Covidestim major model updates July 2022

This major model revision includes changes required following the emergence of the Omicron variant late 2021.

- Hospitalizations: With the Omicron variant, the IFR has decreased relative to previous variants (fewer deaths per 100K infections). This caused problems for fitting the model – with relatively few deaths from many modelled locations, it became difficult to produce precise estimates of COVID-19 burden. The revised model is estimated using data on COVID-19 hospitalizations, instead of COVID-19 deaths. The hospitalizations data are extracted from healthdata.gov. This data set contains weekly data on hospital admissions for each facility in the US. We aggregate these data, first the the Health Service Area level, then to the county level and finally to the state level (detailed description here). 'Confirmed Hospital Admissions (Admissions Confirmed Adults)' is the variable of interest for the covidestim model. This variable is censored for observations of 1-3 hospitalizations per facility. We use the lower bound of the aggregate variables. *i.e.,* if a facility reports censored data in a week, we assume 1 new admission has occurred in that facility.
- Weekly time-step: The hospitalizations data are reported on a weekly basis. Therefore, we have adapted the covidestim model to use a weekly timestep. This change also reduces the computational time required to fit the model. The daily case data are summed to match the weeks of the hospitalizations. All time dependent prior distributions and delay distributions in the model have been converted to match the weekly timescale by multiplying the scale parameter by 7.
- First date of model: the estimation in this update starts on December 1, 2021, to align with the beginning of the Omicron wave. This allows the model to estimate transition probabilities specific to the variant distribution after December 1, 2021.
- Reinfection. New in the covidestim model is the possibility of reinfection. In the previous version of the model, the new infections on each day were subtracted from the susceptible population; implying that an individual could not be reinfected. Currently, three measures of protection are incorporated in the model.
    - Immunity from infection: In this new implementation, we do not assume that an infection prevents reinfection. Rather, we assume that following an infection, immunity wanes progressively as described below ('Waning of immunity').

    - Historic immunity: each covidestim run is initialized with an estimate of the population protection against infection with the Omicron variant on December 1, 2021. This estimate is obtained from the calculations in this pre-print.
    - Booster immunity: the data for any boosters and first vaccinations administered after December 1, 2021, are extracted from the CDC data dashboard for counties and for states. For states where the cumulative boosters or first vaccinations data exceeds the population size (New Hampshire and Rhode Island in the latest inspection), we impute the average of the booster/first vaccinations data from neighboring states. If there are no neighbors, the average of all valid states is used. For counties where the cumulative booster or first vaccination data exceeds the population size, or where the booster data is missing, the state average is imputed. We assume the booster starts at an efficacy of 80% and then wanes exponentially.
- Waning of immunity. All three measures of immunity, have the same associated exponential waning curve associated with them with a median of 120 days (half of the people with an infection will be eligible for reinfection after 120 days).

The county and state detail pages, as well as the landing/overview page, present in addition to the original outcomes, one new outcome variable (cumulative first infections) and a revised outcome (cumulative total infections). The model results are shown from December 1, 2021 onwards. Cumulative counts include the last available cumulative estimates for December 1, 2021, using the previous version of the model (pre-omicron).

- Cumulative first infections (labeled infections_premiere in the model). This is an estimate of the cumulative first infections. This estimate starts at the last estimate from our previous model, that is the cumulative infections on December 1, 2021.
  This measure is calculated by assuming the new infections are proportionally distributed over those never infected and those who have been infected and are susceptible for reinfection.
- Cumulative total infections (labeled infections_cumulative in the model). This is an estimate of the cumulative total infections. This estimate can be larger than the population size, which does not imply that the entire population has been infected, since reinfection is possible.

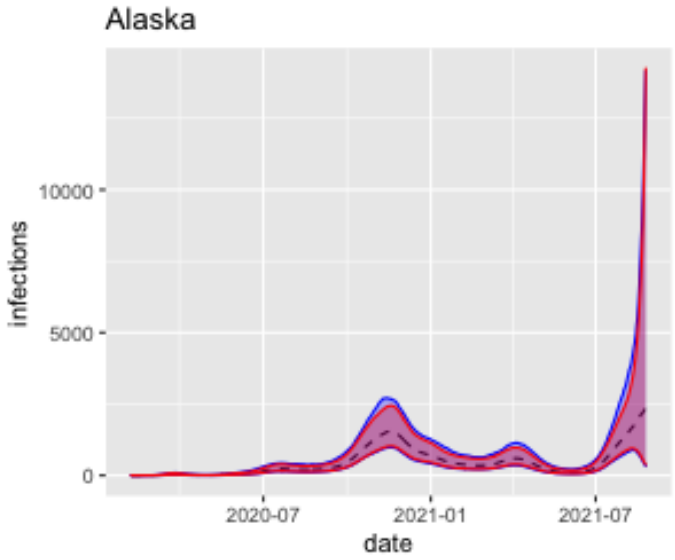
Covidestim model update 8-26-2021

In this update, we include a method that allows us to present uncertainty bounds for all states.

To generate state-level results we have been using a Bayesian sampling method, which produces a large set of epidemic trajectories that are consistent with the available data. We use this set of trajectories to produce a best estimate (the median), and an uncertainty interval (the 2.5% and 97.5% percentiles) for each outcome on any given date. The time limit on our daily runs for state-level estimates is currently 10 hours. If a state is not finished sampling within that time, we produce estimates using an optimization algorithm (see the last bullet on the 2-10-2021 update), which produces a point estimate but no uncertainty intervals. In this update we have introduced a method for computing uncertainty intervals for states run using the optimization routine.

To create these new intervals, a spline regression is estimated using the upper and lower credible bounds for those states that did sample. This creates a prediction equation for the upper and lower intervals, based on the point estimate. This equation is used to compute intervals for the states run using the optimizer.

Illustration: The figure below shows results for Alaska on August 26 with the two approaches for generating intervals. The results of the sampling approach are shown in blue, and the new prediction equation are shown in red. As this example illustrates, both approaches produce similar results.



Covidestim model update 8-16-2021

Major update: Including vaccination data in the model

In this model revision, we have combined multiple data sources on vaccination coverage and include these in the model. The steps are outlined below. As a result of these changes, estimates of infection rates after February 2021 may be revised higher, given that vaccination changes the ratio between infections and COVID-19 deaths.

- Vaccination data at the county level are pulled daily from https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh. The number of fully vaccinated people from the 18+; 65+ and overall population are reported at the county level. We redistribute these counts over more fine grained age groups (age 0-11;12-15;16-17;18-25;25-39;40-49;50-64;65-74;75+) using the national level data (from https://covid.cdc.gov/covid-data-tracker/#vaccinations_vacc-total-admin-rate-total) and the county census data of those age groups (https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html#par_textimage_70769902). The state vaccination coverage is computed as the sum of the number of vaccinated individuals of all counties within that state.
- Mortality adjustment: Vaccination lowers the COVID-19 mortality rate. To accommodate this change in mortality, we compute the relative risk (RR) of dying in each age group for those vaccinated and for those unvaccinated for each day since vaccinations started. This is computed using the age stratified deaths data from December 12, 2020 (before vaccinations started) and the age-specific Covid-19 IFR, https://pubmed.ncbi.nlm.nih.gov/33289900/. We assume that vaccination prevents 87.5% of infections amongst the vaccinated population, and that it prevents 96% of deaths amongst the vaccinated population (this is included in the model as a 68% reduction of deaths amongst those vaccinated and infected), relative to the baseline of no-vaccinations (see https://github.com/covidestim/vaccineAdjust/blob/master/R/run.R).
- IFR adjustment for vaccination coverage: we assume that the probabilities of becoming symptomatic if infected, severe if symptomatic and dying if severe are all reduced proportionally to achieve the 68% reduction in mortality among infected individuals, while allowing for uncertainty in these values.

Additional updates

- Switch from Odds Ratio to Risk Ratio. In the previous version of the model, we adjusted the probability of dying if severe using an odds ratio (OR). The current version of the model adjusts these probabilities using a risk ratio (RR).
- County level death data: Some counties have stopped reporting deaths data (since approximately June 2021). The corresponding data reports '0' counted deaths for those dates, which created unrealistic estimates. For counties that have discontinued reporting of COVID-19 deaths, we now exclude deaths data in the model after the last date of reporting.
- Revised prior on probability of diagnosis if severe: The prior for the probability to die if infected has been altered from a Beta(5, 2) to a Beta(20, 5). This increases the mean probability from .71 to .8, and reduces the variance of the prior distribution.

Configuration changes: We made some small tweaks to the way we run the analyses to improve performance.

Covidestim model updates 4-9-2021

This model update includes a revision to the approach used to estimate IFRs early in the epidemic (early 2020). We have also changed the source of state-level data used in the model.

- Time trends in infection fatality rates in early 2020: there is reason to believe that mortality risks among infected individuals has fallen over time, with earlier identification of disease, and improved clinical management of severe disease. For this reason we have revised IFR assumptions for early 2020 to allow for higher IFRs at this stage of the epidemic. We parameterized this as a rate ratio applied to the probability of death among individuals with severe disease. This rate ratio declined from 2.34 [1.69, 3.19] at the beginning of 2020 to a value of 1.00 by the middle of the year, based on the ratio of reported COVID-19 deaths to hospitalizations prior to May 1 2020 compared to the subsequent 6 months (https://covidtracking.com/data). The time trend in this function was calculated as 1.0 minus the Normal Cumulative Distribution Function, centered on May 1 2020 with a standard deviation of 3 weeks.

- Source of state-level data: given that the COVID Tracking Project has stopped posting new data, we now pull data for state-level estimates from Johns

Hopkins CSSE (https://github.com/CSSEGISandData/COVID-19). This is the same source we use for country data.

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.