THE LANCET
Digital Health

## Supplementary appendix

Supplement to: Myall A, Price JR, Peach RL, et al. Prediction of hospital-onset COVID-19 infections using dynamic networks of patient contact: an international retrospective cohort study. *Lancet Digit Health* 2022; **4:** e573–83.

# Contents

# 1 Methods

## 1.1 Study design and population expanded

**Training and testing (internal)** The study incorporated training and testing data, from 01 April 2020 to 01 April 2021, capturing both the first surge (23 March 2020 - 30 May 2020) and second surge (07 September 2020 – 01 April 2021) of the COVID-19 pandemic in the UK, in line with and our original studies available dates (UK Office for national statistics). In total 51,157 patient admissions were recorded and incorporated over this period.

**Validation (internal)** The internal validation was performed on data from the same 5-hospital group in London in the period immediately following surge 2, to the latest available data (02 April 2021 - 13 August 2021), incorporating 43,375 patient admissions.

**Validation (external)** External validation was also performed on an internationally separate university-affiliated geriatric hospital in Geneva during Switzerland's first surge of COVID-19 (01 February 2020 – 31 May 2020), using 40,057 patient admissions.

**Study population inclusion)** In all datasets, all inpatients with a bed allocation were included in the formation of the dataset (i.e. construction of the contact networks, recording of contextual background data). However, only inpatients spending three or more consecutive days in the hospital were included as samples to predict HOCI versus control.

## 1.2 Infection prevention and control measures

The Infection Prevention and Control (IPC) measures employed by the London hospital trust aligned with national recommendations[1]. A comprehensive strategy of inpatient COVID-19 screening was deployed which included admission screening on day 3-5, day 7 and weekly thereafter (changing to daily for the first seven days following admission date, from 01 December 2021). Throughout the pandemic, a robust surveillance programmev from *Price and Mookerjee et al.*[2] was employed to identify HOCI cases, with HOCIs triggering full IPC investigations, patient, and contact isolation, as well as screening. During the first surge in Geneva, syndromic surveillance of hospitalised patients was conducted, as well as ad-hoc screening of patients during outbreak management. In addition, patients were screened prior to transfer between hospital sites as of April 2020.

## 1.3 Background spatial information

**5-hospital group in London** The 1,200 bed hospital group contained 96 wards, a mixture of nightingale, bay, and side-room layouts. These wards were located over different 17 specialities and 19 buildings[3]. Moreover, comprising of historical buildings, the architecture spans multiple centuries of building design.

## 1.4 Patient pathways as trajectories

We consider the spatial temporal pathway histories of $N$ hospital patients represented by a set of trajectories $\mathscr{T} = \{T_1, T_2, T_3, ..., T_N\}$. Each trajectory $T_n$ is a time-ordered set of spatial-temporal locations

$$T_n = \{l_1, l_2, l_3, ..., l_{k_n}\}, \quad n = 1, \ldots, N, \tag{1}$$

where each element $l_i = (v_i, t_i)$ is a tuple that contains the spatial location (hospital room, hospital ward, or hospital building) $v_i \in V$ visited by the individual at time $t_i$.

We use the spatial-temporal locations to identify patient contacts and construct the contact network that forms the basis of this manuscript.

### 1.5 Contact definition

We extract the set of all contacts, $C$, between $N$ patients. Formally, we define a contact between two patients when they *coincide at a location*, i.e when their trajectories intersect i.e. $T_n \cap T_m \neq \emptyset$. Each contact event then takes the form,

$$c = (n_i, n_j, t), \tag{2}$$

where $n_i$ and $n_j$ indicate patients who came into contact at time $t$.

'Coincidence' of patients can be defined in different ways, and we investigate three alternative measures of contact: (1) The first requires two patients to reside simultaneously in the same room, (2) the second requires two patients to reside simultaneously in the same ward, and (3) the third requires two patients simultaneously in the same building. All definitions were used by the hospitals epidemiology team to capture different potential routes of transmission.

### 1.6 Time-varying contact network

The set of all contacts $C$ forms the basis of a time-dependent contact network $G^{(t)} = (N^{(t)}, E^{(t)})$. Edges $E^{(t)}$ are a subset of the total contacts $E^{(t)} \subset C$, for all contacts $c$ occurring at time $t$. We compute time-dependence of the contact network by including all edges present within a relevant time window [4]. Thus, the windowed time varying contact network $G^{(t_n,\ldots,t_m)}$, captures all contact events occurring between $t_n$ and $t_m$. Moreover, since $G^{(t_n,\ldots,t_m)}$ can capture multiple interactions between two individuals, we consider individual realisations as a weighted graph, with edges weighted by the duration $E_{ij}^{(t_n,\ldots,t_m)} = \{w_{ij}\}$ of contact between individuals $n_i$ and $n_j$ during $t_n$ and $t_m$. For example, for a given time window, if two patients spent three days in the same spatial location then $= \{w_{ij}\} = 3$. For a full introduction into time-varying networks we direct the reader to a review by *Holme and Saramäkid* [5].

### 1.7 Network Centrality measures

Here we outline different measures of network centrality utilised in this study. For an expanded introduction, we direct the reader to *Newman* [6]. Each of the below equations is applied to every windowed time-varying contact network, and the resulting value provides a unique variable of an individual $n_i$ between $t_n$ and $t_m$.

Analysis was primarily performed in Python[7] and *NetworkX*[8], and the final release code also makes use of the *igraph*[9] R package.

**Degree** The degree of a node $n_i$ is the number of edges to other nodes $n_j$ in the network $G$, given as $deg(n_i)$.

**Degree centrality** The degree *centrality* of a node $n_i$ is the fraction of nodes connected to $n_i$ among all other nodes in the network $G$:

$$C_{deg}(n_i) = \frac{deg(n_i)}{N-1}. \tag{3}$$

**Closeness centrality** The closeness centrality[10] of a node $n_i$ is the reciprocal of the average shortest path distance to all other nodes $n_j$ which are reachable:

$$C_{close}(n_i) = \frac{N-1}{\sum_{n_j} d(n_j, n_i)}, \tag{4}$$

where $d(n_j, n_i)$ is the shortest path distance between nodes $n_j$ and $n_i$.

**PageRank centrality** Given the adjacency matrix $A$ of the graph $G$, the PageRank[11] centrality $x_i$ of node $n_i$ is given by:

$$x_i = \alpha \sum_k \frac{a_{k,i}}{d_k} x_k + \beta, \tag{5}$$

where $\alpha$ and $\beta$ are constants and $d_k$ is the out-degree of node $k$ if such degree is positive, and $d_k = 1$ if it is zero.

**Local clustering coefficient** For the local Clustering coefficient of a node $n_i$[12,13], we first consider only its immediate neighbourhood, $\Pi_i$, i,e. all nodes that are directly connected to $n_i$,

$$\Pi_i = \{n_i : e_{ij}, E \wedge e_{ij} \in E\}. \tag{6}$$

The size of $\Pi_i$, is the number of nodes in $\Pi_i$, i.e. $k_i = |\Pi_i|$, and the localised clustering coefficient for an undirected graph is then finally computed as the ratio of number edges among nodes in $\Pi_i$ and the maximally possible number of edges among them:

$$C_{clust}(n_i) = \frac{2|\{e_{jk} : n_j, n_k \in \Pi_i, e_{jk} \in E\}|}{k_i(k_i - 1)}, \tag{7}$$

where $k_i$ is the number of neighbours of a node.

**Betweenness centrality** The Betweenness centrality[14] of a node $n_i$ is the sum of the fractions of all pairs of shortest paths that pass through it:

$$C_{between}(n_i) = \sum_{s,t,\in N} \frac{\sigma(s,t|N_j)}{\sigma(s,t)}. \tag{8}$$

where $\sigma(s,t)$ is the total number of shortest paths and $\sigma(s,t|n_j)$ the number of shortest paths passing through nodes other than $n_i$.

**K-core** We employ a general definition of centrality as given by the $k$-core number of node $n_i$ [15]. The K-core number is obtained via a $k$-core decomposition over $G$. Specifically, this obtained by iteratively removing nodes with degrees smaller than $k$, until the minimum degree in the network is $k$. The K-core value of a node $n_i$ is then given by,

$$kc(n_i) = k, \tag{9}$$

if it belongs to the $k$-core, but not in the $(k+1)$-core. According to this measure, the most central nodes will have the highest $k$-core number.

### 1.8 Network centrality with respect to infectious nodes

In addition to the measures of network centrality described in the previous section, we also introduce two new measures that we adapt from the degree and closeness centrality, respectively. Instead of capturing a node's centrality with respect to all other nodes in the network, we consider their centrality with respect to only the infectious nodes. These adapted centrality measures therefore directly consider the background distribution of infections, as well as the position of an individual within the contact network in relation to potential disease sources.

**Infectious Degree.** We define the infected degree of a node $n_i$ by the number of connected nodes that are also in the infectious set $I$:

$$deg_{n_j \in I}(n_i). \tag{10}$$

**Infectious Degree centrality.** Much like the degree centrality of a node, the infectious degree considers the immediate local connectivity of node $n_i$. However, instead of considering all other immediate contacts, the Infectious Degree only counts connections to immediate contacts $n_j$ if they are in the set of infected patients $I$:

$$C'_{deg}(n_i) = deg_{n_j \in I}(n_i). \tag{11}$$

**Infectious Closeness centrality.** Similarly, the infectious closeness centrality introduces an adaption to the closeness centrality by restricting the measure to consider only the infectious set $I$. Thus, infectious closeness centrality only considers the reciprocal of the average shortest path distance to all other nodes $n_j$ that are reachable and in the infectious set $I$:

$$C'_{close}(n_i) = \frac{N-1}{\sum_{n_j \in I} d(n_j, n_i)}. \tag{12}$$

## 1.9 Global network metrics

Additional to the above node-centric measures we also extracted variables of the entire graphs[16]. Global network metrics can, for example, give insight into how a pathogen will spread over the connections[17].

**Coefficient of variation** We quantified heterogeneity in the number of contacts of each individual by the coefficient of variation (CV) in the degree distribution. The CV can, for example, indicate the presence of 'super-spreaders' (a minority of individuals who infect many others[18]) and can indicate how fast a disease will spread directly across a contact network[19,20]. We compute the CV of a graph $G$ as the ratio of standard deviation, $\sigma$, over mean, $\mu$, of the degree distribution:

$$CV = \frac{\sigma}{\mu}. \tag{13}$$

**Global Clustering coefficient** The global clustering coefficient is a measure of how tightly connected nodes are across the whole graph and is well-known indicator of disease spreading dynamics[21]. For example, a high global clustering coefficient indicates a fast spread of the disease [21]. For an undirected graph $G$, the clustering coefficient is given by,

$$C(G) = \frac{\sum_{i,j,k} A_{ij} A_{jk} A_{ki}}{\sum_i k_i(k_i - 1)}, \tag{14}$$

where $A$ is the adjancy matrix of $G$, and $k_i$ is given by:

$$k_i = \sum_j A_{ij}. \tag{15}$$

## 1.10 Dynamic forecasting framework

To incorporate dynamic variables of contact into a prediction framework, we study rolling 14-day windows (the upper bound incubation period of COVID-19[22]) and forecast patient infection over the subsequent seven days. Over each temporal window, we construct a contact network $G$ capturing all contact between individuals $N$ between $t_n$ and $t_m$.

From the series of successive contact networks $G^{(1)}, \ldots, G^{(T)}$ , we engineer their corresponding variable matrices $X^(1)$, ...,$X^{(T)}$ that include hospital environmental variables, patient clinical variables, and several notions of network centrality computed from the time-varying contact networks. These variable matrices are then used to predict the vectors of detected infections $Y^{(1)}, \ldots, Y^{(T)}$ over the subsequent seven days. The final step is then taking the samples from all time-windows, and aggregating them into a single data-set, $X$ and $Y$, for model construction.

## 1.11 Data samples

A data sample is constructed for each patient present in a given time window and labelled as either HOCI or control, depending on whether a patient did or did not become infected over the subsequent seven days respectively. Since COCI cases were most likely infected prior to hospital admission, we do not include them in the prediction dataset (however, they do exist as infectious patients within the contact network and contributing to background environmental variables). Given the rolling window approach, a single patient can appear multiple times in the final aggregated dataset. Therefore, some HOCI cases will initially be labelled as controls if they have spent time in the hospital without being SARS-CoV-2–positive during the forecasting period.

## 1.12 Univariate analysis statistical correction

Univariate analyses of variables were conducted over the data samples by grouping and averaging variables across patients as a form of statistician correction[23,24]. This correction thus meant that each patient is represented by a single data-point in the analysis, as apposed to multiple, conferring to their appearance across different time windows. Statistical testing was then performed using either the Mann-Whitney[25] or the Chi-squared test[26], and reported using a $p<0.05$ significance threshold. Implementation was performed using the *compareGroups* R package[27].

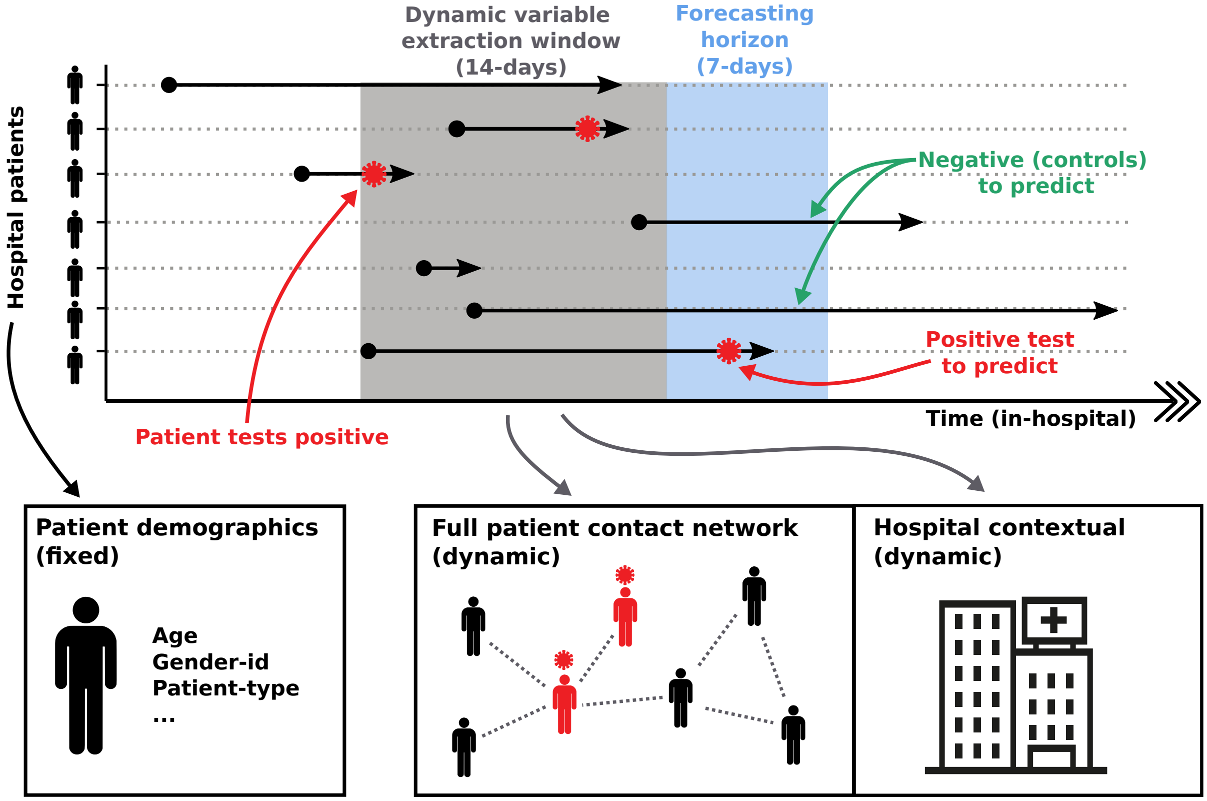

**Figure S1. Overview of forecasting framework.** Patient pathways are extracted from electronic health records which specify the locations each patient has visited over the duration of their hospital stay. Pathways are overlaid with COVID-19 testing results, capturing the space-time positions of patients that tested positive for COVID-19. Forecasting is based on extracting individual patient clinical variables (fixed) and hospital contextual variables (dynamic) during a defined time window, as well as variables capturing the centrality of a patient within the different contact networks (dynamic). We iterate variable extraction over multiple time windows and use the cumulative information for model training and predictions.

### 1.13 Statistical models

In our study, we implemented and compared the performance of the various common machine learning classifier models which have been used in previous studies. Tree based classifiers have proven themselves to be highly predictive across a wide range of medical prediction problems, hence we investigated three tree based classifiers: (i) the XGBoost model[28], (ii) Boosted decision trees, and (iii) a Random forest classifier[29]. We also investigated a Logistic regression model, which was fitted using a LASSO regression framework[29], a support vector machine (SVM)[30], as well as Neural network models[31], which have all shown varying levels of efficacy in healthcare prediction tasks. As further validation and as an additional benchmark we used two simpler, but still effective, prediction models, namely the Naive Bayes classifier and the K-nearest neighbour classifier[29].

All models were implemented using R[32] from the *Caret*[33], and *XGBoost*[34] packages. Visual plotting was done using *ggplot2* [35].

### 1.14 Model tuning

All model hyper parameters where selected using a Bayesian optimiser[36] via optimising 5-fold cross validation AUC-ROC performance over the training dataset. As opposed to an exhaustive, grid, or random search, Bayesian optimisation utilises results from prior evaluations to inform subsequent parameter selections, and identify optimal combinations. We used the *ParBayesianOptimization* in R [37] to implement Bayesian optimisation with a Gaussian Process using the Expected Improvement acquisition function[36]. In order to identify the optimal hyper parameters over models with differing numbers of parameters and complexity of search spaces, we performed subsequent interactions until no performance increase was observed.

For each model, the following parameters below hyper parameters were included in the optimisation:

- **XGBoost**: max_depth, min_child_weight, subsample, colsample_bytree, gamma_param.
- **Random Forest**: mtry, ntree.
- **Boosted decision trees**: n.trees, (interaction.depth, shrinkage, n.minobsinnode.
- **Logistic regression (LASSO)**: alpha, lambda.
- **SVM**: sigma, C.
- **Neural network**: size, decay.
- **Naive Bayes classifier**: fL, usekernel, adjust.
- **K-nearest neighbour**: k.

### 1.15 Collapsed performance metrics

To account for any bias based on the number of predictions made per patient (i.e. the longer a patient stays in hospital, the greater the number of forecasts produced), we collapsed each confusion matrix down, grouping the predictions by single patients. Specifically, any prediction of a HOCI for a patient who had not been a HOCI would be a False Positive; anyone prediction labelled as a HOCI was considered a True Positive if they had tested positive in at least one of the subsequent 7-day forecasting horizons. Similarly, controls consistently predicted controls were a single True Negative, and False Negative was any control predicted HOCI at least once. This operation is, in essence, a collapsing of the original confusion matrix and results in one recorded prediction to evaluate over, even if a patient has been in the hospital over multiple time windows.

Balanced accuracy, Sensitivity, Specificity, Positive Predicted Values, Negative Predicted Values, and both the Positive and Negative Likelihood Ratios could then be computed typically over the reduced confusion matrix.

### 1.16 Data pre-processing

We partitioned our data-set into a 70-30 split, using the initial 70% of samples for model training, and retained 30% for testing. Over the training set, a 5-fold cross validation strategy was used for model fitting and comparison. To ensure no bias in the data, we also performed a patient aware data split, ensuring samples from a single patient could not appear in both the train and test set, and additionally a patient could not appear simultaneously in both the train and validation sets during cross validation.

To address the small proportion of infected samples relative to non-infected samples we employed a sampling strategy, which re-balanced the training data-set. Specifically, we chose an under-sampling strategy, avoiding oversampling as a means to balance our data to prevent over-fitting and to reduce computational overheads [38].

### 1.17 SHAP Values

We computed SHAP values [39] to rank the variable contributions to model prediction outputs (SHAP values computed as specified in *Liu and Just et al.*[40]). SHAP originates in game theory, explaining how predictions change when a particular variable is removed from the model. A SHAP value's magnitude and direction give direct insight into whether a variable contributes to a certain class of predicted labels for each given sample. As well as computing SHAP values for a sample and variables combinations, SHAP values can also estimate variable importances by aggregating results across variables.

### 1.18 Variable selection and ranking

To rank variables and eliminate co-linearity, we employed a recursive variable elimination and cross-validation strategy which performs selection by initially starting with a full model and iteratively removing variables based on their importance[41]. We followed the implementation in ref[41], however, to suit our data we: (i) replaced the Random Forest classifier with an XGBoost model, since it's performance was highest in our exploratory model comparison; (ii) used SHAP values[39], as a less biased method to attribute variable importance. Finally we investigated two selection criteria, firstly one based on the overall maximisation of cross-validation AUC, and secondly, the maximisation of cross-validation AUC for models below a variance inflation factor (VIF)[42] threshold of 5. Since a VIF of above 5 would suggest redundancy in variables[42], a marginal loss in cross validation performance could be tolerated for a more concise and explainable model.

# 2 Results

## 2.1 Patient training and testing inclusion

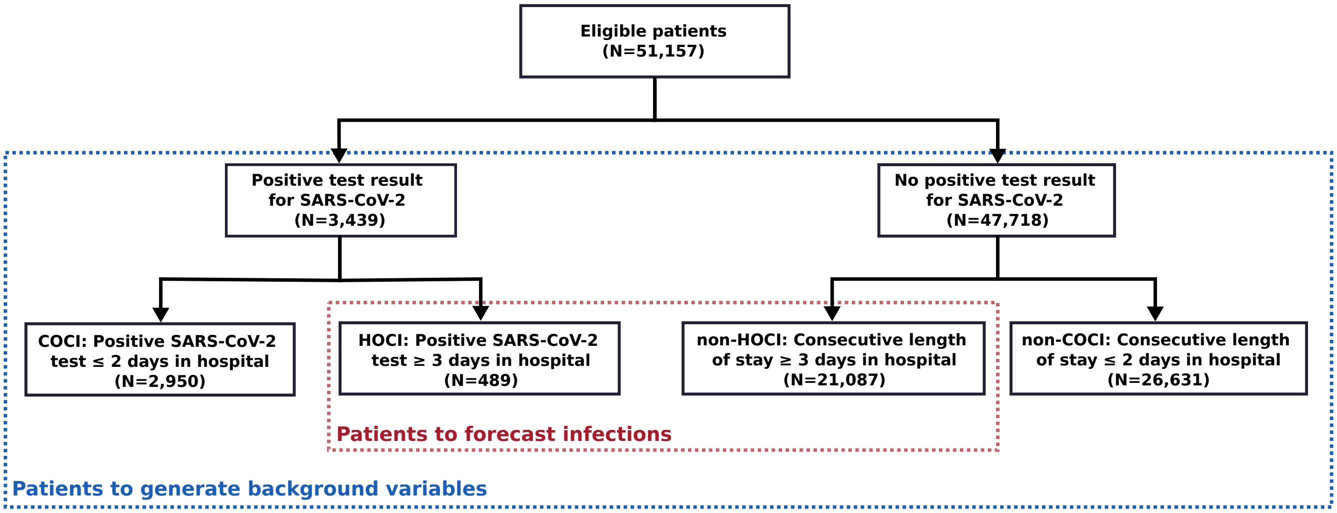
Eligible patients
(N=51,157)
Positive test result
for SARS-CoV-2
(N=3,439)
No positive test result
for SARS-CoV-2
(N=47,718)
COCI: Positive SARS-CoV-2
test ≤ 2 days in hospital
(N=2,950)
HOCI: Positive SARS-CoV-2
test ≥ 3 days in hospital
(N=489)
non-HOCI: Consecutive length
of stay ≥ 3 days in hospital
(N=21,087)
non-COCI: Consecutive length
of stay ≤ 2 days in hospital
(N=26,631)
Patients to forecast infections
Patients to generate background variables

**Figure S2. Patient dataset inclusion diagram.** Patients in hospital two days or fewer only contributed to background variables, whereas patients in hospital three or more days are part of both the generation of background variables and the forecasting dataset.

## 2.2 Study periods comparison

We compared the epideomology of in hospital COVID-19 between UK surges 1, and surge 2 across the 5 hospital group (Table S1). We also provide an incluision of both the data post surge 2 from the UK which was used for internal validation, as well as Geneva surge 1 used for the external validation. Information regarding COVID-19 variants sourced from the Sanger institute.

**Table S1. Comparison of datasets.** Each dataset was analysed and compared between periods. This included a comparison of the case time series, background hospital contextual statistics, and then contact networks (taken as the median local measurements, or overall global metrics).

| Variable | UK Surge 1 | UK Surge 2 | UK post Surge 2 | Geneva |
|---|---|---|---|---|
| | (23/03/2020 – 30/05/2020) | (07/09/2020 – 31/03/2021) | (02/04/2021 – 10/08/2021) | (01/02/2020 – 31/05/2020) |
| **Time series case analysis** | | | | |
| Day length | 72 | 206 | 132 | 121 |
| HOCIs | 167 | 406 | 186 | 138 |
| COCIs | 940 | 1870 | 1260 | 143 |
| Daily HOCI proportion | 15.1% | 17.8% | 12.9% | 51.8% |
| Correlation (Lag 0) | 0.59; 5.8e-08 | 0.79; 1.1e-44 | 0.33; 8.4e-05 | 0.57; 1.7e-05 |
| Correlation (Lag 5) | 0.53; 4.3e-06 | 0.74; 1.5e-36 | 0.31; 4.3e-04 | 0.52; 2.3e-04 |
| Average Alpha variant prevalence | 0% | 59.30% | 27.90% | n/a |
| Average Delta variant prevalence | 0% | 1.10% | 71.10% | n/a |
| **Hospital contextual statistics** | | | | |
| Total patients | 7,321 | 33,975 | 43,375 | 40,057 |
| COVID-19 prevalence | 15.10% | 6.70% | 3.30% | 0.70% |
| Median LoS | 2 | 20 | 23 | 3 |
| **Contact network analysis** | | | | |
| Room | | | | |
| Median degree | 3 | 4 | 5 | 29 |
| Median closeness | 0.01 | 0.06 | 0.6 | 0.7 |
| Clustering coefficient | 0.56 | 0.58 | 0.59 | 0.8 |
| Coefficient of variation | 0.79 | 0.75 | 0.76 | 1.055 |
| Ward | | | | |
| Median node degree | 18 | 23 | 24 | n/a |
| Median node closeness | 0.14 | 0.18 | 0.19 | n/a |
| Clustering coefficient | 0.59 | 0.63 | 0.62 | n/a |
| Coefficient of variation | 0.80 | 0.78 | 0.79 | n/a |
| Building | | | | |
| Median node degree | 57 | 70 | 76 | n/a |
| Median node closeness | 0.29 | 0.30 | 0.31 | n/a |
| Clustering coefficient | 0.65 | 0.66 | 0.70 | n/a |
| Coefficient of variation | 1.02 | 1.13 | 1.12 | n/a |

### 2.3 COCI versus HOCI

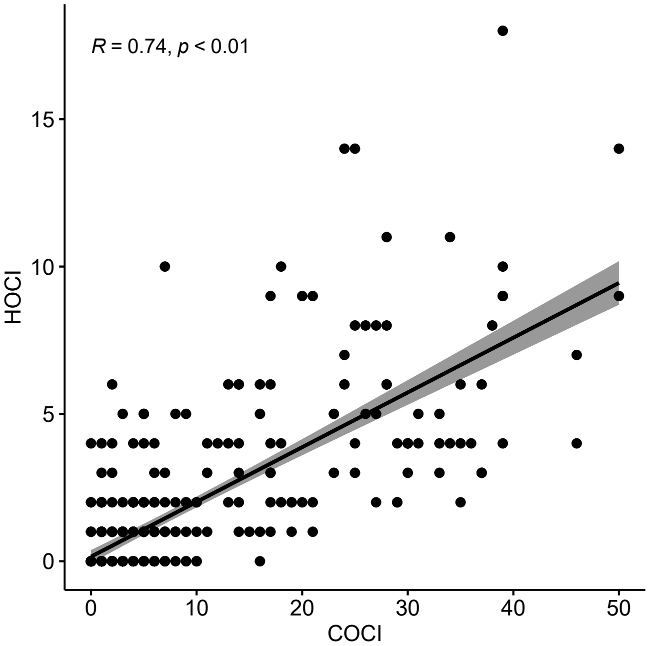

**Figure S3. Correlation between COCI and HOCI time series**. The figure shows the correlation between COCI and HOCI during the training and testing period (23 March 2020- 31 March 2021). Each time series comprises counts of newly identified COVID-19 cases amongst patients and are labelled according to COCI/HOCI definitions.

### 2.4 Time-varying contact network summary

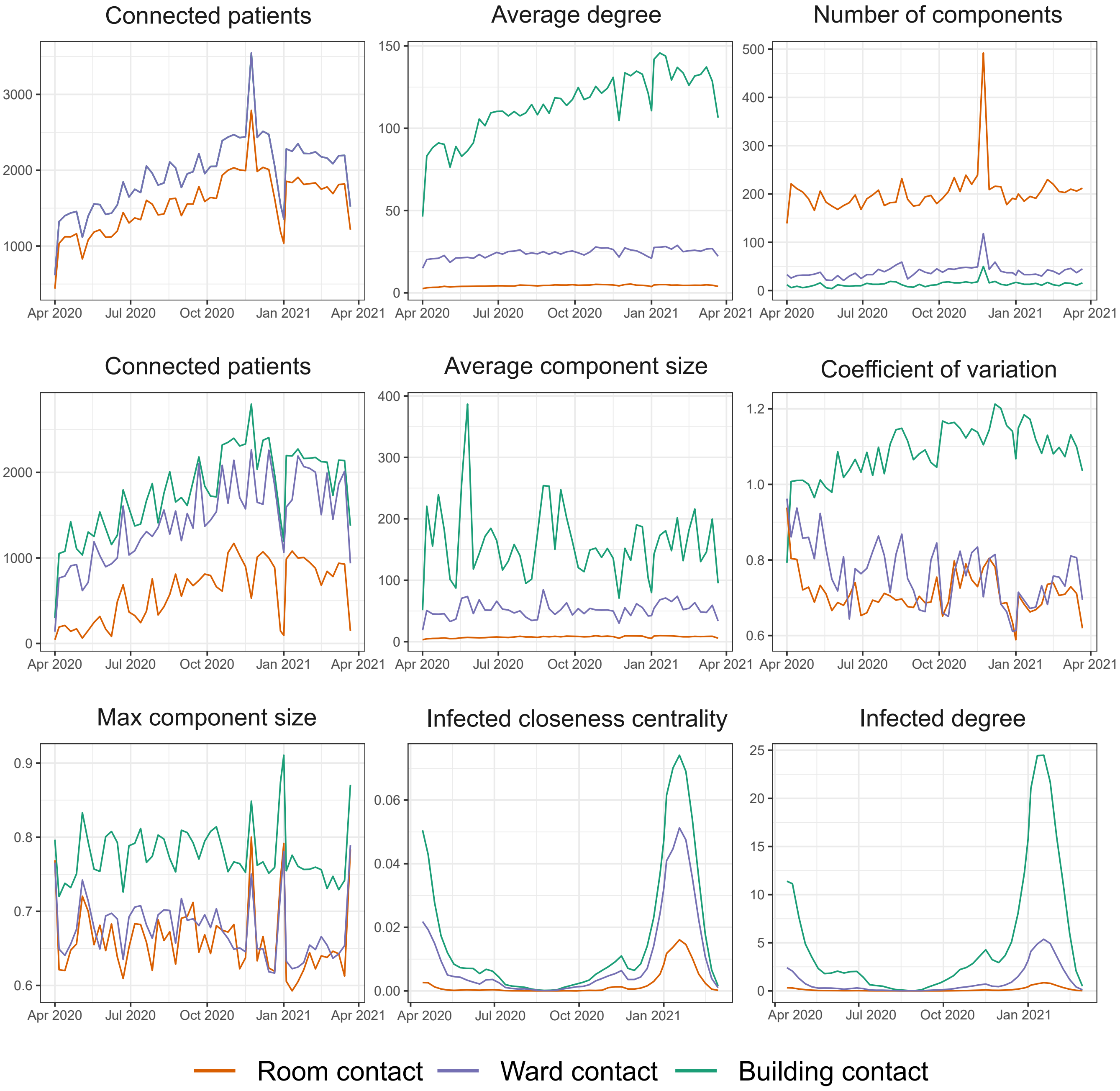

**Figure S4. Summary statistics of time-varying contact networks.** For all contact networks (room/ward/building) we computed summary statistics using a one week window (the total number of connected patients, average degree, number of connected components, maximum component size, average component size, coefficient of variation, and the clustering coefficient).

## 2.5 Statistical model selection

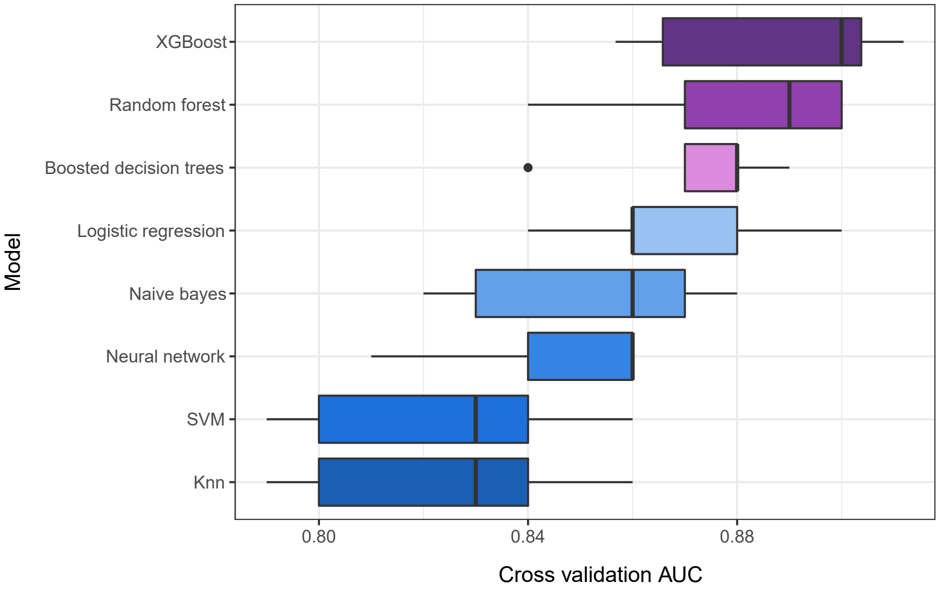

**Figure S5. Performance comparison by statistical models.** Each model (see section 1.13) was examined using 5-fold cross validation over the training data, and tuned with a Bayesian optimisation strategy (see section 1.14). The top three models were all based on decision trees (coloured purple), with the best performing model being XGBoost. For the final reported model in the main manuscript, we thus focus exclusively on XGBoost.

## 2.6 Model calibrations

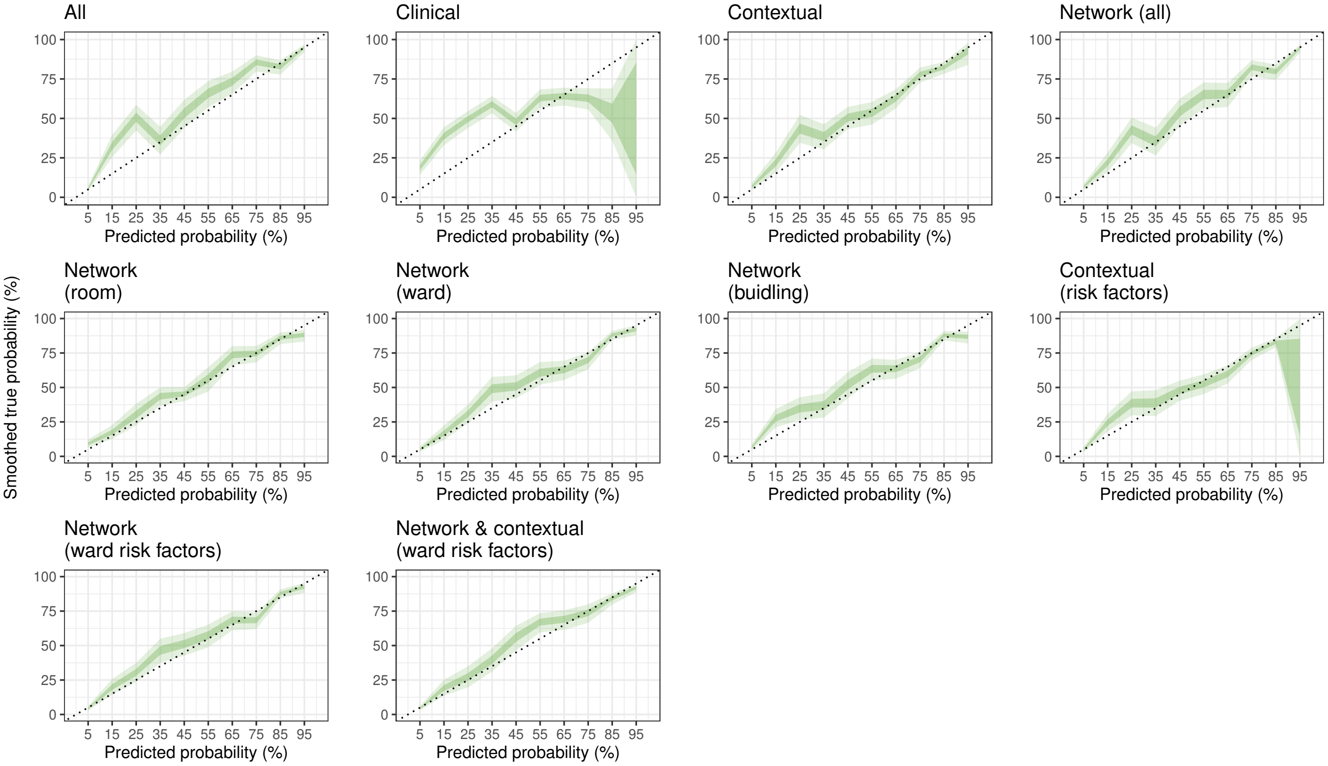

**Figure S6. Callibration plots for test (London) data models by variable set.** The smoothed true probability of actual outcomes of each group is plotted against the predicted probabilities[43]. A perfectly calibrated model would align to the 45-degree line. In terms of performance, the models built on clinical variables alone showed non-accurate calibration. In contrast, the contextual, the network variables (room/ward/building) in addition to the risk factor models showed accurate calibration.

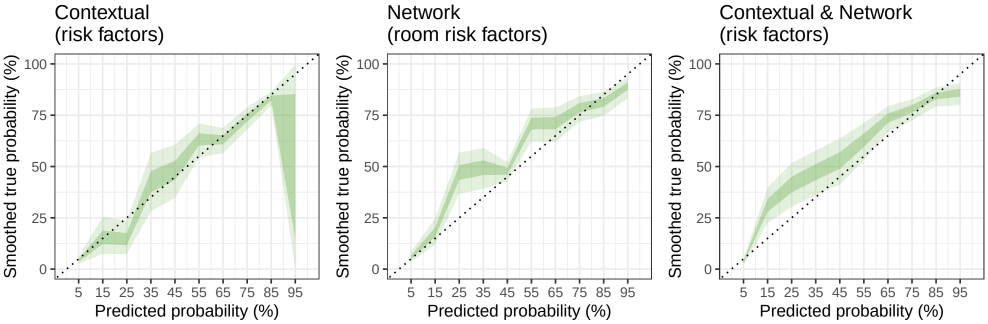

**Figure S7. Callibration plots for validation (Geneva) data models by variable set.** The smoothed true probability of actual outcomes of each group is plotted against the predicted probabilities[43]. A perfectly calibrated model would align to the 45-degree line. Comparatively, the contextual risk factors showed poorer calibration than the network room-risk factors. However, the combined network and contextual risk factor model demonstrated the most accurate calibration in the Geneva dataset.

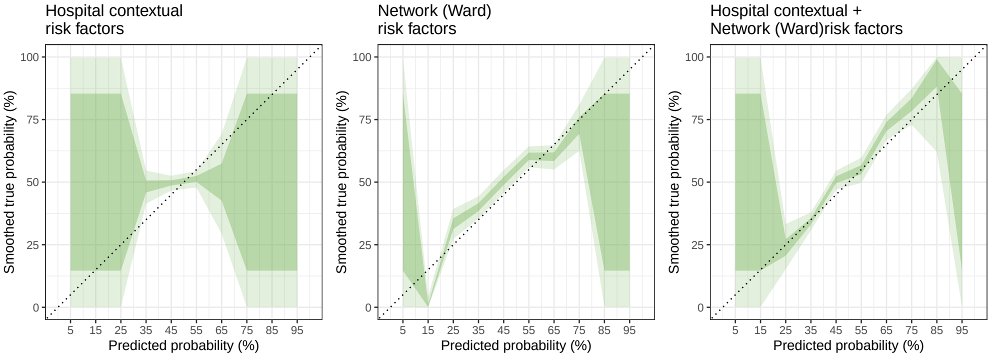

**Figure S8. Callibration plots for validation (London) data models by variable set.** The smoothed true probability of actual outcomes of each group is plotted against the predicted probabilities[43]. A perfectly calibrated model would align to the 45-degree line. The contextual risk factors showed very poor calibration. The network room-risk factors, and the combined network and contextual risk factor models, on the other hand, demonstrated more accurate calibration, with the combined model showing higher accuracy in terms of calibration, as did it also in the London test dataset.

### 2.7 Model performance for different HOCI definitions

For each of the main variable sets (All-variables, clinical variables, contextual variables, and all-networks variables) we broke down the performance by the three HOCI categories (Definite/Probable/Indeterminate) which reflect liklihood of hospital acquisition. Together the 465 HOCIs broke down into 150 definite cases, 122 probable cases, and 193 indeterminate.

Across the full model using all variables, performance stratified by HOCI definition (Indeterminate/Probable/Definite) showed a general increase in the predictability of HOCIs with the confidence in hospital acquisition (Table S2). Performance broken down model variable sets (Clinical, Environmental, and Network) also exhibited an increasing performance by the confidence in HOCI. For example, the clinical variable models only slightly better predicting indeterminate-HOCI than random (0.51 AUC-ROC) but increasing to 0.70 AUC-ROC for definite-HOCI. Similarly, despite the environmental variable set performing well for probable-HOCI and definite-HOCI, it performed no better than random for indeterminate-HOCI (0.5 AUC-ROC).

Across all HOCI definitions, models based exclusively on network variables achieved jointly the highest, or the highest performance. Overall, the consistently higher performance of models using solely network variables suggests that contact network variables alone intrinsically encompass risk non-attributable to more standard variables for disease prediction.

**Table S2.** Summary of London-test set performance by HOCI-type. Performance breakdown is shown using alternative variable sets for predicting all HOCI cases, indeterminate-HOCI, probable-HOCI and definite-HOCI.

| **Variable set** | **AUC-ROC (95% CI)**<br>***(Indeterminate/ Probable/ Definite)*** | **Balanced Accuracy**<br>***(Indeterminate/ Probable/ Definite)*** | **Sensitivity**<br>***(Indeterminate/ Probable/ Definite)*** | **Specificity**<br>***(Indeterminate/ Probable/ Definite)*** |
|---|---|---|---|---|
| All | 0.72 (0.6-0.84)/<br>0.87 (0.82-0.92)/<br>0.91 (0.88-0.94) | 0.65/<br>0.80/<br>0.86 | 0.60/<br>0.89/<br>0.97 | 0.70/<br>0.72/<br>0.75 |
| Clinical | 0.51 (0.39-0.63)/<br>0.6 (0.53-0.67)/<br>0.7 (0.66-0.74) | 0.54/<br>0.66/<br>0.64 | 0.54/<br>0.66/<br>0.64 | 0.31/<br>0.31/<br>0.33 |
| Contextual | 0.5 (0.38-0.62)/<br>0.75 (0.68-0.82)/<br>0.89 (0.86-0.92) | 0.43/<br>0.77/<br>0.81 | 0.43/<br>0.77/<br>0.81 | 0.64/<br>0.66/<br>0.68 |
| Network | 0.73 (0.61-0.85)/<br>0.89 (0.84-0.94)/<br>0.91 (0.88-0.94) | 0.70/<br>0.80/<br>0.86 | 0.70/<br>0.80/<br>0.86 | 0.70/<br>0.72/<br>0.75 |

### 2.8 Variable selection ranking

**Table S3. Variable elimination results over all variables.** We present the top ten variables as ranked in order of inclusion via the variable elimination procedure.

| variable order | variable | Cross validation AUC |
|---|---|---|
| 1 | Background hospital COVID-19 prevalence | 0·81 |
| 2 | Infected closeness centrality (ward) | 0·84 |
| 3 | Infected degree centrality (ward) | 0·84 |
| 4 | Background hospital HOCI prevalence | 0·86 |
| 5 | Infected degree (building) | 0·86 |
| 6 | Infected degree centrality (room) | 0·87 |
| 7 | Infected degree (ward) | 0·88 |
| 8 | Clustering coefficient (ward) | 0·88 |
| 9 | K-core number (building) | 0·88 |
| 10 | Infected closeness centrality (building) | 0·88 |

**Table S4. Variable elimination results over all risk factor variables.** We present the top ten variables as ranked in order of inclusion via the variable elimination procedure.

| Feature order | Feature | Cross validation AUC |
|---|---|---|
| 1 | Background hospital COVID-19 prevalence | 0·82 |
| 2 | Infected closeness centrality (ward) | 0·82 |
| 3 | Infected degree centrality (ward) | 0·84 |
| 4 | Background hospital HOCI prevalence | 0·86 |
| 5 | Infected degree centrality (room) | 0·88 |
| 6 | Infected degree (building) | 0·88 |
| 7 | Clustering coefficient (ward) | 0·88 |
| 8 | K-core number (ward) | 0·88 |
| 9 | Age | 0.89 |
| 10 | Infected degree centrality (building) | 0·89 |

**Table S5. Variable elimination results over hospital contextual risk factors and network-ward risk factors.** We present the top ten variables as ranked in order of inclusion via the variable elimination procedure.

| Variable order | Variable | Cross validation AUC-ROC |
|---|---|---|
| 1 | Background hospital COVID-19 prevalence | 0·81 |
| 2 | Infected degree centrality (ward) | 0·85 |
| 3 | Infected closeness centrality (ward) | 0·85 |
| 4 | Background hospital HOCI prevalence | 0·85 |
| 5 | Infected degree (ward) | 0·87 |
| 6 | Clustering coefficient (ward) | 0·87 |
| 7 | Total hospital bed occupancy | 0·87 |
| 8 | Closeness centrality (ward) | 0·87 |
| 9 | K-core number (ward) | 0·87 |
| 10 | Betweenness centrality (ward) | 0·87 |

## References

1. England, P. H. COVID-19 infection prevention and control guidance (2020).
2. Price, J. R. *et al.* Development and Delivery of a Real-time Hospital-onset COVID-19 Surveillance System Using Network Analysis. *Clin. Infect. Dis.* **72**, 82–89, DOI: 10.1093/cid/ciaa892 (2020).
3. Myall, A. C. *et al.* Network memory in the movement of hospital patients carrying antimicrobial-resistant bacteria. **6**, 1–23, DOI: 10.1007/s41109-021-00376-5. Number: 1 Publisher: SpringerOpen.
4. Holme, P. Modern temporal network theory: a colloquium. *The Eur. Phys. J. B* **88**, 234, DOI: 10.1140/epjb/e2015-60657-4.
5. Holme, P. & Saramäki, J. Temporal networks. *Phys. Reports* **519**, 97–125, DOI: 10.1016/j.physrep.2012.03.001 (2012).
6. Newman, M. *Networks* (Oxford University Press), second edition edn.
7. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).
8. Hagberg, A., Swart, P. & S Chult, D. Exploring network structure, dynamics, and function using networkx. Tech. Rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States) (2008).
9. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695 (2006).
10. Freeman, J. & McGowan, J. E., Jr. Risk Factors for Nosocomial Infection. *The J. Infect. Dis.* **138**, 811–819, DOI: 10.1093/infdis/138.6.811 (1978).
11. Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank citation ranking: Bringing order to the web. Techreport, Stanford InfoLab. Publisher: Stanford InfoLab.
12. Holland, P. W. & Leinhardt, S. Transitivity in structural models of small groups. **2**, 107–124, DOI: 10.1177/104649647100200201. Publisher: SAGE Publications.
13. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks | nature. *nature* **393**, 440–442 (1998).
14. Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry* 35–41 (1977).
15. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. K-core organization of complex networks. *Phys. review letters* **96**, 040601 (2006).
16. Peach, R. L. *et al.* Hcga: Highly comparative graph analysis for network phenotyping. *Patterns* **2**, 100227, DOI: https://doi.org/10.1016/j.patter.2021.100227 (2021).
17. Pastor-Satorras, R. & Vespignani, A. Epidemic Spreading in Scale-Free Networks. *Phys. Rev. Lett.* **86**, DOI: 10.1103/PhysRevLett.86.3200 (2001).
18. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, DOI: 10.1038/nature04153 (2005).
19. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, Oxford, New York, 1992).
20. May, R. M. Network structure and the biology of populations. *Trends Ecol. & Evol.* **21**, 394–399, DOI: 10.1016/j.tree.2006.03.013 (2006).
21. Newman, M. E. J. Properties of highly clustered networks. *Phys. Rev. E* **68**, DOI: 10.1103/PhysRevE.68.026121 (2003).
22. Lauer, S. A. *et al.* The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals Intern. Medicine* **172**, DOI: 10.7326/M20-0504 (2020).
23. Van Belle, G. *Statistical rules of thumb*, vol. 699 (John Wiley & Sons, 2011).
24. Student. The probable error of a mean. *Biometrika* 1–25 (1908).
25. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The annals mathematical statistics* 50–60 (1947).
26. Pearson, K. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, Dublin Philos. Mag. J. Sci.* **50**, 157–175 (1900).
27. Subirana, I., Sanz, H. & Vila, J. Building bivariate tables: The compareGroups package for R. *J. Stat. Softw.* **57**, 1–16 (2014).

**28.** Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794, DOI: 10.1145/2939672.2939785 (Association for Computing Machinery, New York, NY, USA, 2016).

**29.** Friedman, J., Hastie, T., Tibshirani, R. *et al. The elements of statistical learning*, vol. 1 (Springer series in statistics New York, 2001).

**30.** Cortes, C. & Vapnik, V. Support-vector networks. *Mach. learning* **20**, 273–297 (1995).

**31.** Ripley, B. D. *Pattern recognition and neural networks* (Cambridge university press, 2007).

**32.** R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2017).

**33.** Kuhn, M. *caret: Classification and Regression Training* (2020). R package version 6.0-86.

**34.** Chen, T. *et al. xgboost: Extreme Gradient Boosting* (2021). R package version 1.3.2.1.

**35.** Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).

**36.** Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms. **25**.

**37.** Wilson, S. *ParBayesianOptimization: Parallel Bayesian Optimization of Hyperparameters* (2021). R package version 1.2.4.

**38.** Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H. & Santos, J. Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]. *IEEE Comput. Intell. Mag.* **13**, 59–76, DOI: 10.1109/MCI.2018.2866730 (2018).

**39.** Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777 (2017).

**40.** Liu, Y. & Just, A. *SHAPforxgboost: SHAP Plots for 'XGBoost'* (2021). R package version 0.1.0.

**41.** Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* **46**, 389–422, DOI: 10.1023/A:1012487302797 (2002).

**42.** James, G., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning*, vol. 112 (Springer, 2013).

**43.** Crowson, C. S., Atkinson, E. J. & Therneau, T. M. Assessing calibration of prognostic risk scores. *Stat. Methods Med. Res.* **25**, 1692–1706, DOI: 10.1177/0962280213497434 (2016). PMID: 23907781, https://doi.org/10.1177/0962280213497434.