it in the `dynamic_scaling` class from our codebase. To our knowledge, dynamic scaling is not a standard technique, but it matches well with a visual comparison of the graphs.

Fitting the data with knowledge till the end of May

Here are the results of running all three models on data up to May 30 on California:



The thick dashed vertical is the day of the calibration - everything to the left of that was used to make the predictions that appear to the right. The thin dashed verticals are the start of intervention policies announced by the state government, including the “shelter in place” order. The grey line is the smoothed daily death data - as we can see, probably as a result of the implemented policies, California began to “bend the curve” in late April. The outbreak has been in slow decline since.

Imperial CovidSim is shown in red. In the first part of the epidemic it fit fairly well. More recently, it has been undershooting - this is probably because we did not specify any change in the interventions, while on the ground the behavior of the population has relaxed and the state has been opening up.

Overall, IDM Covasim had some initial overshoot but then predicted a lower number of fatalities than Imperial Covidsim, before anticipating the recent rise in cases correctly. MC-19 is also predicting a continued rise of the epidemic. Note that because MC-19 uses Google mobility data, it does actually have evidence that the interventions have been easing up, unlike the other two models.

With the above data in hand, we can compare the models both before and after calibration:

UP TO calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=4119)	+0.51% (+21)	-23.57% (-971)	+26.73% (+1101)
average difference to daily deaths	0.19%	20.05%	8.04%
scoring loss	0.0338	0.2216	0.0725

FROM calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=2984)	-22.92% (-684)	+6.87% (+205)	+86.49% (+2581)
average difference to daily deaths	8.95%	5.8%	71.61%

Let's walk through each of the rows in detail:

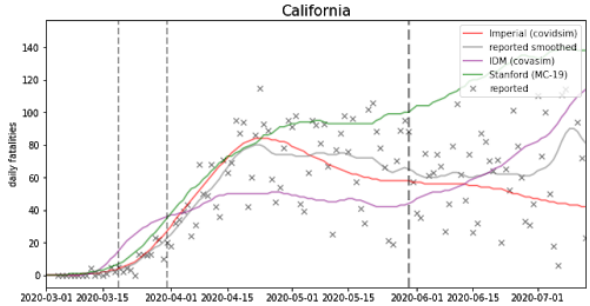
- The first row of each table tells us how closely the model predicted the cumulative number of deaths over the relevant period.
- The second row shows how far the daily predictions are off in relative terms

In summary, we can see that prior to the calibration date, all three models fit the California data really well. After the calibration date, IDM and Imperial CovidSim do a good job predicting the future, with MC-19 being less predictive, but still well within the bounds of credibility.

A full set of results is shown on the next page. Note that for the more advanced epidemics, especially in New York, all models have a really good fit.

2020-05-30

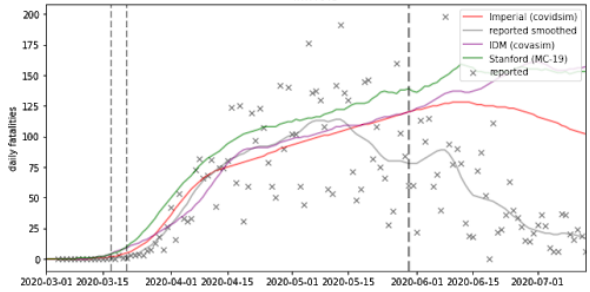
California



UP TO calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=4119)	+0.51% (+21)	-23.57% (-971)	+26.73% (+1101)
average difference to daily deaths	0.19%	20.05%	8.04%
scoring loss	0.0338	0.2216	0.0725

FROM calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=2984)	-22.92% (-684)	+6.87% (+205)	+86.49% (+2581)
average difference to daily deaths	8.95%	5.8%	71.61%

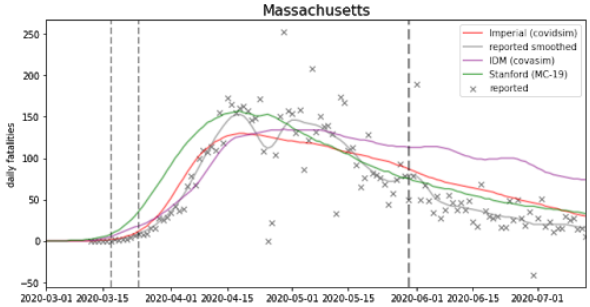
Illinois



UP TO calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=5313)	+4.27% (+227)	+5.21% (+277)	+27.42% (+1457)
average difference to daily deaths	3.44%	3.59%	15.99%
scoring loss	0.1199	0.1952	0.179

FROM calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=2159)	+149.75% (+3233)	+201.34% (+4347)	+214.96% (+4641)
average difference to daily deaths	194.42%	288.29%	294.19%

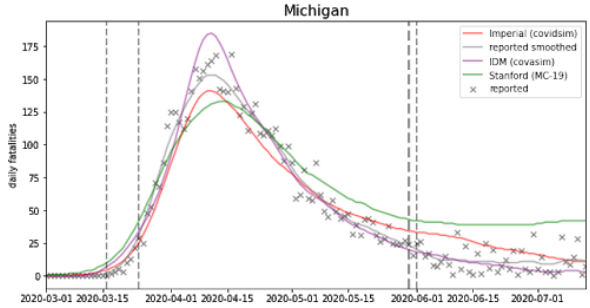
Massachusetts



UP TO calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=6802)	-0.18% (-12)	+4.97% (+338)	+16.79% (+1142)
average difference to daily deaths	5.08%	7.89%	35.5%
scoring loss	0.3447	0.5259	0.8479

FROM calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=1606)	+59.28% (+952)	+171.86% (+2760)	+42.34% (+680)
average difference to daily deaths	38.31%	175.95%	24.96%

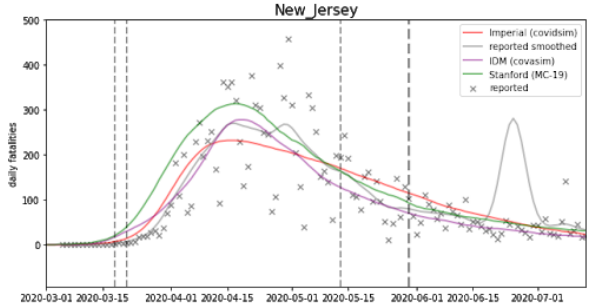
Michigan



UP TO calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=5799)	-6.38% (-370)	+3.14% (+182)	+5.12% (+297)
average difference to daily deaths	1.55%	1.4%	6.46%
scoring loss	0.0919	0.1974	0.1881

FROM calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=547)	+83.00% (+454)	-27.06% (-148)	+231.26% (+1265)
average difference to daily deaths	19.08%	0.0%	147.05%

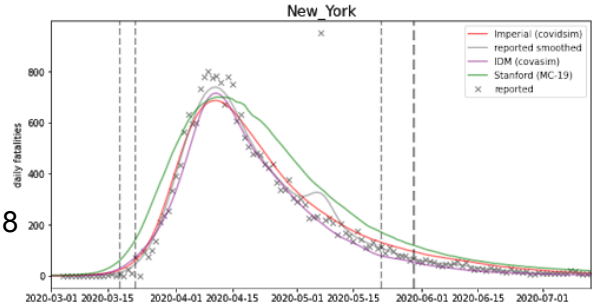
New_Jersey



UP TO calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=11581)	+0.79% (+91)	-5.13% (-594)	+22.45% (+2600)
average difference to daily deaths	11.47%	18.94%	49.21%
scoring loss	0.3492	0.5037	0.6313

FROM calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=4058)	-32.21% (-1307)	-57.74% (-2343)	-39.35% (-1597)
average difference to daily deaths	25.23%	28.88%	17.12%

New_York



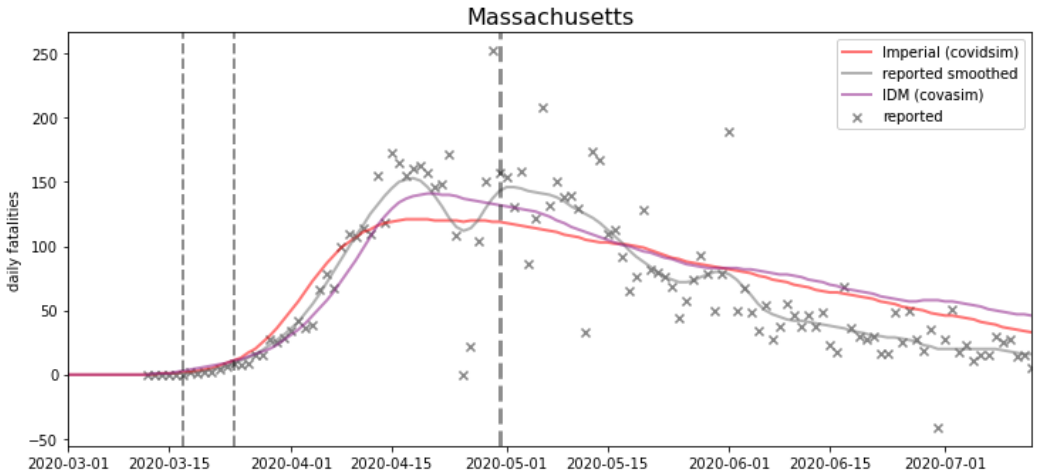
UP TO calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=23834)	+1.36% (+325)	-9.02% (-2151)	+23.76% (+5664)
average difference to daily deaths	6.14%	11.0%	59.71%
scoring loss	0.6001	0.8055	1.8752

FROM calibration date	Imperial	IDM	MC-19
cumulative deaths delta (truth=1221)	+53.40% (+652)	-35.95% (-439)	+95.58% (+1167)
average difference to daily deaths	16.43%	4.15%	54.83%

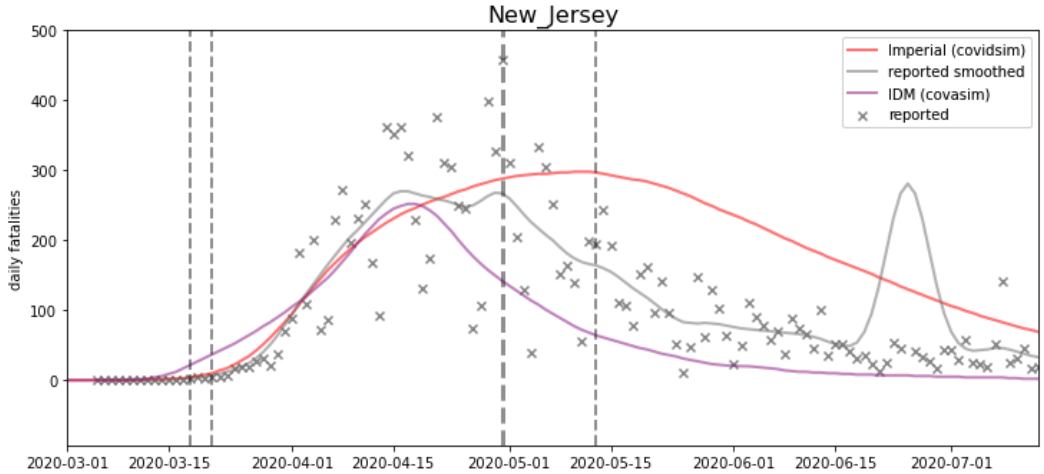
Projecting from the end of April to the present

It's encouraging that all three models gave credible predictions for two weeks in June, with data up to May 30. How about going back further, and seeing how good the predictions were on April 30? So much less was known then, and many states had only recently started "stay at home" orders. We re-ran the above experiment with Imperial CovidSim and IDM Covasim, this time calibrating on April 30 rather than May 30. MC-19 is not included, because its architecture makes such "back testing" a little awkward - by default it always downloads the latest data.

To start with an instructive example, here are the results for Massachusetts:



As we can see, the two models are very close, both correctly anticipating the future trajectory of the epidemic. The projections are quite good for other states as well, but one outlier is New Jersey:



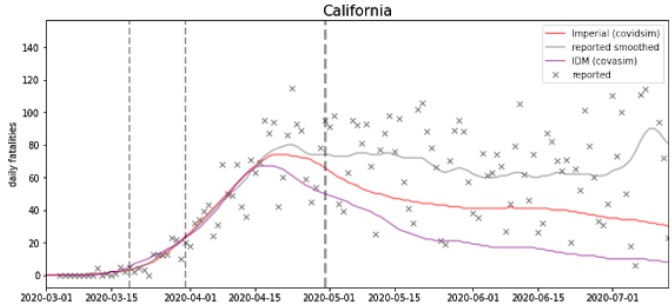
It appears that the fatalities declined rather sharply directly after the data collection ended, which wrong-footed CovidSim. The spike in the smoothed death data is a reporting anomaly, but as it appeared after the calibration it had no effect on the model predictions.

A full set of results for all six states is shown on the next page. As remarked above, the only real outlier is New Jersey, where the poor data quality threw both models off. Illinois also has a fairly poor prediction by Imperial being very substantially off, but overall we can conclude that with good input data, the models do a fair job of predicting six weeks in advance.

Generally, the best fits from the end of April and from the end of May agree on many details. For instance, the CovidSim fit for the basic reproduction number in absence of interventions is each time lowest in California (2.9 for April, 2.6 for May) and highest in New York, New Jersey and Massachusetts (3.75 for both calibrations). Covasim conveys base infectivity with a different parameter called beta. There, too, California is picked out as having the smallest base beta each time, while New York has the highest beta (followed by New Jersey, and then Michigan).

2020-04-30

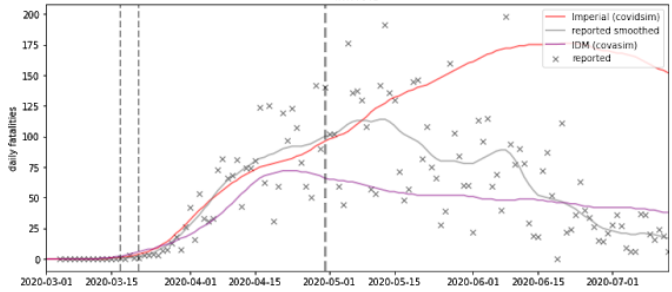
California



UP TO calibration date	Imperial	IDM
cumulative deaths delta (truth=1984)	-2.22% (-44)	-10.94% (-217)
average difference to daily deaths	0.0%	2.26%
scoring loss	0.0193	0.1166

FROM calibration date	Imperial	IDM
cumulative deaths delta (truth=5128)	-37.64% (-1930)	-68.84% (-3530)
average difference to daily deaths	22.0%	53.43%

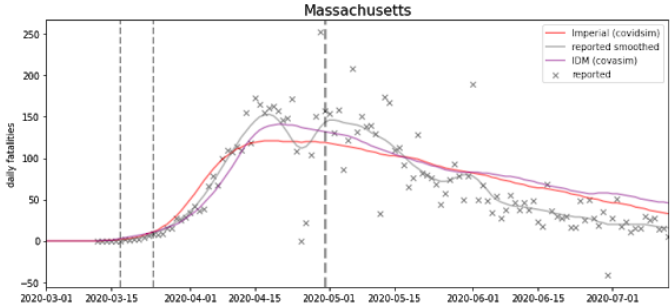
Illinois



UP TO calibration date	Imperial	IDM
cumulative deaths delta (truth=2337)	-4.36% (-102)	-23.53% (-550)
average difference to daily deaths	0.0%	6.03%
scoring loss	0.0403	0.1721

FROM calibration date	Imperial	IDM
cumulative deaths delta (truth=5157)	+121.19% (+6250)	-27.52% (-1419)
average difference to daily deaths	210.12%	27.47%

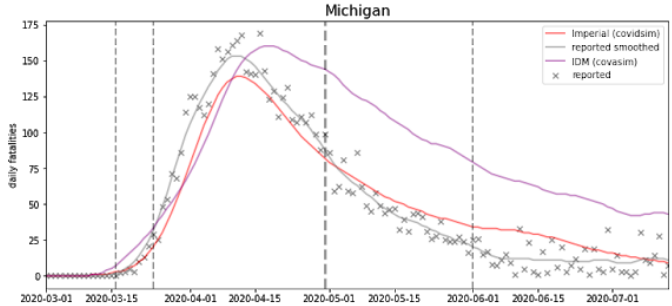
Massachusetts



UP TO calibration date	Imperial	IDM
cumulative deaths delta (truth=3536)	-3.22% (-114)	-4.16% (-147)
average difference to daily deaths	3.85%	2.44%
scoring loss	0.3774	0.3489

FROM calibration date	Imperial	IDM
cumulative deaths delta (truth=4935)	+15.70% (+775)	+24.42% (+1205)
average difference to daily deaths	29.44%	44.05%

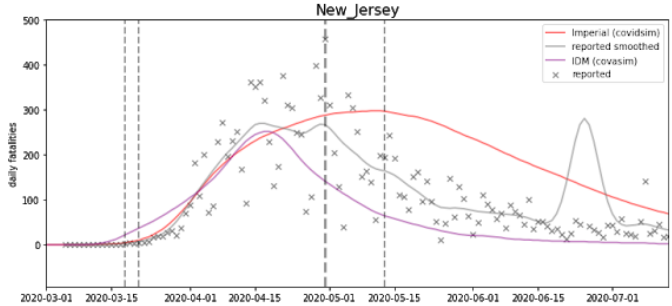
Michigan



UP TO calibration date	Imperial	IDM
cumulative deaths delta (truth=4406)	-13.69% (-603)	+5.31% (+234)
average difference to daily deaths	3.68%	9.81%
scoring loss	0.1123	0.4875

FROM calibration date	Imperial	IDM
cumulative deaths delta (truth=2005)	+34.91% (+700)	+195.16% (+3913)
average difference to daily deaths	16.99%	211.12%

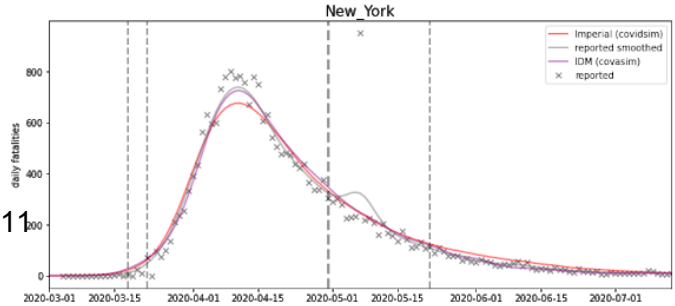
New_Jersey



UP TO calibration date	Imperial	IDM
cumulative deaths delta (truth=7079)	-0.88% (-62)	-7.54% (-534)
average difference to daily deaths	1.88%	46.05%
scoring loss	0.1905	1.231

FROM calibration date	Imperial	IDM
cumulative deaths delta (truth=8747)	+74.88% (+6550)	-71.83% (-6283)
average difference to daily deaths	108.72%	62.54%

New_York



UP TO calibration date	Imperial	IDM
cumulative deaths delta (truth=18255)	-0.28% (-52)	+1.13% (+207)
average difference to daily deaths	2.93%	5.31%
scoring loss	0.3342	0.3669

FROM calibration date	Imperial	IDM
cumulative deaths delta (truth=7058)	+5.10% (+360)	-5.14% (-363)
average difference to daily deaths	9.18%	1.84%

Conclusion

We are now ready to answer the questions we posed at the outset:

- *If we allow the models access to everything we knew on May 30, how well do they fit the data from the past?*
  All three models fit the past data remarkably well.
- *If we only allow the models access to data up until the end of April, how well would their predictions have held up until mid July?*
  The predictions turn out to be very good for four out of six states, and for two the predictions are less accurate.
- *Are the models broadly in agreement over the future, or are the projections wildly different?*
  They do give similar results, and there are no wild variations.

These answers confirm the value and importance of Covid-19 models, even for people (like most of the authors on this note) with no training in epidemiology. The projections give a good indication of what is likely to happen at least for six weeks in advance, assuming that current policies regarding social distancing are kept in place. With more data we may find that the projections can be reliable even further into the future.

The importance of these models, however, is that they are not just for predicting the outcome when we stay the course with current policies: they allow us to experiment with scenarios that have not yet happened. Anyone can freely try out new mitigation strategies, and examine their effects far into the future. Society at large should carefully consider such projections under different hypothetical scenarios. The models had it right for the past ten weeks, so it's quite possible they will be right for the next six months.

That implies non-experts should be able to run a broad spectrum of models for themselves, and fit them to new data as it comes in. Experts examine the differences, and use these comparisons to further refine their models and make the combined projections yet better. Policy makers can share projections with the general public, to have an informed debate weighing health consequences of a given policy against its economic pain.

To enable such self-service experimentation with models by all, we have begun development of the Covid Modeling UI. This is a single UI, where the same scenario of interventions can be run through multiple models. With every model added to the UI, the combined value of the models increases, for experts, for policy makers, and for the public at large. Anyone will be able to run experiments like that reported here, for any number of geographies.

For ourselves, this exercise has proved the need for the community to pull together, and further develop the Covid Modeling UI. We hope you'll join us!

Acknowledgements

We'd like to thank Aditya Sharad for his help in preparing the code and data. We're also grateful to Marko Iskander and Matt Gretton-Dann for their help in running the experiments. Finally we'd like to thank all the other contributors on the Covid Modeling UI, in particular Greg Orzell and Andrew Eisenberg.