Estimating the establishment of local transmission and the cryptic phase of the COVID-19 pandemic in the USA

Jessica T. Davis[1*], Matteo Chinazzi [1*], Nicola Perra [2*,1], Kunpeng Mu[1], Ana Pastore y Piontti[1], Marco Ajelli[3,4], Natalie E. Dean[5], Corrado Gioannini[6], Maria Litvinova[6], Stefano Merler[3], Luca Rossi[6], Kaiyuan Sun[7], Xinyue Xiong[1], M. Elizabeth Halloran[8,9], Ira M. Longini Jr.[5], Cécile Viboud[7], Alessandro Vespignani[1,6,†]

[1]Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston, MA USA
[2]Networks and Urban Systems Centre, University of Greenwich, London, UK
[3]Bruno Kessler Foundation, Trento Italy
[4]Department of Epidemiology and Biostatistics, Indiana University School of Public Health, Bloomington, IN, USA
[5]Department of Biostatistics, College of Public Health and Health Professions, University of Florida, Gainesville, USA
[6]ISI Foundation, Turin, Italy
[7]Fogarty International Center, NIH, USA
[8]Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[9]Department of Biostatistics, University of Washington, Seattle, WA. USA
[*]These authors contribute equally to this work
[†]To whom correspondence should be addressed; E-mail: a.vespignani@northeastern.edu.

We use a global metapopulation transmission model to study the establishment of sustained and undetected community transmission of the COVID-19 epidemic in the United States. The model is calibrated on international case importations from mainland China and takes into account travel restrictions to and from international destinations. We estimate widespread community

transmission of SARS-CoV-2 in February, 2020. Modeling results indicate international travel as the key driver of the introduction of SARS-CoV-2 in the West and East Coast metropolitan areas that could have been seeded as early as late-December, 2019. For most of the continental states the largest contribution of imported infections arrived through domestic travel flows.

The first confirmed case of COVID-19 in the United States (US) was diagnosed in Washington state on January 21, 2020 (*1*). In quick succession other cases were confirmed in Illinois, California, and Arizona (*2–4*). All initial cases reported in the US had a known travel history to mainland China, the epicenter of the pandemic. The first reported *local* transmission on US soil was discovered on January 30 in Illinois (*5*). However, very few COVID-19 cases were detected until the case definition for testing was updated on March 4, 2020 to include symptomatic individuals without travel history (*6*). On April 1, when the US federal government issued social distancing guidelines, $26,655$ new reported cases and $1,050$ deaths were recorded in the United States. In Fig. 1A we include a timeline of initial confirmed cases, testing policies, and early containment and mitigation initiatives concerning the US COVID-19 outbreak.

Given the narrowness of the initial testing criteria, it is expected that the SARS-CoV-2 virus causing COVID-19 was spreading through "cryptic transmission" in January and February, setting the stage for the large epidemic wave experienced in March and April, 2020. In this study, we model the arrival and cryptic phase of the COVID-19 epidemic in the US. We estimate the time frame for the establishment of local transmission in the different states, and provide a statistical analysis of the domestic spread of the COVID-19 epidemic.

To study the spatial and temporal spread of COVID-19, we use the Global Epidemic and Mobility Model (GLEAM), an individual-based, stochastic, and spatial epidemic model (*7–11*). The model was previously used to characterize the early stage of the COVID-19 epidemic in mainland China and the effect of travel restrictions on infections exported to other global re-

gions (*12*). GLEAM generates an ensemble of possible epidemics described by the number of newly generated infections, the time of disease arrival in different regions of the world, and the number of infected travelers. The model divides the global population into more than $3,200$ subpopulations in roughly 200 different countries and territories. The airline transportation data encompass daily origin-destination traffic flows from the Official Aviation Guide (OAG) and the International Air Transport Association (IATA) databases (*13, 14*), whereas ground mobility flows are derived from the analysis and modeling of data collected from the statistics offices of 30 countries on five continents (*7, 8*). The transmission dynamics take place within each subpopulation and assume a classic SLIR-like compartmentalization scheme for disease progression similar to those used in several large scale models of SARS-CoV-2 transmission (*15–20*). Each individual, at any given point in time, is assigned to a compartment corresponding to their particular disease-related state (being, e.g., susceptible, latent, infectious, removed) (*12*). This state also controls the individual's ability to travel (more detail in the supplementary material, SM). Individuals transition between compartments through stochastic chain binomial processes. Susceptible individuals can acquire the virus through contacts with individuals in the infectious category and can subsequently become latent (i.e., infected but not yet able to transmit the infection). The process of infection is modeled by using age-stratified contact patterns at the state level (*21*). Latent individuals progress to the infectious stage at a rate inversely proportional to the latent period, and infectious individuals progress to the removed stage at a rate inversely proportional to the infectious period. The sum of the mean latent and infectious periods defines the generation time. Removed individuals are those who can no longer infect others. To estimate the number of deaths, we use the age-stratified infection fatality ratios from (*22*). The epidemiological model we adopted has a parsimonious structure with respect to others proposed in the literature (*42, 43*). While such details are important in the evaluation of mitigation scenarios, we focused on the very initial phase of the spread where the epidemic was unchecked.

At this stage, the transmission model does not account for heterogeneities due to age differences in susceptibility to the SARS-CoV-2 infection. This is an intense area of discussion at the moment (*44–46*).

We assume a start date of the epidemic in Wuhan, China, that falls between November 15, 2019 and December 1, 2019, with 40 initial infections (*12, 20, 23–25*). The model generates an ensemble of possible epidemic realizations, and is calibrated using approximate Bayesian computation (ABC) methods (*26*) on observations of international importations from mainland China through January 21, 2020 (*12*). Only a fraction of imported cases are detected at the destination (*27*). According to the estimates proposed in (*28*), we stratify the detection capacity of countries into three groups: high, medium and low surveillance capacity according to the Global Health Security Index (*29*) (see SM). The model calibration does not consider correlations among importations (family travel) and assumes that travel probabilities are homogeneous across all individuals in each subpopulation. The ABC calibration using a generation time $T_g = 6.5$ days yields $2,000$ individual realizations of the global evolution of the pandemic that provide information on imported infections, locally generated infections, and deaths in all subpopulations considered in the model (*30*). The model accounts for international travel restrictions according to available data on traffic reduction and government issued policies. The ABC calibration gives a posterior distribution for the basic reproductive number R_0 in the US (median 2.9 [95% CI 2.6-3.1]). The median reproductive numbers for each state range from $2.8 - 3.0$, with doubling times ranging from $3.1 - 3.7$ days, due to the changes in the specific age stratified contact patterns. In the SM we also consider an additional calibration based on the deaths observed in the US. These results do not exhibit major differences and do not change the overall picture presented here.

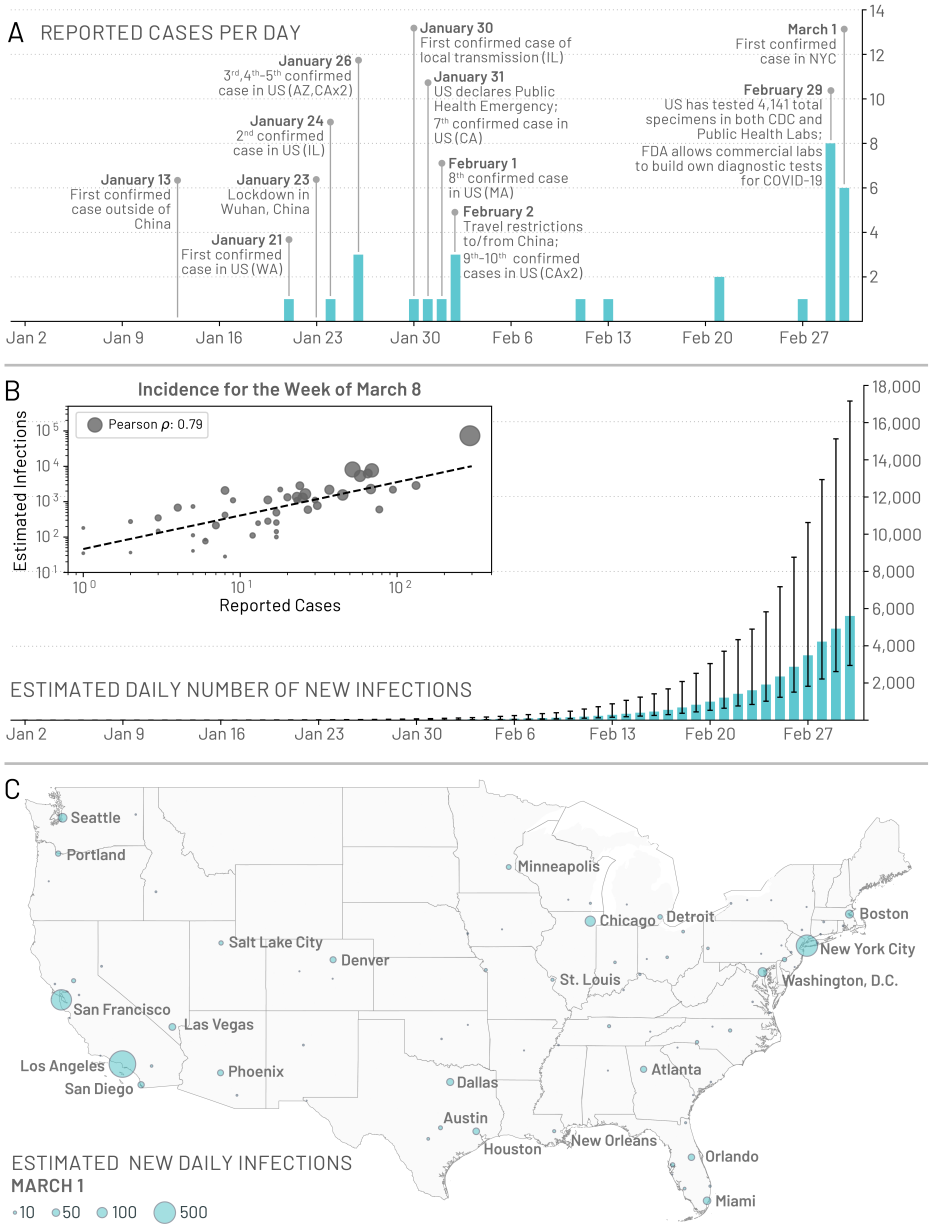

Figure 1: **Early picture of the COVID-19 outbreak in the United States.** (A) A timeline of the daily reported and confirmed cases of COVID-19 in the US including information on the first 10 reported cases and other significant events related to the outbreak up to March 1, 2020. (B) Model-based estimates for the daily number of new infections in the continental US. The inset plot compares the weekly incidence of reported cases with the weekly incidence of infections estimated by the model for the week of March $8 - 14$, 2020 for the 48 continental states that reported at least 1 case. Circle size corresponds to the population size of each state. (C) Model based estimates for the median number of daily infections in the continental US as of March 1, 2020.

Onset of local transmission

Stochastic simulations of the worldwide epidemic spread yield international/domestic infection importations, incidence of infections, and deaths per subpopulation at a daily resolution in the continental US. In Fig.1B we show the model estimates for the median daily incidence of new infections up to March 1, 2020 in the continental US. There is a stark contrast between the model output and the number of officially reported cases by the same date, highlighting the significant number of potential transmission events that may have already occurred before many states had implemented testing strategies independent of travel history. For a model validation we compare our model projections of the number of infections during the week of March 8, 2020 to the number of cases reported during that week within each state that had at least 1 case (shown in Fig.1B inset). We see a strong correlation between the reported cases and our model's projected number of infections, (Pearson's correlation coefficient on log-values 0.79, $p < 0.001$), although many fewer cases had actually been reported by that time. If we assume that the number of reported cases and simulated infections are related through a simple binomial stochastic sampling process, we find that the mean ascertainment rate of detecting an infected individual by March 8, 2020 is 1.3% (90%CI: [.19%-6.3%]). The ascertainment rate is growing quickly as the testing is ramping up, and our estimate is more than doubled by March 15, 2020 (2.9% with 90%CI:[.39%-8.6%]). The Center for Disease Control estimates an ascertainment rate of about 4% to 10% during March and April, 2020 at different locations in the US (*33*). SARS-CoV-2 infections are also distributed heterogeneously across the US. In Fig.1C we show the daily number of new infections on March 1, 2020.

In Fig. 2 we plot the posterior distribution for the earliest date within each state when at least 10 new infections per day occurred in the community. California and New York are the first states with a probability larger than 50% to have experienced local transmission by the

end of January (California) or beginning of February (New York), 2020. However, the wide distribution of dates suggests that SARS-CoV-2 cryptic transmission may have started as early as December, 2019. The posterior distribution for the timing of the onset of local transmission peaks by mid February in Washington, Florida, New Jersey, Texas, Illinois, Massachusetts, Maryland, Nevada, Georgia, Virginia, Arizona, Colorado, Ohio, Pennsylvania, North Carolina, and Michigan. From the posterior distribution of Fig. 2, all states have a median date of the onset of local transmission by early March, with the large majority of them in February, 2020, a critical month for the cryptic spread of SARS-CoV-2 in the continental US. However, during that time, US testing was still focused on returning travelers from China.

International and domestic seeding

As the model allows the recording of the origin and destination of SARS-CoV-2 carriers at the global scale, we can study the possible sources of infection importation for each state. In particular, we are able to record the flow of latent and infectious individuals through international and domestic flight connections. However, the model also considers the effect of possible infections through commuters that may only spend a few hours in a neighboring subpopulation (*7*). It is important to stress that the model's realizations explore the many possible paths of the epidemic. Thus, the analysis provided here must be considered as a statistical description of the potential sources of SARS-CoV-2 importations, rather than providing a specific, single causal chain of events.

In order to identify seeding events relevant to the onset of local transmission in each stochastic realization of the model, we record the number of importations (of latent and infectious individuals) before the local transmission chains were established (defined as 10 daily local transmission events). We visualize the origin of the seeding importations relevant for establishing local transmission by aggregating the importation sources, considering some key geographical

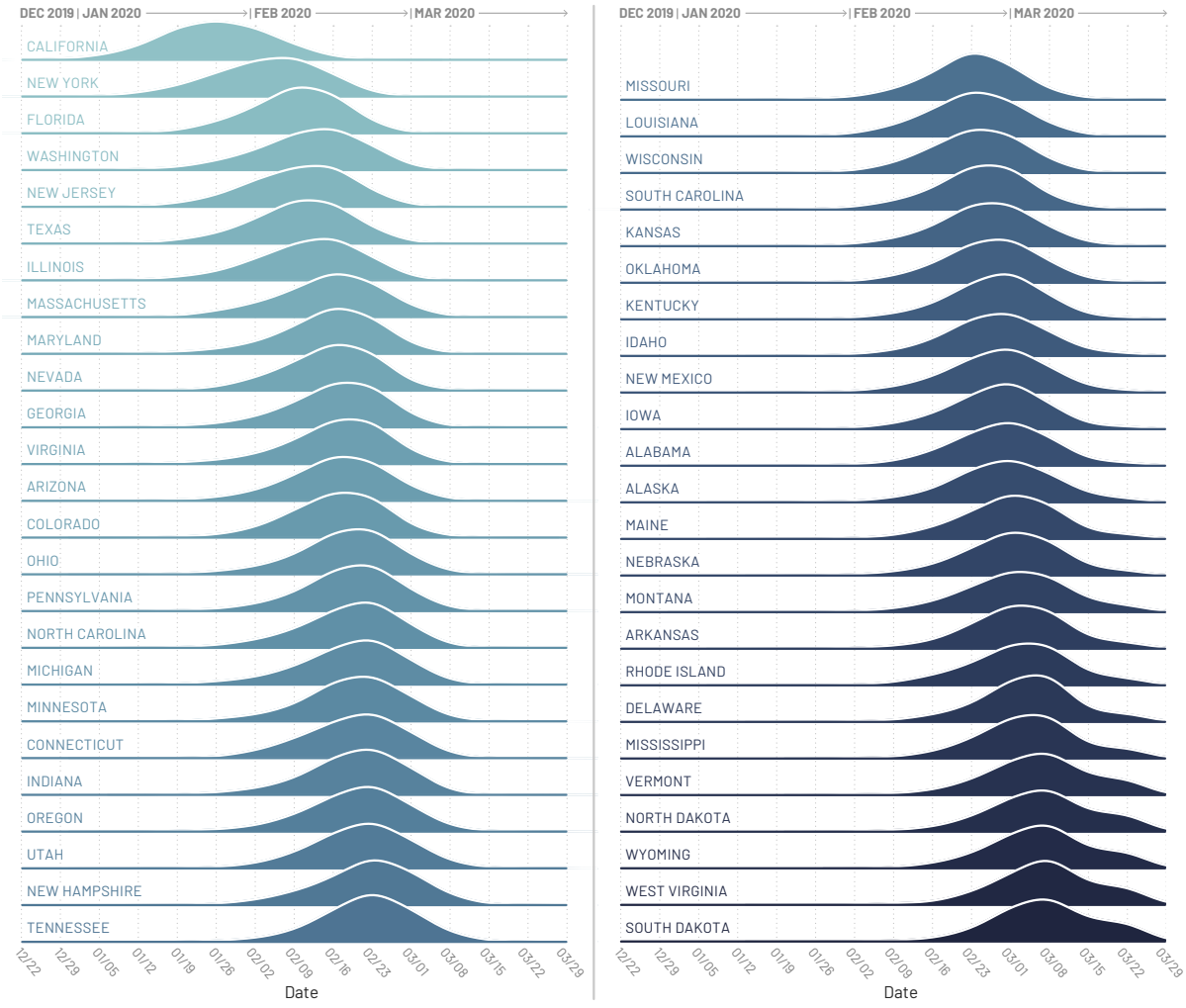

Figure 2: **Timing of the onset of local transmission.** Posterior distributions of the week when each state first reached 10 locally generated SARS-CoV-2 transmission events per day.

regions (e.g. Europe and Asia) while keeping the US and China separate and aggregating all the other countries (i.e. Others), in Fig. 3. It is worth clarifying that seeding importations are different from the actual number of times the virus has been introduced to each state with subsequent onward transmission. Even after local transmission has started, future importation events may give rise to additional onward transmission forming independently-introduced transmission lineages of the virus as seen in the United Kingdom (*31*). Statistics for importations through March 1st, 2020 are reported in the SM file.

Importations from mainland China may be relevant in seeding the epidemic in January (notice the width of the blue arrows from China for the first couple of states), but then play a small role in the COVID-19 expansion in the US because of the travel restrictions imposed to/from mainland China after January 23, 2020. About 42% and 15% of the virus introductions before the onset of local transmission in California and New York State, respectively, were from mainland China. Noticeably, the share of infection importations originating from Europe in California was four times smaller than those in New York State. Among the states for which the model estimates an early onset of local transmission before the third week of February (considering median values), European sources are statistically contributing 18% of SARS-Cov-2 importations for New Jersey, 15% for Massachusetts, 14% for Florida, and only 5% for Washington. While importations from mainland China contribute to early introductions of the virus in the US, our analysis highlights other potential sources of importations, such as Europe, where additional travel advisories and restrictions were implemented a month later at the end February and early March. Interestingly, the domestic importations are, across the board, statistically relevant in seeding the epidemic in many states. Among the states for which we estimated a late onset of local transmission (second half of February), domestic sources account for 85% of the virus introductions in Nebraska, 86% in New Mexico, 86% in Arkansas, and 95% in North Dakota.

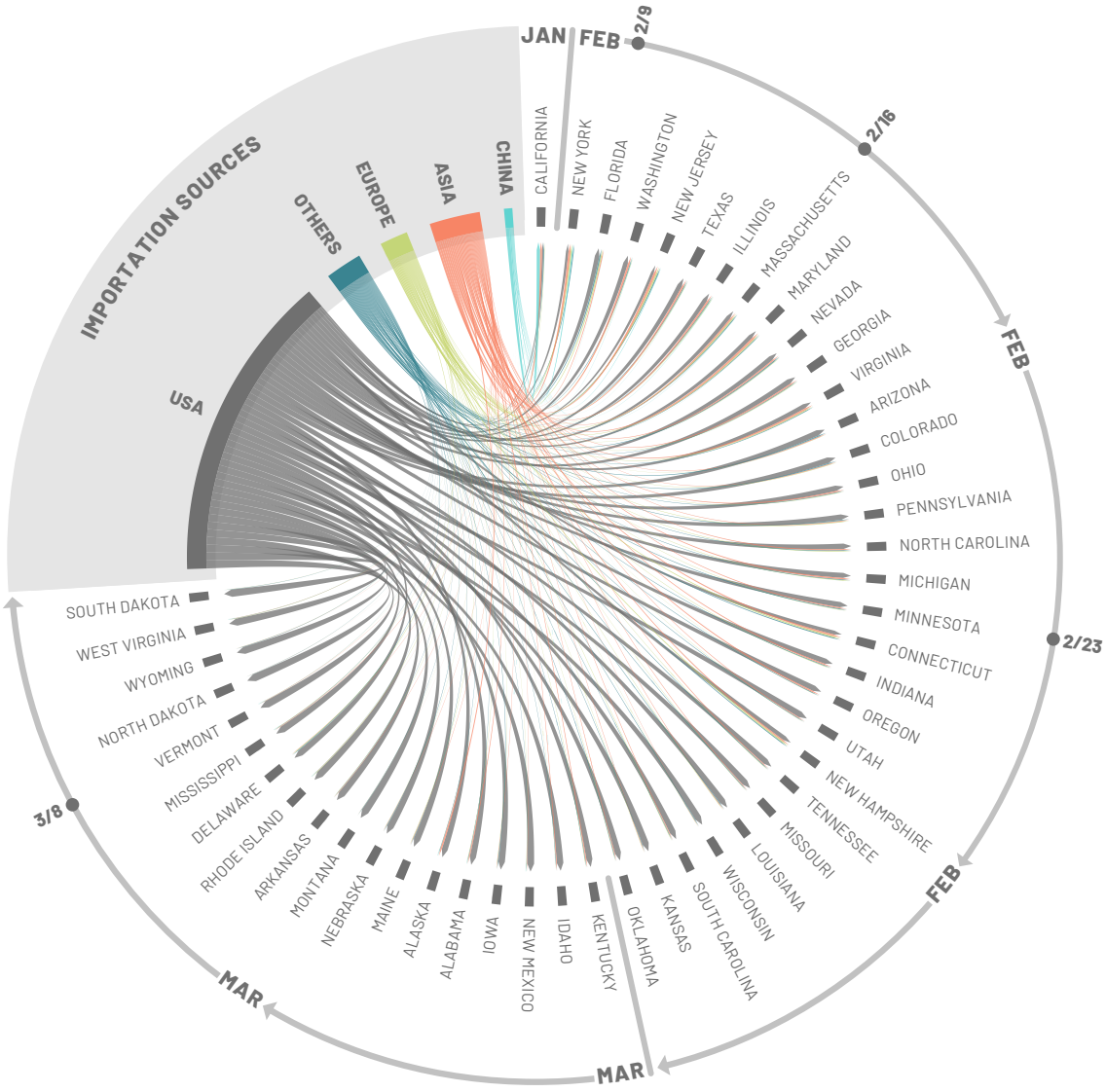

Figure 3: Importation sources. Each state is displayed in a clockwise order with respect to the start of the local outbreak (as seen in Fig. 2). Importation flows are directed and weighted. We normalize links considering the total in-flow for each state so that the sum of importations flows, for each state, is one. In the SM we report the complete list of countries contributing, as importation sources, in each group (i.e., geographical region).

Cryptic spreading phase

From late January to early March, SARS-CoV-2 had been spreading across the US mostly undetected. Cities such as Los Angeles, New York, Chicago, Seattle, and San Francisco experienced sustained local transmission beginning in the first half of February (Fig. 4A). In the time span of one additional month, through March 15, 2020, most large cities in the US had sizeable ongoing outbreaks. It is worth noting that California most likely generated 100 local infections by late January, around the same time as the adoption of draconian containment measures and international travel restrictions in mainland China. The model also allows us to estimate the possible occurrence of several COVID-19 deaths through March 1, 2020 (median 41 [90% CI $7 - 352$]). Although some states launched investigations in search of evidence that COVID-19 was the cause of death as far back as December 2019, it is likely that most deaths were missed because COVID-19 testing guidelines were based on travel history (*32*).

To provide a model consistency check with respect to surveillance data concerning the epidemic, we compared the model-based estimates with observed surveillance data. Particularly, early on in the epidemic, surveillance data were known to be highly unreliable because of underdetection. For each state we compared the order in which they surpassed 100 infections in the model and in the surveillance data gathered from the John Hopkins University Coronavirus Resource Centre (*34*). In Fig. 4B we plot the ordering for states and compute the Kendall rank correlation coefficient τ (see SM for details). The correlation is positive ($\tau = 0.66$, $p < 0.001$) indicating that, despite the detection and testing issues, the expected patterns of epidemic diffusion of the states is largely described by the model. As mentioned before, one possible major driver of the diffusion pattern is air traffic. We compare the ordering of states according to their air travel volume to their epidemic order as previously defined (Fig. 4C). We consider both national and international traffic, and find a positive correlation ($\tau = 0.65$ with $p < 0.001$)

between the epidemic ordering derived from surveillance data and air traffic, suggesting the passenger volume of both international and national traffic are key factors driving the early phases of outbreak across the country. Similar observations have been reported in China, where the initial spreading of the virus outside Hubei was strongly correlated with the traffic to/from the province (*35*). The correlation between the air travel ordering and the simulations is high ($\tau = 0.83$, $p < 0.001$) as actual airline data are used in the model to simulate the mobility of individuals (see SM). It is worth remarking that population size is also correlated with both the traveling flows ($\tau = 0.7$, $p < 0.001$) and the epidemic order of each state ($\tau = 0.66$, $p < 0.001$) as discussed in the SM. In our model it is not possible to exclude increased contacts in highly populated places before social distancing interventions and disentangle this effect from increased seeding due to the correlation between travel volume and population size. As yet another independent test for the model, in Fig. 4D we show the positive correlation between excess deaths, as estimated in (*36*), and the order in which the states reached the threshold of 10 daily transmissions ($\tau = 0.6$, $p < 0.001$).

The model is consistent with the picture emerging from the genomic epidemiology (*37–39*) of an early start of the COVID-19 epidemic in coastal states, followed by the propagation, dominated by the domestic traveling patterns, to the less globally connected regions in the US. The model suggests that COVID-19 spread across the US in about 7 weeks, and that by the end of February many states were experiencing sustained SARS-CoV-2 local transmission. The median time for the onset of local transmission in coastal states is as early as February 2020, and it raises the question of what would the unfolding of the epidemic have looked like in the case of an earlier warning and issuing of travel restrictions to/from China. For this reason we have performed simulations of a counterfactual scenario in which the timeline of all the travel restrictions to/from mainland China is shifted one week earlier compared to the actual one. In Fig. 4E, we show the distribution of probability that the US experiences sustained

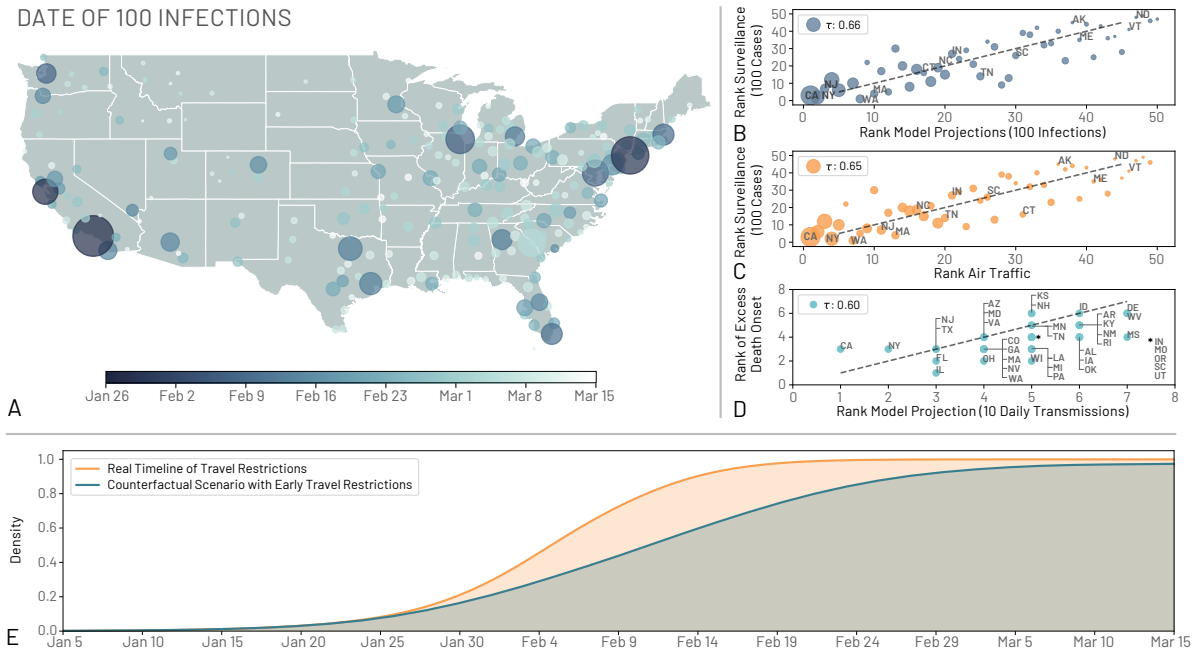

Figure 4: **Comparing model projections to surveillance and air traffic data.** (A) Map of the US showing the date where regions observed at least an average of 100 infections. (B) The correlation between the ordering of each state to reach 100 infections in the model projections and to reach 100 reported cases in the surveillance data. Correlation is computed considering the Kendall rank correlation coefficient, τ. (C) The correlation between the ordering of each state considering the time needed to reach 100 reported cases in the surveillance data and the ranking of the combined international and domestic air traffic. (D) The order correlation of the onset of excess deaths due to pneumonia and influenza and the order of the model projection for the date of 10 transmissions per day. Circle size in A, B, and C correspond to the population size of each state. (E) The cumulative distribution of the probability that the US reached 100 locally generated transmissions per day by a given day for both a scenario using the real timelines of travel restrictions and a counterfactual scenario where the travel restrictions to/from mainland China are shifted one week before. The median dates for the real and counterfactual timelines are February 5, 2020 and February 11, 2020 respectively.

local transmission by a given date in both the counterfactual scenario and the real timeline of interventions. We define sustained national transmission as 100 infections per day. Interventions implemented one week earlier amount to a proportional delay of the onset of local transmission in the continental US. In particular we find that in the real timeline intervention analysis a 50% cumulative probability of reaching the 100 locally generated transmissions mark is reached by February 5, 2020 compared to February 11, 2020 in the counterfactual scenario. This is in agreement with the evidence provided by several studies (*12, 40, 41*) that a considerable number of infections had already traveled from mainland China to international destinations before mid January, thus potentially seeding multiple epidemic outbreaks across the world, and leading to the international expansion of the COVID-19 epidemic, despite the mainland China travel ban. Our analysis however does not consider spontaneous behavioral changes that people might have adopted before the official national and local guidelines were annouced. While certainly some individuals might have taken precautions as a result of the news from China, evidence from surveys of public concern from several countries in Europe suggest that in late February only a very limited fraction of people considered COVID-19 as a concrete threat (*47*). Our model also does not contain any calibration or constraint on the trajectory of the outbreak in the months of March and April. We provide this analysis in the SM showing the consistency of the results.

Discussion

Our study characterizes the cryptic transmission phase during which SARS-CoV-2 spread largely undetected in the US. The results suggest that the first sustained local transmission chains took place as early as the end of January, and by the end of February the infection was spreading to many other national locations. This timeline is shifted several weeks ahead with respect to the detection of cases in surveillance data. This is consistent with the fact that in January and February no country had the capacity to do mass testing. Countries adopted a policy of test-

ing symptomatic individuals with a travel history linked to China, thus, generally missing the cryptic transmission occurring domestically. We find that the order in which the virus initially progressed across states according to our model is highly correlated with the official record. The model highlights that the geographical heterogenities in the observed spreading patterns are explained by the features of the air transportation network and population distributions. The results also indicate that the source of introduction of SARS-CoV-2 infections into the US changed substantially and rapidly through time. While early importations were from international sources, most introductions occurred during February and March 2020. Our results indicate that many states were seeded from domestic sources rather than international. The presented results could be of potential interest in combination with sequencing data of SARS-CoV-2 genomes in order to reconstruct in greater detail the early epidemic history of the US COVID-19 epidemic. The estimated SARS-Cov-2 importation pattern and the cryptic transmission phase dynamic are of potential use when planning and modelling public health policies in the context of international travel.

References

1. Centers for Disease Control and Prevention (CDC), "First Travel-related Case of 2019 Novel Coronavirus Detected in United States" (2020); https://www.cdc.gov/media/releases/2020/p0121-novel-coronavirus-travel-case.html.

2. Business Insider, "2nd case of the Wuhan coronavirus in the US was just confirmed in Chicago" (2020); https://www.businessinsider.fr/us/wuhan-coronavirus-second-confirmed-case-in-the-us-chicago-2020-1.

3. Arizona Department of Health Services, "Public Health Agencies Confirm 2019 Novel Coronavirus Case in Arizona" (2020); https://www.azdhs.gov/director/public-information-

office/index.php#news-release-012620.

4. Los Angeles County Public Health, "Public Health Confirms First Case of 2019 Novel Coronavirus in Los Angeles County" (2020); http://publichealth.lacounty.gov/phcommon/public/media/mediapubhpdetail.cfm?prid=2227.

5. Illinois Department of Public Health, "Second Illinois 2019 Novel Coronavirus Case Identified" (2020); http://www.dph.illinois.gov/news/second-illinois-2019-novel-coronavirus-case-identified.

6. CDC, "Updated Guidance on Evaluating and Testing Persons for Coronavirus Disease 2019 (COVID-19)"; https://emergency.cdc.gov/han/2020/han00429.asp.

7. D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J.J. Ramasco, A. Vespignani, Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*. **106**, 21484-21489 (2009).

8. D. Balcan, B. Gonçalves, H. Hu, J.J. Ramasco, V. Colizza, A. Vespignani, Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. *Journal of Computational Science* **1**, 132-145 (2010).

9. D. Balcan, H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J.J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck, V. Colizza, A. Vespignani, Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC medicine* **7**, 45 (2009).

10. Q. Zhang, K. Sun, M. Chinazzi, A. Pastore y Piontti, N.E. Dean, D.P. Rojas, S. Merler, D. Mistry, P. Poletti, L. Rossi, M. Bray, M.E. Halloran, I.M. Longini, A. Vespignani, Spread of Zika virus in the Americas. *Proceedings of the National Academy of Sciences* **114**, E4334 (2017).

11. A. P. y Piontti, N. Perra, L. Rossi, N. Samay, A. Vespignani, *Charting the Next Pandemic: Modeling Infectious Disease Spreading in the Data Science Age* (Springer, 2018).

12. M. Chinazzi, J.T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A.P. y Piontti, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, E.M. Halloran, I.M. Longini, A. Vespignani, The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020).

13. International Air Transportation Association. https://www.iata.org/.

14. Official Aviation Guide. https://www.oag.com/.

15. M. Gatto,E. Bertuzzo, L. Mari, S. Miccoli, L. Carraro, R. Casagrandi, A. Rinaldo, Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences* **117**, 10484 (2020).

16. S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, M. Lipsitch, Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period *Science* **368**, 860 (2020).

17. R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, J. Shaman, Jeffrey, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489 (2020).

18. J. T. Wu, K. Leung, G. M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet*. **395** 689-697 (2020).

19. S. Lai, N. W. Ruktanonchai, L. Zhou, O. Prosper, W. Luo, J. R. Floyd, A. Wesolowski, M. Santillana, C. Zhang, X. Du, H. Yu, and A. J. Tatem, Effect of non-pharmaceutical

interventions to contain COVID-19 in China. *Nature* (2020). doi.org/10.1038/s41586-020-2293-x

20. N. Imai, A. Cori, I. Dorigatti, M. Baguelin, C. A. Donnelly, S. Riley, N. M. Ferguson, "Report 3: Transmissibility of 2019-nCoV" (Imperial College London, 2020) www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-3-transmissibility-of-covid-19/.

21. D. Mistry, M. Litvinova, M. Chinazzi, L. Fumanelli, M. F. Gomes, S. A. Haque, Q.-H. Liu, K. Mu, X. Xiong, M. E. Halloran, I.M. Longini Jr., S. Merler, M. Ajelli, A. Vespignani. Inferring high-resolution human mixing patterns for disease modeling. *arXiv* [Preprint]. 25 February 2020. https://arxiv.org/abs/2003.01214

22. R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. T. Walker, H. Fu, A. Dighe, J. T. Griffin, M. Baguelin, S. Bhatia, A. Boonyasiri, A. Cori, Z. Cucunubá, R. FitzJohn, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, D. Laydon, G. Nedjati-Gilani, S. Riley, S. van Elsland, E. Volz, H. Wang, Y. Wang, X. Xi, C. A. Donnelly, A. C. Ghani, N. M. Ferguson. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases* (2020). https://doi.org/10.1016/S1473-3099(20)30243-7

23. A. Rambaut, "Preliminary phylogenetic analysis of 11 nCoV2019 genomes, 2020-01-19" (2020); http://virological.org/t/preliminary-phylogenetic-analysis-of-11-ncov2019-genomes-2020-01-19/329.

24. K. Anderson, "Estimates of the clock and TMRCA for 2019-nCoV based on 27 genomes" (2020); http://virological.org/t/clock-and-tmrca-based-on-27-genomes/347

25. T. Bedford, R. Neher, J. Hadfield, E.Hodcroft, M. Ilcisin, N. Müller, "Genomic analysis of nCoV spread. Situation report 2020-01-23" (2020); https://nextstrain.org/narratives/ncov/sit-rep/2020-01-23

26. M. Sunnåker, A.G. Busetto, E. Numminen, J. Corander, M. Foll, C. Dessimoz, Approximate Bayesian Computation, *PLoS Comput Biol*. **9**, e1002803 (2013).

27. P.M. Salazar, R. Niehus, A.R. Taylor, C.O. Buckee, M. Lipsitch. Using predicted imports of 2019-nCoV cases to determine locations that may not be identifying all imported cases. *medRxiv* 2020.02.04.20020495 [Preprint]. 11 February 2020. doi:10.1101/2020.02.04.20020495

28. R. Niehus, P. M. De Salazar, A. Taylor, M. Lipsitch, Using observational data to quantify bias of traveller-derived COVID-19 prevalence estimates in Wuhan, China. *The Lancet Infectious Diseases* (2020). https://doi.org/10.1016/S1473-3099(20)30229-2.

29. Global security index https://www.ghsindex.org/.

30. D. Cereda, M. Tirani, F. Rovida, V. Demicheli, M. Ajelli, P. Poletti, F. Trentini, G. Guzzetta, V. Marziano, A. Barone, M. Magoni, S. Deandrea, G. Diurno, M. Lombardo, M. Faccini, A. Pan, R. Bruno, E. Pariani, G. Grasselli, A. Piatti, M. Gramegna, F. Baldanti, A. Melegaro,S. Merler The early phase of the COVID-19 outbreak in Lombardy, Italy, *arXiv* [Preprint]. https://arxiv.org/abs/2003.09320 (2020)

31. O. Pybus, A. Rambaut, L. du Plessis, A.E. Zarebski, M. U. G. Kraemer, J. Raghwani, B. Gutiérrez, V. Hill, J. McCrone, R. Colquhoun, B. Jackson, Á. O'Toole, J. Ashworth Preliminary analysis of SARS-CoV-2 importation establishment of UK transmission lineages, https://www.oxfordmartin.ox.ac.uk/publications/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/ (2020).

32. Santa Clara County Public Health, "County of Santa Clara Identifies Three Additional Early COVID-19 Deaths", https://www.sccgov.org/sites/covid19/Pages/press-release-04-21-20-early.aspx (2020).

33. CDC Cases, Data Surveillance https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html.

34. John Hopkins University Coronavirus Resource Centre https://coronavirus.jhu.edu/.

35. M. U. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, Open COVID-19 Data Working Group, L. du Plessis, N. R. Faria, R. Li, W. P. Hanage, J. S. Brownstein, M. Layan, A. Vespignani, H. Tian, C. Dye, O. G. Pybus, S. V. Scarpino. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, 493-497 (2020).

36. D. Weinberger, T. Cohen, F. Crawford, F. Mostashari, D. Olson, V. E. Pitzer, N. G. Reich, M. Russi, L. Simonsen, A. Watkins, C. Viboud. Estimation of Excess Deaths Associated With the COVID-19 Pandemic in the United States, March to May 2020. *JAMA Intern Med.* 1 July 2020 doi:10.1001/jamainternmed.2020.3391.

37. "Nextstrain: Real-time tracking of pathogen evolution", https://nextstrain.org/ (2020).

38. M. Worobey, J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, P. Lemey. *bioRxiv* 10.1101/2020.05.21.109322 [Preprint]. 23 March 2020. doi:10.1101/2020.05.21.109322

39. A. S. Gonzalez-Reiche, M. M. Hernandez, M. Sullivan, B. Ciferri, H. Alshammary, A. Obla, S. Fabre, G. Kleiner, J. Polanco, Z. Khan, B. Alburquerque, A. van de Guchte, J. Dutta, N. Francoeur, B. S. Melo, I. Oussenko, G. Deikus, J. Soto, S. H. Sridhar, Y.-C.

Wang, K. Twyman, A. Kasarskis, D. R. Altman, M. Smith, R. Sebra, J. Aberg, F. Krammer, A. Garcia-Sarstre, M. Luksza, G. Patel, A. Paniz-Mondolfi, M. Gitman, E. M. Sordillo, V. Simon, H. van Bakel. Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* (2020). doi:10.1126/science.abc1917.

40. A. Adiga, S. Venkatramanan, J. Schlitt, A. Peddireddy, A. Dickerman, A. Bura, A. Warren, B. D. Klahn, C. Mao, D. Xie, D. Machi, E. Raymond, F. Meng, G. Barrow, H. Mortveit, J. Chen, J. Walke, J. Goldstein, M. L. Wilson, M. Orr, P. Porebski, P. A. Telionis, R. Beckman, S. Hoops, S. Eubank, Y. Y. Baek, B. Lewis, M. Marathe, C. Barrett. *medRxiv* 10.1101/2020.02.20.20025882 [Preprint]. 2 March 2020. doi:10.1101/2020.02.20.20025882

41. C. R. Wells, P. Sah, S. M. Moghadas, A. Pandey, A. Shoukat, Y. Wang, Z. Wang, L. A. Meyers, B. H. Singer, and A. P. Galvani. Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proceedings of the National Academy of Sciences* **117**, 7504 (2020).

42. L. Di Domenico, G. Pullano, C. E. Sabbatini, P.-Y. Boëlle, V. Colizza. Expected impact of lockdown in Île-de-France and possible exit strategies. *medRxiv* 10.1101/2020.04.13.20063933 [Preprint]. 17 April 2020. doi:10.1101/2020.04.13.20063933

43. A. Aleta, D. Martin-Corral, A. Pastore y Piontti, M. Ajelli, M. Litvinova, M. Chinazzi, N. E. Dean, M. E. Halloran, I. M. Longini, S. Merler, A. Pentland, A. Vespignani, E. Moro, and Y. Moreno. Modeling the impact of social distancing, testing, contact tracing and household quarantine on second-wave scenarios of the COVID-19 epidemic. *medArxiv* 10.1101/2020.05.06.20092841 [Preprint]. 18 May 2020. doi:10.1101/2020.05.06.20092841.

44. J. Zhang, M. Litvinova, Y. Liang, Y. Wang, W. Wang, S. Zhao, Q. Wu, S. Merler, C. Viboud, A. Vespignani, M. Ajelli, and H. Yu. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science* (2020). doi:10.1126/science.abb8001.

45. N. G. Davies, P. Klepac, Y. Liu, K. Prem, M. Jit, CMMID COV ID-19 working group, R. M. Eggo, Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nature Medicine* (2020). https://doi.org/10.1038/s41591-020-0962-9

46. Q. Bi, Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S. A. Truelove, T. Zhang, W. Gao, C. Cheng, X. Tang, X. Wu, Y. Wu, B. Sun, S. Huang, Y. Sun, J. Zhang, T. Ma, J. Lessler, T. Feng, Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *The Lancet Infectious Diseases* (2020). https://doi.org/10.1016/S1473-3099(20)30287-5.

47. J. Raude, M. Debin, C. Souty, C. Guerrisi, C. Turbelin, A. Falchi, I. Bonmarin, D. Paolotti, Y. Moreno, C. Obi, J. Duggan, A. Wisniak, A. Flahault, T. Blanchon, V. Colizza, Are people excessively pessimistic about the risk of coronavirus infection? *medArxiv*. 10.31234/osf.io/364qj [Preprint]. 5 March 2020. doi:10.31234/osf.io/364qj

Acknowledgements

M.E.H. acknowledges the support of the MIDAS-U54GM111274. S.M. and M.A. acknowledge support from the EU H2020 MOOD project. C.G. and L.R. acknowledge support from the EU H2020 Icarus project. M.C. and A.V. acknowledge support from Google Cloud and Google Cloud Research Credits program to fund this project. The findings and conclusions in this study are those of the authors and do not necessarily represent the official position of the funding agencies, the National Institutes of Health, or the U.S. Department of Health and Human Services.

Author Contributions

Author contributions: J.T.D., M.C., N.P. and A.V. designed research; M.C., J.T.D. ,N.P., M.A., C.G., M.L., S.M., A.P.P, K.M., L.R., K.S., C.V, X.X., M.E.H., I.M.L., and A.V. performed research; M.C., J.T.D., N.P., A.P.P., K.M. and A.V. analyzed data; and M.C., J.T.D., N.P., M.A., C.G., M.L., S.M., A.P.P, K.M., L.R., K.S., C.V, X.X., M.E.H., I.M.L., and A.V. wrote and edited the paper.

Competing Interests

M.E.H. reports grants from National Institute of General Medical Sciences, during the conduct of the study; A.V. reports grants and personal fees from Metabiota inc., outside the submitted work; M.C. and A.P.P. report grants from Metabiota inc., outside the submitted work. No other relationships or activities that could appear to have influenced the submitted work.

Data and materials availability

Proprietary airline data are commercially available from Official Aviation Guide (OAG) and IATA databases. The GLEAM model is publicly available at http://www.gleamviz.org/.