Covidestim model updates 2-10-2021

This model update includes two revisions primarily focused on strengthening estimates of the fraction of individuals ever infected. There is also a small revision to the approach used to model time-changes in case ascertainment probabilities. Finally, we have changed the way we are fitting the model.

- Location-specific infection fatality rates: the earlier version of the model included a prior distribution for the infection fatality rate (IFR) centered at 0.65%, based on a value provided by the CDC's COVID-19 Pandemic Planning Scenarios in early September. In the revised version of the model, we estimate state- and county-specific IFRs to account for inter-state differences in age distribution, and reconcile our estimates with reported seroprevalence survey data. To create state-specific IFRs, we calculate an age-weighted IFR for each state using the age distribution of deaths in that state (https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-S/9bhg-hcku), and age-stratified IFRs (https://www.nature.com/articles/s41586-020-2918-0). To create county-specific IFRs, we used local-area estimates of the prevalence of risk conditions for severe COVID-19 (https://www.cdc.gov/mmwr/volumes/69/wr/mm6929a 1.htm) to adjust state-specific IFRs by the prevalence of these conditions in each county relative to the state.

- R_t dependent on seroprevalence: the earlier version of the model did not consider immunity conferred by prior COVID-19 exposure, and thus was capable of producing seroprevalence estimates greater than 100%. In the revised version of the model, we included an additional term in the formula for R_t, whereby the original flexible spline is for Rt is adjusted to account for the fraction of the population previously-infected at any given timepoint. This provides a more realistic model for R_t in jurisdictions with higher cumulative disease burden.

- Constraint on spline for case ascertainment in symptomatic, non-severe cases: in the earlier version of the model, we used a cubic b-spline for the logit of the probability of detection for symptomatic non-severe cases. In the revised version of this model, we constrained the first spline parameter to assume the slope is zero at the start of the time series, avoiding implausible trends at a point where there are limited data to inform the model.

- Updates to model fitting: previously, all state and county estimates were produced using a Hamiltonian Monte Carlo (HMC) sampling algorithm. Any geography that could not be fit on a given day was excluded from the model results posted to covidestim.org. We found that this approach has become too computationally intensive and unreliable, making it difficult to successfully use for all geographies each day. In the revised fitting approach, all counties are fit using an optimization routine that reports the *maximum a posteriori* estimate. These estimates represent the mode of the posterior distribution of the model parameters, and **do not have associated credible intervals**. For states, we still use the HMC sampling approach to report point estimates (the median value for each quantity of interest) and equal-tailed 95% credible intervals. If the HMC algorithm does not converge for a state on a

given day, the optimization algorithm is used as a fallback. The .csv estimates we distribute will contain "NA" in the "*.hi", "*.lo" columns for geographies that were optimized that day.

Covidestim model updates 9-28-2020

On 9-28-2020 we introduced an updated version of the model that generates estimates for covidestim. The goals of this update were to make the model faster and more stable, and reflect changes in COVID-19 science and epidemiology since our initial release. We introduced these changes at the same time as we began reporting county-level outcomes, and a number of these changes were needed so that we could successfully estimate outcomes for all these new jurisdictions. The list below describes each of the major changes included in this update. We expect that there will be periodic model updates in the future, and we will document these changes when introduced.

- Change in starting point of model: the initial version of our model specified a flexible function for the number of new infections each day, operationalized as a geometric random walk for the daily change in the number of new SARS-CoV-2 infections. In the revised version of the model this has been replaced by a penalized cubic b-spline for the log of R_t, the effective reproductive number. By replacing the random walk with a spline we have reduced the number of parameters required in the model, which is important for improving run-times as the time-series gets longer. The spline knots are evenly spaced every 4 days, retaining substantial flexibility.

- New approach for estimating R_t: the initial version of our model calculated R_t from the estimated time-series of symptomatic cases, using functions provided in the EpiEstim package (Thompson et al, Epidemics 2019). In the revised version of the model we simulate R_t directly, as described above, and so no longer need to back-calculate R_t from other results.

- More flexible approach to case ascertainment for symptomatic, non-severe cases: the initial version of the model assumed a simple functional relationship for the fraction of symptomatic, non-severe cases that were detected, which was based on time-series data on test-positivity. While this approach worked for most jurisdictions, we found instances where it did not adequately capture the relationship between cases and deaths, particularly as the epidemic progressed. In the new version of the model we have replaced this with a cubic b-spline for the logit of the probability of detection for symptomatic non-severe cases, with knots evenly spaced every 21 days. This probability is bounded between zero and the probability of diagnosis for severe cases, under the assumption that the probability of diagnosis is always higher for severe vs. non-severe cases.

- Allowance for diagnosis of asymptomatic cases: the initial version of the model assumed that diagnosis was only possible for symptomatic cases. In the revised version of the model we have relaxed this assumption to allow for diagnosis of asymptomatic cases. This probability is assumed to be a fraction of the probability of diagnosis of symptomatic, non-

severe cases. This new parameter is operationalized with a Beta(2,18) prior, with mean value of 0.1. While the probability of diagnosis for asymptomatic cases is likely low, these individuals will contribute to case counts (such as through testing of contacts of diagnosed cases), and this revision will make the model more robust in situations where a high fraction of cases are detected.

- Allowance for imported infections: the initial version of the model assumed that all SARS-CoV-2 infections were due to transmission within the modelled jurisdiction. In the revised version of the model we relax this assumption to allow for imported cases, which are given a half-Normal prior distribution equivalent to 0.5 imported infections per day. This addition has no effect in established epidemics, but produced more credible R_t estimates for some early epidemics.

- Assumption about epidemiology prior to the start of the data: the initial version of the model made no assumption about the trajectory of reported cases and deaths in the period preceding the first reported COVID-19 case. While this worked for most epidemics, it produced implausible results in a small number of counties, where observed data appeared to show a declining epidemic at the start of the time series. In the revised model we have added a penalty function to limit the expected number reported cases and deaths arising in the model burn-in period. This enforces the assumption that there were no COVID-19 diagnoses prior to those included in the reported data.

- Revised prior distributions for natural history parameters: in the initial version of the model, the prior distributions for natural history parameters were based on our review of the literature, favoring systematic reviews, local data, and stronger study designs where possible. In the revised version of the model we have revised prior distributions for some parameters to follow the synthesized evidence reported in the CDC's COVID-19 Pandemic Planning Scenarios (https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html).

- New data source: for state-level estimation we have been using data from the COVID-Tracking Project. County-level data are not available from this source, and so for county-level estimates we use data from the Johns Hopkins University COVID-19 Data Repository (https://github.com/CSSEGISandData/COVID-19).

- More efficient code: in the revised model we made multiple small edits to reduce computation time. These included (i) replacing loops with vector operations, (ii) truncating all delay distributions at 60 days, and (iii) setting the shape and scale parameters for reporting delay distributions (time from diagnosis or death to when this event is reported) at fixed values.