

# Forecasting SARS-CoV-2 transmission and clinical risk at small spatial scales by the application of machine learning architectures to syndromic surveillance data

Received: 27 October 2021

Accepted: 30 August 2022

Published online: 21 October 2022

 Check for updates

Thomas Ward<sup>1</sup>✉, Alexander Johnsen<sup>1</sup>, Stanley Ng<sup>1</sup> and François Chollet<sup>2</sup>

Timely and well-informed syndromic surveillance is essential for effective public health policy. The monitoring of traditional epidemiological indicators can be lagged and misleading, which hampers efforts to identify hotspot locations. The increasing predominance of digitalized healthcare-seeking behaviour necessitates that it is fully exploited for the public benefit of effective pandemic management. Using the highest-resolution spatial data for Google Trends relative search volumes, Google mobility, telecoms mobility, National Health Service Pathways calls and website testing journeys, we have developed a machine learning early indicator modelling approach of SARS-CoV-2 transmission and clinical risk at small geographic scales. We trained shallow learning algorithms as the baseline against a geospatial neural network architecture that we termed the spatio-integrated long short-term memory (SI-LSTM) algorithm. The SI-LSTM algorithm was able to—for the assessed temporal periods—accurately identify hotspot locations over time horizons of a month or more with an accuracy in excess of 99%, and an improved performance of up to 15% against the shallow learning algorithms. Furthermore, in public health operational use, this model highlighted the localized exponential growth of the Alpha variant in late 2020, the Delta variant in April 2021 and the Omicron variant in November 2021 within the United Kingdom prior to their spatial dispersion and growth being confirmed by clinical data.

The COVID-19 pandemic has precipitated unprecedented global public health policy interventions and population-level behavioural change. Understanding localized outbreaks of SARS-CoV-2 can be exceedingly difficult due to the inevitable ascertainment bias that occurs through a test-by-request strategy<sup>1</sup>. Furthermore, case, hospitalization and mortality data are lagged indicators due to the delay between infection and the report and clinical outcome date<sup>2,3</sup>. This can be hampered by the tendency for outbreaks to initially cluster in younger-aged demographics where, due to decreased infection

severity and a higher proportion of asymptomatic infections, their representation in clinical data can be more limited<sup>4</sup>. Based on the strength of assortative mixing, it typically takes several weeks or months for infections to bleed into older-aged demographics, the point at which healthcare indicators can detect a substantial outbreak. It is therefore requisite for effective syndromic surveillance to look at data that are auxiliary to clinical outcomes for the early detection and identification of outbreaks so that adequate public health interventions may be able to limit transmission.

<sup>1</sup>UK Health Security Agency, London, UK. <sup>2</sup>Google, California, CA, USA. ✉e-mail: [Tom.Ward@UKHSA.gov.uk](mailto:Tom.Ward@UKHSA.gov.uk)

For a respiratory disease such as COVID-19, understanding the mobility patterns of individuals is central to calculating the transmission rate ( $\beta$ ) and the force of the infection ( $\lambda$ ) by more accurately tracking the effective contacts between individuals. The COVID-19 pandemic has allowed for mobility data to be used by telecoms providers<sup>5</sup>, web-based companies<sup>6,7</sup> and public transport organizations<sup>8</sup> to help understand contact patterns and adherence to non-pharmaceutical interventions (NPIs). Jeffrey et al.<sup>9</sup> used mobility data from O<sub>2</sub> and the Facebook application as a proxy for the actual mobility patterns in the United Kingdom to monitor the adherence to the March 2020 national lockdown. This type of data has been further employed effectively to model dynamic mobility networks to enable the simulation of the SARS-CoV-2 epidemic<sup>10</sup>. There are, however, issues with relying on mobility data as a proxy for effective contacts<sup>11</sup> because increased mobility in itself may not correlate with increased transmission. This is most palpable around the exponential phase of a novel SARS-CoV-2 variant and any analyses employing these data must contextualize changes as only a component of behaviours that can be associated with increases or reductions in transmission<sup>12</sup>. The use of mobility data seems to have had the greatest utility when used for nations or localities where there has been a policy of NPIs and the associated behavioural change can be more effectively interpreted<sup>9</sup>; such data continues to be used for transmission modelling by the Scientific Advisory Group for Emergencies (SAGE)<sup>13,14</sup> for the UK Government.

The use of digital searching and reporting may offer insights for syndromic surveillance that cannot be found from conventional epidemiological indicators. Platforms such as Google Trends offer an interface to analyse relative search volumes (RSVs) for a given locality, providing real-time monitoring of morbidity within populations. Internet reporting behaviour has been explored for emerging pathogens such as Ebola<sup>15</sup>, and found to be of analogous utility to traditional surveillance for monitoring clusters and outbreaks. Google Trends has demonstrated the potential for the monitoring of the respiratory virus H1N1 and the vector-based West Nile virus<sup>16</sup>. Furthermore, Google Trends data have been used to assess adherence to NPIs for the COVID-19 pandemic in Taiwan by looking at, for instance, the demand for face masks<sup>17</sup>. Past research<sup>18–21</sup> has discussed Google Trends as a data source for identifying early increases in national incidence of COVID-19 in much the same way as Google Flu Trends. The Google Flu Trends<sup>22</sup> surveillance experience<sup>23</sup> illustrated the need for secondary data sources, intelligently designed algorithms and constant recalibration as an epidemic evolves to account for temporal changes in searching behaviour. The fusion of Google Trends with multiple data sources (Twitter, hospitalizations and Google Flu Trends) as a surveillance system for influenza<sup>24</sup> showed improved model performance over using Google Trends alone and illustrated the promise for the application of this approach to detect changes in the transmission of SARS-CoV-2.

The interpretation of testing data at local scales is obfuscated<sup>25</sup> by ascertainment bias, where key groups of interest can be excluded through: the geographic impracticality of testing centres, financial implications if compelled to isolate, asymptomatic infection, and test availability at the time of infection or symptom onset. International research has highlighted the increased positivity and lowest testing rates for COVID-19 in the most deprived areas<sup>26–28</sup>. This has been shown to be exacerbated in deprived rural areas with less connectivity to local testing centres, and in minority groups that can be excluded from effective public health messaging. Further work has emphasized the lack of synchronicity that can be observed between testing data and clinical outcomes, which is compounded by the stochasticity of smaller numbers at finer spatial scales<sup>29</sup>. This can be a consequence of spatially clustered testing, which is particularly affected by targeted testing strategies<sup>30</sup> and spatially heterogeneous institutional testing. However, the age-severity infection gradient could also lead to an expected divergence between the levels of test positivity and the number of observed clinical outcomes due to the temporally varying case composition.

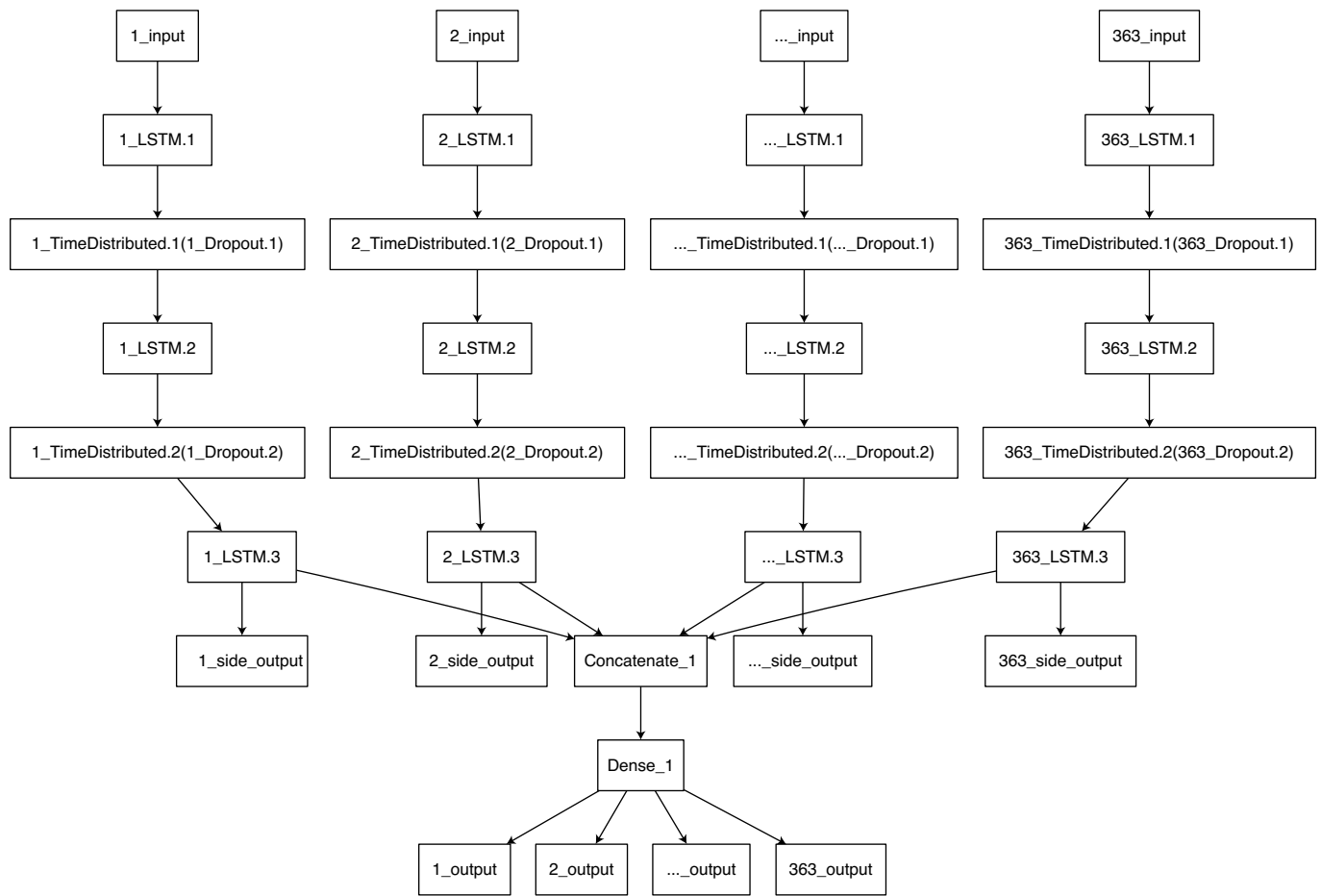
This emphasizes the importance of using epidemiological outcomes that are relevant to the public health policy context and the requirement for early intervention on the basis of the characteristics of novel variants in circulation.

Impactful syndromic surveillance of SARS-CoV-2 requires symptomatic prevalence, healthcare-seeking behaviour, mobility patterns and testing demand to be monitored to map and understand a widespread communicable disease. The interaction of healthcare-seeking behaviour through internet engagement<sup>31</sup> has been shown to be important to understand disease transmission. However, it is the application of this approach (specific to COVID-19 symptomatology) to small spatial scales that has yet to be explored for its potential to function as an early indicator of an outbreak. We evaluated the suitability of Google Trends, Google mobility, telecoms mobility, National Health Service (NHS) Pathways 119 calls and website testing demand as predictive features for modelling outbreaks. We have then employed multiple machine learning models, with sensitivity analysis of temporally lagged features, to assess their predictive performance at capturing localized outbreaks of COVID-19. We have designed a neural network for spatial data, which we have termed a spatio-integrated long short-term memory (SI-LSTM) and a spatio-integrated convolutional long short-term memory (SI-CNN-LSTM) algorithm.

## Leading indicator analysis

We conducted an appraisal of the leading indicators for population-normalized polymerase chain reaction (PCR)-positive tests, hospitalizations and deaths from COVID-19 at local authority district (LAD) in the UK. Our preliminary investigation assessed the feasibility of primary healthcare, social care and secondary healthcare data sources. This included general practitioner calls, clinical staff absences, care home incidence reports and school absences. We further explored 1,108 COVID-19-related Google Trends web search terms, the NHS COVID-19 mobile application<sup>32</sup>, Google mobility, telecoms mobility, COVID-19 test request website journeys and NHS Pathways 119 calls. Leading indicators were assessed using generalized additive models with a negative binomial error structure and dynamic time warping. Data sources were excluded if they did not have full geographic coverage, could not be sourced in a timely fashion (that would be relevant for an operational public health response), or were found to lag or be concurrent with the clinical target variables. Results indicated that the most consistent leading indicators across the epidemic phases were 94 Google Trends terms, telecoms mobility, Google mobility, website test request journeys and NHS Pathways 119 calls.

The Google Trends data were collected hourly for all four nations of the United Kingdom, resulting in data for 4,013 locations; they were scaled to LAD geography using the latitude and longitude coordinates provided by Google to map to the Office of National Statistics (ONS) boundaries<sup>33</sup>. Mobility data were collected from Google<sup>6</sup> and telecoms operators<sup>5</sup>, where it is reported at the LAD and middle layer super output area (MSOA)<sup>33</sup>, respectively. The data are prepared by mapping to the LAD level using the ONS lookups<sup>34</sup> by extracting, among other things, demographic and person category (resident, worker, visitor) information. Website journey test request data were sourced from the Test and Trace Adobe Analytics platform, which measures both symptomatic and asymptomatic journeys through the test booking system. The data are further broken down by whether the journey was complete or incomplete at the final stage. Testing availability was defined as individuals that complete the online journey until the final stage at which they are offered a test and could not proceed relative to individuals that completed the website journey. Adobe geolocates requestors on the basis of their internet protocol and a lookup table was created to aggregate the Adobe locations to LAD level. The 119 number was established as the contact number for the NHS Test and Trace service in May 2020<sup>35</sup>, and provides a way to book a COVID-19 test and enquire about a test result; its scope has since expanded to process vaccination appointments.



**Fig. 1 | The architecture of the SI-LSTM algorithm.** The features from each location are fed in as inputs along network branches that contain LSTM layers, time-distributed dropout layers and a dense layer, producing a side output. The tensors are further concatenated to produce the main model output for each area.

As with the other sources mentioned above, the dataset was aggregated to LAD geography using an ONS lookup table<sup>34</sup>. Only two types of call were selected: calls in which ‘Test enquiry—request a test’ was given as the call reason, and all calls, regardless of reason.

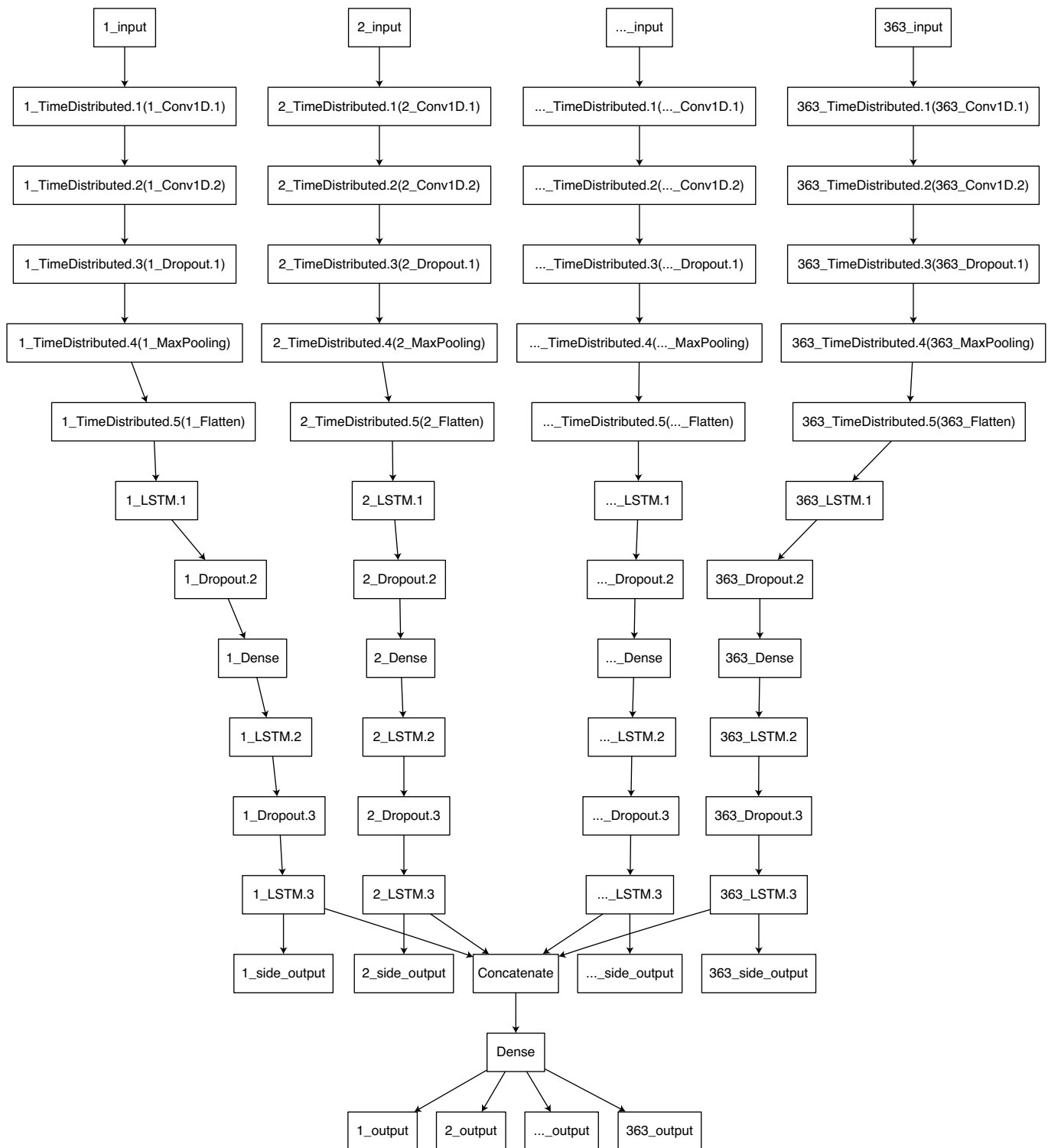
## Modelling outbreak risk

An outbreak risk scoring system was developed for population-normalized COVID-19-positive PCR cases, hospitalizations and mortalities. The PCR-positive case data were sourced through the anonymized combined list collected by the UK Health Security Agency (UKHSA), which is derived from the National Pathology Exchange dataset<sup>36</sup>. The hospitalization data were obtained from the admitted patient care (APC) dataset<sup>37</sup>, which include individuals that tested positive for COVID-19 fifteen days prior to and eight days post admission, and was aggregated from the lower super output areas to the LAD level. Mortality data were obtained from the UKHSA COVID-19 death linelist for England, and the public dashboards for Scotland<sup>38</sup> and Northern Ireland<sup>39</sup> (note that we did not have access to mortality data at LAD geography for Wales). The PCR testing and mortality data that were included for analysis had been evaluated for backfilling (that is, how long it takes before the last complete day of data) over the most recent seven day period prior to inclusion as a target. The hospitalization APC data have defined monthly periods when hospital trusts must declare their admission activity data and the last complete day was included. The daily PCR tests, hospitalizations and mortality data for each LAD were normalized per million and smoothed over a rolling seven day window. The defined thresholds represent equal proportions of these distributions at LAD for a defined temporal window of the epidemic in

the UK. The risk score criteria are dynamic and determined by changes to the daily proportions in cases, hospitalizations and deaths, which are influenced by variant severity, availability of testing within a country, the ascertainment rate and the rate of disease prevalence to be informative indicators of inter-location heterogeneity.

Preliminary univariate analysis was conducted for the risk score targets of COVID-19, using an autoregressive integrated moving average (ARIMA) model fit using a modified Hyndman–Khandakar algorithm at epidemic phase change points. Shallow learning algorithms (Random Forest<sup>40</sup>, XGBoost<sup>41</sup>, GBM<sup>42</sup> and Naïve Bayes<sup>43</sup>) were trained on the leading indicator features, which were lagged from 15 to 40 days relative to risk score target. We did not forecast for greater than these periods as preliminary analysis indicated that model performance quickly deteriorated after 40 days. Random holdouts of up to 40 days were excluded across the epidemic phases to assess the performance of the models. *K*-fold cross-validation was included for each model (*k* = 10) in addition to a primary model that was trained on the entire training dataset. Eleven models were thus trained on the data: ten on each cross-validation split and the primary model on all of the training data. The trained models were then stacked to create an ensemble model using the XGBoost algorithm<sup>43</sup>. The stacking comprises training a second-level learner called a meta-learner to optimize the performance of the base learners.

We developed deep learning algorithms to enhance the algorithm learning from the geospatial data, which have been termed an SI-LSTM (Fig. 1) and SI-CNN-LSTM (Fig. 2) algorithm (please see the Methods for further details). The SI-CNN-LSTM architecture takes advantage of the feature amplification ability of convolutional neural network layers to use a type of weight sharing with local perception to refine



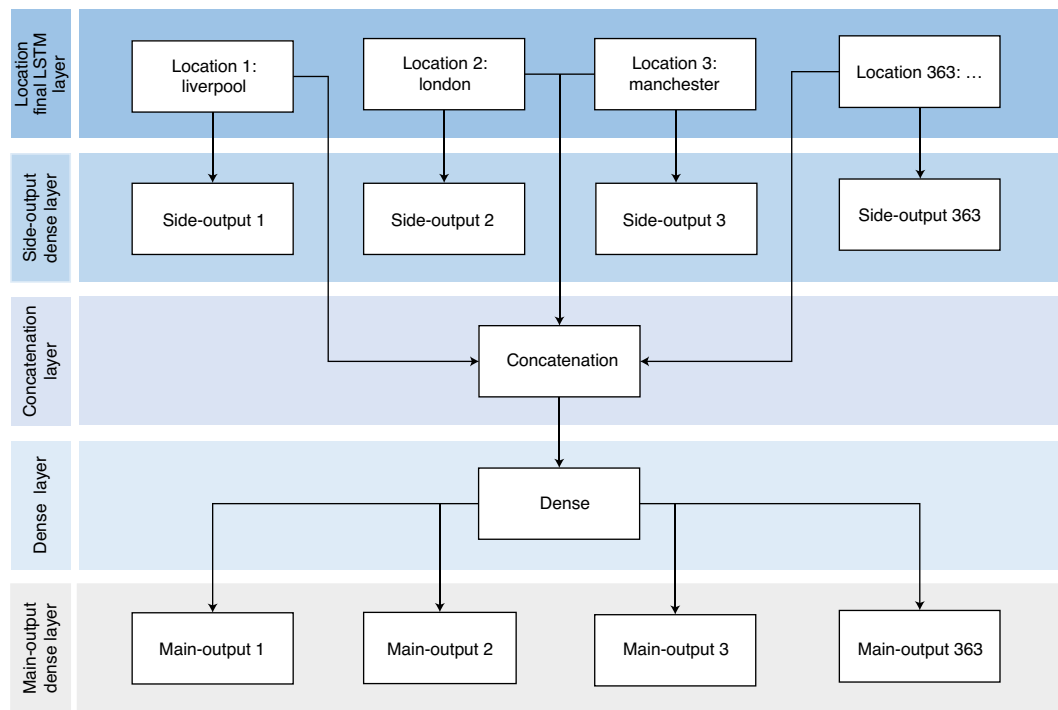
**Fig. 2 | The architecture of the SI-CNN-LSTM algorithm.** The features from each location are fed in as inputs along network branches that contain time-distributed one-dimensional convolution layers, a time-distributed max pooling

layer, a time-distributed flatten layer, LSTM layers, and dense and dropout layers, producing a side output. The tensors are further concatenated to produce the main model output for each area.

and condense the number of parameters that helps to improve the learning efficiency for the LSTM layers<sup>44</sup>. These models were developed using bespoke generator functions<sup>45</sup> for the LAD time-series in the UK and yielded lagged batches of the features for the target variables. The model features were pre-processed using a log transformation to stabilize the variance, and subsequently normalized so that the mean was zero and the standard deviation was one. Due to

the mobility data containing negative values, we employed an offset value before log transformation to ensure that the step produced a real value. This is conducted to speed the process to the global minima of the error surface and mitigate the chance of getting stuck at local optima. The model targets were one-hot encoded to convert the categorical input data into a vector required for the categorical cross-entropy loss function<sup>46</sup>.





**Fig. 3 | The geospatial connection network of the spatially integrated layers across LADs in the UK.** A merging of the 363 independent input branches by combining the list of tensors, from the final LSTM layer for each location, on a single concatenation axis.

The final model architecture included a seven day lookback to capture the weekly trend in the features, a shuffling in the order of the training data, and a decrease in the learning rate for subsequent epochs if an increase in the validation loss was detected. In the final layers of the SI-LSTM and SI-CNN-LSTM, we introduced a connection network between all of the geographic locations so that the model performance can be optimized through intra- and inter-location feature weighting. The 363 independent input branches are merged through combining the list of tensors from the final LSTM layer for each location on a single concatenation axis, which produces a single tensor as described in Fig. 3. The final LSTM layer produces a rank-2 tensor of shape  $(b, u)$ , where  $b$  is the batch size and  $u$  is the number of units in the LSTM layer. After concatenation of tensors from  $L$  locations, the resulting tensor has shape  $(b, Lu)$ . The final dense layer has a softmax activation function to ensure that the output vector  $y_{i \in \{1, \dots, C\}}$  over  $C$  classes is normalized, and that  $y_i$  can be interpreted as the probability that the target is class  $i$ . The cross-entropy loss function is then defined as:

$$\mathcal{L}(y_i, t_i) = - \sum_{i=1}^C t_i \log y_i$$

where  $t_i$  is the one-hot encoded target vector. We then used RMSprop as the optimization function in the back-propagation stage.

## Results

### Univariate forecasting

Univariate ARIMA modelling, using a modified Hyndman–Khandakar algorithm<sup>47</sup> for step wise performance tuning, was conducted using PCR-positive cases, hospitalizations and mortalities from COVID-19 (Extended Data Fig. 1). We can observe that the ARIMA models struggle, particularly at change points in the epidemic wave, to reliably predict the growth trajectory. This is particularly pronounced in the pre-exponential phase, exponential phase and at the turning point of an epidemic peak, which is evidenced by the LAD model results in Extended Data Table 1, in which the models struggled to reach an accuracy of 50% across the Alpha wave.

### Feature importance

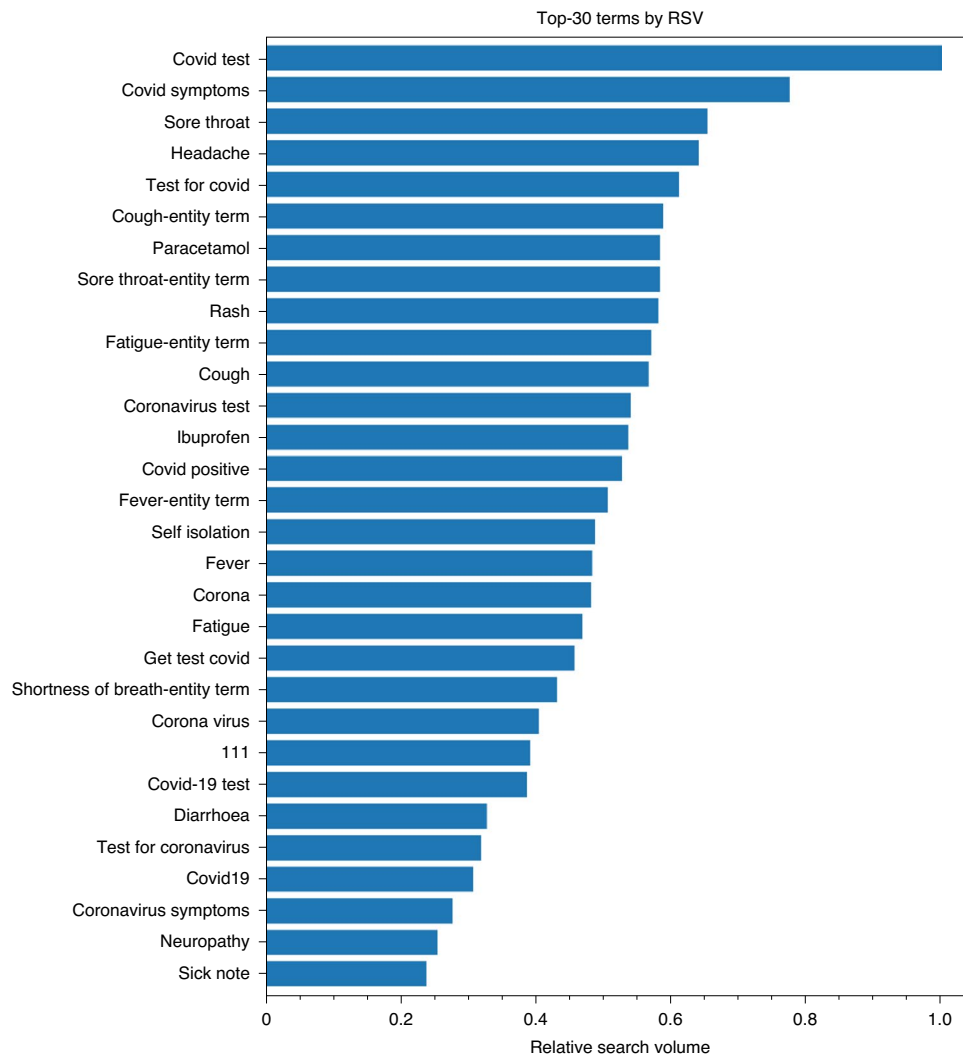
To assess feature importance, we used a Random Forest algorithm, with random temporal holdouts, across the different feature groups included (Extended Data Table 2). The most important feature group for all tasks was Google Trends, followed by Telecoms Mobility, Google Mobility, Website Testing Demand and 119 Calls. A full statistical description of the included model features can be found in Supplementary Table 1. This performance is indicative of the periods assessed in this paper, feature importance has evolved across the COVID-19 epidemic in the UK and has been influenced by extrinsic pressures such as NPIs, changes in testing behaviour/policy, and novel variant patterns of growth (the feature importance for each epidemic phase and variant can be seen in Supplementary Figs. 1–3).

### Google Trends at LAD

The search terms that received highest relative volumes scores across the research period can be seen in Fig. 4. Variations on requests for COVID-19 tests have the highest overall volume observed at LAD level in the United Kingdom. The highest-volume entity terms observed for COVID-19 are ‘sore throat’, ‘cough’, ‘fatigue’, ‘fever’, and ‘shortness of breath’. Although absolute volume is of interest to maintain relevant and timely search terms, it is not necessarily reflective of their overall feature importance in the model spatially and temporally. It is how the terms interact and the auxiliary data that determine their importance for outbreak detection.

### Spatio-temporal modelling

The SI-CNN-LSTM and SI-LSTM algorithms performed better across all temporal periods and for every target relative to shallow learning algorithms assessed (Fig. 5). The greatest performance differential was observed for the mortality risk scores, where the SI-LSTM saw an improvement of up to 15% relative to the best shallow learning algorithm. There was a clear performance improvement from the geospatial concatenation that can be observed in the main-output accuracy relative to the side-output accuracy seen in Extended Data Fig. 2. The peak temporal performance across the modelled targets was observed



**Fig. 4 | The search terms that received highest relative volumes.** A bar chart of the top-30 search terms from the Google Trends data collection in the United Kingdom at LAD, measured by RSV.

in the SI-LSTM: 99.4% accuracy for the case risk score (feature lag of 25 days), 96.3% for the hospitalization risk score (feature lag of 40 days) and 84.8% for the mortality risk score (feature lag of 25 days).

In the deep learning models, training and validation loss convergence was observed after around 20 epochs for case risk scores, 15 epochs for hospitalization risk scores, and 30 epochs for mortality risk scores. The SI-LSTM overall performs slightly better on the test data than SI-CNN-LSTM, and convergence is reached after fewer epochs (Extended Data Fig. 3). Earlier model architectures encountered volatility in the validation loss, which was resolved by decreasing the learning rate and increasing the batch size. Moreover, by providing a dynamic learning rate during training, we found a smaller value of around 0.001 generally produced optimal model convergence. Sensitivity analysis found that the optimizer function RMSprop performed better than stochastic gradient descent and Adamax, as seen in Extended Data Fig. 4. We also found that a larger tensor shape for each LSTM layer—corresponding to 128–160 units—produced higher validation accuracies.

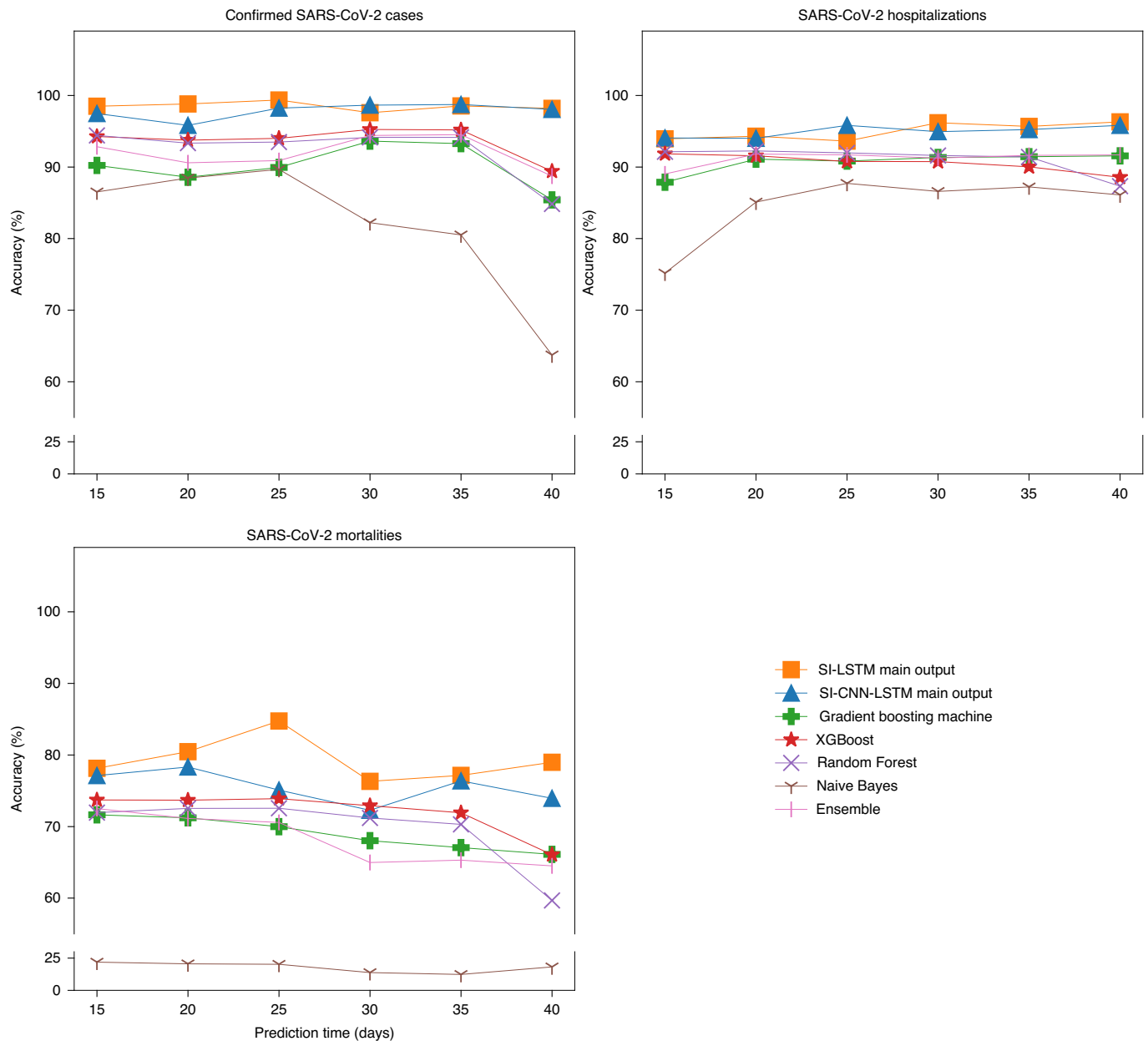
Analysis of the shallow learning models found the highest overall performance of 95.3% on the case risk score for the XGBoost followed by the Ensemble, Random Forest and GBM, which had only slightly reduced accuracy on the test data. The Naïve Bayes model, by contrast, did not perform well for longer projection periods and mortality risk. The log loss across each temporal period and shallow

learning algorithm can be seen in Extended Data Fig. 5. The distinction in performance between the XGBoost, Random Forest and Ensemble relative to the GBM is slightly more pronounced. The shallow and deep learning models performed better on the case risk score and the hospitalization risk score targets. However, all models saw diminished performance on the mortality risk score, which may be related to the increased relevance of the features for capturing transmission in younger demographic groups.

Sensitivity analysis was conducted on the hyperparameters of the XGBoost, Random Forest and GBM algorithms. The number of trees, tree depth and learning rate were varied to establish whether performance could be further optimized. The outcome of this analysis established that for the GBM and XGBoost (Extended Data Fig. 6), providing that the number of trees exceeded 1,000, the tree depth 10 and the learning rate 0.01, the performance was relatively insensitive to the hyperparameters. Moreover, that the Random Forest model performed optimally with a max tree depth of 5 and when the number of trees exceeded 500.

#### Public health operational model outputs

For the purpose of epidemic management, the early spatial identification of the pre-exponential and exponential change points—prior to their identification through traditional epidemiological surveillance—is



**Fig. 5 | The performance of the SI-CNN-LSTM and SI-LSTM algorithms across all temporal periods.** A line graph of the model accuracy for confirmed SARS-CoV-2 case, hospitalization and mortality risk scores for the shallow- and deep-learning algorithms across the temporal delay periods.

important for an effective response to outbreaks of novel variants of concern.

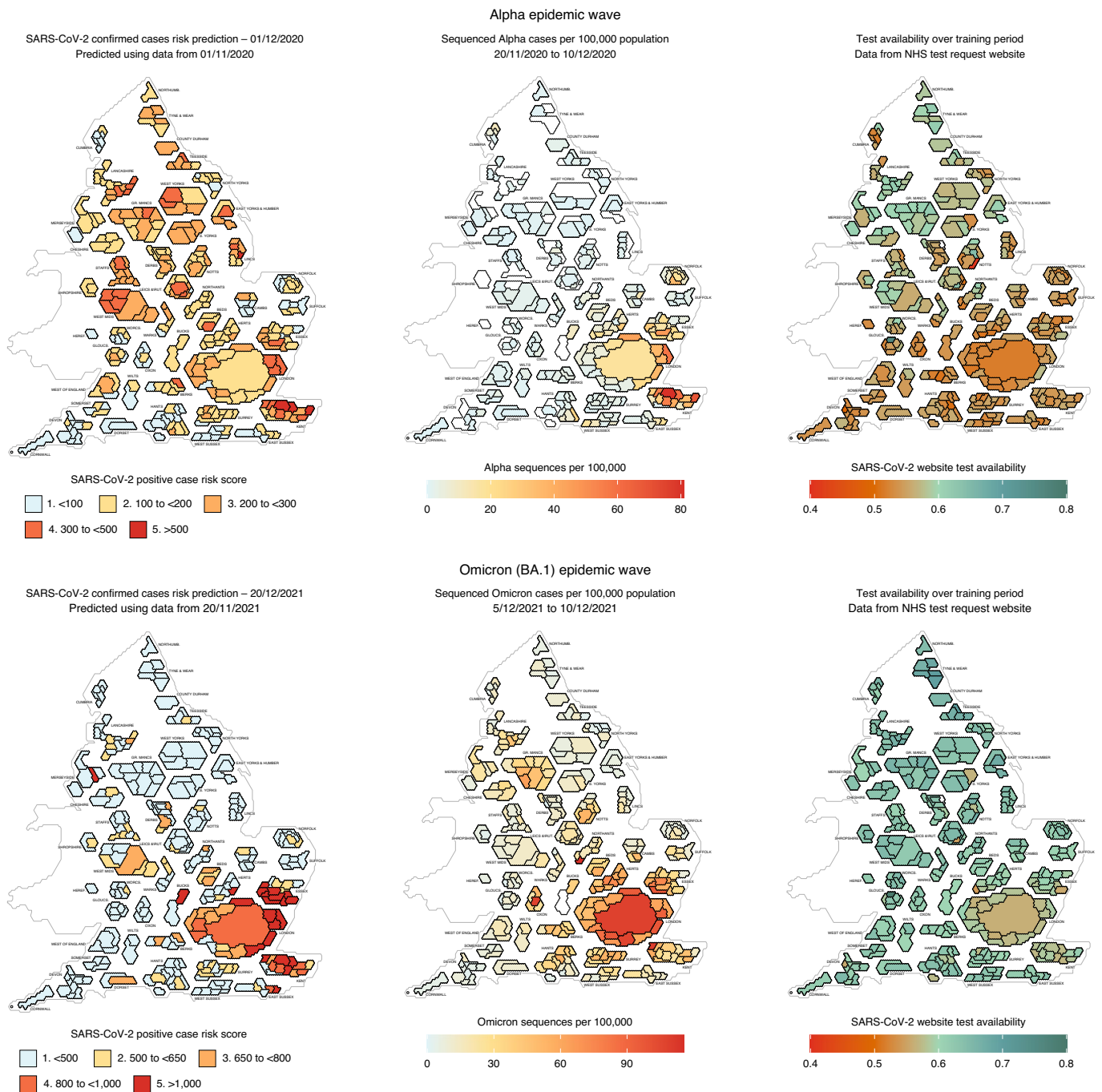
In December 2020<sup>48</sup> the outbreak of the Alpha variant in England began by clustering around the county of Kent. During this period England was also experiencing substantial growth in COVID-19 (D614G mutant of SARS-CoV-2), which had precipitated a lockdown in November 2020<sup>49</sup>. On the 1st of November 2020, the modelling identified the exponential growth of the Alpha variant in Kent (Fig. 6) prior to the identification through sequenced PCR tests. The observed lack of testing availability identified through website test requests in the Alpha wave is noteworthy and may have masked the identification of increased case rates in some local authorities.

The recent outbreak of the Omicron BA.1 variant was initially identified in late November 2021 in England<sup>50</sup>. The clustering of this variant around London and the South East region was detected through the modelling of leading indicator features from 20th November 2021 (Fig. 6).

This was identified in the background of a high prevalence rate for the Delta variant and at this time there had been only eight confirmed sequenced PCR cases of Omicron BA.1 in England. The unprecedented wave of incidence that was observed in December 2021 necessitated a higher tiering in the case data, which can be seen in Supplementary Fig. 4.

## Discussion

The heterogeneous nature of the COVID-19 epidemic, being characterized by localized outbreaks, presents challenges for public health policy in that certain areas may warrant more substantial interventions to contain the spread of SARS-CoV-2. The aim of this modelling approach is to provide policy-makers with an early indicator syndromic surveillance framework for local areas which, when combined with other lines of reporting, can aid in pandemic management. This localized focus has become increasingly more important as importations of SARS-CoV-2 variants of concern become the focus of outbreak response<sup>51,52</sup>. We have



**Fig. 6 | Maps of England showing confirmed case risk predictions, sequenced cases and test availability data for the Alpha and Omicron variants.** In the top panel: sequenced Alpha cases population-normalized per 100,000 averaged from the 20th of November to the 10th of December 2020; confirmed case risk predictions from features on the 1st of November 2020 trained to forecast for

up to 30 days; testing availability over the training period. In the bottom panel: sequenced Omicron cases population-normalized per 100,000 averaged from the 5th to the 10th of December 2021; confirmed case risk predictions from features on the 20th of November 2021 trained to forecast for up to 20 days; testing availability over the training period.

illustrated, akin to the literature on other communicable diseases<sup>31</sup>, that RSV data can be of utility in understanding transmission hotspots when the terms are carefully selected, and further clinical and non-clinical data are included in model development.

The SI-LSTM geospatial architecture design allowed for specific intra-location learning while also benefitting from inter-location information sharing. This model architecture achieved the highest overall performance of above 99% accuracy on the unseen data for the case risk score at the local authority level in the UK. We found that a smaller learning rate and larger batch size were important in reducing

validation loss volatility, despite research that LSTMs work well with larger learning rates<sup>53</sup>, because they push the output gate to zero. The inclusion of convolutional neural network (CNN) layers and regularization in the dense layers produced comparable performance for each temporal delay period assessed in this paper. We discovered in early model development that the performance of the SI-CNN-LSTM and SI-LSTM models were more improved relative to the shallow learning algorithms with a longer time-series of training data; therefore, when dealing with a shorter time-series there may be a preference towards a shallow learning algorithm approach.



The willingness or ability to opt into the testing system<sup>54</sup> substantially impacts insights from conventional epidemiological data for epidemic surveillance. The motivation to seek or report a test has been found to be related to symptom severity and a lack of understanding with regards to the main symptoms of COVID-19, which has been observed to a greater extent in older age groups<sup>55</sup>. This will be further impacted by socio-economic vulnerabilities, the ability to acquire a test and location feasibility. Due to the vulnerability of the confirmed case risk score model target to this ascertainment bias, we normalized positive test counts and defined epidemiologically important ranges that would be more robust to these fluctuations. We further adjusted the model target ranges to reflect the spatio-temporal variation in testing availability and observed that the inclusion of testing availability as a model feature improved performance for some local authorities. In locations that have limited testing coverage, particularly relevant as public health policy evolves in response to the pandemic, the modelling framework proposed may be better suited to the further clinical targets of COVID-19 infection included in this study.

The study found mobile and telecoms mobility data to be a robust predictive feature of the increased transmission of SARS-CoV-2. The novel application of this data to disease modelling in the COVID-19 pandemic has allowed for a greater understanding of movement patterns that can help to identify locations of concern, importations between local authorities and behavioural responses to the easing of NPIs<sup>9</sup>. However, the importance of the mobility data as a leading indicator evolves with the temporal epidemic phases and extrinsic factors. In later periods when NPIs were more limited, the mobility data, in isolation, were a better predictor of transmission when the virus showed patterns of endemicity. Models that have been developed<sup>56</sup> to primarily focus on mobility proxies may therefore be limited in their ability to accurately capture novel variant growth. This can be explained by fluctuations in transmission being determined by mobility patterns when a variant is more established and growth is more stable, but this data independently will have less utility at recognizing the pre-exponential and exponential phase growth from the introduction of a new variant, particularly if pre-pandemic contact patterns have returned. However, this study finds that—in combination with proxies of symptomatic prevalence—mobility data can be an effective leading indicator across the epidemic phases.

For the use of Google RSV data in an operationally relevant environment, it is essential to monitor the relative frequency of the terms (see Extended Data Fig. 7) as behaviours<sup>57,58</sup> and worldwide government directives evolve over the course of the pandemic. This is to preclude monitoring terms that are no longer relevant as healthcare-seeking habits change<sup>59</sup> or those that are likely to be unduly driven by extrinsic pressures such as media reporting<sup>60</sup>, and to capture novel behaviours that may be important. Novel variants have presented diverse symptomology profiles<sup>61</sup> and therefore it is important to keep a broad spectrum of symptoms included in the data collection. Further research of Google RSV data at the local authority level should investigate locations for post-acute COVID-19 (long COVID<sup>62</sup>) in areas disproportionately impacted and that have had stubbornly high transmission of COVID-19. Moreover, there may be further insight gleaned from the use of this data to assess the impact on mental health between locations that have been under longer-term local lockdowns<sup>63</sup>.

Digitalized web-based data sources (Google Trends, and Test and Trace website test requests) included in the analysis have a bias towards younger-aged demographics. However, these groups were the focus of the analysis, because an epidemic wave of a respiratory infection such as COVID-19 is predominantly driven by younger age groups (<65 years old), which have higher effective contact rates<sup>64,65</sup>. Moreover, further research has identified that resurgent epidemic waves of the SARS-CoV-2 virus have been driven largely by working-age adults<sup>66</sup> and that the 18–39 age group led the replacement of Delta by Omicron BA.1 (ref. <sup>67</sup>) in the UK. A preliminary assessment of leading indicators from

primary health and social care data sources that exclusively target the oldest age groups were found to have limited geographic coverage in the UK, were difficult to source in an operationally useful manner, and found to lag community transmission. The 119 telephonic requests for PCR and lateral flow tests included in the modelling were found to have a slightly older age composition relative to online test requests, which may have aided in the identification of increased transmission for these ages.

The difficulty of identifying change points in an epidemic curve has been a consistent modelling challenge across the pandemic<sup>68</sup>. This has been frequently observed for widely developed transmission models<sup>69,70–72</sup> that are reliant on historical data to fit the model and transmission simulations of prescribed parameters (which are difficult to quantify<sup>70</sup>) to develop projections. The parameter space for transmission models evolves for each new variant, with the collection of data required to update these parameters too lagged for early epidemic management. For instance, the estimates required for the generation time, serial interval, incubation period and the time to a clinical event<sup>2,73</sup> have required usually, at minimum, a month or longer for an adequate sample to be collected from contact tracing. Different choices in these parameter spaces lead to a great divergence in the modelling projections fit to the same data. Machine learning approaches<sup>74–76</sup> and statistical forecasting models<sup>77</sup> that are univariately trained on confirmed cases are limited in an operationally responsive space to provide a meaningful window for interventions as they will struggle to identify a signal until incidence is in clear exponential growth or decay. This will be further compounded by confirmed tests being a lagged indicator of increased incidence, which is exacerbated by times of heightened ascertainment bias. Moreover, at a small spatial scale, models trained solely on case data will suffer with a great deal of false signals especially if confirmed cases are not adjusted for some measure of testing availability or the rate of ascertainment.

We propose a novel modelling approach that has been developed for public health response organizations that has wider relevance for modelling outbreaks of COVID-19 outside of the United Kingdom. This study is designed to provide a modelling framework and data sources that can be effectively employed to create early warning indicators of changes in transmission and to project the hospital and mortality burden at small spatial scales. The defined modelling approach is designed to be adaptable to different stages in the pandemic and the risk scoring system should be tailored to the current rate of prevalence and the severity profile of a variant for a specific population. This approach focuses on trends and changes in those trends that would provide spatial insights on a novel outbreak and the epidemic trajectory.

## Conclusion

Timely and well-informed syndromic surveillance is essential to inform effective public health policy over the SARS-CoV-2 pandemic. The monitoring of traditional clinical indicators can be lagged and misleading, which hampers efforts to identify hotspot localities. We have coalesced the most meaningful leading indicator data currently available in the UK to identify local authorities of concern. The models described are used as part of the UK's coordinated response to the COVID-19 pandemic with a suite of other data sources to inform public health policy and identify areas with concerning levels of transmission.

This study found that the SI-LSTM algorithm design was able to, for the assessed temporal periods, accurately predict hotspot locations over time horizons of a month or more with a high degree of accuracy. The novel architecture described in this paper provides a framework for modelling temporally variable geospatial data. We anticipate that this model architecture has uses beyond the epidemiological application described in this paper.

In public health operational use, the models accurately forecast the exponential increase in the Alpha variant in December 2020, the Delta Variant in April 2021 and the Omicron Variant in November 2021



within local authorities in the UK. The evolution of the pandemic may render certain data sources less important for modelling purposes and, due to extrinsic pressures, modelling RSV trends must be conducted with careful design, relevant auxiliary features and meaningful clinical targets.

## Methods

The section will first outline the steps taken to collect and prepare the data sources for modelling. The development of the models is described at the end of this section.

### Data collection and preparation

**Google Trends.** Google Trends data provides RSV by search term and location over time which can be accessed via the public website [trends.google.com](https://trends.google.com). The data are normalized by total search volume<sup>78</sup>, and reflect the relative importance of terms over time and space. Both national- and highly localized city-level data were analysed for this work. The city-level data can be found under the 'Interest by city' panel within the user interface. We collected hourly RSVs for all four nations of the United Kingdom, resulting in data for 4,013 locations.

The project had the support of Google Trends' editorial team throughout the project, whom facilitated data acquisition and provided a Google Health Trends API key. A preliminary analysis was conducted on the daily relative values provided by Google for each city location. The daily relative value was found to be of limited utility due to the high proportion of zero values reported. Further exploration discovered that data collected at an hourly frequency resolved this issue. We therefore executed hourly requests to collect the Google search trends data.

At the outset of the project, the collection included 108 terms to capture the most frequently observed symptoms of COVID-19<sup>79</sup>, NHS medical advice seeking behaviour, COVID-19 testing, and common over-the-counter treatments for COVID-19. These terms were supplemented with a further 1,000 search items found to be the most commonly employed phrases used in NHS Pathways 111 telephonic COVID-19 triages<sup>80</sup>. We excluded certain words and phrases for their lack of overall relevance in the context of a search term and their relative occurrence at a national level in the Google Trends user interface. Preliminary analysis conducted at a national level involved generalized additive models with a negative binomial error structure and dynamic time warping to assess the selected terms' relevance as a predictive feature of COVID-19 incidence and clinical outcomes. The analysis highlighted 94 important terms that were relevant for further analysis and seven primary symptoms of COVID-19 were included as Google entity terms.

The Google data were then processed to match geographically, by date, the recorded SARS-CoV-2 case, hospitalization and mortality data at LAD. Google estimates search locations using sources including the GeoIP and, where available, the GPS coordinates of the device<sup>81</sup>. Lookups were therefore developed using the latitude and longitude provided by Google to map the data to the ONS<sup>33</sup> designated LAD geographies. This was not possible for central London and as a result a group of LADs was created to match Google's London location.

**Mobility data.** Mobility data were collected from Google<sup>6</sup> and telecoms operators<sup>5</sup> where it is reported at LAD and MSOA<sup>33</sup>, respectively. The Google mobility data measures change in the visits and length of stay at six different place categories compared with a baseline period between the 3rd of January and the 6th February 2020<sup>6</sup>. The categories are grocery and pharmacy, parks, transit stations, retail and recreation, residential, and workplaces. Locations provided—based on the ISO 3166 standard—are 'country\_region\_code', 'sub\_region\_1' and 'sub\_region\_2'. The telecoms mobility data contain counts of the number of people and their number of journeys over time at MSOA geography. The data are prepared by mapping to LAD using the ONS lookups<sup>34</sup> by extracting,

among other things, demographic and person category (resident, worker, visitor) information. The absolute numbers in this dataset are challenging to interpret, but, as with other sources presented in this paper, it is the trends rather than absolute numbers that are important.

**Website COVID-19 testing journey data.** Website COVID-19 testing journey data were sourced from the Test and Trace Adobe Analytics platform, which measures both symptomatic and asymptomatic journeys through the test booking system. The data are further broken down by whether the journey was complete or incomplete. An incomplete booking journey is one in which a person does not proceed at the final stage of the online journey to book the test. Adobe geolocates requestors on the basis of their internet protocol and a lookup table was created to aggregate the Adobe locations to LAD level.

**NHS Pathways 119 Data.** The 119 number was established as the contact number for the NHS Test and Trace service in May 2020<sup>35</sup> and provides a way to book a coronavirus test and enquire about a test result; its scope has since expanded to process vaccination appointments. The dataset includes the call date and reason along with the geographic location of the caller. As with the other sources mentioned above, the dataset was aggregated to LAD geography using an ONS lookup table<sup>34</sup>. Only two types of call were selected: calls in which 'Test enquiry—request a test' was given as the call reason and all calls, regardless of reason.

**Testing availability.** The greatest quantity of diagnostic tests conducted for COVID-19 are through website requests. Testing availability was defined as individuals that complete the online journey until the final stage at which they are offered a test and could not proceed relative to individuals that completed the website journey. This may be due to lack of available RT-PCR tests, because the testing centre location was not accessible, or the requestor chose not to proceed.

Due to the temporal and geographic disparity over testing availability throughout the pandemic we calculated testing availability, as a function of location  $l$  and time  $t$ . A completion denotes an individual that finished the website test request journey and that a test was conducted. It is defined by the following equation:

$$\text{Availability}(l, t) = \frac{\text{Symptomatic completions}(l, t) + \text{Asymptomatic completions}(l, t)}{\text{Symptomatic final stage users}(l, t) + \text{Asymptomatic final stage users}(l, t)}$$

Availability  $(l, t) = 1$  corresponds to an area where all of those who request a test receive one,

Availability  $(l, t) = 0$  corresponds to an area where testing is entirely unavailable on request

Testing availability was employed as a feature when modelling the case rates for a locality. Testing data coverage is heterogeneous, and the ascertainment bias is time varying therefore, for the operational presentation of modelling results that were trained on case data we included testing availability scores to understand gaps in local coverage that the model may not identify.

**Outbreak risk score.** The primary purpose of this modelling approach is to highlight areas of concern before a substantial outbreak occurs within a LAD. An outbreak risk score was therefore developed for confirmed SARS-CoV-2 PCR-positive cases, hospitalizations and mortalities (Supplementary Fig. 5). The PCR-positive case data were sourced through the anonymized combined list collected by the UKHSA, which is derived from the National Pathology Exchange dataset<sup>36</sup>. The hospitalization data were obtained from the APC dataset<sup>37</sup>, including individuals that tested positive for COVID-19 fifteen days prior to and eight days post admission, and was aggregated from the lower super output area to the LAD level. Mortality data were obtained from the UKHSA COVID-19 death linelist for England, and the public dashboards

for Scotland<sup>38</sup> and Northern Ireland<sup>39</sup> (we did not have access to mortality data at LAD geography for Wales).

The PCR testing and mortality data that were included for analysis had been evaluated for ‘backfilling’ (how long it takes before the last complete day of data) over the most recent seven day period prior to inclusion as a target. The hospitalization APC data has defined monthly periods when hospital trusts must declare their admission activity data and the last complete day was included. The daily PCR tests, hospitalizations and mortality data for each LAD was normalized per million and smoothed over a rolling seven-day window.

The thresholds for the risk scores were determined by analysis of the population-normalized daily distribution of cases, hospitalizations, and mortalities, at LAD. The defined thresholds represent equal proportions of these distributions at LAD for a defined temporal window of the epidemic in the UK. These thresholds were, in a public health operational response setting, initially informed by the localized interventions in the United Kingdom through the tiering system<sup>47</sup>. The risk score criteria are dynamic and determined by changes to the daily proportions in cases, hospitalizations and deaths, which are influenced by variant severity, availability of testing within a country, the ascertainment rate, and the rate of disease prevalence to be informative indicators of inter-location heterogeneity.

### Model development

The data used for analysis in this work were collected from the 1st of October 2020 and the model performance was measured up to July 2021. The software used for model development included Python v.3.10.0 and R v.4.2.0. The targets for the machine learning modelling were defined as the daily confirmed case risk score, hospitalization risk score and the mortality risk score. The features used for the machine learning modelling included Google Trends search data, Google mobility, telecoms mobility, NHS Pathways 119 call categories, testing availability, location, and asymptomatic and symptomatic website testing request journeys. The features, analogous to the targets, were smoothed over a rolling seven-day window due to the erratic nature of this time-series data when analysed daily. For our modelling purposes, and its operational use case, we sought to identify trends and not the precise value on a given day to highlight an area of concern.

Time-series analyses of the data was conducted using shallow learning and deep learning algorithms and the features were lagged relative to the target from 15 to 40 days to assess their predictive temporal relationship with the clinical indicators. Forecasting was not attempted for longer than these periods as preliminary analysis found that model performance quickly deteriorated after 40 days. This project ran a total of 2,057 models including the sensitivity analysis of hyperparameters.

### Univariate forecasting

To understand the difficulty of the predictive task and where the proposed models are likely to struggle, a univariate forecasting approach was developed for population-normalized cases, hospitalizations and mortalities at the LAD level. An ARIMA model was fit using a modified Hyndman–Khandakar algorithm<sup>82</sup> for step wise performance tuning using unit root tests and the Akaike information criterion. Model performance was further measured by the risk scoring criteria developed for cases, mortalities and hospitalizations.

### Shallow learning

**Model design.** With the features lagged from 15 to 40 days, we trained Random Forest<sup>40</sup>, XGBoost<sup>41</sup>, GBM<sup>42</sup> and Naïve Bayes<sup>42</sup> algorithms on the risk score target. Log loss was the defined loss metric for the Random Forest, XGBoost and GBM with a stopping tolerance of 0.001 (full model hyperparameter specifications can be found in Supplementary Table 2). Random holdout outs of up to 40 days of data were excluded from the training sample and used to assess model performance. *K*-fold cross-validation was also included for each model (*k* = 10) in addition to a primary model that

was trained on the entire training dataset. Eleven models were therefore trained on the data: ten on each cross-validation split, and the primary model on all of the training data. The trained models were then stacked to create an ensemble model using the XGBoost algorithm<sup>43</sup>. The stacking comprises of training a second-level learner called a meta-learner, which combines the base learners to optimize performance.

**Feature importance and sensitivity analysis.** Sensitivity analysis was conducted to find the optimal hyperparameter combinations for each shallow learning algorithm across the assessed temporal periods. This included the tree depth, number of trees and the learning rate. To illustrate the relative importance of each data source at predicting the risk score targets, a Random Forest algorithm was trained on each source’s features in turn and the performance was evaluated. We measured the performance at a 15 day lag in the features for the PCR-positive case, 20 day lag for hospitalization and a 25 day lag for mortalities. The delays were selected as the optimal performance periods of the Random Forest algorithm. The results provided are the overall performance across the assessed periods however, these relationships change across epidemic phases. Therefore, feature importance was assessed across every epidemic phase for each replacing variant of SARS-CoV-2 using an XGBoost algorithm.

### Deep learning

In the following section we discuss the data pre-processing for the deep learning algorithms, the preliminary sensitivity analysis, and the final model architectures.

**Data pre-processing.** The model features were pre-processed using a log transformation to stabilize the variance and subsequently normalized, so that the mean was zero and the standard deviation was one. Due to the mobility data containing negative values we employed an offset value prior to log transformation to ensure that the step produced a real value. This is conducted to speed the process to the global minima of the error surface and mitigate the chance of getting stuck at local optima. The model targets were one-hot encoded to convert the categorical input data into a vector required for the categorical cross-entropy loss function<sup>46</sup>.

The model utilized a generator function<sup>45</sup> for every LAD and yielded lagged batches of the features for the target variables. The arguments of the generator function included:

- Lookback (how many time steps of features to include for each target)
- Lag (how many time steps in the past are the features relative to the target)
- Shuffle (whether to shuffle the order of the training data)
- Batch size (how many samples are used per batch)
- Minimum and maximum indices (the portion of the overall time-series to use for each location)

**Preliminary analysis.** Preliminary exploratory analysis was conducted on the defined lookback period, shuffling of the training order, the number of LSTM and CNN layers, L1 and L2 regularization on dense layers, the shape of the tensor for each layer, and the use of dropout layers. We also assessed the relative impact of different optimization functions: RMSprop<sup>83</sup>, stochastic gradient descent<sup>84</sup> and Adamax<sup>85</sup>.

**Model design.** The final model design included a seven day lookback on the delay periods 15, 20, 25, 30, 35 and 40 days. This determined that the algorithm would, for a target on a given day, utilize the past seven days of features. This was included to capture the weekly trend in the features for a defined risk score of confirmed SARS-CoV-2 cases, hospitalizations or mortalities. Following the sensitivity analysis, we included a shuffling in the order of the training data and developed a model structure that allowed the learning rate to decrease for subsequent epochs if an increase in the validation loss was detected, which is a proxy metric for overfitting.

At the final layers of the SI-LSTM and SI-CNN-LSTM we introduced a connection network between all geographic locations so that the model can learn from the intra and inter-location feature weighting. We merge the 363 independent input branches by combining the list of tensors, from the final LSTM layer for each location, on a single concatenation axis and to produce a single tensor as described in Fig. 3. The final LSTM layer produces a rank-2 tensor of shape  $(b, u)$  where  $b$  is the batch size and  $u$  is the number of units in the LSTM layer. After concatenation of tensors from the  $L$  locations, the resulting tensor has shape  $(b, Lu)$ .

The final dense layer has a softmax activation function, which ensures that the output vector  $y_{i \in \{1, \dots, C\}}$  over  $C$  classes is normalized and that  $y_i$  can be interpreted as the probability that the target is class  $i$ . The cross-entropy loss function is then defined as:

$$\mathcal{L}(y_i, t_i) = - \sum_{i=1}^C t_i \log y_i$$

where  $t_i$  is the one-hot encoded target vector. We then used RMSprop as the optimization function in the back-propagation stage.

**SI-LSTM.** The model has an initial input layer for each location followed by two LSTM layers with a time distributed dropout layer, which helped to prevent overfitting in the early model epochs. There is a final LSTM layer before the model forks, as seen in Fig. 1, to produce a dense side-output layer for each location and a concatenation layer followed by a dense layer. The final output layers have a softmax activation function due to the probabilistic categorical cross-entropy loss function.

**SI-CNN-LSTM.** The SI-CNN-LSTM architecture takes advantage of the feature amplification ability of CNN layers to use a type of weight sharing with local perception to refine and condense the number of parameters that helps to improve the learning efficiency for the LSTM layers<sup>44</sup>. Due to the dimensional size of the features after the one-dimensional CNN layers, a time-distributed dropout layer, a one-dimensional max pooling layer and a flatten layer are included. The model structure then includes three LSTM layers, with the first LSTM layer followed by a dropout layer and a dense layer, with a further dropout layer on the second LSTM layer. The model then branches out to a dense side-output layer and a concatenation layer before the final dense layer.

### Reporting summary

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

### Data availability

Google mobility data are available from <https://www.google.com/covid19/mobility> and Google Trends data can be queried at <https://www.google.com/covid19/mobility>. SARS-CoV-2 cases and deaths data can be found at the required spatial scales on the UK's coronavirus dashboard at <https://coronavirus.data.gov.uk>, as well as on the devolved administration dashboards (<https://www.health-ni.gov.uk/articles/covid-19-daily-dashboard-updates>, [https://www2.nphs.wales.nhs.uk/CommunitySurveillanceDocs.nsf/61c1e930f9121fd080256f2a004937ed/c84f742604ce56f0802586b600374b49/\\$FILE/Rapid%20COVID-19%20surveillance%20data.xlsx](https://www2.nphs.wales.nhs.uk/CommunitySurveillanceDocs.nsf/61c1e930f9121fd080256f2a004937ed/c84f742604ce56f0802586b600374b49/$FILE/Rapid%20COVID-19%20surveillance%20data.xlsx) and <https://www.gov.scot/publications/coronavirus-covid-19-trends-in-daily-data/>). An application can be made to the UK Health Security Agency for the PCR cases and deaths data, and all other data used in this study. Data requests can be made to the Office for Data Release (<https://www.gov.uk/government/publications/accessing-public-health-england-data/about-the-phe-odr-and-accessing-data>) and by contacting [odr@phe.gov.uk](mailto:odr@phe.gov.uk). All requests to access data are reviewed by the ODR and are subject to strict confidentiality provisions in line with the requirements of: the common law duty of confidentiality; data protection legislation (including the General Data Protection Regulation); the eight Caldicott

principles; the Information Commissioner's statutory data sharing code of practice; and the national data opt-out programme.

### Code availability

Supplementary Software 1 and 2 have been included for the deep- and shallow-learning models in R, respectively. Python and PyTorch code for the SI-CNN-LSTM and SI-LSTM models can be made available on reasonable request to the corresponding author.

### References

1. Wu, S. L. et al. Substantial underestimation of SARS-CoV-2 infection in the United States. *Nat. Commun.* **11**, 4507 (2020).
2. Ward, T. & Johnsen, A. Understanding an evolving pandemic: an analysis of the clinical time delay distributions of COVID-19 in the United Kingdom. *PLoS ONE* **16**, e0257978 (2021).
3. Linton, N. et al. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *J. Clin. Med.* **9**, 538 (2020).
4. Davies, N. et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat. Med.* **26**, 1205–1211 (2020).
5. Mobile-powered data and insights (O<sub>2</sub>, 2022); <https://www.o2.co.uk/business/solutions/mobile/data-mobile/o2-motion>
6. COVID-19 Community Mobility Reports (Google, 2021); <https://www.google.com/covid19/mobility/>
7. FACEBOOK Data for Good (Facebook, 2021); <https://dataforgood.fb.com/docs/covid19/>
8. *Coronavirus (COVID-19) Mobility Report* (Greater London Authority, 2021); <https://data.london.gov.uk/dataset/coronavirus-covid-19-mobility-report>
9. Jeffrey, B. et al. Anonymised and aggregated crowd level mobility data from mobile phones suggests that initial compliance with COVID-19 social distancing interventions was high and geographically consistent across the UK. *Wellcome Open Res.* **5**, 170 (2020).
10. Chang, S. et al. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **589**, 82–87 (2020).
11. Gatalo, O., Tseng, K., Hamilton, A., Lin, G. & Klein, E. Associations between phone mobility data and COVID-19 cases. *Lancet Infect. Dis.* **21**, e111 (2020).
12. Grantz, K. et al. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nat. Commun.* **11**, 4961 (2020).
13. Birrell, P., Blake, J., van. Leeuwen, E., Gent, N. & Angelis, D. D. Real-time nowcasting and forecasting of COVID-19 dynamics in England: the first wave. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **376**, 2021 (1829).
14. *Scientific Evidence Supporting the Government Response to Coronavirus (COVID-19)* (SAGE, 2021); <https://www.gov.uk/government/collections/scientific-evidence-supporting-the-government-response-to-coronavirus-covid-19>
15. Cleaton, J., Viboud, C., Simonsen, L., Hurtado, A. & Chowell, G. Characterizing Ebola transmission patterns based on internet news reports. *Clin. Infect. Dis.* **62**, 24–31 (2015).
16. Carneiro, H. A. & Mylonakis, E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clin. Infect. Dis.* **49**, 1557–1564 (2009).
17. Husnayaina, A., Fuad, A. & Su, E. C.-Y. Applications of Google Search trends for risk communication in infectious disease management: a case study of the COVID-19 outbreak in Taiwan. *Int. J. Infect. Dis.* **95**, 221–223 (2020).
18. Venkatesh, U. & Gandhi, P. Prediction of COVID-19 outbreaks using google trends in India: a retrospective analysis. *Healthc. Inform. Res.* **26**, 175–184 (2020).



19. Jurić, T. Google Trends as a method to predict new COVID-19 cases. *Athens J. Med. Stud.* **8**, 67–92 (2021).
20. Jimenez, A., Estevez-Reboredo, R., Santed, M. & Ramos, V. COVID-19 symptom-related Google searches and local COVID-19 incidence in Spain: correlational study. *J. Med. Internet Res.* **22**, e23518 (2020).
21. Kurian, S. et al. Correlations between COVID-19 cases and google trends data in the United States: a state-by-state analysis. *Mayo Clin. Proc.* **95**, 2370–2381 (2020).
22. Ginsberg, J. et al. Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014 (2009).
23. Butler, D. When Google got flu wrong. *Nature* **494**, 155–156 (2013).
24. Santillana, M. et al. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput. Biol.* **11**, e1004513 (2015).
25. Mahase, E. Covid-19: the problems with case counting. *Brit. Med. J.* **370**, m3374 (2020).
26. Vandentorren, S. et al. The effect of social deprivation on the dynamic of SARS-CoV-2 infection in France: a population-based analysis. *Lancet Public Health* **7**, e240–e249 (2022).
27. Coronavirus (COVID-19) Infection Survey, Characteristics of People Testing Positive for COVID-19, UK: 25 May 2022 (Office for National Statistics, 2022); <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveycharacteristicsofpeopletestingpositiveforcovid19uk/25may2022>
28. Hendricks, B. et al. Coronavirus testing disparities associated with community level deprivation, racial inequalities, and food insecurity in West Virginia. *Annals Epidemiol.* **59**, 41–49 (2021).
29. Sherratt, K. et al. Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of COVID-19 in England. *Phil. Trans. R. Soc.* **376**, 20200283 (2021).
30. Surge Testing for New Coronavirus (COVID-19) Variants (UK Health Security Agency, 2021); <https://www.gov.uk/guidance/surge-testing-for-new-coronavirus-covid-19-variants>
31. Pelat, C. et al. More diseases tracked by using Google Trends. *Emerging Infect. Dis.* **15**, 1327–1328 (2008).
32. NHS COVID-19 App (UK Health Security Agency, 2022); <https://www.gov.uk/government/collections/nhs-covid-19-app>
33. Local Authority Districts (April 2020) Names and Codes in the United Kingdom (Office of National Statistics, 2021); [https://geoportal.statistics.gov.uk/datasets/fe6bcee87d95476abc84e194fe088abb\\_0](https://geoportal.statistics.gov.uk/datasets/fe6bcee87d95476abc84e194fe088abb_0)
34. Output Area to Lower Layer Super Output Area to Middle Layer Super Output Area to Local Authority District (December 2020) Lookup in England and Wales (Office of National Statistics, 2021); <https://geoportal.statistics.gov.uk/datasets/output-area-to-lower-layer-super-output-area-to-middle-layer-super-output-area-to-local-authority-district-december-2020-lookup-in-england-and-wales/explore>
35. Get a Free PCR Test to Check if you Have Coronavirus (COVID-19) (GOV.UK, 2021); [www.gov.uk/get-coronavirus-test](http://www.gov.uk/get-coronavirus-test)
36. NPEx: A National Scale Solution for the COVID-19 Crisis (NPEx, 2021); <https://www.npex.nhs.uk/news/200409>
37. Secondary Uses Service (NHS, 2022); <https://digital.nhs.uk/services/secondary-uses-service-sus/secondary-uses-service-sus-what-s-new>
38. Coronavirus (COVID-19): Daily Data for Scotland (Scottish Government, 2021); <https://www.gov.scot/publications/coronavirus-covid-19-daily-data-for-scotland/>
39. COVID-19—Daily Dashboard Updates (Department of Health, 2021); <https://www.health-ni.gov.uk/articles/covid-19-daily-dashboard-updates>
40. Breiman, L. Random Forests. *Mach. Learn.* **25**, 5–32 (2001).
41. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, 2016)*.
42. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* 339 (Springer, 2001).
43. Van der Laan, M., Polley, E. & Hubbard, A. Super learner. *Stat. Appl. Genet. Mol. Biol.* **6**, 25 (2007).
44. Sainath, T., Vinyals, O., Senior, A. & Sak, H. Convolutional, long short-term memory, fully connected deep neural networks. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (IEEE, 2015)*.
45. Chollet, F. & Allaire, J. *Deep Learning* (Manning, 2018).
46. *Probabilistic Losses* (Keras, 2021); [https://keras.io/api/losses/probabilistic\\_losses/](https://keras.io/api/losses/probabilistic_losses/)
47. *Review of Local Restriction Tiers* (GOV.UK, 2020); <https://www.gov.uk/government/speeches/review-of-local-restriction-tiers-17-december-2020>
48. Kraemer, M. et al. Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* **373**, 889–895 (2021).
49. *Prime Minister announces new national restrictions* (Prime Minister's Office, 2020); <https://www.gov.uk/government/news/prime-minister-announces-new-national-restrictions>
50. *First UK cases of Omicron Variant Identified* (Department of Health and Social Care, 2021); <https://www.gov.uk/government/news/first-uk-cases-of-omicron-variant-identified>
51. *Risk Related to Spread of New SARS-CoV-2 Variants of Concern in the EU/EEA* (ECDC, 2021); <https://www.ecdc.europa.eu/sites/default/files/documents/COVID-19-risk-related-to-spread-of-new-SARS-CoV-2-variants-EU-EEA.pdf>
52. *SARS-CoV-2 Variants of Concern and Variants Under Investigation in England* (Public Health England, 2021); [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/984274/Variants\\_of\\_Concern\\_VOC\\_Technical\\_Briefing\\_10\\_England.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/984274/Variants_of_Concern_VOC_Technical_Briefing_10_England.pdf)
53. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *MIT Press* **9**, 1735–1780 (1997).
54. Atchison, C. et al. Early perceptions and behavioural responses during the COVID-19 pandemic: a cross-sectional survey of UK adults. *BMJ Open* **11**, e043577 (2020).
55. Graham, M. S. et al. Knowledge barriers in a national symptomatic-COVID-19 testing programme. *PLOS Glob. Public Health* **2**, e0000028 (2022).
56. Vahedi, B., Karimzadeh, M. & Zoraghein, H. Spatiotemporal prediction of COVID-19 cases using inter- and intra-county proxies of human interactions. *Nat. Commun.* **12**, 6440 (2021).
57. Fischer, I. et al. The behavioural challenge of the COVID-19 pandemic: indirect measurements and personalized attitude changing treatments (IMPACT). *R. Soc. Open Sci.* **7**, 201131 (2020).
58. Naughton, F. et al. Health behaviour change during the UK COVID-19 lockdown: findings from the first wave of the C-19 health behaviour and well-being daily tracker study. *Health Psychol.* **26**, 624–643 (2021).
59. Elliot, A. et al. The COVID-19 pandemic: a new challenge for syndromic surveillance. *Epidemiol. Infection* **148**, e122 (2020).
60. Elliot, A. et al. The potential impact of media reporting in syndromic surveillance: an example using a possible Cryptosporidium exposure in North West England, August to September 2015. *Eurosurveillance* **21**, 30368 (2016).
61. What are the symptoms of omicron? ZOE (7 February 2022); <https://joinzoe.com/learn/omicron-symptoms>
62. Greenhalgh, T., Knight, M., Buxton, M. & Husain, L. Management of post-acute covid-19 in primary care. *Brit. Med. J.* **370**, m3026 (2020).

63. Pierce, M. et al. Mental health responses to the COVID-19 pandemic: a latent class trajectory analysis using longitudinal UK data. *Lancet Psychiatry* **8**, 610–619 (2021).
64. Mossong, J. et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5**, e74 (2008).
65. Jarvis, C., Gimma, A., Wong, K. & Edmunds, J. *Social Contacts in the UK From the CoMix Social Contact Survey* (GOV.UK, 2022); <https://cmmid.github.io/topics/covid19/reports/comix/Comix%20Weekly%20Report%20101.pdf>
66. Monod, M. et al. Age groups that sustain resurging COVID-19 epidemics in the United States. *Science* **371**, eabe8372 (2021).
67. Paton, R. S., Overton, C. & Ward, T. The rapid replacement of the Delta variant by Omicron (B.1.1.529) in England. *Sci. Transl. Med.* **14**, eabo5395 (2022).
68. Dolton, P. The statistical challenges of modelling COVID-19. *Nat Institute Econ. Rev.* **257**, 46–82 (2021).
69. Moein, S. et al. Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan. *Sci. Rep.* **11**, 4725 (2021).
70. *Reproduction Number (R) and Growth Rate (r) of the COVID-19 Epidemic in the UK: Methods of Estimation, Data Sources, Causes of Heterogeneity, and Use as a Guide in Policy Formulation* (The Royal Society, 2020); **Royal Society publishes rapid review of the science of the reproduction number and growth rate of COVID-19.**
71. Roda, W., Varughese, M., Han, D. & Li, M. Y. Why is it difficult to accurately predict the COVID-19 epidemic? *Infect. Dis. Modell.* **5**, 271–281 (2020).
72. Ioannidis, J., Cripps, S. & Tanner, M. Forecasting for COVID-19 has failed. *Int. J. Forecast.* **2**, 423–438 (2022).
73. Overton, C. & Ward, T. Omicron and Delta serial interval distributions from UK contact tracing data (2021); [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/1046481/S1480\\_UK\\_HSA\\_Omicron\\_serial\\_intervals.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1046481/S1480_UK_HSA_Omicron_serial_intervals.pdf)
74. Niazkar, H. R. & Niazkar, M. Application of artificial neural networks to predict the COVID-19 outbreak. *Glob. Health Res. Policy* **5**, 50 (2020).
75. Alali, Y., Harrou, F. & Sun, Y. A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models. *Sci. Rep.* **12**, 2467 (2022).
76. Kumar, R. L. et al. Recurrent neural network and reinforcement learning model for COVID-19 prediction. *Front. Public Health* **9**, 744100 (2021).
77. Lin, Y. T. et al. Daily forecasting of regional epidemics of coronavirus disease with bayesian uncertainty quantification, United States. *Emerging Infectious Diseases* **3**, 810–821 (2021).
78. *FAQ About Google Trends Data* (Google, 2021); <https://support.google.com/trends/answer/4365533>
79. *COVID-19: Epidemiology, Virology and Clinical Features* (Public Health England, 2021); <https://www.gov.uk/government/publications/wuhan-novel-coronavirus-background-information/wuhan-novel-coronavirus-epidemiology-virology-and-clinical-features>
80. *NHS 111* (NHS, 2021); <https://www.england.nhs.uk/urgent-emergency-care/nhs-111/>
81. *Google, Privacy & Terms* (Google, 2021); <https://policies.google.com/technologies/location-data>
82. Hyndman, J. & Khandakar, Y. Automatic time series forecasting: the forecast package for R. *J. Stat. Softw.* **27**, 1–22 (2008).
83. Tieleman, T. & Hinton, G. *Lecture 6.5-rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude*. COURSERA: *Neural Networks for Machine Learning* Vol. 4, 26–31 (Scirp, 2012).
84. Kiefer, J. & Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* **23**, 462–466 (1952).
85. Kingma, D. & Ba, J. Adam: a method for stochastic optimization. International Conference on Learning Representations (ICLR) (2015).

## Acknowledgements

We give special thanks to S. Hall for his faith and determination to support this work, and to S. Feeney, R. Stainer and H. Shamsi for their contribution at the onset of this project. We would also like to thank B. Pinnington and S. Rogers from the Google Trends editorial team for their eagerness to help with the acquisition of the Trends data required.

## Author contributions

T.W. conceived, designed and led the study. T.W. and F.C. designed the deep learning models. T.W. and A.J. wrote the model code. A.J., S.N. and T.W. developed and tuned the shallow learning models. T.W. conducted the sensitivity analysis for the deep learning models. A.J., S.N. and T.W. developed the visualizations. T.W. analysed and interpreted the results. T.W. and A.J. wrote the original manuscript. T.W. reviewed the manuscript and wrote the revisions.

## Competing interests

The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s42256-022-00538-9>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s42256-022-00538-9>.

**Correspondence and requests for materials** should be addressed to Thomas Ward.

**Peer review information** *Nature Machine Intelligence* thanks Xuhong Zhang, Ioanna Miliou and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

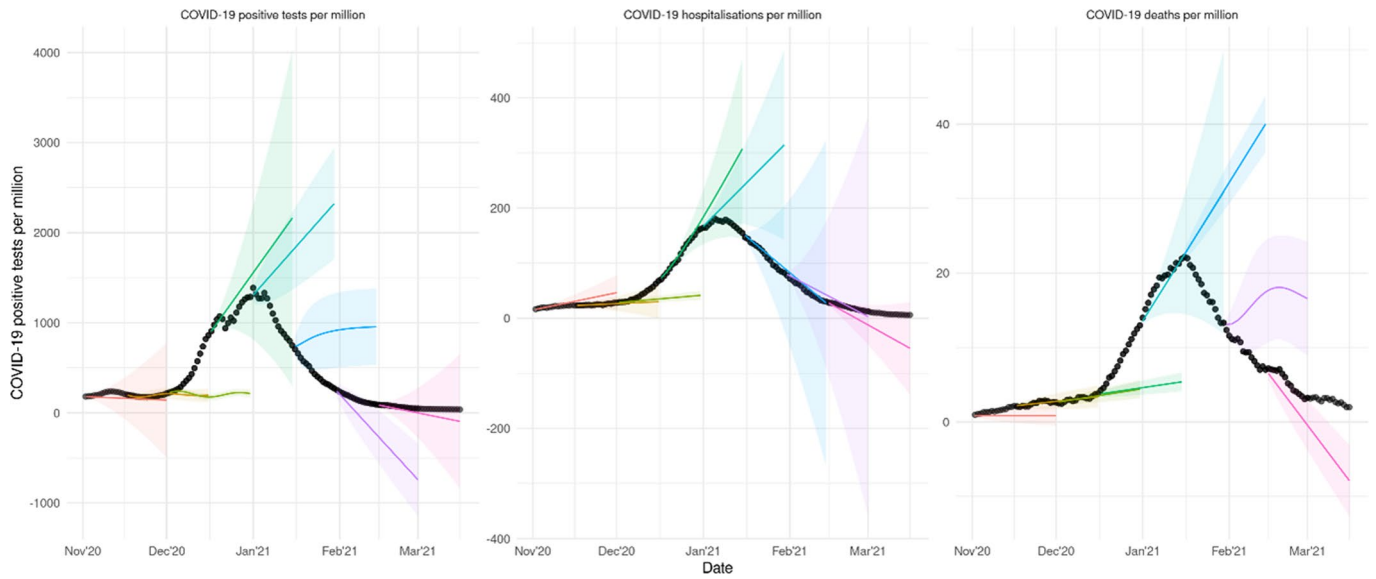
**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

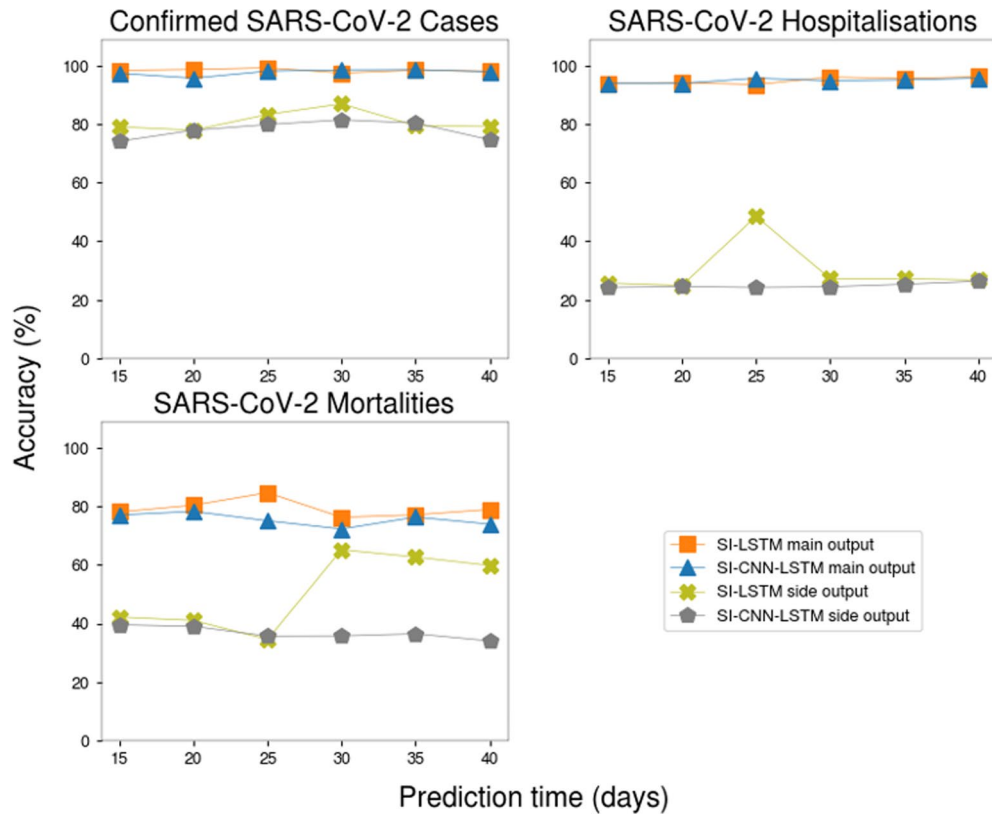
Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2022

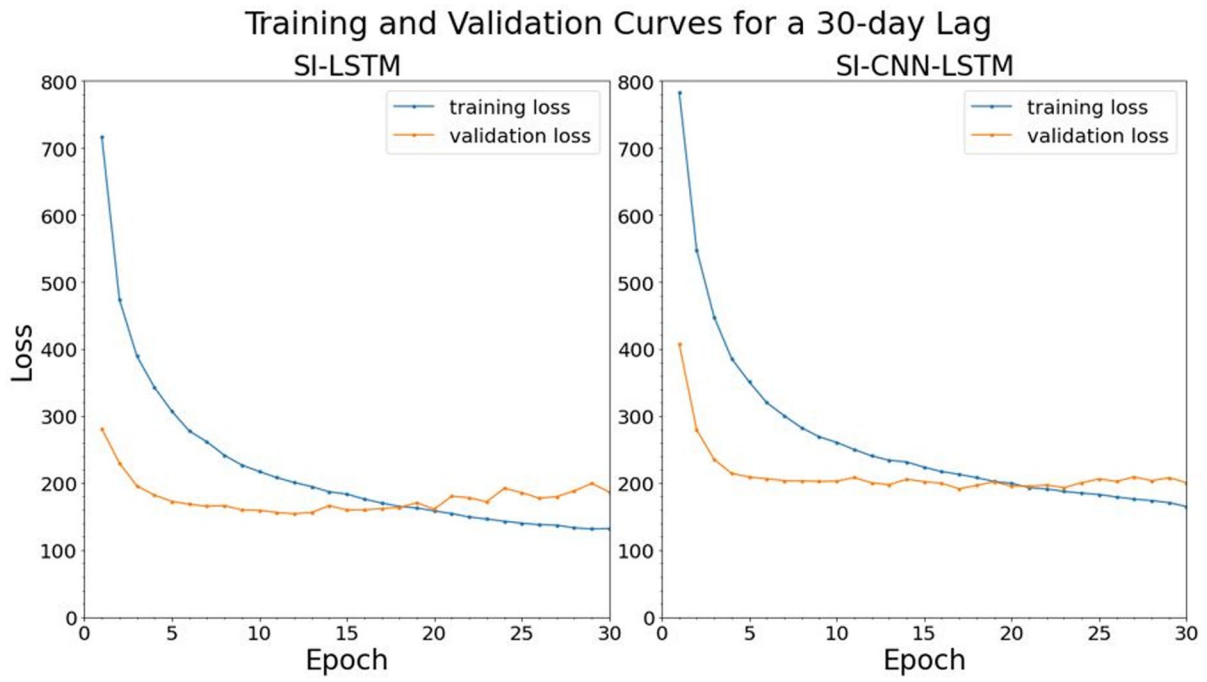




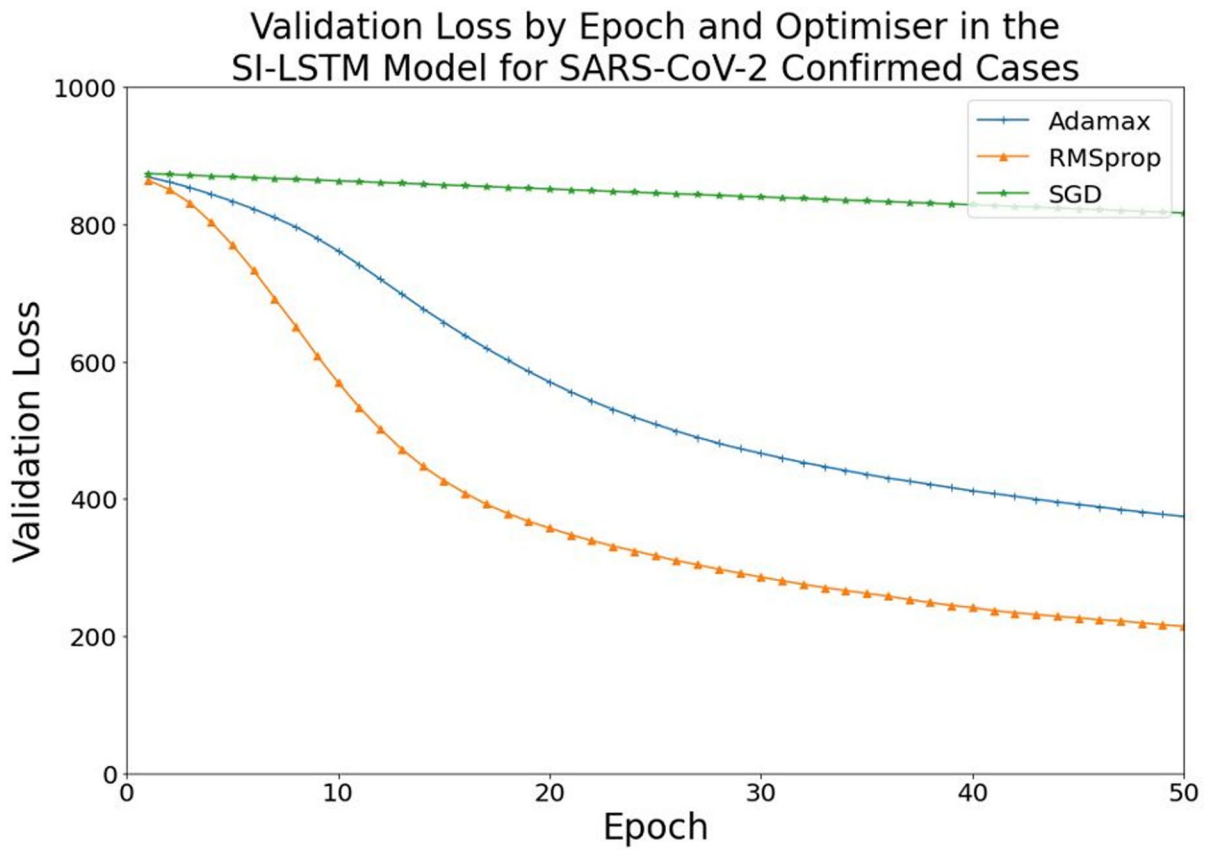
**Extended Data Fig. 1 | ARIMA modelling forecasts for Positive PCR tests, hospitalizations, and mortalities from COVID-19 using a Hyndman–Khandakar algorithm across the Alpha wave from 1<sup>st</sup> November 2020 – 15<sup>th</sup> February 2021.**



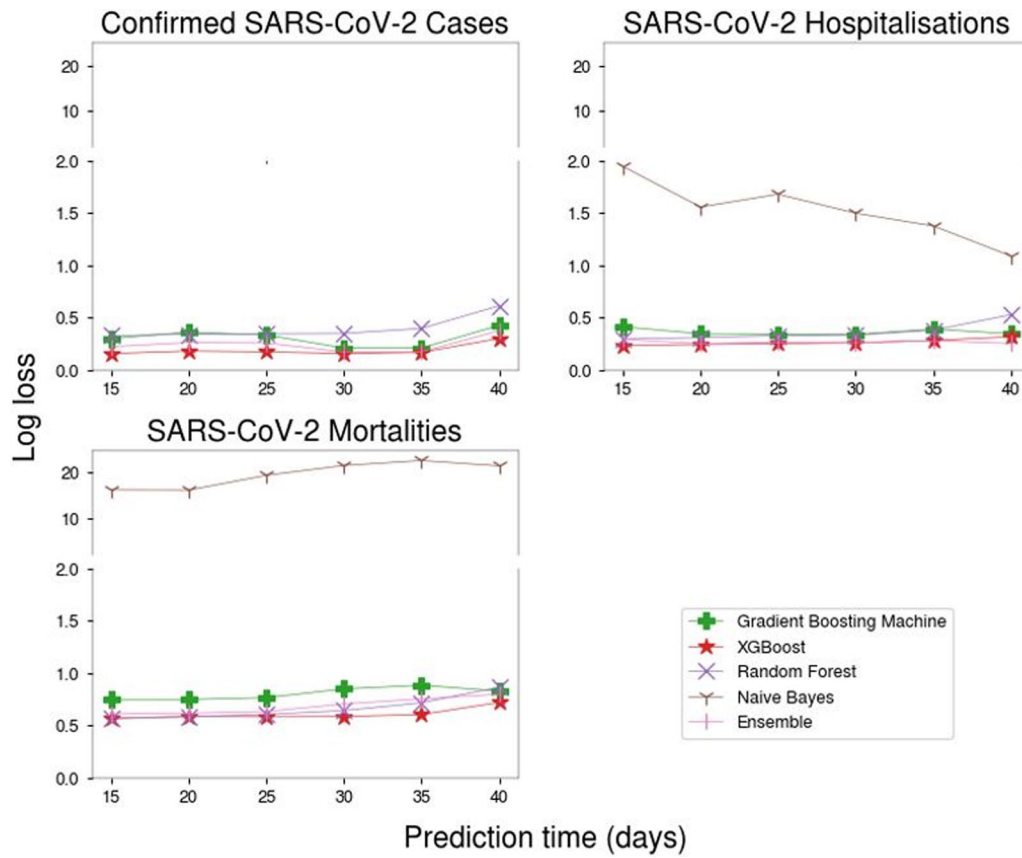
**Extended Data Fig. 2 | A line graph of the side output and main output model accuracy for SARS-CoV-2 Case, Hospitalization, and Mortality Risk Scores for the SI-LSTM and SI-CNN-LSTM algorithms across the temporal delay periods.**



**Extended Data Fig. 3 | A line graph of the training and validation loss for the SI-LSTM and SI-CNN-LSTM models with a 30-day target lag for the Confirmed Case Risk Score.**

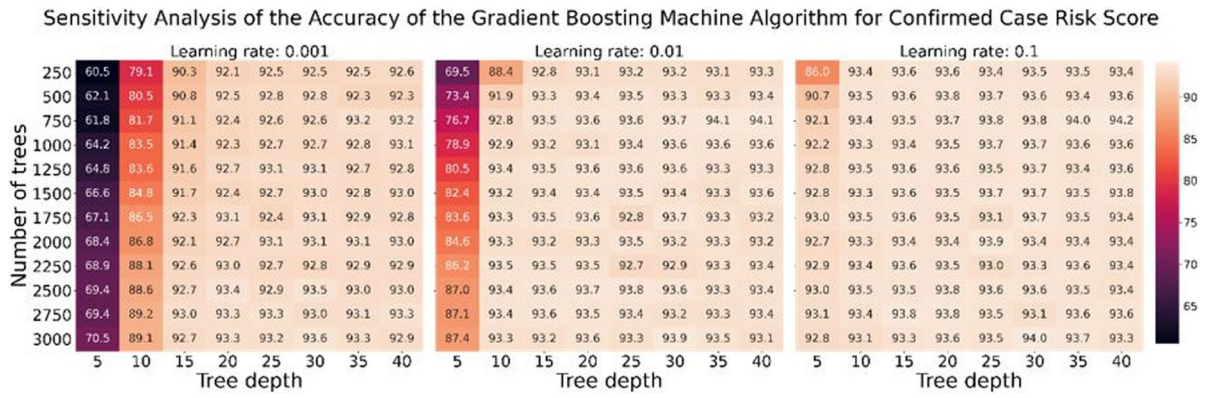


**Extended Data Fig. 4** | A line graph of the validation loss for the optimizer functions Adamax, RMSprop, and stochastic gradient descent for the Confirmed Case Risk Scores.



**Extended Data Fig. 5 |** A line graph of the log loss results for SARS-CoV-2 Confirmed Case, Hospitalization, and Mortality Risk Scores for the shallow learning algorithms across the temporal delay periods.





Extended Data Fig. 6 | Sensitivity analysis for the Gradient Boosting Machine (GBM) algorithm for the confirmed case risk score.



Extended Data Table 1 | ARIMA risk score forecasting performance for Positive PCR tests, hospitalizations, and mortalities from COVID-19 across the Alpha wave from 1st November 2020 – 15th February 2021

<b>Risk Category</b>	<b>Accuracy</b>
PCR Positive Cases	31.4%
Hospitalisations	50.1%
Mortalities	39.8%

**Extended Data Table 2 | A table of the overall model feature importance for Google Trends, telecoms mobility, Google mobility, 119 calls, and website test request journeys using a Random Forest algorithm across the assessed periods**

<i>Features</i>	<i>RF 15-day PCR confirmed cases Accuracy / %</i>	<i>RF 20-day hospitalisations Accuracy / %</i>	<i>RF 25-day mortalities Accuracy / %</i>
<i>All (Google Trends, Google Mobility, Telecoms Mobility, 119 Calls, Website Testing Demand, Location)</i>	90.4	92.7	72.5
<i>Google Trends</i>	87.4	89.9	65.4
<i>Google Mobility</i>	71.6	81.4	59.2
<i>Telecoms Mobility</i>	84.4	89.1	64.8
<i>119 Calls</i>	35.4	69.3	54.2
<i>Website Testing Demand</i>	49.9	74.4	54.5

## Reporting Summary

Nature Portfolio 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 Portfolio policies, see our [Editorial Policies](#) 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   |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of all covariates tested  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> 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) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> 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

Data analysis

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Google mobility data is available from the public website <https://www.google.com/covid19/mobility> and Google Trends data can be queried at <https://www.google.com/covid19/mobility>. SARS-CoV-2 cases and deaths data can be found at the required spatial scales on the UK's coronavirus dashboard at <https://coronavirus.data.gov.uk> and on the devolved administration dashboards (<https://www.health-ni.gov.uk/articles/covid-19-daily-dashboard-updates>, [https://www.2.nphs.wales.nhs.uk/CommunitySurveillanceDocs.nsf/61c1e930f9121fd080256f2a004937ed/c84f742604ce56f0802586b600374b49/\\$FILE/Rapid%20Report%20-%2020220722.pdf](https://www.2.nphs.wales.nhs.uk/CommunitySurveillanceDocs.nsf/61c1e930f9121fd080256f2a004937ed/c84f742604ce56f0802586b600374b49/$FILE/Rapid%20Report%20-%2020220722.pdf))



20COVID-19%20surveillance%20data.xlsx, <https://www.gov.scot/publications/coronavirus-covid-19-trends-in-daily-data/>).

For the PCR cases and deaths data and all other data used in this study, an application can be made to the UK Health Security Agency. Data requests can be made to the Office for Data Release (<https://www.gov.uk/government/publications/accessing-public-health-england-data/about-the-phe-odr-and-accessing-data>) and contacting [odr@phe.gov.uk](mailto:odr@phe.gov.uk). All requests to access data are reviewed by the ODR and are subject to strict confidentiality provisions in line with the requirements of:

- the common law duty of confidentiality
- data protection legislation (including the General Data Protection Regulation)
- 8 Caldicott principles
- the Information Commissioner's statutory data sharing code of practice
- the national data opt-out programme

## Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

### Reporting on sex and gender

*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data where this information has been collected, and consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*

### Population characteristics

*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*

### Recruitment

*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*

### Ethics oversight

*Identify the organization(s) that approved the study protocol.*

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

## 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](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

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

Sample size	A total of 82264 rows of data. Each row corresponding to one day in each Local Authority District in the United Kingdom.
Data exclusions	The Isles of Scilly were excluded from the UK analysis due to poor quality data for NHS pathways, mobility and Google relative search volume. Mortality risk did not include Wales as the data at the time was not available at Local Authority District.
Replication	The model runs for each algorithm were replicated during sensitivity analysis of the hyperparameters.
Randomization	The model accuracy was measured at randomly selected time series (not included in the training of the algorithm).
Blinding	Not relevant to this study.

## Behavioural & social sciences study design

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

Study description	<i>Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).</i>
-------------------	--

Research sample	State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.
Sampling strategy	Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.
Data collection	Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.
Timing	Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Non-participation	State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.
Randomization	If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

## Ecological, evolutionary & environmental sciences study design

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

Study description	Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.
Research sample	Describe the research sample (e.g. a group of tagged <i>Passer domesticus</i> , all <i>Stenocereus thurberi</i> within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.
Sampling strategy	Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.
Data collection	Describe the data collection procedure, including who recorded the data and how.
Timing and spatial scale	Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Reproducibility	Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.
Randomization	Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.
Blinding	Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Did the study involve field work?  Yes  No

## Field work, collection and transport

Field conditions	Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).
Location	State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).
Access & import/export	Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in

Access & import/export *compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*

Disturbance *Describe any disturbance caused by the study and how it was minimized.*

## 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                                 | Involvement in the study                               |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies                    |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data                 |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern  |

### Methods

- | n/a                                 | Involvement in the study                        |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq               |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |

## Antibodies

Antibodies used *Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Validation *Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

## Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s) *State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*

Authentication *Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*

Mycoplasma contamination *Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*

Commonly misidentified lines (See [ICLAC](#) register) *Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

## Palaeontology and Archaeology

Specimen provenance *Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*

Specimen deposition *Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods *If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

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

## Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals	<i>For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.</i>
Wild animals	<i>Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.</i>
Reporting on sex	<i>Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.</i>
Field-collected samples	<i>For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.</i>
Ethics oversight	<i>Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.</i>

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

## Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	<i>Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.</i>
Study protocol	<i>Note where the full trial protocol can be accessed OR if not available, explain why.</i>
Data collection	<i>Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.</i>
Outcomes	<i>Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.</i>

## Dual use research of concern

Policy information about [dual use research of concern](#)

### Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Public health
<input checked="" type="checkbox"/>	<input type="checkbox"/>	National security
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Crops and/or livestock
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Ecosystems
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Any other significant area

## Experiments of concern

Does the work involve any of these experiments of concern:

- | No                                  | Yes                      |
|-------------------------------------|--------------------------|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> |
- Demonstrate how to render a vaccine ineffective
  - Confer resistance to therapeutically useful antibiotics or antiviral agents
  - Enhance the virulence of a pathogen or render a nonpathogen virulent
  - Increase transmissibility of a pathogen
  - Alter the host range of a pathogen
  - Enable evasion of diagnostic/detection modalities
  - Enable the weaponization of a biological agent or toxin
  - Any other potentially harmful combination of experiments and agents

## ChIP-seq

### Data deposition

- Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

#### Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

#### Files in database submission

Provide a list of all files available in the database submission.

#### Genome browser session

(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

## Methodology

#### Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

#### Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

#### Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

#### Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

#### Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

#### Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

## Flow Cytometry

### Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

#### Sample preparation

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

#### Instrument

Identify the instrument used for data collection, specifying make and model number.

Software *Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*

Cell population abundance *Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*

Gating strategy *Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

## Magnetic resonance imaging

### Experimental design

Design type *Indicate task or resting state; event-related or block design.*

Design specifications *Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*

Behavioral performance measures *State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

### Acquisition

Imaging type(s) *Specify: functional, structural, diffusion, perfusion.*

Field strength *Specify in Tesla*

Sequence & imaging parameters *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*

Area of acquisition *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI  Used  Not used

### Preprocessing

Preprocessing software *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

Normalization *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

### Statistical modeling & inference

Model type and settings *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis:  Whole brain  ROI-based  Both

Statistic type for inference (See [Eklund et al. 2016](#)) *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

## Models & analysis

- n/a | Involved in the study
- Functional and/or effective connectivity
  - Graph analysis
  - Multivariate modeling or predictive analysis

Functional and/or effective connectivity

*Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

Graph analysis

*Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

Multivariate modeling and predictive analysis

*Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*