nature

https://doi.org/10.1038/s41586-020-2284-y

Accelerated Article Preview

Population flow drives spatio-temporal distribution of COVID-19 in China

Received: 18 February 2020

Accepted: 21 April 2020

Accelerated Article Preview Published online 29 April 2020

Cite this article as: Jia, J. S. et al. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* https://doi.org/10.1038/s41586-020-2284-y (2020).

Jayson S. Jia, Xin Lu, Yun Yuan, Ge Xu, Jianmin Jia & Nicholas A. Christakis

Article

Population flow drives spatio-temporal distribution of COVID-19 in China

https://doi.org/10.1038/s41586-020-2284-y

Received: 18 February 2020

Accepted: 21 April 2020

Published online: 29 April 2020

Jayson S. Jia[1], Xin Lu[2,3], Yun Yuan[4], Ge Xu[5], Jianmin Jia[6,7]✉ & Nicholas A. Christakis[8]

Sudden, large-scale, and diffuse human migration can amplify localized outbreaks into widespread epidemics.[1–4] Rapid and accurate tracking of aggregate population flows may therefore be epidemiologically informative. Here, we use mobile-phone-data-based counts of 11,478,484 people egressing or transiting through the prefecture of Wuhan between 1 January and 24 January 2020 as they moved to 296 prefectures throughout China. First, we document the efficacy of quarantine in ceasing movement. Second, we show that the distribution of population outflow from Wuhan accurately predicts the relative frequency and geographic distribution of COVID-19 infections through February 19, 2020, across all of China. Third, we develop a spatio-temporal "risk source" model that leverages population flow data (which operationalizes risk emanating from epidemic epicenters) to not only forecast confirmed cases, but also to identify high-transmission-risk locales at an early stage. Fourth, we use this risk source model to statistically derive the geographic spread of COVID-19 and the growth pattern based on the population outflow from Wuhan; the model yields a benchmark trend and an index for assessing COVID-19 community transmission risk over time for different locations. This approach can be used by policy-makers in any nation with available data to make rapid and accurate risk assessments and to plan allocation of limited resources ahead of ongoing outbreaks.

Tracking population flows is especially exigent in the context of China's COVID-19 outbreak, which began in Wuhan (a prefecture-city in the province of Hubei) in the run-up to Chinese Lunar New Year eve on January 24, 2020 with its annual *chunyun* mass migration (which can involve as many as 3 billion trips). The potential scale and range of the outbreak's diffusion was particularly alarming given Wuhan's position as a central hub in China's rail and aviation networks and given the severity of COVID-19.

We used nationwide mobile phone data to track population outflow from Wuhan and linked this to COVID-19 infection counts by location – at the prefecture level. Our data include 296 prefectures in 31 provinces and regions in China (average population 4.40 million, 94.07% of China's population). Mobile phone geolocation data, which can reliably quantify human movement, provide precise, verifiable, and real-time information.[5–11] We conceptualize epidemiological morbidity and mortality as a function of human population movement from a disease origin. We thus normalize disease risk by population inflow from Wuhan rather than the size of local population.

Our approach differs from prior work linking individual mobility and disease spread[1–4,12] in terms of: our use of real-time data about actual movement; our focus on aggregate population flows rather than individual tracking; and our particular modeling approach. That is, other recent research on COVID-19 has used *historical* population flow data (e.g., previous years' *chunyun* migrations) to estimate case exportation during the current outbreak.[14–18] But the benefits of observing rather than estimating population movements are substantial since inaccurate predictions can have important consequences for policy-making: under-reaction can result in disease spread, and over-reaction can lead to medically, socially, and economically inefficient policies. Moreover, distinct from prior approaches to epidemiological modelling,[12–18] we take advantage of detailed data about population flow originating at the source of the outbreak to develop a population-flow-based "risk source" model to test the extent to which population flow data can capture the spatio-temporal dynamics of the spread of the SARS-CoV-2 virus.

To measure total aggregate population outflow from Wuhan prior to its quarantine on January 23, 2020, we used country-wide data, provided by a major national carrier, tracking all movement out of Wuhan between January 1 and January 24, 2020. The symptom onset of the first recorded case in Wuhan was December 1, 2019; by February 19, the end of our study period, 74,279 infection cases had been verified in China.[19–22] Our time period includes the time that news about the outbreak initially appeared (on December 31, 2019 and January 9, 2020) and the annual Lunar New Year migration (which culminated on January 24, 2020). The dataset included any mobile phone user who had spent at least 2 hours in Wuhan during this period, and it tracked the total daily flow of such individuals to all other prefectures throughout

[1]Faculty of Business and Economics, The University of Hong Kong, Hong Kong SAR, China. [2]College of Systems Engineering, National University of Defense Technology, Changsha, China. [3]Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden. [4]School of Economics and Management, Southwest Jiaotong University, Chengdu, China. [5]School of Management, Hunan University of Technology and Business, Changsha, China. [6]Shenzhen Finance Institute, School of Management and Economics, The Chinese University of Hong Kong, Shenzhen, China. [7]Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China. [8]Yale Institute for Network Science, Yale University, New Haven, CT, USA. ✉e-mail: jmjia@cuhk.edu.cn

China. Locations were detected when users simply had their phones on. The dataset includes two measures of population outflow: the carrier's own customer count and their extrapolated count of total population movement. We use the latter in our primary analyses and the former as a robustness check (see Supplementary Information).

We defined population flow as the total aggregate count of people entering any given prefecture from Wuhan during the whole observation period (January 1 to 24). Since Wuhan (population 11.08 million in 2018) is a major transportation hub, many of these people were through travelers rather than residents. The definition is also weighted by number of transits through Wuhan since some people may have entered and exited Wuhan on several occasions in January (especially if they lived in neighboring prefectures). This can be thought of as a linear weighting of additional infection and transmission risk from repeated transits. There were 11,478,484 counts of movements from Wuhan: 8,685,007 to other prefectures within Hubei and 2,793,477 to prefectures in other provinces.

Key dates during this period were January 24, Lunar New Year's Eve (outbound holiday travel is typically completed before this evening), and January 23, when Wuhan was quarantined. We observed the efficacy of the quarantine (Fig. 1b, c), which was manifested in a 52% (38%) drop of inter- (intra-) provincial population outflow on January 23 compared to January 22 (when there were 546,324 and 141,208 counts of intra- and extra- provincial travel, respectively), and a further of 94% (84%) drop on January 24 compared to January 23. With the imposition of the quarantine – first with respect to Wuhan (and two neighboring prefectures) at 10 a.m. on January 23, and then with respect to 12 other prefectures in Hubei by the end of the day on January 24 – population outflow from Wuhan almost completely stopped (the average daily outflow thereafter was just 1,087 people to all prefectures outside of Hubei, probably government workers).

We combined the population flow dataset with the count and geographical location of COVID-19 confirmed cases nationwide (Fig. 1), which used consistent and stringently enforced case ascertainment during this period. As of February 19, 2020, there were 74,279 infection cases in China; 29,867 cases occurred outside of Wuhan; and there were 2,006 fatalities.[22]

Population flow from Wuhan may be hypothesized to export the virus to other locations, where it causes local outbreaks (i.e., either by importation or "community transmission"[19–22]). And indeed, we find a strong correlation between total population flow and number of infections in each prefecture (Fig. 2a, b). Consistent with our hypothesis, the cumulative number of infections is highly correlated with aggregate population outflow from Wuhan from January 1 to 24, and the correlation increases over time from $r = 0.522$ on January 24 to 0.919 on February 5, and further to 0.952 on February 19 ($p < .001$ for all) (Fig. 2a, b, c). Since there is little travel throughout the country during this period, the population outflow variable is comparable to a lagged variable in a time series. The correlation exhibited the same robust pattern even when using different time windows of population outflow (Extended Fig. 1). The correlation between population outflow from Hubei province (excluding Wuhan itself) and number of infections in each prefecture (Fig. 2c) followed a similar pattern but was substantially weaker, $r = 0.365$ on January 24 to 0.583 on February 19.

For completeness, we compared the predictive strength of aggregate population outflow to certain other factors – such as the relative frequency of Baidu search for virus-related terms in each prefecture (e.g., novel coronavirus, flu, SARS, atypical pneumonia, surgical mask),[23–25] each prefecture's GDP and population, and also other movement variables. Each of these factors became *less* predictive of local outbreak size over time, either for cumulative or daily reported cases (Fig. 2c, d, Extended Fig.2-3).

We also evaluated a gravity model.[4,13] Gravity models were originally developed to model flow volumes or other interactions between geographical areas based simply on distance between two locales and their populations. Here, we use a special case of the gravity model with only the "recipient" prefecture's population variable since Wuhan is always the "donor" and thus a constant value (Supplementary Information 4.1). This model (with a significantly negative parameter for distance) predicts the high quantity of travel from Wuhan to other prefectures in Hubei and to geographically proximate provinces (Fig. 1). But it does not explain the high traffic of population outflow to more distant coastal cities. That outflow does not strictly follow a gravity model is not surprising given the rationales for *chunyun* migration patterns, which are primarily based on social connections.[8,26]

We also tested a gravity model to predict the infection count. Although "recipient" population size and distance were significant predictors ($p < .001$), a mediation analysis shows that population flow from Wuhan mediates the effect of distance. Fig. 2c and 2d intuitively illustrate why this is the case. Aggregate population flow from Wuhan exhibits a high and progressively stronger correlation with infection prevalence in destination locations over time. In contrast, the predictive strength of prefecture's distance from Wuhan, population size, and GDP (an alternative source of "gravity") declines over time. There is no advantage to estimating population flow and to estimating infection spread using estimated population flow when actual population flow is observable, as in our case.

Next, we use two sets of models – one cross-sectional and the other dynamic – to statistically model and benchmark the extent to which aggregate population outflow from Wuhan predicts the spread and distribution of COVID-19 infections across China. We develop what we call a "risk source" model that leverages observed population flow data to operationalize the risk emanating from the epidemic source.

We first modeled the effect of outflow on infection by using the following multiplicative exponential model:

$$y_i = c \prod_{j=1}^{m} e^{\beta_j x_{ji}} e^{\sum_{k=1}^{n} \lambda_k I_{ik}} \quad (1)$$

where y_i is the number of the cumulative (or daily) confirmed cases in prefecture i (depending on the model); x_{1i} is cumulative population outflow from Wuhan to prefecture i from January 1 to 24; x_{2i} is the GDP of prefecture i; x_{3i} is the population size of prefecture i; m is the number of variables included; and c and βi are parameters to estimate. And λ_k is the fixed effect for province k; n is the number of prefectures considered in the analysis; I_{ik} is a dummy for prefecture i and $I_{ik} = 1$, if $i \in k$ (prefecture i belongs to province k), otherwise $I_{ik} = 0$. (See Supplementary Information for more details.)

We applied a supervised machine learning approach with confirmed cases as the dependent variable to estimate the parameters of a model with aggregate Wuhan population outflow from January 1-24 as the sole variable ($R^2 = 0.772$ on January 24 to 0.946 on February 19) and a model with population size and GDP as co-variates ($R^2 = 0.809$ on January 24 to 0.967 on February 19) (Supplementary Tables 1-2). Although these additional variables improve fit, the parameter for population flow from Wuhan becomes increasingly dominant, while a prefecture's GDP and population become increasingly less predictive over time. Overall, the models' performance continuously improved as more infection cases were confirmed, suggesting that the spreading pattern of the virus gradually converges to the distribution of the population outflow from Wuhan to other prefectures in China. As a robustness check, we evaluate a model using daily confirmed cases and find consistent results (Supplementary Tables 3-4).

The logic behind this convergence over time, as well as the model's predictive strength, is that population flow from Wuhan to other prefectures fundamentally determines the eventual distribution of total infections in China. During the earliest phase of the outbreak, before the quarantine of Wuhan, there was a relative lack of awareness of the virus and few countermeasures preventing its spread. SARS-CoV-2 should thus have spread relatively randomly across the

entire prefecture of Wuhan; that is, our results imply that the number of infected people was uniformly distributed (statistically speaking) in the population outflowing from Wuhan into different prefectures across the country.

Using the daily predicted cases in model (1), we are also able to calculate a daily risk score for prefectures based on the difference between their predicted and confirmed cases on any given date (see Supplementary Information). A higher-than-expected level of infection suggests more community transmission (i.e., "underperforming" compared to the benchmark derived from the outflow population from Wuhan). On the other hand, "over-performing" prefectures, with fewer cases than expected are also noteworthy, since they could have implemented highly successful public health measures (or be prone to inaccurate data reporting). Extended Fig. 4 identifies prefectures with transmission risk index values over the upper bound of the 90% confidence interval on January 29, for example, and this was indeed associated with imminent quarantine. The predictive strength of aggregate population flow from Wuhan and the overall fit of model (1) over time can also act as an early warning index of an epidemiological transition; they reflect the degree to which imported infections are dominant at any point in time. If model strength declines significantly at any location, this may indicate that community transmission may be overtaking imported cases.

We next developed a spatio-temporal model to explore changes in distribution and growth of COVID-19 across all prefectures over time (rather than on individual dates) (Supplementary Information 3.2). We use a Cox proportional hazards framework and replace the constant scaling parameter of model (1) with a time-varying hazard rate function $\lambda_0(t)$,which typically has an S-shaped property (e.g., logistic, generalized logistic, or Gompertz functions[27,28]) that epidemics typically follow:

$$\lambda(t|x_i) = \lambda_0(t)\left(\prod_{j=1}^{m} e^{\beta_j x_{ji}}\right)e^{\sum_{k=1}^{n}\lambda_k l_{ik}} \qquad (2)$$

where $\lambda(t|x_i)$ is the hazard function describing the number of cumulative confirmed cases at time t given population outflow from Wuhan to prefecture i, and other variables $x_i=\{x_{1i}, x_{2i}, \ldots x_{mi}\}$ are the realized values of the covariates for prefecture i; and the other notation is the same as model (1).

This model extends our risk source model to a dynamic context; it incorporates all infection cases across all locales and dates to statistically derive the COVID-19 epidemic curve and growth pattern across China. We used the same machine learning method as before to estimate the parameters (see Supplementary Information). When using only the single variable of total population outflow from Wuhan (from January 1-24) to each other prefecture, we observe $R^2 = 0.927$ for the exponential-logistic model (Fig. 3a); and the inclusion of local population and GDP increases R^2 to 0.957 (alternate models are in Supplementary Table 5).

We use a similar logic as before in contrasting expected and observed outcomes to gauge epidemiological risk. Here, model predictions serve as reference patterns across time (Extended Fig. 5, 6). The differences in the growth trends between predicted and confirmed cases can signal higher levels of SARS-CoV-2 community transmission. We use the integral of the differences over time to create a total transmission risk index (normalized by subtracting the mean and dividing by the standard deviation) and identify a list of prefectures above and below the 90% confidence interval (Extended Fig. 7, Supplementary Table 11). Indeed, our model identifies a list of statistically significant "underperformers"; in most of these cases, we observed the subsequent imposition of quarantine (see the Supplementary Information, including Supplementary Table 12 and Extended Fig. 8 and 9). On the other hand, prefectures with lower trends than expected might have had more successful public health measures. Fig. 3b depicts the dynamic shifts in risk index score for selected prefectures, which allows monitoring which prefectures performed better in controlling transmission risk over time.

In sum, using detailed mobile phone geolocation data to compute aggregate population movements, we track the transit of people from Wuhan to the rest of China through January 24, 2020. The geographic flow of people anticipates the subsequent location, intensity, and timing of outbreaks in the rest of China through February 19, 2020. These data outperform other measures, such as population size, wealth, or distance from the risk source. We modeled the epidemic curves of COVID-19 across different locales using population flows and showed that deviations from model predictions served as tools to detect the burden of community transmission.

The logic of our population-flow-based "risk source" model differs from classic epidemiological models that rely on assumptions regarding population mixing, population compartment sizes, and viral properties. By assuming that risk arises from human population movements, our "risk source" model is able to parsimoniously capture the distribution of the epidemic. The model has several advantages: it makes no assumptions regarding travel patterns or effective distance effects; allows for non-linear estimations; generates a non-arbitrary, source-linked risk score; and is easily adapted to other empirical contexts. Importantly, the multiplicative functional form can also accommodate multiple risk sources – for example in countries where there are multiple disease epicenters. As an example, we evaluated the distinct impact of population flow from Hubei (excluding Wuhan) as an alternative risk source in our models, and indeed find that it had little impact on COVID-19 spread and growth in the country (Supplementary Tables 6 and 10).

We have focused on the relative strength of the outbreak in each area, rather than the absolute number of cases, though one can predict the number of cases by using reported data to calibrate the parameters of the model. A key contribution of our approach is to robustly characterize the structure or relative distribution of cases across different geographic areas and over time, which is driven fundamentally by the cumulative outflow from Wuhan. Moreover, another benefit is that non-systematic inaccuracy of COVID-19 case-finding is relatively unimportant as long as we capture the *distribution* of population flow accurately over time, which we do.

Our approach is generalizable to any dataset that captures population movements (e.g., train ticketing or car tolling data). This method can also be implemented in a live fashion (if suitable data are available) to facilitate policy decisions – for example the allocation of resources and manpower across specific geographic locales based on the predicted strength of the epidemic. This could also yield a dynamic performance metric when contrasted against real-time reports of infections, and, as we show, identify which areas have higher virus transmission risk or more effective measures.

Other techniques to forecast the levels of an epidemic in defined populations in advance have, of course, been proposed – whether the use of online searching[23–25] or the use of network sensors (i.e., the monitoring of people who are at heightened risk for falling ill given their network position).[29] Our approach relies on data regarding population flow. Indeed, historical (i.e., baseline) information about population flows – undisturbed by the imposition of quarantines or by publicity regarding outbreaks, both of which happened here – could also be valuable to public health experts and government officials when new outbreaks occur.

When people move, they take contagious diseases with them. Their movements are thus a harbinger of the future status of an epidemic, and this offers the prospect of using data-analytic techniques to control an epidemic before it strikes too hard.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2284-y.

1. Colizza, V., Barrat, A., Barthélemy, M. & Vespignani, A. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc. Natl. Acad. Sci. USA* **103**, 2015-2020 (2006).
2. Halloran, M. E., Vespignani, A., Bharti, N., Feldstein, L. R., Alexander, K. A., Ferrari, M., & Del Valle, S. Y. Ebola: mobility data. *Science* **346**, 433-433 (2014).
3. Brockmann, D. & Helbing, D. The Hidden Geometry of Complex, Network-Driven Contagion Phenomena. *Science* **342**, 1337-1342 (2013)
4. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J. & Vespignani, A. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. USA* **106**, 21484-21489 (2009).
5. Brockmann, D., Hufnagel, L. & Geisel, T. The scaling laws of human travel. *Nature* **439**, 462–465 (2006).
6. González, M. C., Hidalgo, C. A. & Barabási, A. L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
7. Onnela, J. P., Arbesman, S., Gonzalez, M. C., Barabási, A. L. & Christakis, N.A. Geographic Constraints on Social Network Groups. *Plos One* **6**, e16939 (2011)
8. Lu, X., Bengtsson, L. & Holme, P. Predictability of population displacement after the 2010 Haiti earthquake. *Proc. Natl. Acad. Sci. USA* **109**, 11576-11581 (2012).
9. Yan, X., Wang, W., Gao, Z. & Lai, Y. Universal model of individual and population mobility on diverse spatial scales, *Nat. Commun*. **8**, 1639 (2017).
10. Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J. C., Huens, E., Van Dooren, P., & Blondel, V. D. Exploring the mobility of mobile phone users. *Physica A* **392**, 1459-1473 (2013).
11. Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. Quantifying the impact of human mobility on malaria. *Science* **338**, 267-70 (2012).
12. Adda, J. Economic activity and the spread of viral diseases: Evidence from high frequency data. *Q. J. Econ*. **131**, 891-941 (2016).
13. Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., & Grenfell, B. T. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312**, 447-451 (2006).
14. Wu, J. T., Leung, K., & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet*. **395**, 689-697 (2020).
15. Wu, J.T., Leung, K., Bushman, M. et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat. Med*. (2020).
16. Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., ... & Viboud, C. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* (2020).
17. Du, Z., Wang, L., Cauchemez, S., Xu, X., Wang, X., Cowling, B. J. & Meyers, L. A. Risk of 2019 novel coronavirus importations throughout China prior to the Wuhan. *Lancet* **361**, 1761-6 (2020).
18. Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science* (2020).
19. Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., ... & Yuan, M. L. A new coronavirus associated with human respiratory disease in China. *Nature*, 1-8 (2020).
20. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., ... & Niu, P. A novel coronavirus from patients with pneumonia in China, 2019. *New Engl. J. Med*. (2020).
21. Chan, J. F. W., Yuan, S., Kok, K. H., To, K. K. W., Chu, H., Yang, J., ... & Tsoi, H. W. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*. (2020).
22. China Center for Disease Control and Prevention. (2020).
23. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012-1014 (2009).
24. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google Flu: traps in big data analysis. *Science* **343**, 1203-1205 (2014).
25. Viboud, C. & Vespignani, A. The future of influenza forecasts. *Proc. Natl. Acad. Sci. USA*, **116**, 2802-2804 (2019).
26. Massey, D. S. & García España, F. The Social Process of International Migration, *Science* **237**, 733-738 (1987).
27. Bürger, R., Chowell, G., & Lara-Díıaz, L. Comparative analysis of phenomenological growth models applied to epidemic outbreaks. *Math Biosci Eng*. **6**, 4250–4273 (2019).
28. Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J. M., Yan, P., & Chowell, G. Short-term Forecasts of the COVID-19 Epidemic in Guangdong and Zhejiang, China: February 13-23, 2020. *J Clin Med* **9**, 596 (2020).
29. Christakis, N. A. & Fowler, J. H. Social network sensors for early detection of contagious outbreaks. *PloS One* **5**, e12948 (2010).

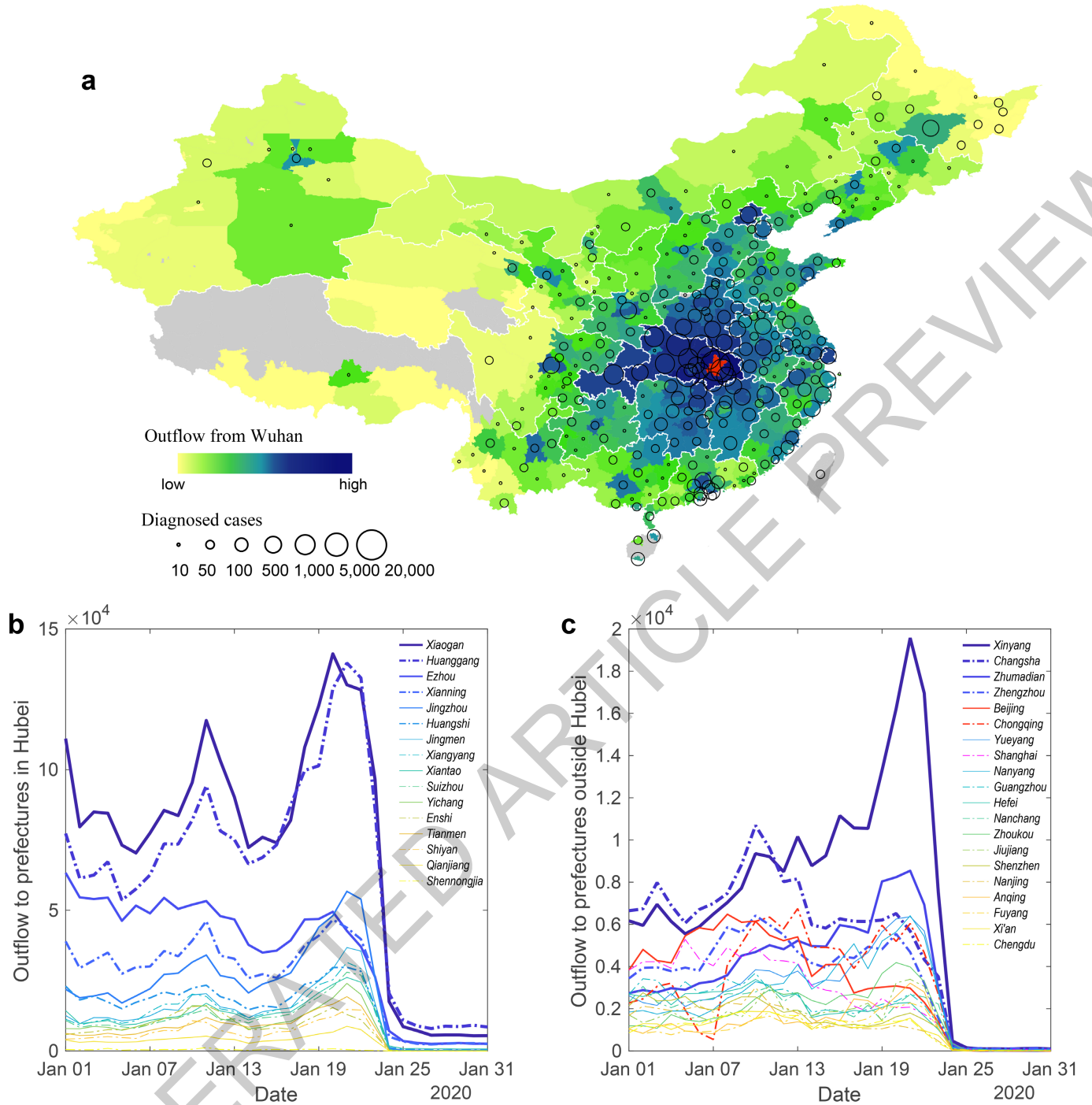


Fig. 1 | Geographical distribution of population outflow and confirmed COVID-19 cases as of February 19, 2020. a, there is a high overlap between the geographic distribution of aggregate population outflow from Wuhan through January 24, 2020 (in red) and confirmed cases of COVID-19 in other Chinese prefectures (N=296, map source: National Catalogue Service For Geographic Information). Gray areas lack population outflow data. **b, c**, During what is historically the peak period for outbound Lunar New Year holiday travel, total population outflow from Wuhan to other parts of Hubei (**b**) is approximately 6.5 times population outflow to outside provinces (**c**). Upon implementation of the quarantine at 10:00 a.m. on January 23, 2020, population outflow from Wuhan became minimal, except to the adjacent prefectures (**b**). In **b**, the first peak possibly corresponds to the start of winter break of (roughly one million) college students in Wuhan and the second peak is outbound *chunyun* travel.

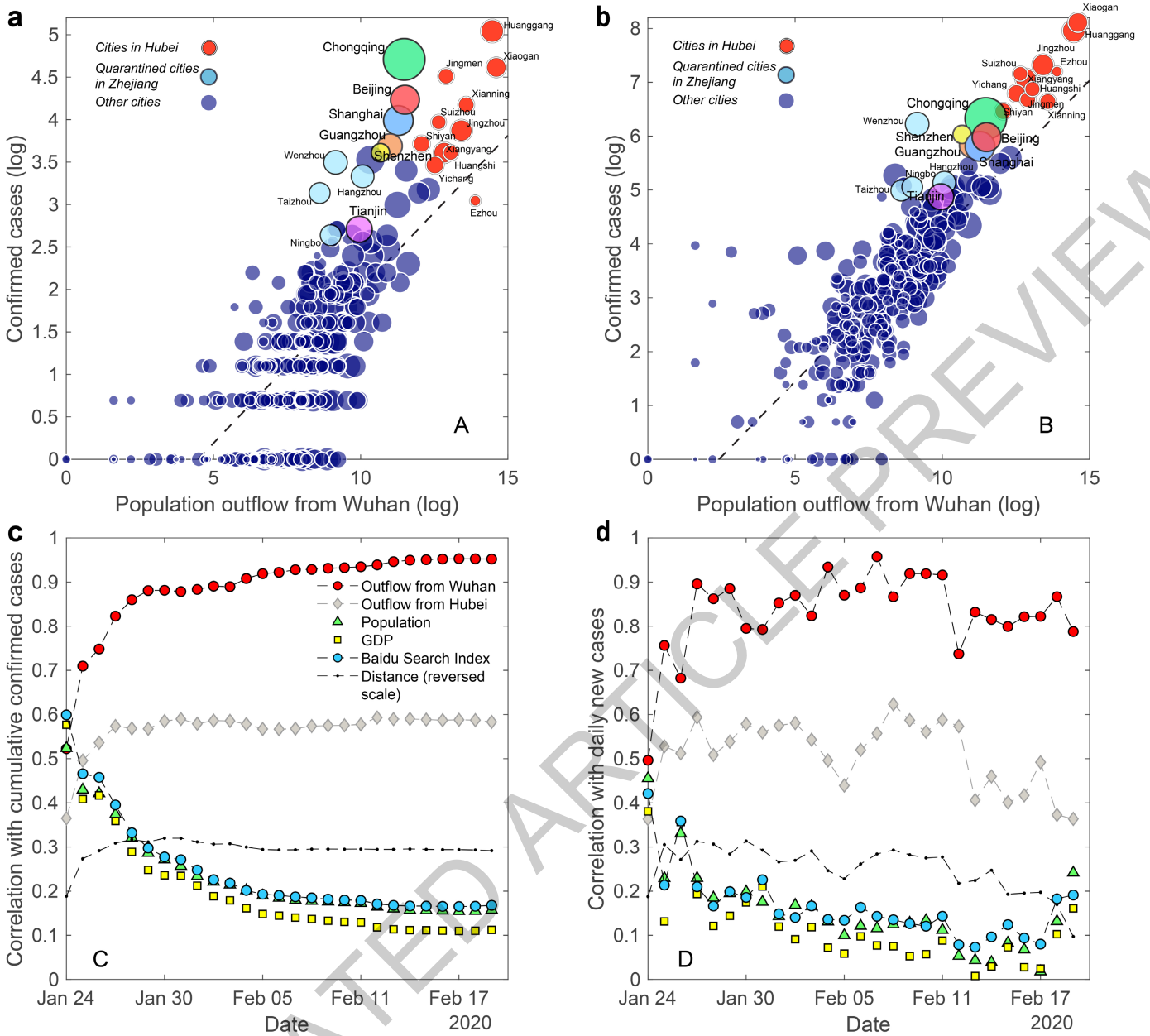


Fig. 2 | Factors correlated with confirmed COVID-19 cases. a, b, The relationship between aggregate population outflow from Wuhan (up to January 24) and confirmed cases by prefecture on January 26 (**a**) and February 19 (**b**). Red circles are prefectures in Hubei; light blue circles are four quarantined prefectures in Zhejiang (including Wenzhou); and the six largest prefectures in China are indicated with unique colors. **c**, Relationship over time between number of confirmed cases (**c**, cumulative through February 19 on x-axis) and prefectures' (i) cumulative population inflow (up to Jan. 24) from Wuhan, (ii) cumulative inflow from Hubei province excluding Wuhan, (iii) frequency of Baidu search terms related to the virus, (iv) GDP, (v) population, and (vi) distance from Wuhan. Over time, the correlation between population outflow from Wuhan and the number of infection cases increases from Pearson's $r = 0.522$ on January 24 to $r = 0.952$ (N=296). The decline in the predictive strength of online search behavior might reflect information saturation, while the decline in predictive strength of GDP, population size, and distance suggests that late-stage *chunyun* migration from Wuhan was to a more diverse set of prefectures (and not merely to the closet, largest, and most developed prefectures) and/or that community transmissions began to predominate. The correlation with daily infections (**d**) is consistent, with Pearson's r ranging from 0.496 on January 24 to a peak of 0.926 on February 4 (N=296). Fluctuations are likely lags in case reporting (that are smoothed in **c**); weaker correlations on the last few days reflect that >90% of prefectures outside of Hubei reported no new cases.

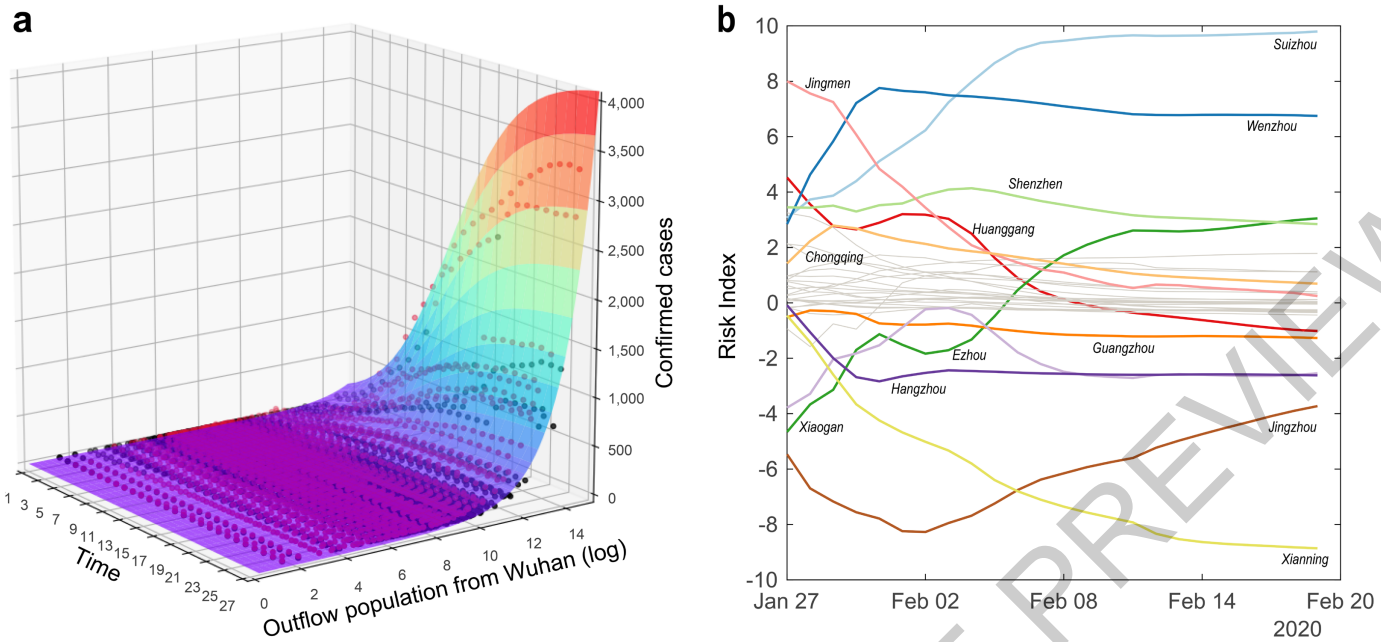


Fig. 3 | Predictive model based on population outflow. a, The surface indicates the fitted performance of our epidemiological model (see Supplementary Information, model (3)) with just a single variable x_{1i} indicates outflow population from Wuhan to prefecture i (*log* transformed), for all prefectures, with t as the number of days after *chunyun* is over (i.e., $t = 1$ is January 24). The dots represents the actual number of comfired cases under a given x_{1i} and t. Red dots represent prefectures where the reported number of confirmed cases is greater than the model's predicted values; black dots are all other cases, $R^2 = 0.930$ (N=7,992). **b**, Risk scores over time provide a dynamic picture of shifting transmission risks in different prefectures.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data necessary to reproduce the primary results of this study are included in this published article.

Code availability

Code necessary to reproduce the primary results of this study is included in this published article.

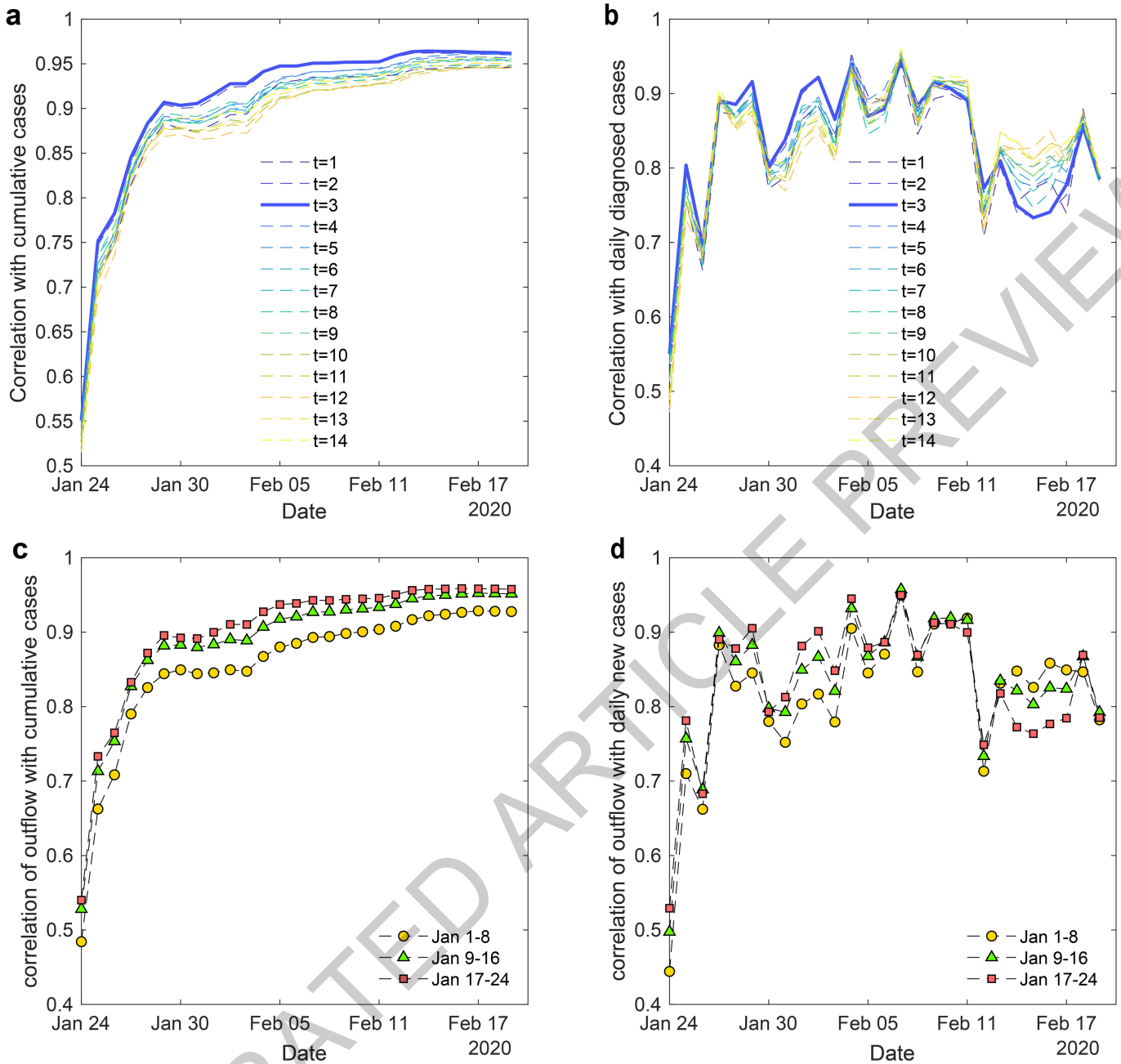
Acknowledgements We thank a major unnamed national carrier for providing the anonymized and aggregated data allowing the computation of population movements.

JJ is supported by the National Natural Science Foundation of China (72042009, 71490722) and Shenzhen Institute of Artificial Intelligence and Robotics for Society (2020-INT001). JSJ is supported by the Research Grants Council of Hong Kong (14505217). XL is supported by the National Natural Science Foundation of China (82041020, 91846301, 71771213, 71901067, 61773120) and the Science and Technology Department of Sichuan Province (2020YFS0007). GX is supported by the National Natural Science Foundation of China (71704052). We thank staff at the telecom carrier for their assistance in data preparation. This work was deemed exempt from IRB review.

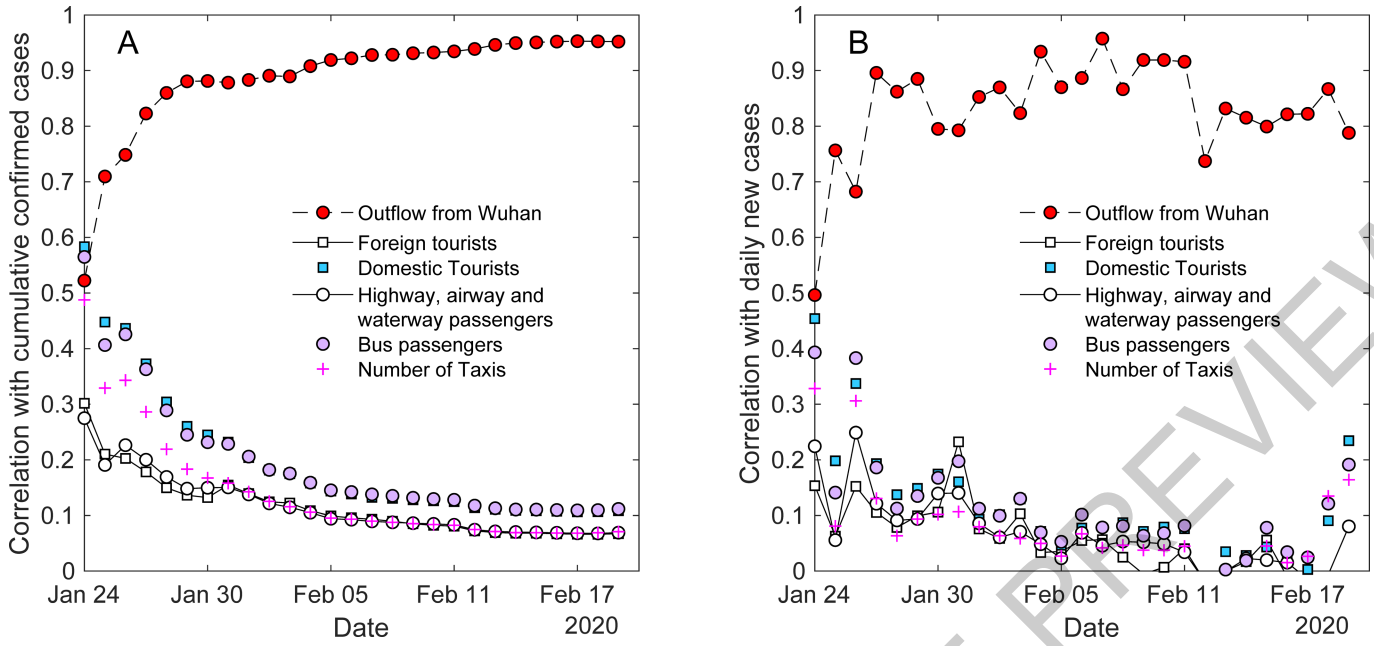
Author contributions All authors made equal contributions to the paper. JSJ, JMJ, and NAC conceived the research. JMJ, YY, XL, JSJ, and GX analysed the data. JSJ and NAC wrote the paper. JMJ, JSJ, and XL obtained funding. All authors contributed to research design, analytic development, and critical revisions.

Competing interests The authors declare no competing interests.

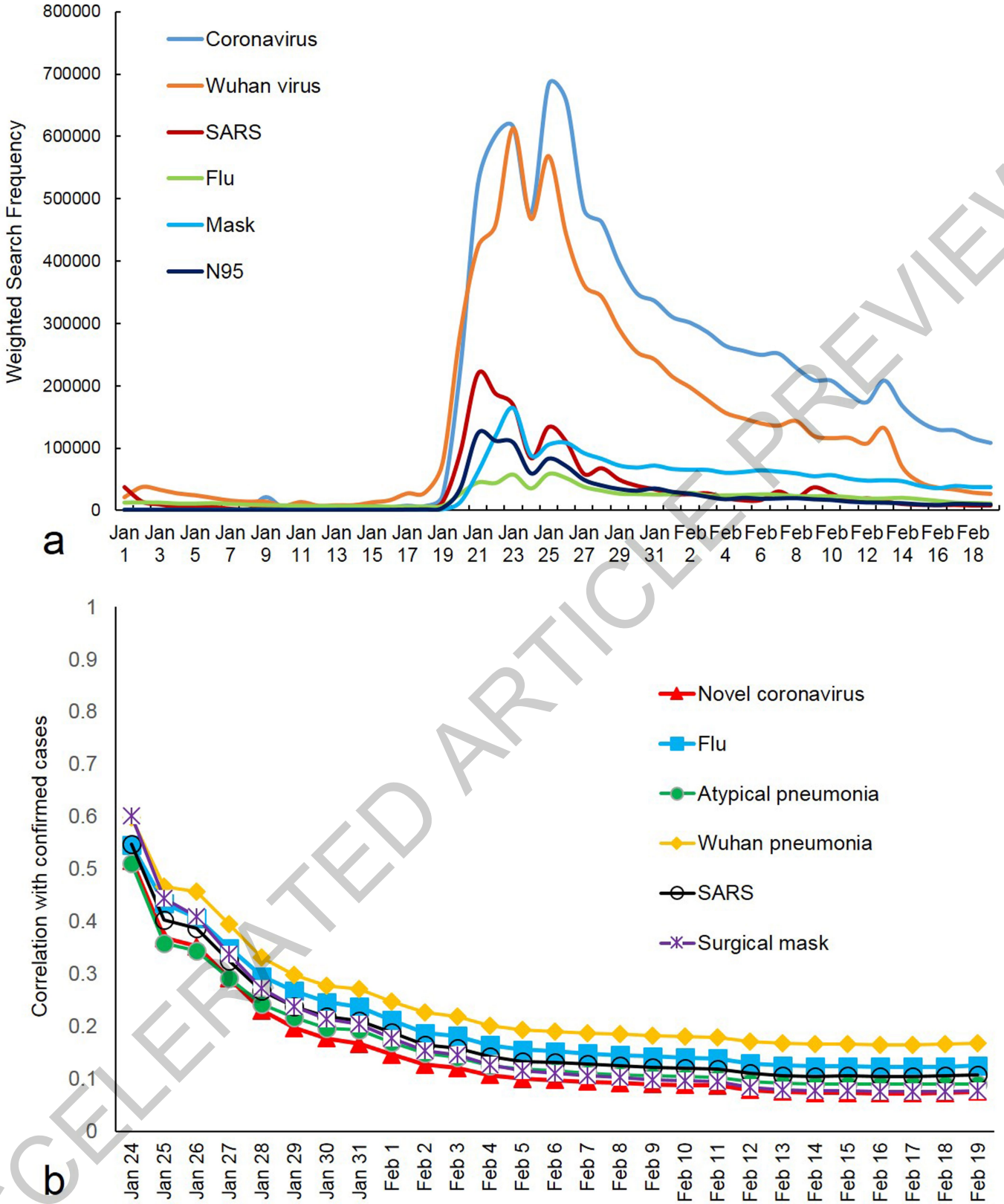
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2284-y.
Correspondence and requests for materials should be addressed to J.J. or J.J.
Reprints and permissions information is available at http://www.nature.com/reprints.



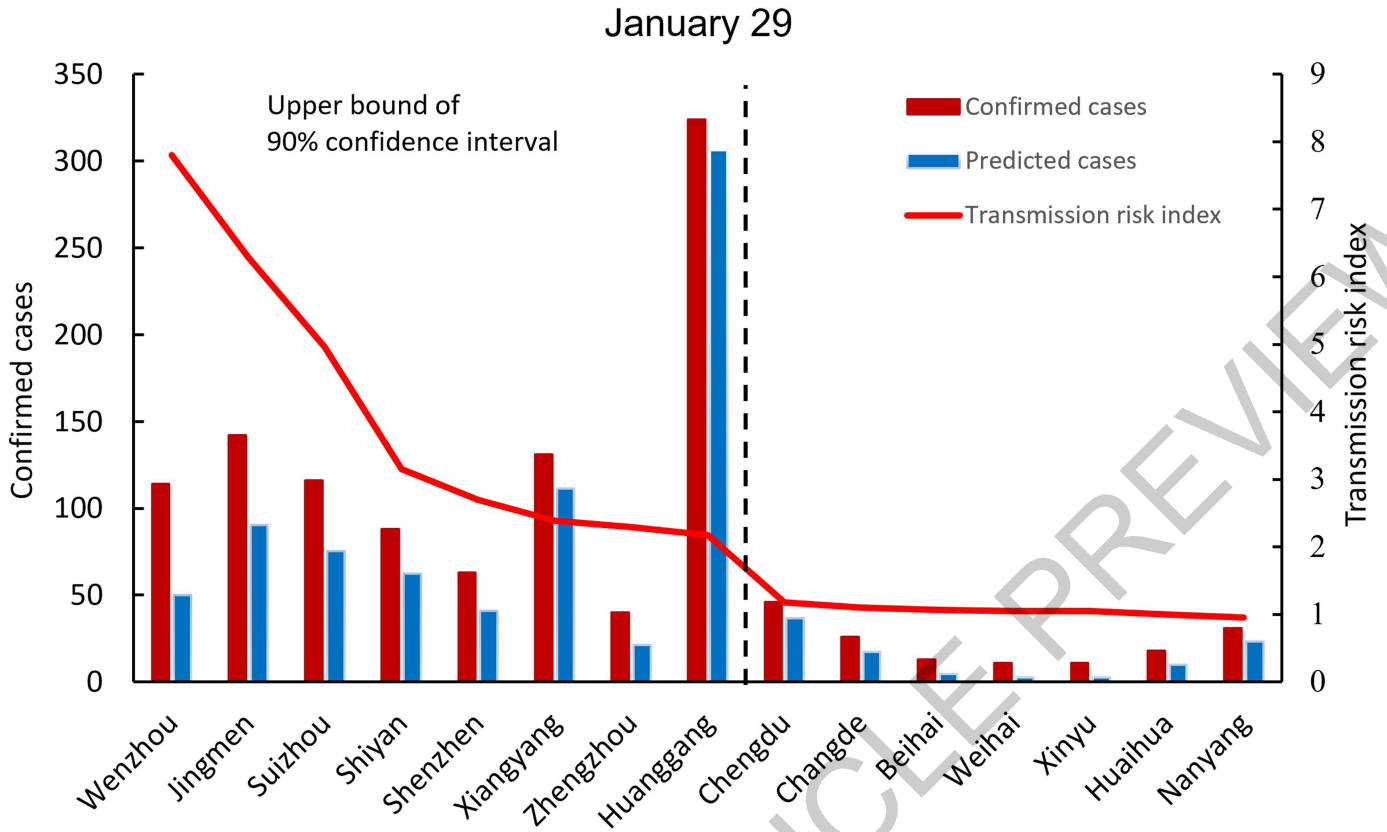
Extended Data Fig. 1 | Time window sensitivity test for the correlational analysis. a, b, Pearson's correlation (N = 296 prefectures) between the number of cumulative confirmed cases and population outflow from Wuhan on different days ranging from 1 to 14 days before January 24, for (**a**) the cumulative number of diagnosed cases over time, and (**b**) the number of newly diagnosed (daily) cases over time. Daily outflow is used for the calculation, e.g., $t = 3$ indicates that the correlation is measured by daily outflow from Wuhan on January 21 with cumulative confirmed cases from January 24 onwards. **c, d**, Pearson's correlation (N = 296 prefectures) during three different (8-day) time periods from January 1 to 24, 2020 between population outflow and (**c**) the cumulative number of diagnosed cases over time, and (**d**) the number of newly diagnosed (daily) cases over time.



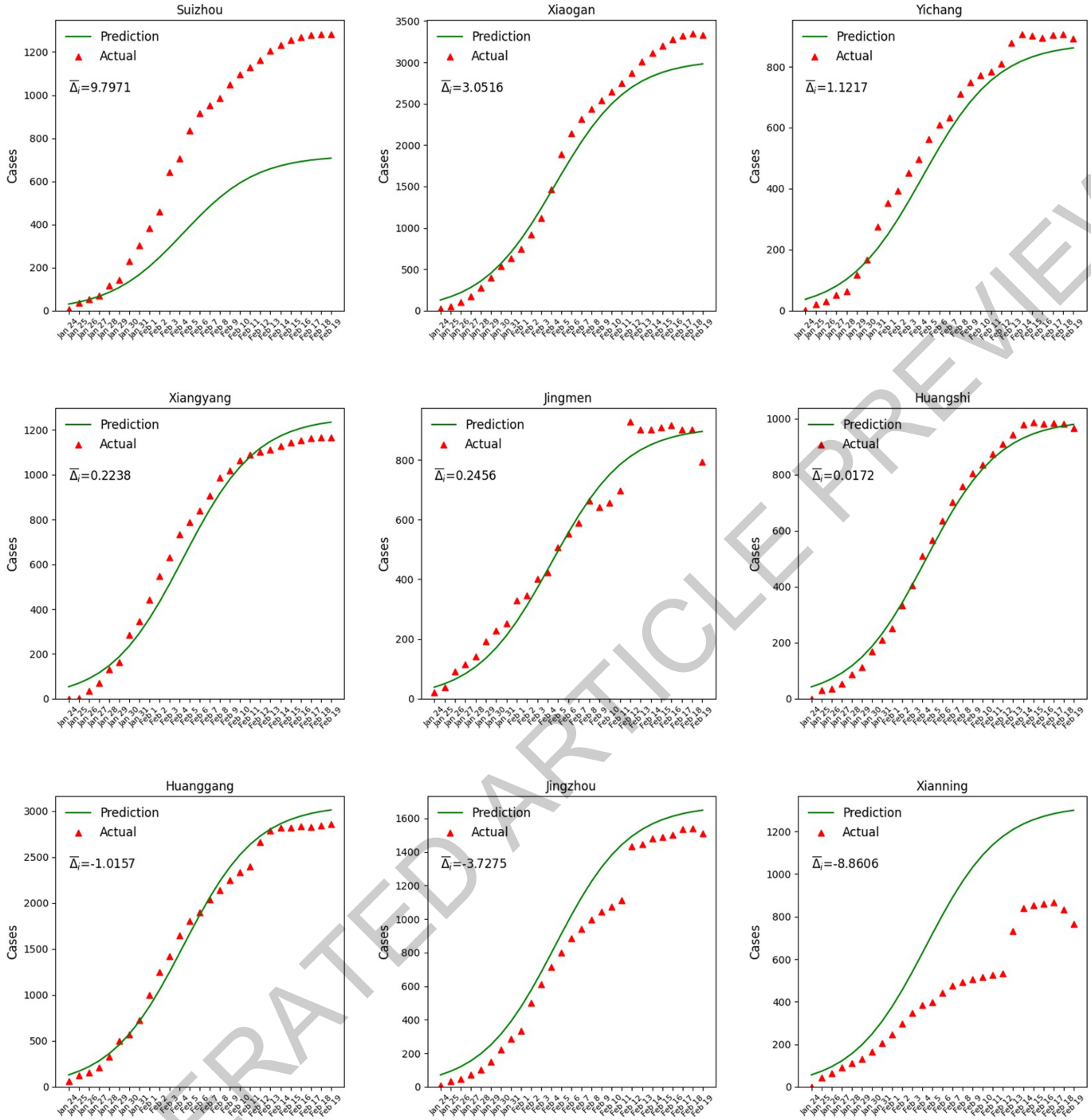
Extended Data Fig. 2 | Correlation with alternative population movement measures. Pearson's correlation (N = 296 prefectures) between alternative publicly available movement measurements from the 2018 City/Prefectures Statistical Year Book of China (with aggregate population outflow from Wuhan from January 1-24, 2020 as a reference) and COVID-19 count using (**a**) cumulative confirmed cases over time, and (**b**) for daily confirmed cases over time. Foreign tourist, domestic tourist, and "highway, airway, and waterway passenger" numbers reflect inter-prefecture travel, while bus passengers and number of taxis reflect local travel.



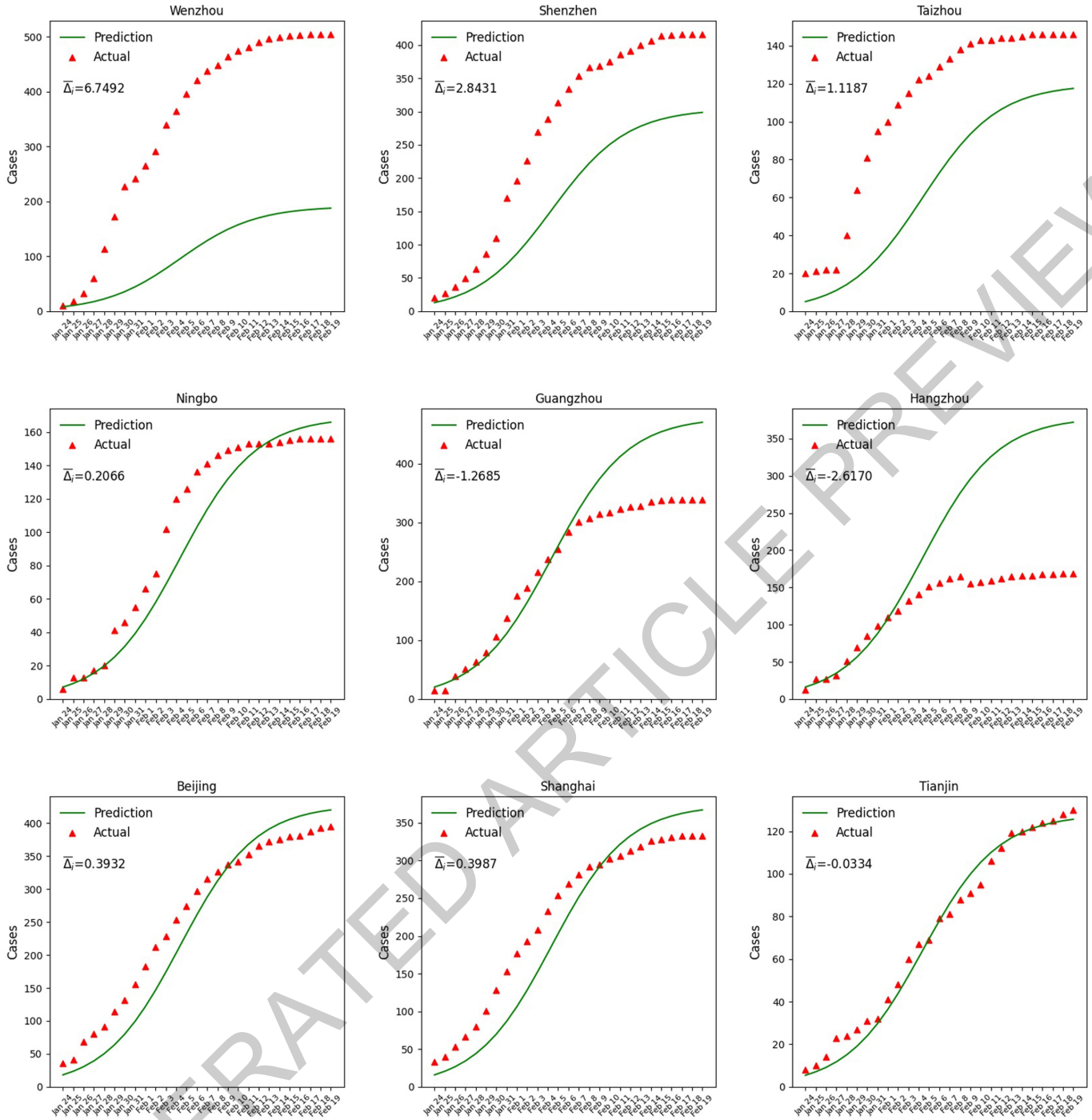
Extended Data Fig. 3 | Search terms and correlation with confirmed cases. **a**, Search frequency of Baidu search terms related to COVID-19. **b**, Pearson's correlation (N = 296 prefectures) between Baidu search terms and (cumulative) confirmed cases of COVID-19 over time. The initially high and then declining predictive strength of search may reflect the fact that initially high volumes of information search about the virus signaled stronger risk perception in any given prefecture (e.g., because of early reported cases, having more relatives in Wuhan, etc.), but that, over time, information saturation reduced the impetus for specific search.



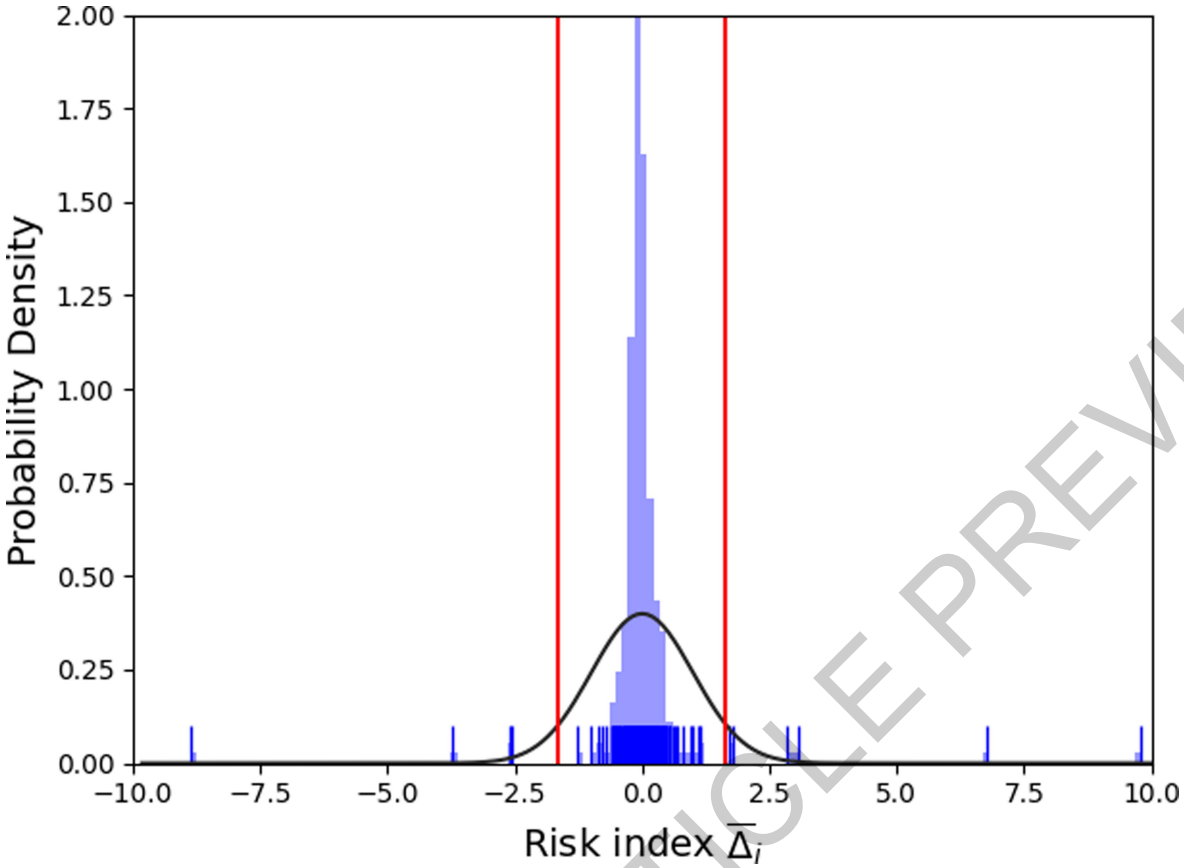
Extended Data Fig. 4 | Prefectures with high transmission risk index on January 29, 2020. The predicted structure of the spread of the SARA-CoV-2 virus can be used as a benchmark to identify which locales deviate significantly. Since model (1) predicts the number of cases in a prefecture based on the population outflow from Wuhan (i.e., imported cases and the initial local community transmission of the virus), a greater difference between predicted and confirmed cases suggests a higher level of community transmission. Prefectures to the left of the dashed line have community transmission risk index values over the upper bound of the 90% confidence interval. Our model identified Wenzhou as having the most severe community transmission risk on January 29, 2020. And the government announced a full quarantine of the prefecture on February 2, 2020.



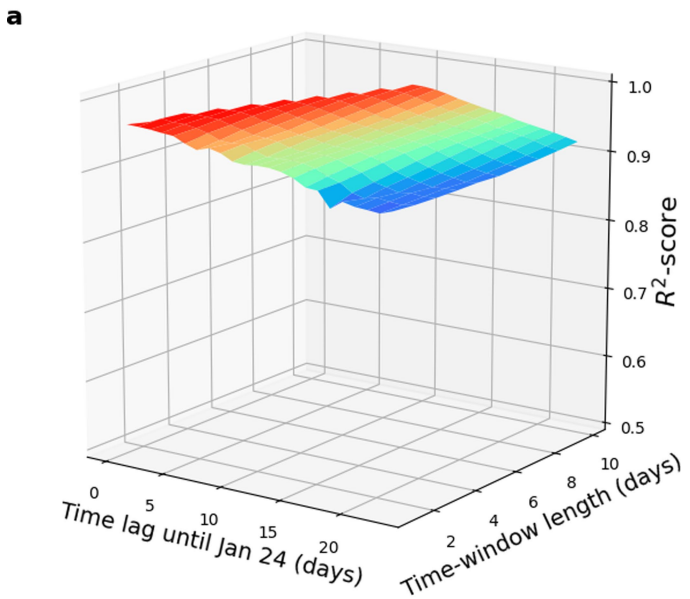
Extended Data Fig. 5 | Benchmark (predicted) versus actual virus growth in Hubei's prefectures. Model (2) used aggregate population outflow from Wuhan from January 1-24, 2020 to provide a reference growth *pattern* (i.e., epidemic curves) for COVID-19's spread across time and space, without making *a priori* assumptions of growth pattern or mechanism. Differences in the growth trends between predicted and confirmed cases can signal higher levels of COVID-19 transmission (Supplementary Table 11). The discrete jumps in confirmed cases in some prefectures after Feb 13 reflected a change in the local governments' infection count criteria; clinically diagnosed cases came to be included in total confirmed case counts in those prefectures (within Hubei province).

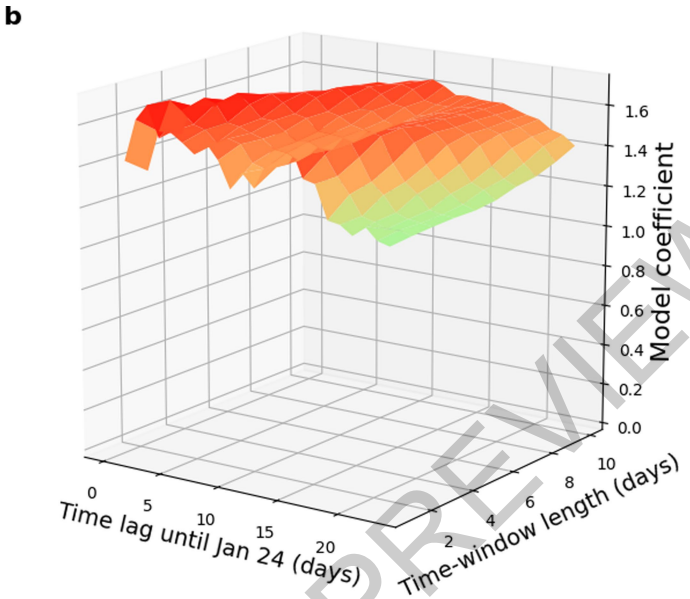


Extended Data Fig. 6 | Benchmark (predicted) versus actual virus growth in selected prefectures outside of Hubei. Model (2) used aggregate population outflow from Wuhan from January 1-24, 2020 to provide a reference growth *pattern* (i.e., epidemic curves) for COVID-19's spread across time and space, without making *a priori* assumptions of growth pattern or mechanism. Differences in the growth trends between predicted and confirmed cases can signal higher levels of COVID-19 transmission (Supplementary Table 11).



Extended Data Fig. 7 | The distribution of transmission risk index $\overline{\Delta}_i$. The transmission risk index is the normalized score of the integral of the differences between actual confirmed infection cases and predicted numbers in our model. Prefectures above the 90% confidence interval of the index are likely experiencing more local community transmission than imported cases, and prefectures below the 90% confidence interval may have a better performance in the control of the virus (see Supplementary Table 11).





Extended Data Fig. 8 | Robustness check of model (2) with different time lags and time window lengths. We explore which time window and time lags of aggregate population outflow best explain the spread and intensity of COVID-19. "Time window" refers to how many days of outflow data were used; "time lag" (0 to 23) is how many days before January 24 the time window starts. For example, time lag = 1 and time window = 2 is using outflow data between January 23-24. The surfaces show that a more recent time lag improves (**a**) the R^2 as well as (**b**) the parameter value of the population outflow coefficient in model (2).

nature research

Corresponding author(s): Jianmin Jia

Last updated by author(s): April 16, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Aggregated mobility data extracted from mobile phones are provided by one of the largest operators in China (regarding the total number of mobile phone users in China). The data on population flows and other key covariates used in the primary analyses will be made available upon publication. The daily infection data is public information that the government releases in China.
Data analysis	We used the Levenberg–Marquardt (LM) algorithm for model estimation, and the code was from Newville et al. (2016); Software includes Matlab (R2018a, The MathWorks, Inc.), Python (V.3.7.4, Python Software Foundation), and ArcGIS (V.10.2, Esri). We will make the code available upon publication (and for review).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data and code necessary to reproduce the primary results of this study are included in this published article for release online by Nature (and in the supplementary information files).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We use data regarding outflow population from Wuahn to different prefectures in China (ascertained with mobile phone records) to explore the spread of Coronavirus, ascertained by the Chinese CDC, and to assess transmission risk in difference areas.
Research sample	We used aggregate population outflow data of all people transiting through Wuhan, China between Jan 1-24, 2020; data was provided by a major national carrier. Types of data described in SI.
Sampling strategy	We used all available population outflow data in analyses (and conducted robustness checks using all different variants/alternative measures of the population outflow data provided). N=296 prefectures based on available covariate data (for GDP and population) in statistical yearbook published by National Bureau of Statistics of China, which covered 94% of the population. Any prefectures not covered was due to lack of data availability from this official government source.
Data collection	We obtained the aggregated mobile data via our industry partner in China and linked these records, at the level of 289 Chinese prefectures, to publicly available coronavirus cases in these areas.
Timing	The mobility data was collected during the period January 1 to January 24, 2020; and the confirmed case data was collected starting from January 24 up to February 19, 2020.
Data exclusions	All data that can be matched with the China Prefectures (City) Statistical Year Book have been included in the analysis (to provide covariates for our model); smaller sparsely populated prefectures not covered by the official Statistic Bureau's yearbook were excluded.
Non-participation	NA We used aggregated data of all customers of the carrier that traveled through or were in Wuhan during the study period.
Randomization	NA This study was not an experiment, and it did not have experimental conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	We used the aggregated mobile data of the Chinese phone users transiting through Wuhan in January 2020.
Recruitment	NA Population flow data was provided in aggregate form by a major Chinese carrier.
Ethics oversight	This work has been supported by the National Natural Science Foundation of China for the urgent policy research (given the pandemic). We do not use individual-level data, only anonymized aggregate flows, and this work is exempt from IRB review in China. An email to this effect was also obtained from the Yale IRB, and it has been shared with Clare Thomas, Senior Editor, Nature.

Note that full information on the approval of the study protocol must also be provided in the manuscript.