PMC U.S. COVID-19 Case Estimation and Forecasting Model: Report for November 25, 2024, pmc19.com/data

Michael Hoerger, PhD, MSCR, MBA, Pandemic Mitigation Collaborative (PMC)

Technical Appendix

Cite as follows or any of the weekly reports if more recent than the Technical Appendix updates: Hoerger, M. (2024, Nov 25). *PMC U.S. COVID-19 Case Estimation and Forecasting Model: Report for November 25, 2024*. Pandemic Mitigation Collaborative. http://www.pmc19.com/data

Overview

This Technical Appendix outlines the core assumptions of the dashboard and report. Expect minor updates periodically. Major updates will be flagged more prominently and noted in the weekly report. The dashboard is “participatory action research,” meaning a collaborative action-oriented project first that is designed to help people with the ongoing pandemic, and secondarily an empirical scientific project to be documented through publication and research grants. In general, we describe the methodology in a way that should be understandable to non-scientists and with the maximum level of detail that will still allow us to protect the ownership of this work. This is a free service, and we would much prefer a government entity to make their own case estimates and forecasting models with more substantial resources. Please provide feedback on Twitter so we can continue to improve our dashboard in ways that are helpful. Thank you.

Summary of the Changes from Version 2.0.3 to Version 2.0.4

Version 2.0.3 ran from November 4 through November 18. Version 2.0.4 launched on November 25. There is one statistical change to a variable in the forecasting model. This does not affect current/prior case estimates. The forecasting model uses a combination of recent data (past 4 weeks) and historical data (median level of transmission for a particular day of the year). The latter variable has been changed; instead of using the simple median, it now uses the rolling average of medians for +/- 3 months. In the historical data, this has the effect of -- by default -- assuming waves will be less spiky and valleys less deep, allowing more recent data (past 4 weeks) to push the forecasted estimates higher or lower. It’s basically a more conservative way of using the historical data. There are several approaches along these lines, such as artificially setting the beta for that variable, but we found this approach less arbitrary. The overall R^2 of the model remains .98 (shifting from .97997 to .97945), so it is basically just giving more weight to the recent data than historical data. We considered the possibility of making a switch along these lines sometime in the future, such as March 2025, in anticipation of a summer wave, as the summer wave timing is less grounded to the calendar. However, given the recent patterns of subvariant evolution, it is possible we may start to see greater differences in the winter too (slight shifts in timing of the peak). In terms of concrete examples in the forecast, when a wave is smaller than typical for that time of year, instead of seeing the forecast slowly push the wave down (as new forecasts come out each week), the forecast will already start lower. However, if a wave is bigger than usual (or even close to the historical median), you will essentially see the forecast looking like it is catching up (bigger and bigger peak as the peak nears and the recent data “catch up”). In giving more weight to recent data, the model will also become more dependent on real-time data coming in, which are prone to retroactive corrections. In weeks where the real-time data are a bit “off,” this will make forecasts a bit bouncier or variable. We will point out these nuances periodically. Expect the model to do better with waves departing from the “usual” calendar and slightly worse when a wave lines up right with the historical median.

Summary of the Changes from Version 2.0.2 to Version 2.0.3

Version 2.0.2 ran from September 30 through October 28. Version 2.0.3 launched on November 4. There are two main changes. One, we have re-added Biobot data back into the model, weighting it at 20% for current case estimation, relative to 80% for the CDC data. We had been using Biobot at 40%, discontinued after they stopped reporting without notice for 3 weeks, and are reintroducing because their reporting has been consistent since, correlated near-perfectly with CDC data of late (r=.95), and will help fill in the gaps it the CDC experiences reporting delays during the winter

surge. Two, we have added 95% confidence intervals to figure 3 (far right) on a trial basis; too many lines tend to confuse lay audiences and scientists in areas other than data science, but we are hopeful we can present these bands in ways that are useful. Updates are noted in brown text.

Summary of the Changes from Version 2.0.1 to Version 2.0.2

Version 2.0.1 ran from September 9 through September 23. Version 2.0.2 launched on September 30. In the 3rd graph on our dashboard (upper right) we have added models that show what a sizeable error in the CDC real-time reporting would do to the forecasted estimates (very minor update).

Summary of the Changes from Version 2.0.0 to Version 2.0.1

Version 2.0 ran from August 12 through September 2. Version 2.0.1 launched on September 9.

What's New?

- **Sunsetting Biobot.** We will continue to put forth the highest-quality models available. Although Biobot discontinued their public dashboard in May, they continued to provide weekly national and regional graphs publicly through August 14, reflecting data collected from approximately August 7. They have not updated their public data nor Twitter account since that time to indicate a rationale for the delay and plans going forward. Discontinuing public data access for 3 weeks without explanation during the height of surge is a serious problem in terms of organizational ethics and strategy. Pragmatically, it prevents us from gaining useful data for the modeling. Their public data continue to inform the historical picture of the pandemic, which is of vital public health significance. However, we have downgraded them from 40% weight in estimating daily cases on July 1 (with CDC accounting for the remaining 60% of the weight) in a linear trend to 0% weight on August 7 (final average data collection date, shifting to 100% CDC that day). This is an unnecessary setback for public health. Biobot did excellent – perhaps the best – wastewater surveillance data collection and analysis of the pandemic thus far, and the CDC non-renewal a year ago was and continues to appear problematic in terms of public health impact. From a modeling perspective, having two extremely high-quality surveillance data sources increases real-time estimates of accuracy substantially, so expect more frustration anytime CDC substantially revises real-time estimates of transmission, as it will no longer be tempered or counterbalanced against Biobot. Biobot will continue to perform excellent work in wastewater surveillance, and if they or another high-quality real-time data source emerges, they will be inform real-time estimates in the model. In posting on social media over the weekend that we were officially sunsetting Biobot, many expressed concern because they did not know the CDC was also collecting surveillance data or did not know these data were of high quality. The CDC collects excellent surveillance data and has a nice dashboard (https://www.cdc.gov/nwss/rv/COVID19-nationaltrend.html). The data are of excellent quality (correlating near-perfect with Biobot and IHME estimates, r = .94 to .96). Although public health institutions often draw dubious and politicized interpretations from their own data, in the present case, CDC data are highly sound, and any bias would be readily detectable. We have very high confidence in the ongoing model. Others have expressed concern that relying on the CDC too centrally leaves the model vulnerable if the CDC data are discontinued. There will always be data sets to correlate with transmission (WastewaterSCAN, Walgreens, Google Trends, behavioral indicators, and more). Lower-quality data require more data sources to

distinguish the signal from the noise. These are not necessary at the moment but would be compiled if needed. We intend to continue the PMC model through 2030.

- **Minor Updates.** We have updated information on the CDC heat map, example sources using the PMC model, and useful data sources.

Summary of the Changes from Version 1 to Version 2.0

Version 1 ran from August 2, 2023, through August 2, 2024. All reports are archived at pmc19.com/data. Version 2 launched on August 12, 2024.

What's New?

In short, the new model has substantial data quality improvements by combining multiple data sources for estimating transmission in unique ways that will hopefully increase forecasting accuracy, provide a truer representation of what has happened and is happening during the pandemic, and linkages to some statistics you will find helpful in day-to-day decision making.

Here is a deeper dive into the changes (skip to next section if desired). The new model is designed to provide a "true" picture of what has happened during the pandemic. It integrates three main data sources: the IHME true case estimation model (https://covid19.healthdata.org/united-states-of-america?view=infections-testing&tab=trend&test=infections), Biobot SARS-CoV-2 wastewater surveillance data (https://biobot.io/data/), and the current CDC NWSS SARS-CoV-2 wastewater data (https://www.cdc.gov/nwss/rv/COVID19-nationaltrend.html). IHME provided a comprehensive true case estimation model through April 1, 2023. Biobot was the CDC wastewater subcontractor through last fall and continues to do extensive non-CDC wastewater work. The CDC NWSS data are currently subcontracted with Verily, a subsidiary of Alphabet, which is the parent company of Google. Over the past year, we have seen Biobot scale back their public data and visualizations, and Verily has made steady improvements in their work with the CDC.

We previously relied solely on Biobot for forecasting and a Biobot-IHME data linkage for case estimation. It was a Biobot-heavy model. The current model is not tied strictly to any data set, but rather the PMC's best estimate of the truth, a true-case model that uses multiple data sources in the spirit of IHME's original work in this area. Essentially, we link all three data sources, which have been active over different points of the pandemic to derive a composite "PMC" indicator of true levels of transmission. The indicator is weighted based on which data sources were available and their perceived quality at each point in time. We scale this composite PMC indicator to the metric the CDC uses when helpful for comparisons with their website, and scale it with the true case estimates of the IHME otherwise, as true cases are more relevant than arbitrary wastewater metrics.

A great feature of the model is that it continues to integrate real-time data from Biobot and the CDC. From the perspective of Classical Test Theory, this is a huge advantage, as it provides a much more reliable indicator of what is currently happening with transmission. Both sources often make retroactive corrections for the most recent week's data, sometimes sizable, and pitting the two indicators against one another reduces measurement error on average, which offers vital improvements in forecasting.

What are the Biggest Improvements in the Model?

- ***Accuracy in Real-Time Data*** – In integrating two active surveillance data sources, the real-time data will be more accurate. The biggest predictor of next week's transmission levels, and the shape of how transmission is increasing or decreasing, accelerating or decelerating, is the current week's real-time data. If the real-time data are off by 5% or 10%, the big-picture take on the forecast will still be reasonable, but a more precise estimate allows for greater accuracy in estimating the height and timing of waves.
- ***Regional Statistics*** – We are already integrating some regional data. Like you, we miss the vast and high-quality regional data and visualizations Biobot provided. We are hoping to take back some of those advantages through the new model and will improve them over time.
- ***Credibility*** – Although Biobot and CDC have unique strengths and limitations, a clear strength of adding the current CDC data set is that many people prefer to defer to the credibility of the CDC. The PMC model can be characterized fairly as a "CDC-derived case estimation and forecasting model," which should lend more credence with those who are not deep enough in the weeds to evaluate the data as critically and prefer appeals to authority. We also provide some statistics that will allow you to draw more useful inferences from the CDC website.

What's the Same in the Current Model?

The analytic assumptions underlying the forecasting model remain the same. It uses regression-based techniques common across all industries, using a combination of historic data (median levels of transmission for each day of the year) and emerging data from the past four weeks to characterize how transmission is growing or shrinking. Holidays and routine patterns of behavior that map on well to a calendar are "baked in" to the historic data. "New variants" and atypical patterns of behavior are baked into the data on recent patterns of transmission. It's a top-down big picture model.

What are the Biggest Drawbacks of the New Model?

- ***Disruptions in Longitudinal Comparisons*** – You will notice some inconsistencies between the current and prior model that use additional data to form more accurate estimates, which is sometimes frustrating. A few examples. In the early pandemic, we estimated cases linking Biobot to IHME case estimates. Biobot transmission estimates were a bit "hotter" than others during that time period, the IHME estimates "cooler." Our composite model depicts each of the first 4 waves somewhat smaller, which we believe provides a better picture of the "truth" as we can estimate it, but it is annoying psychologically to re-envision what has happened. This also throws off some of the big-picture statistics; for example, as of August 12, 2024, we estimated that Americans had about 3.3 infections on average. A few months ago, we estimated nearly 3.5, so this is consistent with "cooler" picture of early-pandemic transmission. As of August 12, 2024, the CDC transmission estimates wre running much hotter than those of Biobot, leading to estimates of a larger and earlier peak in the wave than Version 1 had predicted. We would have preferred the CDC re-up with Biobot at the potential contract renewal to promote continuity in the data, but these sorts of changes in model estimation are the expected consequences of such a transition.
- ***Constantly-updating Historical Data*** – The CDC updates all of their historical estimates of transmission frequently, any time a new site comes on board, and twice annually to standardize the data longitudinally. This can sometimes create weird issues, where transmission is going up, but real-time values are lower than what was reported in real time the prior week because recent data were corrected downward. It will also throw off some of the helpful statistics we provide. These are minor nuisances, but be aware of them in case you spot something that seems strange.
- ***Documentation of Accuracy*** – We have excellent data on the accuracy of the prior model and will submit a report for publication shortly. All prior reports are publicly available. Many report quick facts on longitudinal accuracy, international comparisons, use in news articles, and references to use in peer-reviewed scientific journal articles. We cannot document the real-time accuracy of the new model yet, but know

that when using historical data, the model accounts for 98% of the variability in wastewater transmission 1-week into the future, which is 2% higher than our prior model. The vast majority of forecasting errors have been and will continue to be based on inaccuracies in the real-time data wastewater surveillance companies report, and the model changes reduce those issues. We hope you will trust our history and that the methodologic changes represent improvements.

Data Integration

The dashboard involves integrating data from IHME, Biobot, and CDC. Each has different timelines of existence and has varied in quality. Present cases are estimated using a composite indicator that gives 20% weight to Biobot data and 80% weight to CDC data, and the particular weights given to IHME, Biobot, and CDC have varied over time. The data sources have also differed in the level of lag in reporting, relative to labeled dates in their data files or graphs. The first step is to clean the data. We would not be able to do this as effectively if not knee-deep in the data the past year, as that allows immense opportunities to spot inconsistencies in how data are being reported and the lag between the publicly-noted date and the average actual date data were derived from. Many readers working with Covid data are familiar with the various peculiarities across different data sets. Much of the “real-time” data people think they are receiving are actually a week outdated, which is why forecasting is particularly important. In comparing data files, we noted that the lag phase varied marginally over time in some data sets (e.g., 7 day lag versus 2 day lag) and corrected the files accordingly. This allowed the longitudinal transmission estimates to line up closely. Then, we developed conversion multipliers to go from the metrics of one data set to another using a 10% trimmed mean. The intercorrelations among IHME, Biobot, and CDC ranged from .93 to .98 (all near perfect), indicating that they all are getting at the same thing. Correlations >.70 would be desirable, and the >.90 values indicate extremely high validity across multiple teams with different methodologies, basically that they all are getting at the same thing (construct validity). The minor discrepancies are likely reasonable given different assumptions or geographic coverage. All three data sources were converted to a single PMC transmission metric, which was then standardized to both the CDC levels as well as the IHME true case estimate. More details on the date, weights, and conversion multipliers will appear in an eventual publication; however, this should suffice to demonstrate the general methodologic approach. Overall, the intercorrelations among data sets were extremely high, near-perfect, and much more encouraging than what we would have imagined or been willing to integrate effectively.

Case Estimation

The PMC composite estimate of transmission correlates near-perfectly (r=.98) with the IHME model’s estimate of “true” cases (not merely reported or counted cases), with the intercorrelations among individual data sets all >.93. A simple multiplier is used to convert the PMC composite indicator to IHME estimated daily cases. The multiplier uses a 10% trimmed mean because accurate case estimation is more challenging at the peaks and valleys of transmission. Others may use the simple mean or median, scale the peak of one data set to another, or use regression with or without an intercept. All of these techniques are perfectly reasonable and may yield case estimates within +/-15% if very closely matched to what PMC is doing, and more like +/-30% if using much different methods. Since “true” cases are actually magnitudes higher than what much of the general public not monitoring wastewater tends to believe, these differences across modelers are quite negligible in terms of their implications for personal, organizational, and policy decision making. To put more concretely, much of the public may believe there are hundreds of infections a day, while modelers debate whether it is 0.6 or 0.8 million at a given time point. Most modeling is done privately; however, on the final page we have noted various sources for helpful dashboards, some of which also include forecasting.

Case estimation can ultimately be used to estimate the percentage of the population actively infectious. As documented in the international section of the archived reports from Version 1 of the case estimation model, the PMC estimates of the percentage of the population infectious correlated very closely with estimates from Dr. Moriarty's lab in Canada that also uses wastewater surveillance data as well as the UK's winter 2023-24 study that used testing-based surveillance. Overall, the difference in estimates were often quite negligible (fractions of percentage points) and reasonable given international variation in transmission. If the UK testing-based surveillance study was estimating statistics 10x or 100x different, that would obviously be cause for concern.

Wastewater data are extremely valuable for tracking transmission. Among individuals who are new to tracking wastewater surveillance data, a common concern is that new subvariants could lead to differences in the quantity of virus individuals excrete into the wastewater. Such concerns are reasonable when considering a particular local wastewater tracker with unknown methodology or even in the WastewaterSCAN dashboard (which we do not presently use), where the waves get unrealistically bigger and wider. However, in the CDC and Biobot data, such a critique would be a bit Dunning-Kruger as it would assume that one has a vastly more sophisticated understanding of how to standardize wastewater data than the environmental scientists and epidemiologists who are trained and experienced in that exact niche. In fact, the transmission estimates from different high-quality wastewater surveillance systems correlate near-perfectly and correspond closely with the international estimates noted above. In the past, we have noted that reported cases, as compiled by BNO, for example, tended to be about 1/25th the true case estimates, with predictable fluctuations entering and exiting waves; as reporting continues to decline, that multiplier is now much steeper, but those who are curious about the impact of subvariants could examine it several months apart, before and after a new subvariant dominates, and will likely see little change. It is something we monitor. At a more simplistic level, we simply do not see true case estimates (or percent infectious estimates) bouncing around from absurdly low to impossibly high values, and certainly not at random time points throughout the year when a new subvariant become dominant. Thus, surveillance and case estimation should be rigorous, and any outright dismissal viewed as specious, and sometimes driven by substantial financial conflicts of interest where people are making considerable advertising and subscription revenue by feeding people pleasant denialism. Dozens of publications show the importance of wastewater surveillance for estimating community transmission and other metrics, which can be found through a Google Scholar search for *covid wastewater cases* and using similar terms (e.g., Hill et al., 2023, Infectious Disease Modelling; McMahan et al., 2021, Lancet Planetary Health; Shah et al., 2022, Science of the Total Environment; Varkila et al., 2023, JAMA-NO).

Percentage Infectious

The percentage of the population infectious is directly tied to the case estimates noted in the previous section. The estimated new daily infections are divided by the size of the U.S. population on the last day IHME reported true cases on April 1, 2023 (Census estimate of 334,565,848). This indicates the percentage of the population infected *per day*. Then, we multiply that number by 7, approximately the average number of days that individuals are infectious (Hakki et al., 2022, Lancet Respiratory Medicine, "Onset and window…"). The majority of modelers we talk to also use 7 days for the infectious window, though some use slightly shorter or longer (6 to 8.3 days), or try to tie to particular subvariants. Given the landscape of viral evolution, reinfections, vaccinations, and treatments, we view the latter approach as pseudo-specific but will update from 7 days if compelling evidence emerges. Those unfamiliar with the literature are often concerned that 7 days is "too short." However, that perspective confuses the *average* infectious window with a more cautious public health approach of encouraging people to isolate longer to avoid forward transmission. While a 10- or 14-day isolation period may be a reasonable heuristic (though not as useful as 3 consecutive days of testing negative) for reducing forward transmission, many people are infectious longer, and many are infectious only very briefly, leading to the average of about 7

days. We are aware of only one modeler using a 10-day average infectious window, which would lead to a 10/7=1.43 or 43% overestimation of the percentage of the population infectious, but even that type of error is small given the magnitudes difference in what the public believes is happening with transmission as well as other differences of opinion in underlying model assumptions that may cancel out. Those who prefer a longer or shorter estimate of infectiousness can simply revise the PMC statistics accordingly, which is why the 7-day assumption is noted. For example, if one believes people are only infectious for an average of 5 days, they could reduce our estimates of the percentage of the population infectious proportionately. Wastewater-derived estimates should not be confused with non-surveillance (i.e., opt-in or seek-out) estimates of positivity from Walgreens or local hospitals that use non-random samples and might lead one to overestimate true transmission. As noted in the prior section, these wastewater surveillance estimates were extremely close to the testing-based surveillance estimates from the winter UK study.

Regional Estimates of the Percentage Infectious

We have simplified the analyses in the previous section to provide a “PMC Multiplier” to convert CDC levels to an estimate of the percentage of a population infectious. This multiplier only works for the CDC levels, not other random dashboards. In the PMC weekly report, we provide the estimates of the percentage infectious for each region (a) using the Multiplier to work directly from the CDC website, and (b) corrected using the PMC forecast to account for the lag in reporting. The latter is our best estimate, but we also wanted to include the values one would obtain if hand calculting directly from the CDC website.

The value of the Multiplier will vary marginally over time, as the CDC updates historical levels of tranmission estimates. It is typically about .330 at present. Take the value of the Multiplier and multiply it by the CDC level on the CDC website for the nation or region to get a rough estimate of the percentage of the population infectious. For example, if the most up-to-date version of the Multiplier is .326 and the CDC level is 8, just take .326 x 8 = 2.6, referencing that an estimated 2.6% of the population is infectious. These are approximate estimates based on the assumptions underlying the model. These are crude indicators. Estimates around the proportion of a population get less precise when focusing on smaller geographic entities, so take them with a grain of salt. We had previously provided such conversion calculations for Biobot data when they were offering county-level data on their dashboard, so this process is familiar to many readers. Note, the Multiplier may vary marginally from week to week as the CDC updates historical data on transmission levels, but even if using a slightly “stale” few-week-old Multiplier, it should still be in the ballpark. The ballpark is what matters because these estimates are likely magnitudes greater than what people not monitoring wastewater surveillance likely believe is happening with transmission.

Note that many people are uncomfortable dealing with decimals or percentages. It can be helpful to translate these percentages into “1 in ___ people” statistics. Just take 100 and divide it be the percentage to do so. For example, if the percentage was 4%, then 100 / 4 = 25, which would mean 1 in 25 people are infectious.

Regional graph:
https://www.cdc.gov/nwss/rv/COVID19-nationaltrend.html

State selector:
https://www.cdc.gov/nwss/rv/COVID19-statetrend.html

Regional Heat Map

The “heat” map of transmission merely uses the CDC state-level data to indicate levels of transmission, with deeper red indicating higher transmission. Gray refers to a lack of available data. Diagonal lines indicate limited data. The map is derived by taking the CDC state-level map as follows: hue +150, -75 brightness, and +100 contrast in Photoshop. We previously only changed the hue; however, the CDC switched from using 11 colors to just using 6 and condensed the level of contrast so that all states would appear what we would describe as tones of medium blue. Most coursework in geographic analysis recommends using red to indicate when something is “hot” and blue to indicate when something is “cool,” so the CDC map goes against conventional norms.

Forecasting Model

The forecasting model is very similar to Version 1 and uses the same regression-based techniques any business or industry would use to forecast key outcomes, from energy usage to pork consumption. The main difference from Version 1 is that the underlying data are different, using a composite indicator of transmission, derived from three data sets. The model itself is near identical. To summarize for a general audience, it uses a mix of historical data on transmission patterns and emerging data on transmission. The historical data include the median level of transmission on a particular day (calculated as the average of the medians for +/- 3 months) as well as the year of the pandemic (1, 2, 3…). It also uses the most recent 4 weeks of data on transmission from the CDC, with the more recent data useful for indicating current levels, changes, changes in changes, and changes in changes in changes. In non-calculus terms, the recent data get at the trajectory of transmission, whether it is picking up or slowing down. In Version 1, we captured the historical data in a few different ways over time, first using month, and then moving to half-month. The benefit of the current approach of using the median transmission for a particular day, as opposed to a longer stretch of time, is that it will help more with forecasting predictable peaks (winter surge). It would have less precision with the summer wave, which has a more variable peak. Using the rolling average of the historical medians across a range of +/- 3 months has the effect of softening the peaks and valleys. This makes good use of the historical data, but not overuse, which will improve the summer forecast where peak-timing is more variable, and help when winter waves depart from the historical norms. The strengths and limitations of a model are worth noting so people can augment their qualitative interpretations.

The model itself is quite simple, very few terms, just as described above: median transmission for the given date (based on the composite indicator), year, and the past 4 weeks of (weekly) data (presently, solely from the CDC). That’s it. Anyone with training in analytic modeling could readily derive such a model, and intuitively the inputs should make sense to most non-scientists. This description may be less useful to scientists who have limited experience with analytic modeling. If someone has a statistical background but not in analytic modeling, please feel free to post public questions on Twitter. We share as much as possible and believe we are the most transparent modeling group in the U.S., with the caveat that we need to protect our intellectual property from ethically dubious scientists who would wish to claim our work publicly, in articles, or in grant applications; we have dealt with several instances on this and other projects the past few years, and it's a pain. Please do reach out publicly on social media if you have questions not answered here.

Starting with version 2.0.3, in the 3rd graph on our dashboard (upper right) we have added models that show 50% and 95% confidence bands. The 50% confidence band shows the typical “normal” variation, and the 95% confidence band shows what would be expected approximately 95% of the time, barring a highly atypical event, so the upper and lower bounds give a sense of so-called worst and best-case scenarios. The bands are

comprised of two factors: 1) levels of error in historical data during our first year of forecasting, plus 2) additional error added in that assume values of +/- 8.33% for the accuracy of real-time reports (based on tracking errors in real-time reporting errors during year 1 of our model). The intervals are the sum of both components. The real-time reporting errors matter more for near-term forecasting, whereas the modeling errors (after real-time data are corrected) get wider the farther out the forecast. Note that with current reporting lags (typically 7 days for CDC data and 9 days for Biobot data) and the lag of preparing this report (3 days), the reporting lag is 10-12 days. This means that if you see estimates for “November 4,” this actually reflects a 10+ day forecast from the most recent data sources. It’s a bit like describing today’s expected weather when the best weather data are 10 days old. Our “1-month” forecasts are really about 40 days out, somewhat like a 40-day weather forecast. We will update the graphical presentation as needed to try to make the information more informative across audiences (lay audiences, scientists who are not data scientists, data scientists, modelers). We do not include confidence intervals in other figures to avoid making them too complex for lay audiences.

Individuals often ask whether the model includes holidays, flight data, information on new subvariants, and other considerations. No. Just the indicators noted above. That said, holidays and routine patterns of behavior that map on well to a calendar are “baked in” to the historic data. “New variants” and atypical patterns of behavior are baked into the data on recent patterns of transmission. It’s a high-efficiency top-down big picture model. Dr. Eastman has a model completely the opposite, ground-up, data heavy, and granular, which provides a similar forecast through a completely different approach. It shows high concordance, and we link to his model and other dashboards as well.

Long COVID Cases

The PMC model assumes that 5-20% of people who are infected with SARS-CoV-2 will develop long Covid from their infection. These values are based on the range reported in the scientific literature. The lowest value we have seen discussed is 1%, but may be influenced by financial conflicts of interest. The higher values are in the range of 40% and tend to be more inclusive of symptoms, including those that are less likely to cause a known functional impairment. We would suggest taking the 5% estimate for more severe cases that lead to a newly diagnosed health condition, whereas 20%+ may be for any Long COVID. This may be frustrating to readers, but unfortunately it is the norm in descriptive epidemiology, as even topics that have been studies extensively (e.g., rates of depression among people with cancer) often have a broader range of estimates than might be expected. The recent Al-Aly study of course noted a 3.5% estimate (Xie et al., 2024, NEJM, “Postacute Sequelae of SARS-CoV-2…”), which was likely a slight underestimate as based on medical record data on newly documented diagnoses. These are likely underestimates because they only include pre-specified conditions that have been diagnosed and documented in the medical record (consistent with the 5%+ value). Further, these estimates of 5-20% are likely underestimates to the extent that we are still in the “early” pandemic in 2024, and longer-term sequelae will pile on top of these estimates by 2030-40 (sleeper effects, like chronic conditions that take years or decades to become more observable). The analysis does not account for repeat instances of Long COVID, and we do not make projections surrounding recovery. Many Long COVID symptoms remit eventually, but some of the more troubling symptoms (i.e., cognitive) tend to be the more persistent, and we are still not accounting for the cumulative reinfection impact and much longer-term sequelae noted above. Our gist representation is that many people are getting new Long COVID or new phenotypes of Long COVID on top of existing cases, many people recover from various symptoms over time, useful treatments are quite limited, and the very long-term picture is likely quite bad. We hope the estimates are helpful, and we keep the estimates very simple so that one can adjust them as desired based on their own interpretation of the Long COVID literature.

Conflicts of Interest

Dr. Hoerger and his family have no conflicts of interest, including no ownership stake or investment in any pandemic-related company or product. He and his research team complete the university's annual conflict-of-interest paperwork, and he is aware of no conflicts of interest among his team. Any product recommendations, websites, promo codes, or other potential endorsements are unsolicited and will change as the evidence or available products shift over time. Dr. Hoerger intentionally avoids advertising and subscription revenue from his websites and social media accounts. Readers should be clear that many COVID minimizers generate substantial revenue from blog subscriptions and social media and website advertisements, very serious financial conflicts of interest, sometimes >$100,000 annually. Dr. Hoerger has never accepted a payment, meal, or other compensation from a pharmaceutical company, biomedical company, mask company, air purifier company, or similar organization.

Permission to Re-Use

Example Sources Using the PMC Models

- JAMA Oncology: https://jamanetwork.com/journals/jamaoncology/fullarticle/2813585
- BMC Public Health: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-023-16787-1
- TODAY: https://www.today.com/health/news/covid-wave-2024-rcna132529
- Forbes: https://www.forbes.com/sites/judystone/2023/12/01/cdc-improves-their-covid-19-reporting-with-a-new-wastewater-dashboard
- Salon: https://www.salon.com/2023/10/19/a-lapse-in-wastewater-detection-is-worrying-scientists-about-distorted-data/
- The New Republic: https://newrepublic.com/article/177849/biden-democrats-covid-pandemic-2024
- Yahoo! News: https://news.yahoo.com/us-sees-largest-covid-wave-165811252.html
- Washington Post: https://www.washingtonpost.com/health/2024/01/12/covid-surge-january-2024/
- Time: https://time.com/6554340/covid-19-surge-2024/
- Stateline: https://stateline.org/2024/01/23/wastewater-tests-show-covid-infections-surging-but-pandemic-fatigue-limits-precautions/
- PRISM: https://prismreports.org/2024/01/29/covid-surges-senate-hearing-california/
- SELF Magazine: https://www.self.com/story/cdc-new-covid-19-isolation-guidelines
- TODAY: https://www.today.com/health/coronavirus/covid-wastewater-monitoring-rcna143158
- Institute for New Economic Thinking: https://www.ineteconomics.org/perspectives/blog/from-long-covid-odds-to-lost-iq-points-ongoing-threats-you-dont-know-about
- TODAY: https://www.today.com/health/coronavirus/states-with-highest-covid-rates-2024-rcna163403
- NEWSMAX: https://www.newsmax.com/health/health-news/covid-summer-surge/2024/07/25/id/1173945/
- Yahoo! News: https://uk.news.yahoo.com/lifestyle/covid-is-surging-here-are-8-articles-im-reading-to-stay-informed-as-a-health-editor-135558875.html

- People: https://people.com/massive-covid-spikes-in-21-states-cdc-8683478
- Truthout: https://truthout.org/articles/the-us-government-has-abandoned-us-to-endless-covid-we-can-do-better/
- San Francisco Chronicle: https://www.sfchronicle.com/health/article/mask-recommendation-covid-san-francisco-19599948.php
- JAMA-NO: https://jamanetwork.com/journals/jamanetworkopen/article-abstract/2821699
- MSN: https://www.msn.com/en-gb/health/other/masking-policies-prevalent-in-top-cancer-centers-amid-winter-covid-wave/ar-BB1qZWnr
- PRISM: https://prismreports.org/2024/08/06/covid-data-tracking-disappears/
- CBS: https://www.wwltv.com/article/news/health/new-orleans-free-home-air-filters-for-cancer-patients-covid-cases-special-kit-safe/289-5d873151-7069-478a-ab03-2260cd08c22a
- NBC: https://www.wdsu.com/article/new-orleans-cancer-covid-prevention-kit/61899479
- FOX: https://x.com/michael_hoerger/status/1826479530124456205
- OBR Oncology: https://www.oncologynewscentral.com/article/controversy-over-cancer-center-masking-policies-as-covid-surge-looms
- Grants funded intramurally by Tulane University (twice) and externally by the American Cancer Society, and additional grant submission with PCORI and the UAB CCTS under review

Useful Data Sources

- CDC wastewater surveillance data:
  - https://www.cdc.gov/nwss/rv/COVID19-currentlevels.html
- Biobot wastewater surveillance data (click ‘Latest Report’ link at bottom):
  - https://biobot.io/data/
- UK testing-based surveillance data (not active during summer 2024, may come back):
  - https://www.gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports-2023-to-2024-season
- WastewaterSCAN – good early alert for local changes in transmission, but beware that the data are not well standardized longitudinally, so it can be misleading to compare current data against prior waves (not presently used in the PMC model for that reason)
  - https://data.wastewaterscan.org/
- Walgreens dashboard – This gives an estimate of the chances someone has Covid if they are feeling very sick. It’s opt-in testing, not surveillance testing, so it does not give a good sense of community rates. Also, be careful when comparing winter versus summer rates. In the summer, there is much less else circulating. Also, note that some states have very limited levels of testing, especially during lulls.
  - https://www.walgreens.com/healthcare-solutions/covid-19-index
- Dr. Moriarty’s dashboard – click #5 and #10 on the left side:
  - https://covid19resources.ca/covid-hazard-index/
- Dr. Eastman’s dashboard – long-range forecasting:
  - https://josepheastman.substack.com/
- World Health Network (WHN) forecasting model by Dr. Eastman and colleagues, with light consultation from Dr. Hoerger on the big-picture analytic approach:
  - https://whn.global/estimation-of-infections-based-on-wastewater-data-us/
- Ms. Willette’s dashboard – comprehensive data visualization:

    - https://iowacovid19tracker.org/
- IHME dashboard – Active through April 1, 2023, still useful for examining the historical landscape
    - https://covid19.healthdata.org/united-states-of-america?view=infections-testing&tab=trend&test=infections