U.S. STATE-LEVEL COVID-19 TRANSMISSION INSIGHTS FROM A MECHANISTIC MOBILITY-INCIDENCE MODEL

EDWARD W. THOMMES[1,2,3]

ZAHRA MOHAMMADI[2]

DARREN FLYNN-PRIMROSE[2,4]

SARAH SMOOK[2]

GABRIELA GOMEZ[1,5]

SANDRA S. CHAVES[1]

LAURENT COUDEVILLE[1]

ROBERTUS VAN AALST[1,6,7]

CEDRIC MAHÉ[1]

MONICA G. COJOCARU[2]

[1] *Sanofi*

[2] *University of Guelph, Guelph, Ontario, Canada*

[3] *York University, Toronto, Ontario, Canada*

[4] *McMaster University, Hamilton, Ontario, Canada*

[5] *London School of Hygiene and Tropical Medicine, London, U.K.*

Date: March 2022.

[6]*Brown University School of Public Health, Providence, Rhode Island, USA*

[7]*University of Groningen, Groningen, Netherlands*

Summary

Background. Throughout the COVID-19 pandemic, human mobility has played a central role in shaping disease transmission. In this study, we develop a mechanistic model to calculate disease incidence from commercially-available US mobility data over the course of 2020. We use it to study, at the US state level, the lag between infection and case report. We examine the evolution of per-contact transmission probability, and its dependence on mean air temperature. Finally, we evaluate the potential of the model to produce short-term incidence forecasts from mobility data.

Methods. We develop a mechanistic model that relates COVID-19 incidence to time series contact index (CCI) data collected by mobility data vendor Cuebiq. From this, we perform maximum-likelihood estimates of the transmission probability per CCI event. Finally, we retrospectively conduct forecasts from multiple dates in 2020 forward.

Findings. Across US states, we find a median lag of 19 days between transmission and case report. We find that the median transmission probability from May onward was about 20% lower than it was during March and April. We find a moderate, statistically significant negative correlation between mean state temperature and transmission probability, $r = -.57$, $N = 49$, $p = 2 \times 10^{-5}$. We conclude that for short-range forecasting, CCI data would likely have performed best overall during the first few months of the pandemic.

Interpretation. Our results are consistent with associations between colder temperatures and stronger COVID-19 burden reported in previous studies, and suggest that changes in the per-contact transmission probability play an important role. Our model displays good potential as a short-range (2 to 3 week) forecasting tool during the early stages of a future pandemic, before non-pharmaceutical interventions (NPIs) that modify per-contact transmission probability, principally face masks, come into widespread use. Hence, future development should also incorporate time series data of NPI use.

1. Introduction

As of end of early June 2022, the global COVID-19 pandemic has produced 530M recorded cases and 6.3M recorded deaths worldwide[1]. Throughout its course, the complex epidemiology of the disease has been shaped, above all, by the changes in human behavior it has elicited. Indeed, in a counterfactual world that took no measures against it, the course of the pandemic would have been simple and catastrophic; it is estimated [1][2] that about 90% of the world's population would have been infected in a single massive wave lasting roughly two months, with a death toll of about 40 million.

[1] `https://covid19.who.int/`

[2] The original report published by the Collaborating Centre for Infectious Disease Modelling and collaborators: `https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-12-global-impact-covid-19/`

The most immediate response consisted of the near-universal lock-downs which began in rapid succession around the world in spring of 2020. The publication of freely-available worldwide human mobility data by Google[3], Apple[4] and Facebook[5], and the re-purposing of business intelligence mobility data from vendors such as Cuebiq[6] and Safegraph[7], has made it possible to trace changes in mobility with high spatial and temporal resolution, and to directly observe the results of mobility-related measures enacted to counter disease transmission. Numerous studies have examined the connection between mobility and COVID-19 epidemiology. Some use statistical models to characterize associations between measures of mobility and measures of disease burden (e.g. [2], [3], [4], [5]). Others use hybrid approaches that combine statistical and mechanistic models (e.g. [6] [7] [8]), in some cases with the help of artificial intelligence (e.g. [9]). Many further examples are given in the systematic review of Zhang et al.[10].

Our approach here is almost entirely mechanistic. Changes in human mobility affect disease transmission by modifying the rate of person-to-person contacts. Most available mobility data is in the form of indices that are indirect proxies for contact rate, and which require additional work (e.g. [11]) to infer contact rate itself. Here, we use data from US mobility data provider Cuebiq, which probes person-to-person contact rate more directly (see Section 3), thus lending itself better to use in a mechanistic model. This allows us to estimate, within a proportionality constant, the per-contact transmission probability of the disease. We restrict our analysis to 2020 in order to avoid complication due to i) emergence of new variants, ii) vaccination and iii) significant accumulation of post-infection natural immunity in the population.

In 2020, prior to the availability of COVID vaccines, the evolution of per-contact transmission probability over time within a given region reflected the time-varying practice of non-pharmaceutical interventions (NPIs), notably mask-wearing and short-range social distancing (i.e. maintaining a minimum separation of e.g. 6ft among people). Note that the latter is technically encompassed within the contact rate, however the mobility data we use does not have a high enough spatial resolution to discern the degree to which distancing on the scale of a few meters is practiced.

An association of colder temperatures/climates with different measures of COVID-19 burden has been reported in multiple studies (see [12] for a 2020 systematic review; more recent studies include [13], [14] and [15]). Using our model results, we compare the transmission probability across US states during spring of 2020, at the onset of the pandemic.

Finally, we take an exploratory look at the potential of Cuebiq mobility data for short-term forecasting.

2. Methods

In this section we provide an overview of our model; the full derivation is given in Appendix A.

We assume that the disease dynamics are adequately described by a Susceptible-Infected-Recovered (SIR) compartmental model [16] of a homogeneously mixed population. The

[3] https://www.google.com/covid19/mobility/

[4] https://covid19.apple.com/mobility

[5] https://dataforgood.facebook.com/dfg/covid-19

[6] https://www.cuebiq.com/

[7] https://www.safegraph.com/

equation for the rate of change of disease prevalence is then

$$\frac{dI}{dt} = \frac{\beta(t)S(t)I(t)}{N} - \gamma I(t), \tag{1}$$

where $\beta(t)$ is the (time-dependent) rate of effective contacts, γ is the recovery rate from the infectious state, and N is the size of the total population. Effective contacts are ones which would transmit disease if they involved an infectious person. As long as a small enough proportion of the population has been infected that $S/N \approx 1$—as was the case in 2020 for COVID throughout the US—the SIR model solution for the prevalence is

$$I(t) = I_0 e^{(\beta(t)-\gamma)t}, \tag{2}$$

see e.g. [17]. The incidence, i.e. the rate of new cases, is given by the first term on the right-hand side of Equation 1 alone. Substituting, we obtain

$$inc(t) = inc_0 \left(\frac{\beta(t)}{\beta_0}\right) e^{(\beta(t)-\gamma)t} \tag{3}$$

Furthermore, can decompose $\beta(t)$ into

$$\beta(t) = P_{trans}(t) \cdot cr(t), \tag{4}$$

where $cr(t)$ is the contact rate, while $P_{trans}(t)$ is the transmission probability per contact.

Suppose we have time series data of incidence, $inc_0, inc_1, ..., inc_n$, and contact rate, $cr_0, cr_1, ..., cr_n$ at evenly-spaced times $t_0, t_1, ..., t_n$. Suppose further that this time interval is sufficiently short that we can consider P_{trans} to be approximately constant throughout. With some more manipulation (see Appendix A), we obtain an expression for the incidence at time t_n in terms of the contacts occurring between times t_0 and t_n:

$$(\ln inc_n - \ln inc_0) = (\ln cr_n - \ln cr_0) + P_{trans} \sum_{i=1}^{n} cr_i - \gamma(t_n - t_0) \tag{5}$$

In reality, reporting delays and the incubation and latent periods of the disease will together impose a distribution of delays between the time that transmission occurs and the time that the resulting cases are captured by surveillance. We can account for this by replacing the time series of cr_i with an appropriately lagged version (see Appendix A).

3. Data

For incidence, we use the US COVID-19 surveillance data compiled by the New York Times, available at `https://github.com/nytimes/covid-19-data`, aggregated at the state level.

We obtain contact data from Cuebiq, a vendor of US mobility data sourced from mobile phone users who have opted into sharing location data through a California Consumer Privacy Act (CCPA) compliant process, hereafter referred to as Cuebiq users. Several previous studies have assessed the representativeness of the data by calculating the correlation between the spatial distribution Cuebiq user home locations, and the spatial distribution of the population as captured by US Census data. The studies found high correlations at the census tract level in Washington State [18], the Boston metropolitan area [19] and Philadelphia [20], and U.S.-wide at the county level [21], with Pearson correlation coefficients of 0.91, 0.8, 0.72 and 0.94, respectively.

We make use of the so-called Cuebiq contact index, hereafter CCI. The CCI is a 7-day rolling average of the daily number of encounters that a Cuebiq user has with other Cuebiq users in a given county. An encounter is registered for every instance of two devices occupying the same 50-foot geohash region within the same 5 minute interval; hereafter we refer to this as a CCI encounter. The CCI index is described in more detail on Cuebiq's website[8,9].

A CCI encounter thus amounts to a *contact opportunity* rather than an actual contact; in practice only a fraction $f_{contact}$ of them will consist of two Cuebiq users encountering each other at a small enough separation to be meaningful for disease transmission (e.g. < 6 feet for COVID-19). At the same time, only a fraction of the population are Cuebiq users. We thus calculate (Equation B.2) an adjusted index, $\mathrm{CCI}_{100\&}$, which is the estimated rate of Cuebiq encounters if the entire population were Cuebiq users, under the assumption (see above) that Cuebiq users constitute a representative sample of the population. The relationship between contact rate and CCI is then

$$cr = f_{contact}\mathrm{CCI}_{100\%}, f_{contact} < 1, \tag{6}$$

and so we can write Equation 5 as

$$(\ln inc_n - \ln inc_0) = (\ln cr_n - \ln cr_0) + f_{contact}P_{trans}\sum_{i=1}^{n}\mathrm{CCI}_{100\%_i} - \gamma(t_n - t_0) \tag{7}$$

Appendix B describes the details of how a time series of P_{CCI} is obtained for each state by fitting Equation 5 to COVID-19 case reports.

4. Results

Using Equation 5, we first determine the best-fit lag between $\mathrm{CCI}_{100\%}$ and incidence for each state, as described in Appendix B. Figure 1 shows the time series of $\mathrm{CCI}_{100\%}$ B.2) together with its lagged version for four example states. Using this lag, we then perform maximum-likelihood fits of the scaled transmission probability ($f_{contact}P_{trans}$) and initial incidence inc_0 to observed incidence over successive 6-week intervals, again using Equation 5. Results are shown in Figure 2, while Figure 3 shows the model-derived incidence using the maximum-likelihood values, together with the observed incidence. Fits for all 51 states are presented in the Supplementary Material.

Figure 4 shows the distributions across all states of ($f_{contact}P_{trans}$) averaged over March and April 2020, $(f_{contact}P_{trans})_{early}$, together with the average across the rest of 2020, $(f_{contact}P_{trans})_{RoY}$. The former reflects a largely pre-mask measure of the transmission probability, while for the latter, transmission probability is modified by subsequent widespread yet heterogeneous adoption of masks across US. This figure also shows the distribution of best-fit mobility-transmission lags; the median is 19 days.

Figure 5 shows $(f_{contact}P_{trans})_{early}$ versus mean spring temperature for the states (excluding the District of Columbia)[10] Computing the Pearson product-moment correlation coefficient of the two quantities, we find a moderate, statistically significant negative correlation, $r = -.57$, $N = 49$, $p = 2 \times 10^{-5}$. That is, colder temperatures tended to be associated with higher transmission probabilities in the initial stage of the pandemic, before differences in mask adoption among the states obscured the picture.

[8] `https://www.cuebiq.com/visitation-insights-contact-index/`

[9] `https://help.cuebiq.com/hc/en-us/articles/360041285051-Mobility-Insights-Mobility-Index-CMI`

[10] Data taken from `https://www.currentresults.com/Weather/US/average-state-weather.php`

Figure 6 shows a simple demonstration of how the model developed here could be used for short-range forecasting, using Florida as an example. Additional examples are shown in the Supplementary Material. To perform a forecast starting from a given date T forward, we first fit our model to the previous six weeks, $[T - 6w, T]$, of incidence data and lagged $\mathrm{CCI}_{100\%}$ data to obtain a maximum-likelihood estimate of $(f_{contact}P_{trans})$. The best-fit lag between incidence and CCI is 15 days for Florida. This means that at time T, we still have lagged CCI data up to date $T + 15d$. Using this data, together with the estimate of $(f_{contact}P_{trans})$, we are thus able to run the model 15 days into the future. We retrospectively perform forecasts from dates T_1 = 1 April 2020, T_2 = 1 June, T_3 =1 July and T_4 = 25 August, each date chosen to come just before a turnover in incidence from growth to decay or vice versa. The quality of each forecast can be visually assessed by comparing it to the actual incidence of cases reported over the forecast horizon.

5. Discussion

Across the 51 states, the scaled transmission probability averaged over the first two months of the pandemic (March and April 2020), $(f_{contact}P_{trans})_{early}$, has a median value of 0.0039. The transmission probability across the rest of the year, $(f_{contact}P_{trans})_{RoY}$, has a median value of 0.0031, about 20% lower (Figure 4, top). A marked decrease in transmission probability is consistent with the overall increasing level of NPI adoption, principally mask-wearing, as the year progressed. Seasonality may also play a role. However, looking at the individual state time series of $(f_{contact}P_{trans})$ (see Figure 2 for four examples, and the Supplementary Material for the remaining 47 states) reveals significant heterogeneity, with some states (e.g. New Jersey, Florida) showing a clear reduction in P_{CCI} after spring of 2020, yet others (e.g. Alaska) showing no clear time trend. This may be reflective of the heterogeneity in the practice of NPIs that reduce per-contact transmission probability, primarily the adoption level of masks and the level of compliance with social distancing (which, since it occurs on a scale of a few meters, is not resolved in Cuebiq's mobility data). Given the evidence of temperature dependence in COVID transmissibility, it may also be reflective of heterogeneity in seasonal weather patterns among states.

Across all states, the median best-fit time lag between $\mathrm{CCI}_{100\%}$ and observed incidence is 19 days, though here, again, there is significant heterogeneity (Figure 4, bottom). The time between infection and case report is the sum of the incubation period of a disease, the diagnostic delay and the reporting delay. In the hypothetical case of instantaneous diagnosis and reporting, the delay would be due to the incubation period alone. Thus the smallest lag we observe, 11 days, constitutes an upper limit to the median incubation period. Meta-analyses have variously reported a mean COVID-19 incubation period of 6.5 (95% CI: 5.9–7.1) days[22], 5.8 (95% CI: 5.0-6.7) days[23], 5.6 (95% CI: 5.2–6.0) days or 6.7 (95% CI: 6.0–7.4) days[24], 5.74 (95% CI: 5.18-6.30) days[25], and 6.2 (95% CI 5.4, 7.0) days[26], all of which fall below 11 days. One source of variability may be heterogeneity in state-level reporting practices. Also, since the latent period is determined by in-host interaction, it may vary systematically by population characteristics (age distribution, comorbidity profile etc.), which may also contribute to the heterogeneity.

Multiple studies have reported associations between colder temperatures and various metrics of COVID burden (see Introduction). Our results suggest, specifically, an association between temperature and per-contact transmission probability. Although prior studies have found evidence that the COVID-19 virus half-life is reduced at higher temperatures ([27],

[28], [29]), it is important to note that our results do not by themselves imply that this particular mechanism is responsible. Behavior could also contribute: People in warmer climates tend to spend a larger proportion of their time outdoors, thus a larger proportion of daily contacts will occur outdoors in, for example, springtime California versus springtime Alaska. And there is strong evidence to suggest that the outdoor risk of COVID transmission is substantially lower than the indoor risk (see [30] for a systematic review). Both these causal pathways, and more, could be operating together.

In the forecasting demonstration shown in Figure 6, the sharp downturn in Florida incidence just after 1 April 2020 is reasonably well predicted, as is the return to incidence growth after 1 June. However, neither the downturn after 1 July nor the upturn beginning in late August are predicted. Indeed, Florida's $\mathrm{CCI}_{100\%}$ (Figure 1) varies much less after the beginning of July than it does before. However, widespread mask mandates started coming into effect in Florida on 23 June; with a 15-day lag (i.e. 8 July) this is close to Florida's second incidence peak. This suggests that at later time, variations in mask use, rather than in contact rate, may have played the dominant role in driving changes in transmission. We leave to future work a more sophisticated forecast model that incorporates mask use, where such data is available.

The work presented here is subject to a number of limitations: i) Both our model and the data we use lack any stratification, thus any effects arising from heterogeneous demography, health status etc. within a given state are not accounted for. ii) Though previous studies all found high correlation between the geographic distribution of Cuebiq users and that of the population as a whole, this does not fully guarantee the represenativeness of Cuebiq users. iii) In comparing states to each other, we have made the assumption that the scaling (Equation B.1) between true contact rate and adjusted Cuebiq contact index is the same across all states. iv) We have assumed that within a given state, the lag between mobility and incidence, which we estimate using only the first four months of the pandemic, remains constant. v) We have argued that transmission probability changes more slowly over time than mobility, and thus approximated P_{trans} as constant within successive 6-week periods. However, this approximation may not always hold well, in particular when a change in mask mandates falls within a given period. Also, during the phase of gradual relaxation after the initial lock-downs, mask use may have increased at a similar rate to mobility as businesses, public spaces etc. re-opened while at the same time requiring masking. vi) In our forecasting experiment, we have unfairly granted ourselves fore-knowledge of the state's mobility-incidence lag, which was actually fit using the first four months of data. In practice, in the very beginning of a pandemic we would have to resort to using a range of plausible lags, the lower bound being the mean incubation period of the disease.

6. CONCLUSION

Using a mobility index that can be considered a direct proxy of contact rate has allowed us to construct a fully mechanistic model that derives disease incidence from this data. As a result, we have been able to get direct insight into the variability of per-contact COVID-19 transmission in the U.S. both by state and by date. Our findings are consistent with associations between colder temperatures and stronger COVID-19 burden reported in previous studies, and suggest that it is specifically changes in the per-contact transmission probability which play a role. As a forecast tool, the model would have performed best before NPIs that modified per-contact transmission probability—principally masks—came

into widespread use. To lift this limitation, future development should also incorporate time series data of NPI use. Our methodology is also readily extensible to other respiratory diseases such as influenza or RSV, contingent on the availability of good-quality surveillance data. Indeed, in a non-pandemic setting forecasting will be aided by the (likely) absence of NPIs, and by mobility following more predictable seasonal patterns rather than being driven by reaction to epidemiology. The availability of mobility data that even more directly probes person-to-person contacts, e.g. through Bluetooth proximity detection of the sort used in COVID exposure-notification apps, would also benefit the performance of this model.

Contributors

All authors were involved in the conception and design of the study. EWT, ZM and MGC developed the methodology, with input from all other authors. CM acquired the funding to purchase the commercial (Cuebiq) data used. EWT and MGC accessed, verified and collected the data. EWT, ZM and MGC contributed to the analysis, including the development of the software used therein. EWT wrote the original draft. All authors critically reviewed and edited the manuscript for scientific content. All authors have access to the data and software used, and are thus able to validate the analysis.

Funding

The mobility data used in this study was purchased by Sanofi.

Declaration of interests

EWT, SSC, LC, RVA and CM are employees of Sanofi and may hold stock options. GG was an employee of Sanofi during part of the time over which this manuscript was prepared, and may hold stock options. MGC has received funding from Sanofi for an unrelated project. All other authors declare no conflicts of interest.

References

[1] Walker PG, Whittaker C, Watson OJ, Baguelin M, Winskill P, Hamlet A, et al. The impact of COVID-19 and strategies for mitigation and suppression in low-and middle-income countries. Science. 2020;369(6502):413-22.

[2] Basellini U, Alburez-Gutierrez D, Del Fava E, Perrotta D, Bonetti M, Camarda CG, et al. Linking excess mortality to mobility data during the first wave of COVID-19 in England and Wales. SSM-Population Health. 2021;14:100799.

[3] Kartal MT, Depren Ö, Depren SK. The relationship between mobility and COVID-19 pandemic: Daily evidence from an emerging country by causality analysis. Transportation Research Interdisciplinary Perspectives. 2021;10:100366.

[4] Ilin C, Annan-Phan S, Tai XH, Mehra S, Hsiang S, Blumenstock JE. Public mobility data enables covid-19 forecasting and management at local and global scales. Scientific reports. 2021;11(1):1-11.

[5] Sadowski A, Galar Z, Walasek R, Zimon G, Engelseth P. Big data insight on global mobility during the Covid-19 pandemic lockdown. Journal of big Data. 2021;8(1):1-33.

[6] Cot C, Cacciapaglia G, Sannino F. Mining Google and Apple mobility data: temporal anatomy for COVID-19 social distancing. Scientific reports. 2021;11(1):1-8.

[7] Chang S, Pierson E, Koh PW, Gerardin J, Redbird B, Grusky D, et al. Mobility network models of COVID-19 explain inequities and inform reopening. Nature. 2021;589(7840):82-7.

[8] Liu M, Thomadsen R, Yao S. Forecasting the spread of COVID-19 under different reopening strategies. Scientific reports. 2020;10(1):1-8.

[9] Fritz C, Dorigatti E, Rügamer D. Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly COVID-19 cases in Germany. Scientific Reports. 2022;12(1):1-18.

[10] Zhang M, Wang S, Hu T, Fu X, Wang X, Hu Y, et al. Human mobility and COVID-19 transmission: A systematic review and future directions. Annals of GIS. 2022:1-14.

[11] Mohammadi Z, Cojocaru M, Thommes E. Human behaviour, NPI and mobility reduction effects on COVID-19 transmission in different regions of the world. BMC Public Health. 2022, submitted.

[12] Mecenas P, Bastos RTdRM, Vallinoto ACR, Normando D. Effects of temperature and humidity on the spread of COVID-19: A systematic review. PLoS one. 2020;15(9):e0238339.

[13] Diao Y, Kodera S, Anzai D, Gomez-Tames J, Rashed EA, Hirata A. Influence of population density, temperature, and absolute humidity on spread and decay durations of COVID-19: a comparative study of scenarios in China, England, Germany, and Japan. One Health. 2021;12:100203.

[14] Loché Fernández-Ahúja JM, Fernández Martínez JL. Effects of climate variables on the COVID-19 outbreak in Spain. International Journal of Hygiene and Environmental Health. 2021;234:113723. Available from: `https://www.sciencedirect.com/science/article/pii/S1438463921000389`.

[15] Smith TP, Flaxman S, Gallinat AS, Kinosian SP, Stemkovski M, Unwin HJT, et al. Temperature and population density influence SARS-CoV-2 transmission in the absence of nonpharmaceutical interventions. Proceedings of the National Academy of Sciences. 2021;118(25). Available from: `https://www.pnas.org/content/118/25/e2019284118`.

[16] Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proceedings of the royal society of london Series A, Containing papers of a mathematical and physical character. 1927;115(772):700-21.

[17] Ma J. Estimating epidemic exponential growth rate and basic reproduction number. Infectious Disease Modelling. 2020;5:129-41.

[18] Wang F, Wang J, Cao J, Chen C, Ban XJ. Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example. Transportation Research Part C: Emerging Technologies. 2019;105:183-202.

[19] Aleta A, Martin-Corral D, Pastore y Piontti A, Ajelli M, Litvinova M, Chinazzi M, et al. Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. Nature Human Behaviour. 2020;4(9):964-71.

[20] Nande A, Sheen J, Walters EL, Klein B, Chinazzi M, Gheorghe AH, et al. The effect of eviction moratoria on the transmission of SARS-CoV-2. Nature communications. 2021;12(1):1-13.

[21] Deng H, Du J, Gao J, Wang Q. Network percolation reveals adaptive bridges of the mobility network response to COVID-19. PloS one. 2021;16(11):e0258868.

[22] Alene M, Yismaw L, Assemie MA, Ketema DB, Gietaneh W, Birhan TY. Serial interval and incubation period of COVID-19: a systematic review and meta-analysis. BMC Infectious Diseases. 2021;21(1):1-9.

[23] McAloon C, Collins Á, Hunt K, Barber A, Byrne AW, Butler F, et al. Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research. BMJ open. 2020;10(8):e039652.

[24] Quesada J, López-Pineda A, Gil-Guillén V, Arriero-Marín J, Gutiérrez F, Carratala-Munuera C. Incubation period of COVID-19: A systematic review and meta-analysis. Revista Clínica Española (English Edition). 2021;221(2):109-17.

[25] Rai B, Shukla A, Dwivedi LK. Incubation period for COVID-19: a systematic review and meta-analysis. Journal of Public Health. 2021:1-8.

[26] Dhouib W, Maatoug J, Ayouni I, Zammit N, Ghammem R, Fredj SB, et al. The incubation period during the pandemic of COVID-19: a systematic review and meta-analysis. Systematic Reviews. 2021;10(1):1-14.

[27] Ijaz M, Brunner A, Sattar S, Nair RC, Johnson-Lussenburg C. Survival characteristics of airborne human coronavirus 229E. Journal of General Virology. 1985;66(12):2743-8.

[28] Morris DH, Yinda KC, Gamble A, Rossine FW, Huang Q, Bushmaker T, et al. Mechanistic theory predicts the effects of temperature and humidity on inactivation of SARS-CoV-2 and other enveloped viruses. Elife. 2021;10:e65902.

[29] Riddell S, Goldie S, Hill A, Eagles D, Drew TW. The effect of temperature on persistence of SARS-CoV-2 on common surfaces. Virology journal. 2020;17(1):1-7.

[30] Bulfone TC, Malekinejad M, Rutherford GW, Razani N. Outdoor transmission of SARS-CoV-2 and other respiratory viruses: a systematic review. The Journal of infectious diseases. 2021;223(4):550-61.

Appendix A: The model in detail

In an SIR model, the equation for rate of change of disease prevalence is

$$\frac{dI}{dt} = \frac{\beta(t)S(t)I(t)}{N} - \gamma I(t), \tag{A.1}$$

where $\beta(t)$ is the time-dependent average number of effective contacts per person per unit time, γ is the recovery rate from the infectious state, and N is the size of the total population. Effective contacts are ones which would transmit disease if they involved an infectious person. As long as only a small fraction of the total population has become infected—as was the case in the US and most of the world throughout 2022—$S/N \approx 1$, and the SIR model solution for the prevalence is

$$I(t) = I_0 e^{(\beta(t)-\gamma)t}, \tag{A.2}$$

see e.g. [17]. The incidence, i.e. the rate of change of cumulative cases $C(t)$, is given by the first term on the right-hand side of Equation A.1 alone:

$$inc(t) = \frac{dC}{dt} = \frac{\beta S(t)I(t)}{N} \approx \beta(t)I(t), \tag{A.3}$$

where the approximate equality holds when, again, $S/N \approx 1$. Substituting, we obtain

$$inc(t) = \beta(t)I_0 e^{(\beta(t)-\gamma)(t)} = inc_0 \left(\frac{\beta(t)}{\beta_0}\right) e^{(\beta(t)-\gamma)t} \tag{A.4}$$

where

$$inc_0 = \beta_0 I_0. \tag{A.5}$$

Since for an SIR model the instantaneous effective reproduction number is

$$R_{eff}(t) = \frac{\beta(t)}{\gamma}\frac{S(t)}{N} \tag{A.6}$$

thus we can also write the incidence in terms of the reproduction number:

$$inc(t) = inc_0 \left(\frac{R_{eff}(t)-1}{R_{eff,0}-1}\right) e^{\left(R_{eff}(t)-1\right)\gamma t} \tag{A.7}$$

Taking the log of Equation A.4, we have

$$(\ln inc(t) - \ln inc_0) = (\ln \beta(t) - \ln \beta_0) + (\beta(t) - \gamma)\, t \tag{A.8}$$

Considering now an infinitesimally small time interval, dt, this becomes

$$d\ln inc(t) = d\ln \beta(t) + (\beta(t) - \gamma)\, dt$$

or

$$\frac{d\ln inc(t)}{dt} = \frac{d\ln \beta(t)}{dt} + \beta(t) - \gamma \tag{A.9}$$

Integrating with respect to t, we obtain, over a time interval $[t_0, t]$,

$$(\ln inc(t) - \ln inc(t_0)) = (\ln \beta(t) - \ln \beta(t_0)) + \int_{t_0}^{t} \beta(t')dt' - \gamma(t - t_0) \tag{A.10}$$

We can decompose $\beta(t)$ into

$$\beta(t) = P_{trans}(t)cr(t), \tag{A.11}$$

where $cr(t)$ is the contact rate, while $P_{trans}(t)$ is the transmission probability per contact. Note that we can then express the reproduction number as:

$$R_{eff}(t) = \frac{P_{trans}(t)cr(t)}{\gamma} \tag{A.12}$$

In general, both the contact rate and the transmission probability change over time, the latter due to changes in the practice of non-pharmaceutical interventions (NPIs) such as mask-wearing, as well as changes intrinsic to the disease, e.g. emergence of new variants. Since widespread changes in NPIs and in the relative distribution of variants are usually gradual, whereas contact patterns can change significantly from one day to the next (for example between a weekday and the weekend, or as the result of a mass gathering event), we expect $P_{trans}(t)$ to generally vary more slowly than $cr(t)$. If P_{trans} can be considered constant over the time interval $[t_0, t]$, then Equation A.10 becomes

$$(\ln inc(t) - \ln inc(t_0)) = (\ln cr(t) - \ln cr(t_0)) + P_{trans}\int_{t_0}^{t} cr(t')dt' - \gamma(t - t_0) \tag{A.13}$$

Suppose we have time series data of incidence and contact rate reported with constant time interval δt, so that cr_i and inc_i are the contacts per person and the total number of new cases, respectively, occurring within the time interval $t_{i-1} < t \leq t_i$. We can then apply Equation A.13 in discrete form to obtain the change in incidence between a time t_0 and time t_n in terms of the contacts occurring during this time:

$$(\ln inc_n - \ln inc_0) = (\ln cr_n - \ln cr_0) + P_{trans}\sum_{i=1}^{n} cr_i - \gamma(t_n - t_0) \tag{A.14}$$

In practice, disease incidence captured by surveillance will be subject to under-reporting, i.e. the reported incidence is

$$inc = f_{rep} \cdot inc_{true} \tag{A.15}$$

where inc_{true} is the true underlying incidence, and f_{rep} is the fraction of cases reported. If f_{rep} can be considered constant over the time interval $[t_a, t_b]$, then if we now replace the reported incidence with the true incidence in Equation A.14, the left-hand side becomes

$$\ln\left(\frac{inc_b}{f_{rep}}\right) - \ln\left(\frac{inc_a}{f_{rep}}\right)$$

or

$$\ln\left(\frac{f_{ref} \cdot inc_b}{f_{rep} \cdot inc_a}\right),$$

thus f_{rep} cancels out and we recover the original equation. Therefore, when considering time intervals over which the degree of under-reporting can be considered constant, the relationship described by Equation A.13 is independent of under-reporting.

A simplification we have made thus far is to assume that there is no lag between contacts and their effect on reported incidence. In reality, reporting delays and the incubation and latent periods of the disease will together impose a distribution of delays between the time that transmission occurs and the time that the resulting cases are captured by surveillance. If cases reported at time t_i depend on contacts occurring between times t_{i-q} and t_{i-p}, with $q > p$, then we can account for this by replacing cr_i with an appropriately lagged version,

$$cr'_i = \alpha_{i-q}cr_{i-q} + \alpha_{i-q+1}cr_{i-q+1} + ... + \alpha_{i-p}cr_{i-p}. \tag{A.16}$$

Appendix B: Fitting the model to data

The CCI of a given region is the daily number of instances of two Cuebiq users occupying the same 50ft x 50ft geohash grid cell within the same 5 minute interval, divided by the total number of Cuebiq users within that region. From this, we want to estimate the rate of encounters between Cuebiq users at distances $\leq d_{trans}$, where d_{trans} is the maximum distance for potentially disease-transmitting contacts. Assuming an average movement speed v, the time to pass through a $d_{trans} \times d_{trans}$ cell is approximately d_{trans}/v. The proportionality constant between CCI and the contact rate within distance d_{trans} between Cuebiq users is given by the ratio of their associated space-time volumes. And, assuming that v is approximately constant, the contact rate is linearly proportional to CCI:

$$cr_{Cuebiq} \approx \left(\frac{d_{trans}}{50ft}\right)^2 \left(\frac{d_{trans}/v}{5min}\right) CCI = f_{contact} CCI \tag{B.1}$$

Since only a fraction of the total population are Cuebiq users, the CCI only captures a fraction of the total contacts experienced by a person per day. In order to estimate $\text{CCI}_{100\%}$, the hypothetical CCI which would be measured if the entire population were Cuebiq users, we additionally obtain from Cuebiq the time series of total number of Cuebiq user devices seen on day i across a given region, n_{Cuebiq_i}. Insofar as the Cuebiq users can be considered a representative sample of the general population, we can then estimate $\text{CCI}_{100\%}$ on day i across a given region by rescaling the CCI as follows:

$$\text{CCI}_{100\%_i} = \text{CCI}_i \cdot \frac{N_{pop}}{n_{Cuebiq_i}} \tag{B.2}$$

where N_{pop} is the population size of the region. Our estimate for the total contact rate is then

$$cr = f_{contact} \text{CCI}_{100\%} \tag{B.3}$$

We can then write Equation 5 as

$$(\ln inc_n - \ln inc_0) = (\ln cr_n - \ln cr_0) + f_{contact} P_{trans} \sum_{i=1}^{n} \text{CCI}_{100\%_i} - \gamma(t_n - t_0) \tag{B.4}$$

where $P_{CCI} = f_{contact} P_{trans}$ is the transmission probability per Cuebiq encounter.

We conduct our analysis at the state level, and thus aggregate incidence and Cuebiq data (both of which are provided at the county level) accordingly. In order to calculate the lagged version of the time series of $\text{CCI}_{100\%_i}$ as per Equation A.16, we make the simplifying assumption that $q = p + 7$. Given that the CCI is already computed as a 7-day rolling average, we then only need to find p for each state. To do so, we perform a two-step optimization. First, we select a time window $t_j < t \leq t_k$ within the available data. We then lag the time series of $\text{CCI}_{100\%_i}$ by values of $L = 0, 1, 2, ..., 50$. For each value of L, we use the R optimization function `optim()` to find the value of P_{CCI} which produces the best fit of Equation B.4 to the observed incidence over the time window, in the sense of minimizing the negative log likelihood (NLL).

We take as t_j the day on which cumulative cases first reached or exceeded 10 in a given state. We repeat the above fitting procedure with $t_k = 1$ May 2020, $t_k = 1$ June 2020, and $t_k = 1$ July 2020, each time computing a best-fit lag. We then take the average of these three as the best-fit lag p for the given state. Using p, we compute the lagged CCI time series for the state as per Equation A.16, which we then use to fit Equation B.4 to reported incidence.

We do so over successive 6-week time intervals, going from t_j until the end of 2020, thus obtaining a best-fit P_{CCI} for each interval.

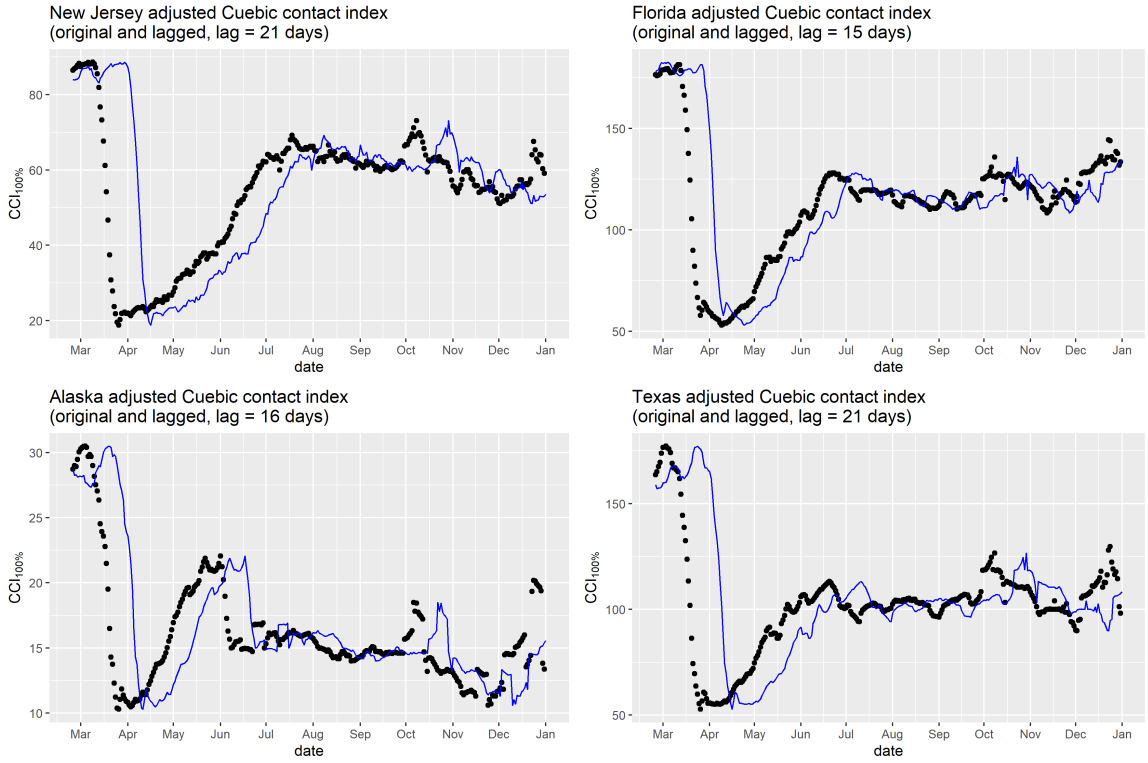

FIGURE 1. The rescaled seven-day rolling average Cuebiq contact index, $\mathrm{CCI}_{100\%}$, for four states during 2020 (black points). Also shown is the same data lagged by the best-fit mobility-incidence delay for the given state, obtained as described in Appendix B

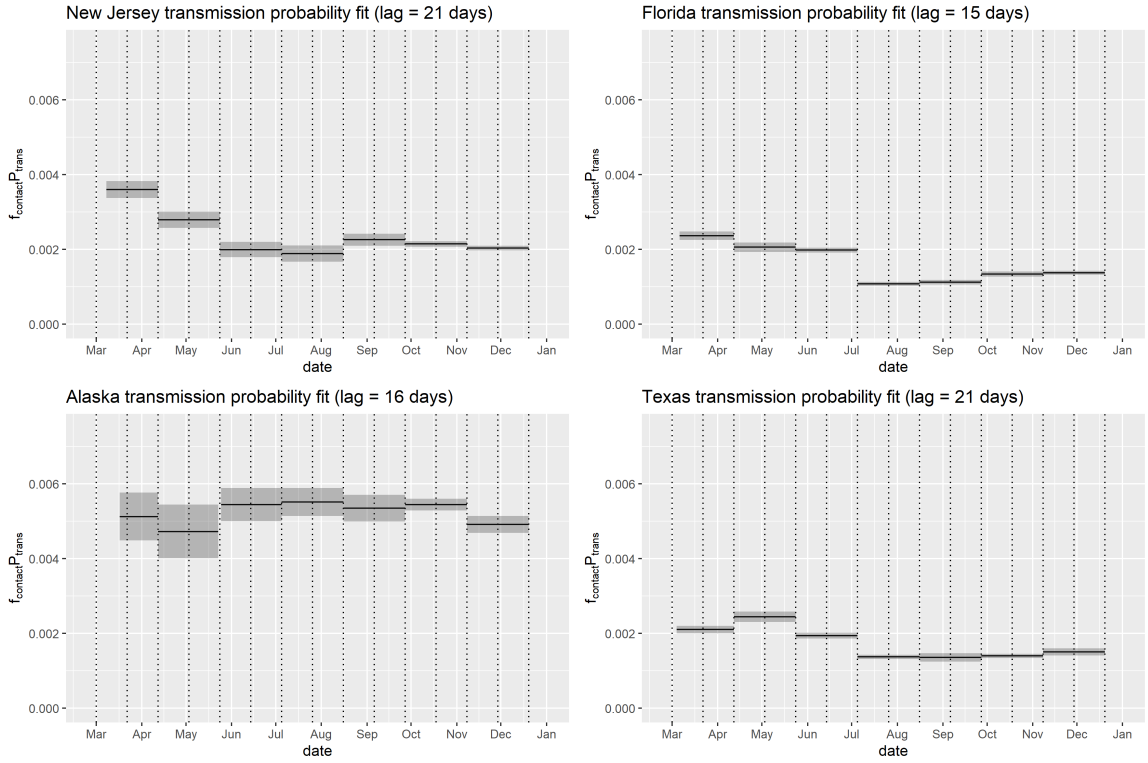

FIGURE 2. P_{CCI}, the probability of an effective Cuebiq contact (see Equation 6) from the same rolling fits computed computed at six-week intervals (gray dotted lines) for Figure 3. Gray bands denote the 95% confidence interval for each interval.

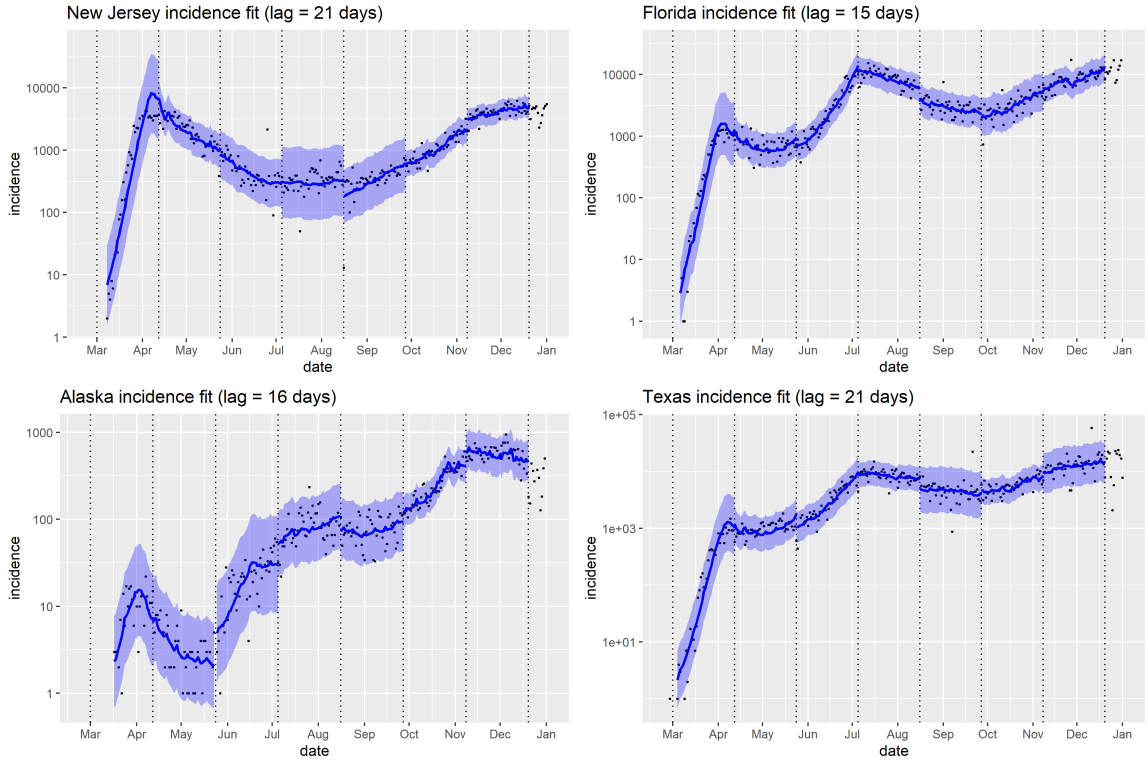

FIGURE 3. Model incidence (blue curve segments; blue bands show 95% confidence interval) from Equation 5, using the maximum-likelihood fits of P_{CCI} (see Figure 2) and initial incidence inc_0 to the time series of observed incidence (black dots) and $\mathrm{CCI}_{100\%}$ (see Figure 1) for four states. Fits are performed over successive six-week intervals (delimited by dotted vertical lines).

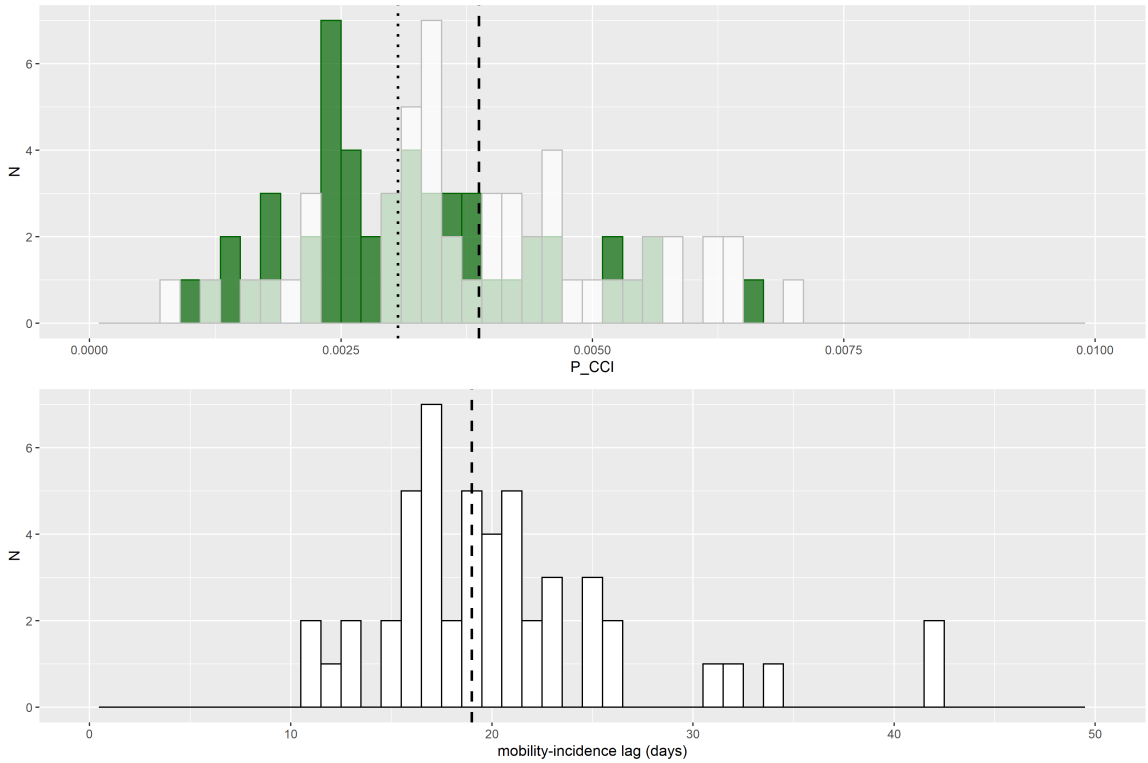

FIGURE 4. Top: Distribution across all 51 states of early (March 1 to April 30, 2020) average P_{CCI} (white; dashed line shows median = 0.0039), and rest-of-year (May 1 to December 31, 2020) average P_{CCI} (green, dotted line shows median = 0.0031). Bottom: distribution across all 51 states of best-fit lag between reported incidence and mobility (dashed line shows median = 19 days).

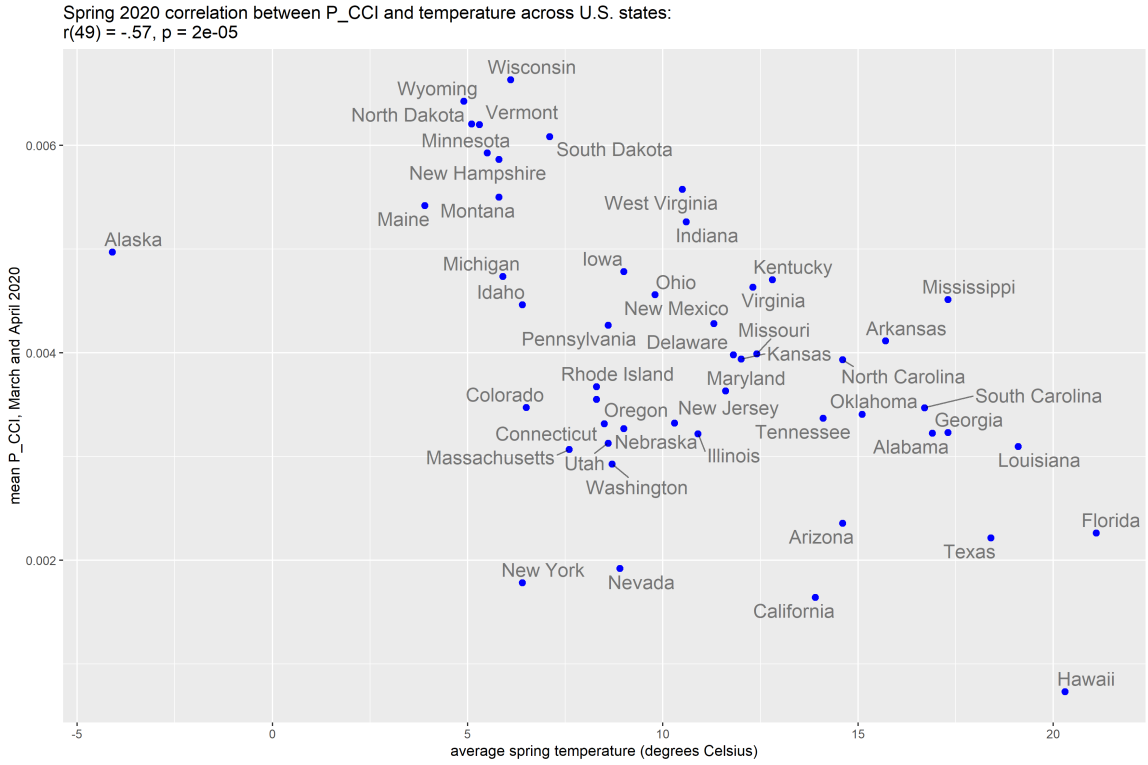

FIGURE 5. Early (1 March to 30 April) P_{CCI} versus average winter temperature by state (excluding DC). A moderate, statistically significant negative correlation exists, $r(49) = -.57$, $p = 2 \times 10^{-5}$, i.e. lower temperatures tend to be associated with higher P_{CCI}.

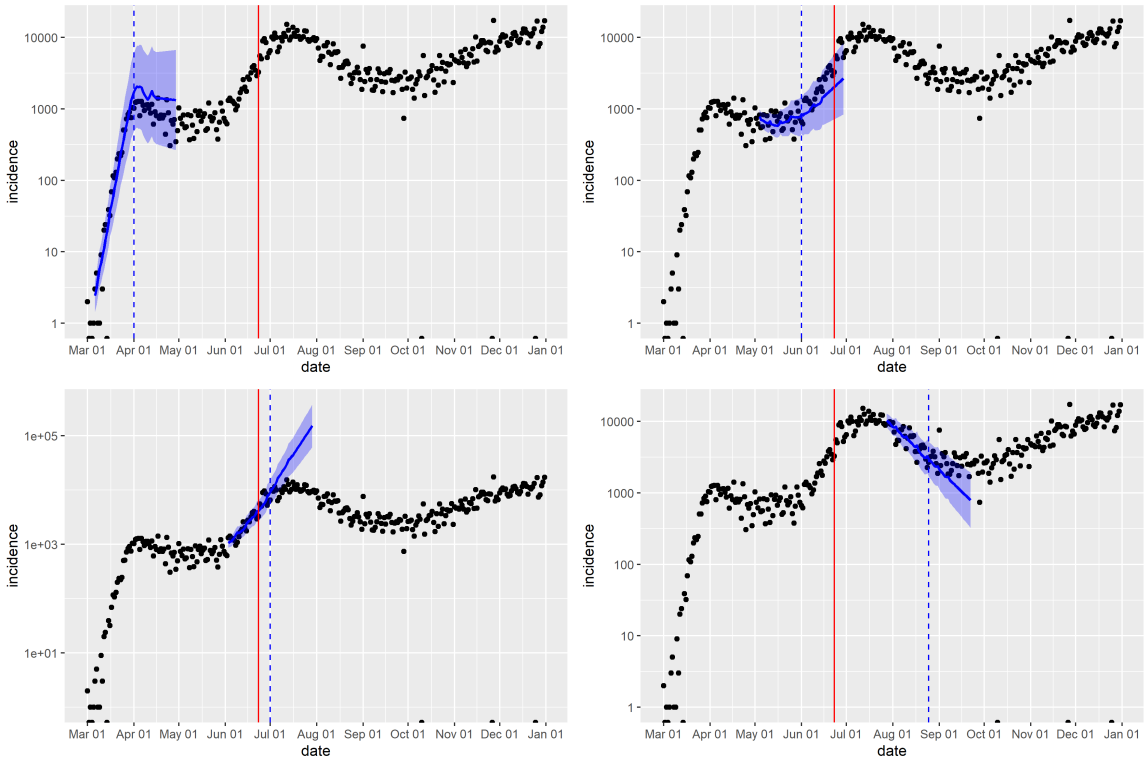

FIGURE 6. Illustration of the application of our fitting methodology to short-range forecasting, using the state of Florida as an example. Forecasts are performed at different times (dashed vertical lines), each time using a fitting time window of the previous six weeks to estimate P_{CCI}. The forecast time horizon is equal to the mobility-transmission lag, which for Florida is estimated as 15 days. Also shown for comparison is 23 June (solid vertical line), the date on which widespread mask mandates started coming into effect in Florida, starting with Miami-Dade County.