Estimating nosocomial infection and its outcomes in hospital patients in England with a diagnosis of COVID-19 using machine learning

Flavien Hardy,[1] Johannes Heyl,[1] Katie Tucker,[2,3] Adrian Hopper,[1,3] Maria J. Marchã,[4] Annakan V. Navaratnam,[1,5] Tim W.R. Briggs,[1,6] Jeremy Yates,[4] Jamie Day,[1] Andrew Wheeler,[1,6] Sue Eve-Jones,[1] William K. Gray[1]

1. Getting It Right First Time, NHS England and NHS Improvement, London, UK,
2. Department of Women's and Children's Health, School of Life Course Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom,
3. Guy's and St Thomas' NHS Foundation Trust, London, UK,
4. Science and Technology Facilities Council Distributed Research Utilising Advanced Computing High Performance Computing Facility, London, UK,
5. University College London Hospitals NHS Foundation Trust, London, UK,
6. Royal National Orthopaedic Hospital NHS Trust, London, UK,

Correspondence: William K. Gray, Getting It Right First Time programme, email: William.gray5@nhs.net

Word count: Abstract: 292 words, Main body: 3,808 words, Figures: 5, Tables: 1, Supplementary figures: 3, Supplementary tables: 2, References: 44

Running title: COVID-19 nosocomial infections in England

Keywords: COVID-19, coronavirus, mortality, hospital acquired infection, nosocomial infection

Summary

Background. COVID-19 nosocomial infections (NIs) may have played a significant role in the dynamics of the pandemic in England, but analysis of their impact at the national scale has been lacking. Our aim was to provide a comprehensive account of NIs, identify their characteristics and outcomes in patients with a diagnosis of COVID-19 and use machine learning modelling to refine these estimates.

Methods. From the Hospital Episodes Statistics database all adult hospital patients in England with a diagnosis of COVID-19 and discharged between March 1st 2020 and March 31st 2021 were identified. A cohort of suspected COVID-19 NIs was identified using four empirical methods linked to hospital coding. A random forest classifier was designed to model the relationship between acquiring NIs and the covariates: patient characteristics, comorbidities, frailty, trust capacity strain and severity of COVID-19 infections.

Findings. In total, 374,244 adult patients with COVID-19 were discharged during the study period. The four empirical methods identified 29,896 (8.0%) patients with NIs. The random forest classifier estimated a mean NI rate of 10.5%, with a peak close to 18% during the first wave, but much lower rates thereafter and around 7% in early spring 2021. NIs were highly correlated with longer lengths of stay, high trust capacity strain, greater age and a higher degree of patient frailty. NIs were also found to be associated with higher mortality rates and more severe COVID-19 sequelae, including pneumonia, kidney disease and sepsis.

Interpretation. Identification of the characteristics of patients who acquire NIs should help trusts to identify those most at risk. The evolution of the NI rate over time may reflect the impact of changes in hospital management practices and vaccination efforts. Variations in NI rates across trusts may partly reflect different data recording and coding practice.

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Research in Context

Evidence before this study: We searched PubMed on September 16th, 2021 for articles that documented associations between nosocomial infections (NIs) and COVID-19 using search terms ("SARS-CoV-2" OR "COVID-19") AND ("nosocomial" OR "hospital acquired"). Of the 1,034 papers identified, 137 were from the United Kingdom and 34 were specific to England. Two of these were reviews, 15 were single site studies and 11 were multi-site studies; none covered the entirety of the first and second pandemic waves.

Added value of this study: We present data for 374,244 patients admitted to hospital for the first 13 months of the COVID-19 pandemic in England. A cohort of suspected NIs was identified, and a machine learning model was trained on this dataset to identify the characteristics of these infections. The NI rate was estimated to be 10.5%, with a peak close to 18% during the first wave, but much lower rates thereafter and around 7% in early spring 2021. The model allowed us to overcome apparent under-reporting of these infections in clinical coding. Extended hospital stay, service strain, patient frailty, severity of COVID-19 (pneumonia, sepsis, kidney disease) and deprivation were found to be strongly associated with higher risks of NI. 10.5%, with a peak close to 18% during the first wave, but much lower rates thereafter and around 7% in early spring 2021.

Implications of all the available evidence: COVID-19 related NIs declined from the first to the second wave. Our findings should inform practice with regard to identifying patients at risk of NI for COVID-19. Our methods could also be applied to other settings and other conditions (e.g., winter flu, norovirus and highly transmissible bacterial infections).

INTRODUCTION

The Coronavirus Disease 19 (COVID-19) pandemic placed significant strains on healthcare systems internationally, with sudden surges in hospitalised cases prompting urgent hospital infrastructure adaptations.[1] In some settings, cohorting strategies were introduced to protect non-COVID-19 patients from exposure, although there is evidence that this was not fully effective in preventing nosocomial infection (NI).[2] As well as patient-to-patient transmission, healthcare workers are likely to be important in the transmission of infection within hospitals.[3] The burden on healthcare services of healthcare worker infection is particularly important given the potential to spread infection to large numbers of patients and result in work absence at times of greatest pressure.

In England, evidence regarding the scale and role of COVID-19 NIs in the dynamics of the pandemic is starting to emerge.[4-6] There is some evidence that patients with NI may be at risk of poorer outcomes,[7] though this has been debated and disputed.[8] The Getting It Right First Time (GIRFT) programme is funded by the Department of Health and Social Care in England to investigate variation in practice and patient outcomes across the National Health Service (NHS). GIRFT has a particular interest in assessing the variability of outcomes for COVID-19 during the pandemic to identify lessons to be learned. Understanding the extent and pattern of COVID-19 NI is an important part of the approach to minimising future NIs in COVID-19 and other diseases. The aim of this study was to assess the capacity of the Hospital Episode Statistics (HES) dataset for England to identify COVID-19 NIs, and to understand the profile of patients with NIs. In particular, we aimed to develop a machine learning model capable of identifying highly probable NIs from administrative data of cases for COVID-19 in NHS hospital in England.

METHODS

Ethics

Consent from individuals involved in this study was not required for this analysis of the HES administrative dataset. The analysis and presentation of data follows current NHS Digital guidance for the use of HES data for research purposes. Reported data are anonymised to the level required by ISB1523 Anonymisation Standard for Publishing Health and Social Care Data.[9]

Study design and data collection

This was a retrospective exploratory analysis of HES data. HES data are collected by NHS Digital for all NHS-funded patients admitted to hospitals in England. Hospital trusts run all NHS hospitals in England. A hospital trust is an administrative unit typically covering one-four hospitals which provides secondary and/or tertiary care for all people living in a geographically defined catchment area. Data collection and reporting is mandatory, and data are entered by clinical coders at each trust.

Timing, case ascertainment, inclusion and exclusion criteria

Data were taken from HES for all patient discharges during the period 1st March 2020 to 31st March 2021. COVID-19 was identified using the International Classification of Disease and Function, tenth revision (ICD-10) codes U071 and U072 in any position in the diagnostic record for a completed hospital spell. A hospital spell is defined as a continuous period in hospital and may include multiple smaller episodes of care in various hospital settings and under different consultants (see **Supplementary material Figure S1**). Patients were excluded if they were < 18 years of age and only the final admission during the study period for each patient was included. The full data extraction process is summarised in **Figure 1**.

NI case identification

Several methods were considered to establish a baseline of patients having acquired COVID-19 NIs. The following two methods were used to identify definite NIs:

Method 1: Use of the ICD-10 code Y95 (nosocomial condition) in either the first, second or third diagnosis code position following U071 or U072 or in the diagnosis code position immediately prior to U071 or U072 in the HES record.

Method 2: First appearance of U071 or U072 in the episode level record > 14 days after the start of the spell.

To these were added two additional methods to identify probable NIs:

Method 3: First appearance of U071 or U072 in the episode level record 8 to 14 days after the start of the spell.

Method 4: Emergency readmission to hospital with U071 or U072 in the diagnostic record within 8 days of a hospital discharge from a prior admission where the ICD-10 code Z208 (exposure to communicable diseases) appeared in the diagnostic record but U071 or U072 did not appear in the diagnostic record and where 8 days or more separated the two admission dates.

These methods were defined to align with the European Centre for Disease Prevention and Control (ECDC) definitions of definite (diagnosis > 14 days post admission) and probable (diagnosis 8-14 days post-admission) NI.[10]

The additional ECDC definition of possible NI (diagnosis 3-7 days post admission) was used to define an upper limit for the model (see **Data analysis** below) and was operationalised as any elective admissions with length of stay > 2 days or any emergency admission without a COVID-19 diagnosis during the first episode, with a length of stay > 2 days.

Covariates and data features

Patient characteristics: sex, age, ethnicity (White, Black or Black British, South Asian or South Asian British, Other Asian, Mixed, other), comorbidities (Charlson Comorbidity Index[11]), frailty (Hospital Frailty Risk Score HFRS[12] and the Global Frailty Score[13]), and deprivation (Index of Multiple Deprivation IMD scores).[14]

Features of hospital stay: in-hospital deaths, length of stay, hospital trust. Deaths during hospital stay are reported based on UK Office for National Statistics (ONS) data with an in-hospital death recorded if the date of death was ±1 day of the recorded day of discharge in the HES record. Length of stay are reported as the difference between the spell admission and discharge date.

Strain on healthcare services: From the features directly extracted from HES, two metrics were created to serve as proxy for the pressure placed by COVID-19: the number of patients admitted to the same trust on the same day, and this number scaled to the maximum number of admissions on any day in that trust during the same wave of the pandemic. The first wave was defined as ending on 31st August 2020 and the second as starting on 1st September 2020.

Severity of COVID-19 during hospital stay: ICD-10 codes in the diagnostic record for the entire spell were used to identified common, severe COVID-19 related complications: pneumonia, kidney disease, blood clotting, issues related to cardiology or circulation, neurology, digestive system, and sepsis. **Supplementary material Table S1** details the ICD-10 codes used to identify these markers for the severity of the COVID-19 sequalae.

Categorical variables - such as ethnicity, region, or the severity of the infection - were hot encoded into binary features for analysis. Patients with a particular trait (e.g., Black ethnicity) were coded as 1 and all other patients 0 to create a dummy variable for that trait and this was repeated for all traits for that variable.

Outcome (target) variable

The target is described by a binary flag indicating whether the infection was empirically identified as NI.

Data analysis

Data were extracted onto a secure encrypted server controlled by NHS England and NHS Improvement. Analysis within this secure environment took place using Alteryx 2019.3 (Alteryx Inc., Irvine, CA, USA), Python 3.9.6 and the scikit-learn machine learning library 0.24.2 (Python Software Foundation, Beaverton, OR, USA).[15]

Modelling procedure: The data were modelled using a random forest classifier, a data analysis technique which falls under the broad suite of machine learning methods. In order to identify the characteristics that were most likely to be associated with NIs, a cohort of definite and probable NIs was identified using methods 1 to 4 described above. A withheld dataset was then randomly selected as a test dataset containing 20% of COVID-19 infections. The remaining 80% of the dataset was further split in an 80:20 ratio into a training and validation dataset respectively. A stratified splitting procedure was used to preserve the proportion of NIs in the test dataset. Missing numerical or categorical values were replaced by the means or modes within each class. A random forest classifier fits a chosen number of individual decision trees on multiple sub-samples of the dataset. Such models have been shown to be robust against over-fitting[16]—particularly useful in our case, given the likelihood of mislabelled NIs in our dataset—and benefits from the ability to highlight important features after being trained. We then used the trained classifier to derive a model-based estimate for the NI rate and identify important characteristics of patients with NI.

Model optimisation and evaluation: The trained model was optimized on the validation dataset by progressively removing the lower ranked features until removal led to a noticeable change (1%) in performance on the validation set, as measured by the area under the receiver operating characteristic (AUROC) curve. Features were further removed if they were highly

correlated (Pearson coefficient > 0.8) to other features to avoid any redundant information being used when training the classifier. Final optimization used a grid search with cross-validation; the hyper-parameters that were found to be the most influential are listed in **supplementary material Table S2**, along with their corresponding values. The default values from scikit-learn were used for the remaining parameters.[17]

Model predictions and interpretations: The lower bound of the model estimation of the NI rate was constrained by NIs identified by method 1-2 (definite NIs). An upper bound for the model was defined for patients diagnosed with COVID-19 using the following criteria which aimed to identify all definite, probable and possible NIs: method 1 or 4 above (definite and probable NIs) and the additional operationalisation of the possible NI definition described above.

To minimise bias towards features with high-cardinality when assessing the importance of features,[18] we determined the difference in prediction accuracy before and after permuting each predictor variable. This procedure has been shown to be more robust than considering variations in impurity when building the trees, though it can lead to under-representing correlated features.[19]

Plotting NI rates over time: As HES date are collated and defined at the point of discharge, discharge date was used as a marker for time when developing the model. This is consistent with our previous work.[20,21] However, when presenting the data over times, the discharge date was found inappropriate due to much longer stay in patients with NI. Admission date was also unsuitable at it represented a pre-infection date for those with NI and a post-infection date for those with community-acquired infection. As such, we estimated infection at a population level for patients identified as NIs based on median time from admission to first mention of COVID-19 in the episode-level diagnostic record using empirical methods 2-3 (14 days). For community-acquired infection we used estimates of time from symptom onset to diagnosis

from previous reports (2 days for < 20 years old patients; 7 days for 20-60 years old patients; 5 days for 60-80 years old patients; 3 days for > 80 years old patients).[22,23]

RESULTS

Cohort of identified NIs using empirical methods

Data were available for 374 244 patients (**Figure 1**), 29 896 (8.0%) of who were identified with NIs using the four empirical methods of identifying definite and probable NIs. The weekly numbers of discharged patients having been diagnosed with COVID-19 are shown in **Supplementary material Figure S3**. The degree of overlap in the cases identified by each method is presented as a Venn diagram in **Figure 2**. **Table 1** summarises the profile of NI and community-acquired infection patients based on these methods. Patients with NIs were older, more likely to be of White ethnicity, have a much longer hospital stay and had a higher mortality rate. The variation in NI rate across trusts and regions as identified by methods 1-4 is shown in **Figure 3**: the regions with the highest identified NI rate were the North West (10.9%) and the South West (9.9%); the lowest rate was identified in London (5.8%).

Using discharge date to define time, the evolution of the NI rate across the study period is summarised in **Supplementary material Figure S2**. Method 1, method 1-2 and method 1-4 showed a very similar trend. The apparent peak in the NI rate in summer 2020 is largely a function of longer length of stay with later discharge for patients with NI compared to those with community-acquired infection with low number of community-acquired infections at this time and relatively large number of NI patients being discharged.

Model calibration and predictions

The random forest classifier was trained and optimized before being applied to the withheld test set, the AUROC curve was 0.89 (**Supplementary material Figure S4)**. The model was

then calibrated by ensuring that the predictions were bounded by the upper and lower limits, as illustrated in **Supplementary material Figure S2**.

The time evolution of the NI rate predicted by the final model based on estimated date of infection is shown in **Figure 4**. The model estimates a mean NI rate of 10.5% over the entire time period; conservative lower and upper bound values of 9.3% and 11.8% can be estimated by varying the classification threshold within the limits of the model calibration (threshold ranging from 0.24 to 0.30). The model estimates that the NI rate reached a peak of ~ 18% during the first wave and stabilised to ~ 10%, before reaching a second peak at ~ 14% during the second wave of the pandemic. The rate is then estimated to have dropped to ~ 7% in early 2021.

Feature importance

The most important features identified by the random forest classifier are shown in **Figure 5**. Length of stay was the most important feature, followed by covariates related to the patient numbers and service strain, patient frailty, severity of COVID-19 and deprivation. The most important features related to disease severity were pneumonia, sepsis and kidney disease.

DISCUSSION

To our knowledge, this is the largest study of COVID-19 NIs conducted globally to date. The proportion of cases identified as NIs was found to increase during the peak of COVID-19 hospital admissions for each wave. The differences in estimated NI rates between the machine learning model predictions and empirical methods (reliant on clinical coding of COVID-19 through methods 1 to 4) were particularly marked in the first wave, which may reflect reduced case findings due to limited testing. NI rates were much lower during the second wave of the pandemic in England despite high case numbers and were closer to the case rate identified by methods 1-4 than during the first wave. The NI rate declined further during early spring 2021.

This is likely to reflect lower patient numbers, greater understanding of how COVID-19 is transmitted and how best to mitigate against in-hospital transmission.

Our NI rates are similar to those reported in a number of other UK-based studies, although most cover only the first few months of the pandemic. The ISARIC4C group reported an overall NI rate of 11.3% from February to July 2020, with a peak of 15.8% in May across 72,175 patients.[5] Most other previous studies are much smaller. A study of 2,354 patients across Wales reported an NI rate of 17.3% from 1st March to 1st July 2020 based on documentation of NI and 16.4% based on an interval of > 14 days between admission and diagnosis. A multi-centre observational study of 1564 patients admitted up to 28th April 2020 across 10 UK and one Italian hospitals estimated a rate of 12.5%.[8] Other single-site studies from spring 2020 report similar estimates.[24,25]

Variation in NI rates across time and between settings

The characteristics of COVID-19 infections—its highly contagious nature, its viability on surfaces for up to three days,[26] an incubation period of up to 14 days,[27] high proportions of asymptomatic infections[28]—facilitate transmissions in healthcare settings. The risks are compounded by rapidly changing and relatively poorly understood transmission dynamics, notably during the first wave. In England, this was reflected in the evolution of the Public Health England COVID-19 guidance for healthcare providers.[29-31] Understanding that COVID-19 could be transmitted from human-to-human and asymptomatically evolved over time, and testing capacity and the availability of personal protective equipment (PPE) was limited during the first wave. Symptomatic staff testing was introduced in some trusts in mid-March[32,33] and universal asymptomatic staff testing was introduced in November 2020[34] Antibody testing of staff in a large teaching hospital in North-West England between 29 May and 4 July 2020 found the highest rates of COVID-19 in nursing staff on COVID wards but with limited PPE.[35] However, intensive care unit staff also exposed to COVID patients but

fully equipped with PPE had positivity rates similar to staff working on non-COVID wards.[35] During March 2020 many hospitals also restricted or stopped allowing visitors, making subsequent NIs likely to be almost exclusively due to patient-to-patient or staff-to-patient transmissions.

Increased patient and hospital staff testing, greater availability of PPE, improved infection control strategies (e.g. cohorting COVID-19 patients, reduction of staff exposure through the national SPACES strategy[36]) all appear to have been successful in suppressing NI rates as cases surged during winter 2020/21 in England. Some London trusts implemented systems of triaging all patients prior to admission based on a screen for likelihood of COVID-19 infection whilst test results were awaited.[2,37] This is similar to the widely praised Traffic Control Bundling approach employed in Taiwan during an easier SARS outbreak, where patients were cohorted prior to admission.[38] A better understanding of atypical presentations of COVID-19 infections (e.g. gastrointestinal symptoms) over time will have improved the effectiveness of such initiatives.[39] During early 2021, the vaccination programme, which targeted healthcare workers and older people initially, may also have contributed to the low rates of NIs.

Rates varied substantially between hospital trusts and regions of England. However, this is unsurprising. Greater patient numbers were an important predictor of higher NI rates and as infection and hospitalisation rates varied across regions, so will NI rates. Local outbreaks will also be an important driver for the variation seen, as will local approaches to infection control. Likewise, patient profiles will have varied by trust type and by region. Noticeably lower NI rates in London reflects the lower age structure of COVID-19 patients in London hospital trusts.[40] The UK-wide ISARIC4C study noted substantially higher NI rates in residential and mental health trusts where patients will have much longer stays, supporting our own findings.[5]

Identification of NIs

Variation in NI rates across trusts is also likely to reflect variations in data recording practice between trusts. Whilst the most obvious cause of data recording practice-related variation will be due to differences in how ICD-10 codes are recorded in HES by each trust, clinical coders can only record NI where it is clearly recorded in the patient notes by a clinician. As such, trusts where clinicians are more aware of the need to record NI accurately will appear to have higher rates. We tried to overcome these limitations by using a number of methods to identify NIs. The modest overlap between patients identified by each of the four methods suggests that this was worthwhile, but also with some degree of under-identification, hence the need for modelling.

NI-specific outcomes

A higher mortality rate for NI compared to community-acquired infection has been reported elsewhere, and this is likely to reflect the older age profile and greater frailty of people with NI.[6] Older patients are more likely to have relatively long hospital stay, and COVID-19 is more likely to present atypically (e.g. with delirium rather than with cough or fever).[41] This limited the effectiveness of infection control measures, particularly during the first wave when testing was limited. The older age of patients with NI is also reflected in apparently higher rates of NI in the White ethnic group, chich has an older age structure than other ethnic groups. The association between NI and measures of patient numbers and strain on services may reflect the difficulty in managing NIs—isolating and maintaining infection control—at times of greatest activity.

Our study has several strengths including the size of the dataset and the ability to use multiple methods for case-ascertainment. We used a broader NI case definition and leveraged this uncertainty within our modelling procedure. We were also able to use time trends across an extended period to constrain the model. Our study also has some limitations, the most important of which have already been acknowledged in relation to coding practice and

reported inter-trust and inter-regional variability. The use of the ICD-10 code for a nosocomial condition (Y95) was also found to be insufficient when used in isolation. The underuse of this code is likely to be even greater, since we used a more relaxed use of the Y95 code than described in the official COVID-19 National Clinical Coding Standards and Guidance.[42] An addition limitation is that, unlike many other studies, we did not have the precise date of in-hospital diagnosis of COVID-19 but had to rely on the fact that in most cases this will have triggered the start of a new episode of care where COVID-19 was recorded. In cases where a new episode was not recorded upon diagnosis cases will have been missed. However, some of these cases should have been identified through other methods. The similarity of our estimates with previous reports suggests that this under-reporting was relatively limited. Although there is some uncertainty in our model, we were able to constrain the model predictions within an upper and lower bound. Nevertheless, identification of many COVID-19 NIs is challenging, even in a prospective study, given limited testing capacity early in the pandemic, variation in incubation periods and presentation.

In summary, using a random forest classifier we have been able to estimate the rate of NIs in hospitals in England, how this has varied over time and we have detailed the characteristics of patients with NIs. Although there was significant between trust variation in NI rates and some of which will be due to avoidable local outbreaks, much of this will be due to variation in patient numbers and patient profiles. More importantly, it may simply reflect variations in data recording practice. Efforts should be made to improve the clinical coding of NIs for all conditions with awareness raising of its importance conducted for both clinicians and clinical coders.

COVID-19 NIs have become less common with time as trusts have changed management approaches. The early vaccination of healthcare workers and older adults is likely to have

contributed to the relatively low estimated NI rates during the first three months of 2021. Nevertheless, there is scope to reduce COVID-19 NI rates further. Reviewing and learning from strategies that were successful in minimizing COVID-19 NIs in individual trusts is likely to be a key step in supporting all trusts in England to help manage NIs successfully. Such an approach may also help trust manage NIs beyond COVID-19 and target probable high-risk patients for other conditions, including winter flu, norovirus and highly transmissible bacterial infections. Similar modelling work to that conducted here will help support this approach.

Contributors

This study was designed and organised by FH, JH, WKG, KT, MM, AH and JY. Data cleaning, and analysis was by FH, supported by JH, MM, KT, AH, AW, SEJ and WKG. Writing of the first draft was by FH and WKG. All authors critically reviewed the manuscript and agreed to submission of the final draft.

Declaration of interests

The authors declare that there is no conflict of interest.

Acknowledgements

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

We acknowledge NHS Digital for permission to use their data in this report. The GIRFT programme is providing a framework for examining contemporary clinical practice in unprecedented detail and breadth. We also thank all staff within individual NHS trusts who collected and entered the data used in this study and GIRFT Clinical leads for advice: Michael

Jones, Philip Dyer, Chris Moulton, Anna Batchelor, Michael Swart, Christopher Snowden, Martin Allen, Partha Kar, Gerry Rayman.

Data availability statement

This report does not contain patient identifiable data. Consent from individuals involved in this study was not required. Requests for any underlying data cannot be granted by the authors because the data were acquired from data under licence/data sharing agreement from NHS Digital, for which conditions of use (and further use) apply. Individuals and organisations wishing to access HES data can make a request directly to NHS Digital.

Informed consent

Informed consent was not sought for the present study because it was an analysis of routine administrative data.

Ethical approval

Ethical approval was not sought for the present study because it did not directly involve human participants. This study was completed in accordance with the Helsinki Declaration as revised in 2013.

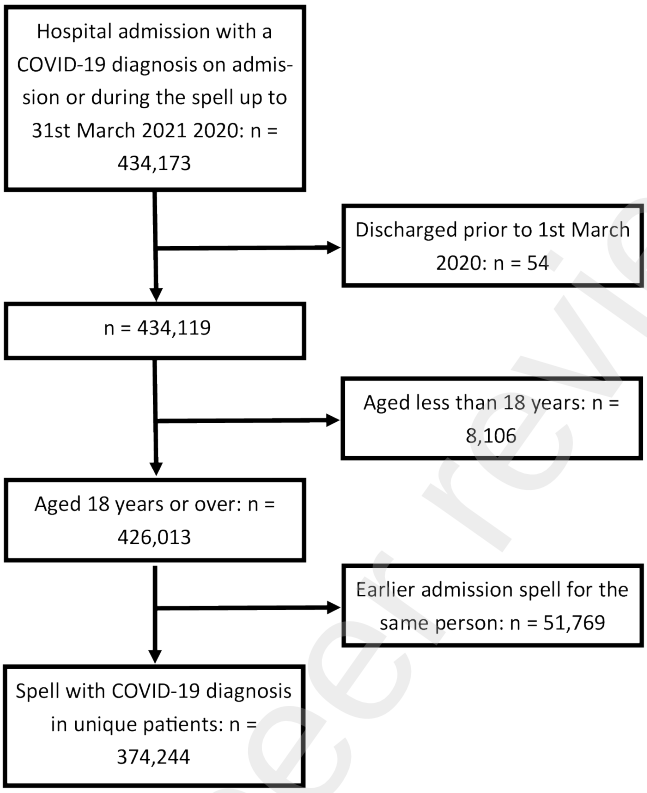


Figure 1: Flow diagram of the data extraction process.

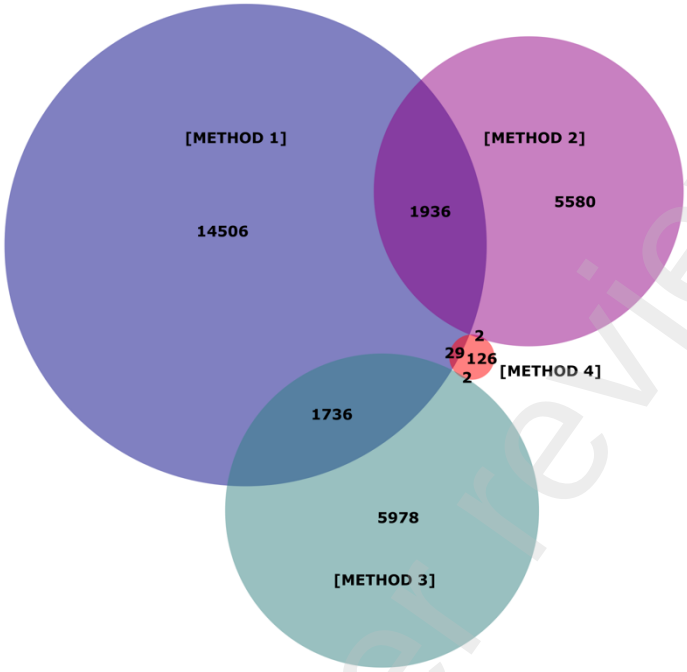


Figure 2: Venn diagram illustrating the overlap between the NIs identified by methods 1 to 4. Generated using *DeepVenn*.[43] [METHOD 1] = Use of code Y95; [METHOD 2] = First use of U071/U072 after 15 days in hospital; [METHOD 3] = First use of U071/U072 after 8-14 days in hospital; [METHOD 4] = Use of Z208 followed by emergency readmission.

Table 1: Profiles of suspected nosocomial infections and community infections identified using empirical methods 1 to 4.

	Suspected nosocomial infection	Suspected community infection
Number	29896	344348
Median age (IQR)	80 (70 to 87)	70 (54 to 82)
Males (%)	15775 (52.8%)	182862 (53.1%)
Ethnicity (%)		
White	27609 (92.3%)	278905 (81.0%)
Black or Black British	717 (2.4%)	15736 (4.6%)
South Asian or South Asian British	724 (2.4%)	25684 (7.4%)
Other Asian	266 (0.9%)	8134 (2.4%)
Mixed	117 (0.4%)	3256 (0.9%)
Other Ethnic Groups	463 (1.6%)	12633 (3.7%)
Median length of stay in days (IQR)	28 (17 to 44)	7 (3 to 14)
In-hospital mortality (%)	12152 (40.6%)	80992 (23.5%)
Median IMD score (IQR)	19 (10-30)	21 (11-34)

IQR =Inter-Quartile Range. IMD = Index of Multiple Deprivation.[14]

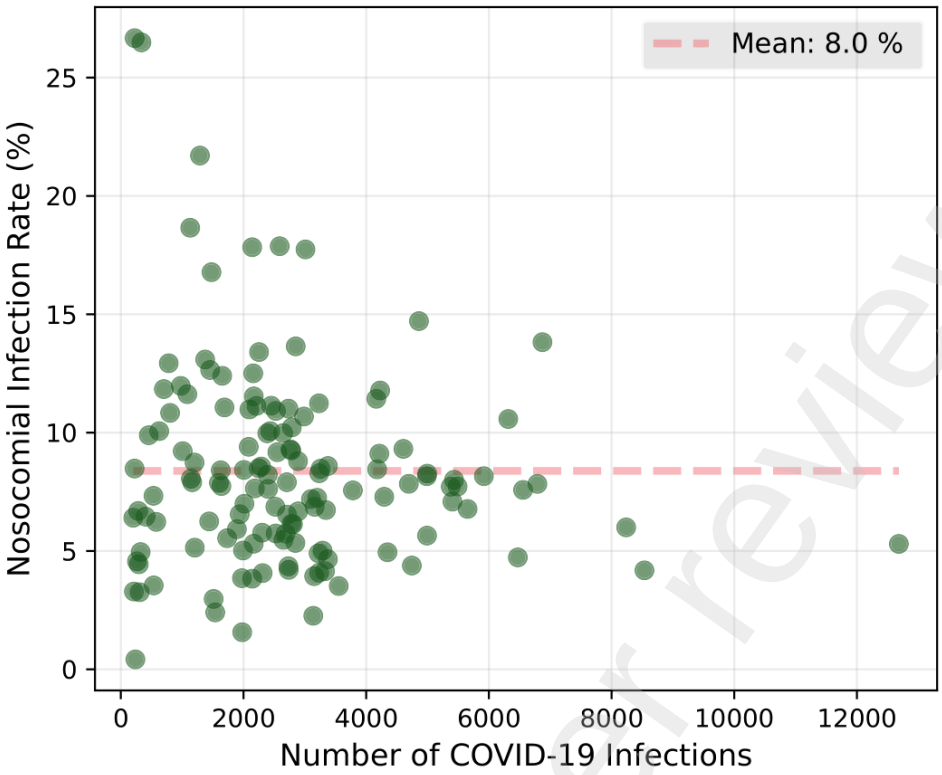


Figure 3: Distribution of NI rates across trusts as identified by methods 1 to 4. The orange dotted line (8.0%) corresponds to the average rate in England and is likely to be an under-estimate. The model trained on these data points and presented in Figure 4 estimates an average NI rate of 10.5% from 1st March 2020 to 31st March 2021.

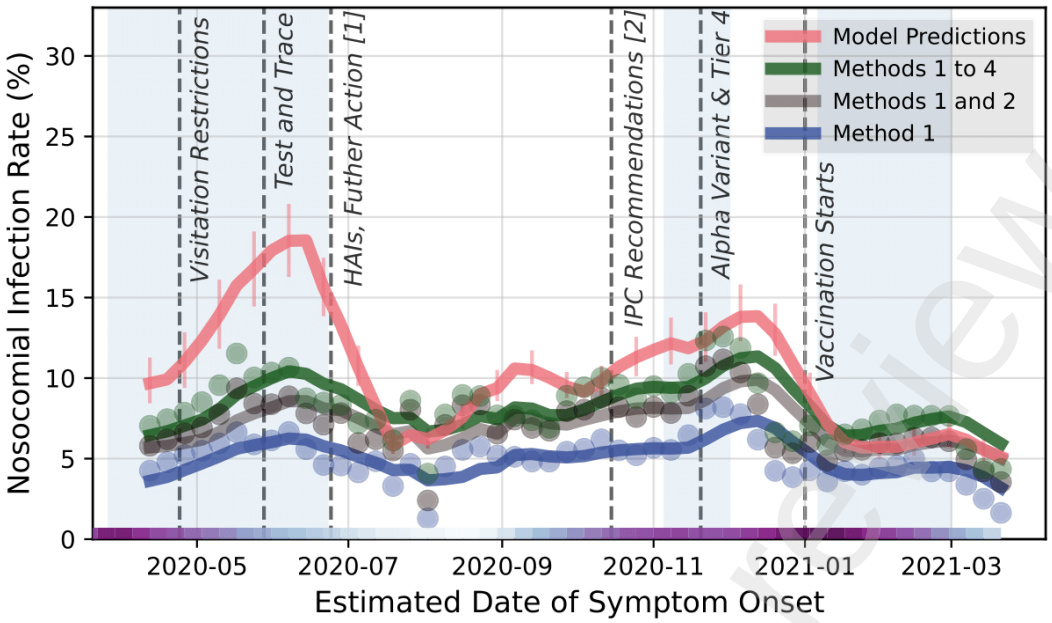


Figure 4: One-month rolling average of the NI rate using empirical and modelling methods based on estimated date of infection. Infections identified by method 1 (blue), methods 1 and 2 (black) and methods 1 to 4 (green). The predictions of the calibrated model are shown in orange; it predicts a mean NI rate of 10.5% over the entire time period. The blue shaded areas correspond to the first, second and third lockdowns in England. The colour-bar at the bottom of the plot indicates the number of recorded infections on a log scale (see also Supplementary material Figure S3). HAIs = Hospital Acquired Infections; IPC = Infection Prevention and Control. [1] and [2]: see [44,45].

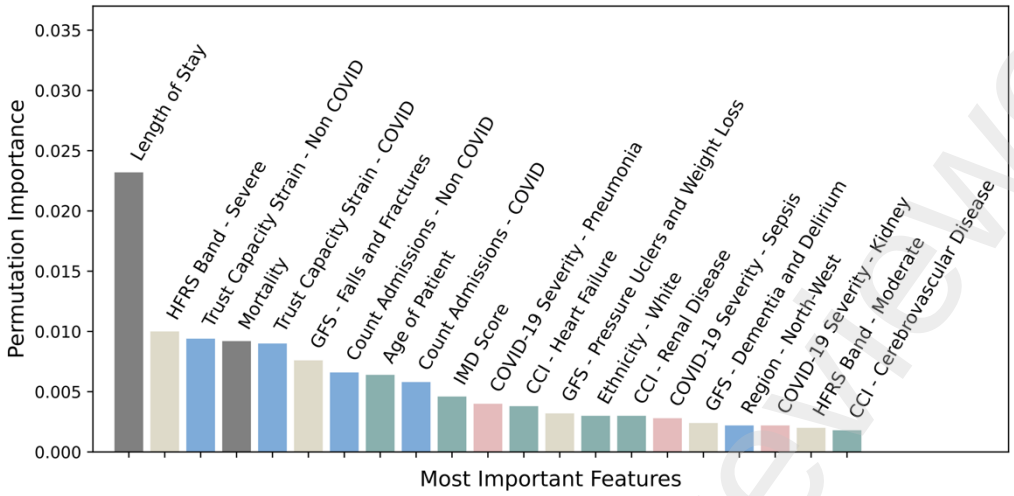
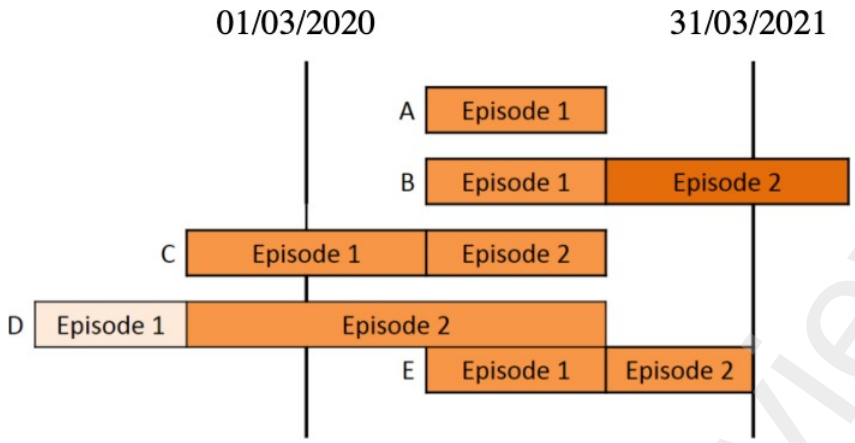


Figure 5: Most important features identified by the random forest classifier, using feature permutation importance. The outcomes (Length of Stay and Mortality) are coloured in black and features explicitly related to NIs are coloured in blue. In yellow, green and red are shown features related to frailty, patient demographics and severity of the COVID-19 infection respectively. Descriptions of the features are given in Section 2. HFRS = Hospital Frailty Risk Score,[12] GFS = Global Frailty Score,[13] CCI = Charlson Comorbidity Index,[11] IMD = Index of Multiple Deprivation.[14]

Supplementary material Table S1: ICD-10 codes used to identify severe COVID-19 infection

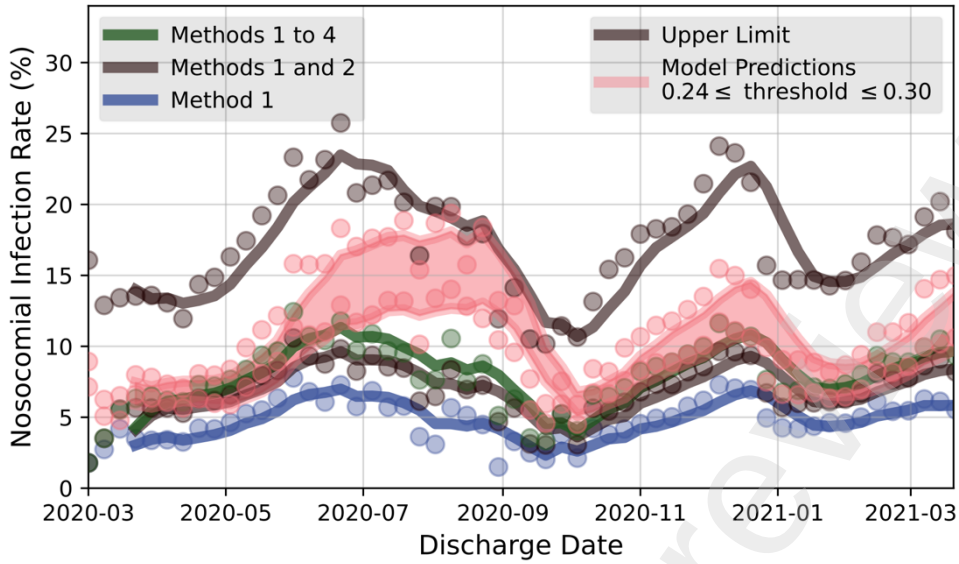
Disease type	Condition	Codes
Pneumonia	Pneumonia	J12-, J15-, J168, J17-, J18-
	Aspiration pneumonia	J690, J698
	Respiratory failure (acute and unspecified)	J960, J969
	Pulmonary embolism	I26-
	Adult respiratory distress syndrome (ARDS)	J80
	Pulmonary fibrosis	J841+B972
	Acute upper respiratory conditions	J00, J040, J042, J069
Kidney disease	Renal failure (acute)	N17-
Blood clotting	Thrombocytopaenia	D695, D696, D688, D689
	Disseminated intravascular coagulation syndrome (DIC)	D65
	Venous sinus thrombosis	G08, G951
	Other blood clotting	I74-, I80-, I81, I82
Cardiology/ circulation	Myocarditis	I40-, I411, I514, I520, I521
	Acute myocardial infarction (MI)	I21-, I22-
	Cardiomyopathy	I430
	Vasculitis	I776
Neurology	Stroke (due to Covid-19)	I60-, I61-, I62-, I63-, I64
	Brain injury	A858, A86, G048, G049, G931
Digestive system	Intestinal ischaemia	K550
Sepsis	Sepsis	R572, A40-, A41-



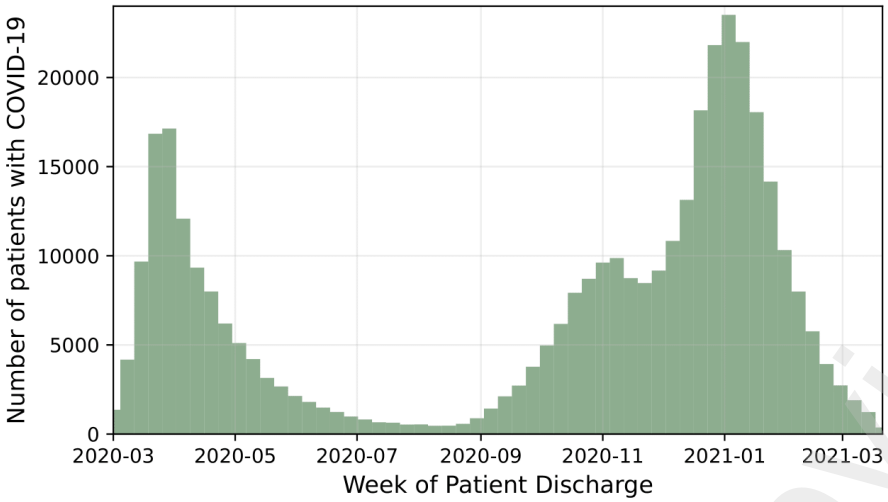
Supplementary material Figure S1: Example of the relationship between episodes and spells in the Hospital Episodes Statistics dataset**.** Diagram of spells extracted from Hospital Episodes Statistics dataset, with a discharge date between March 1st 2020 and March 31st 2021: The extracted spells are labelled A-E and are made up of individual episodes of care. A spell may be a single episode of care under a single consultant (spell A) or made up of multiple episodes (spells B-D). Methods 2 and 4 rely on identifying the first episode of a spell with a COVID-19 diagnosis. Adapted from *Methodology to create provider and CIP (continuous in-patient) spells from HES APC (Admitted Patient Care) data*.[46]

Supplementary material Table S2: Optimised hyperparameters of the random forest obtained through grid search. The remaining parameters were fixed to their default scikit-learn values.[17]

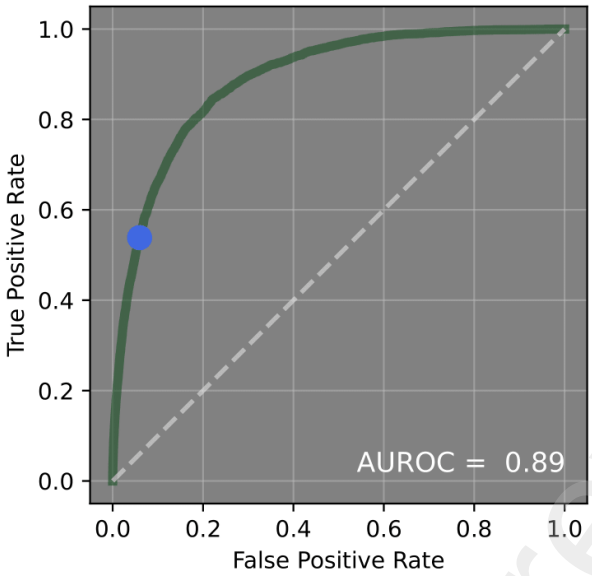
Hyperparameters	Values
Number of Trees	283
Maximum Depth	81
Bootstrap	TRUE
Class weights	None



Supplementary material Figure S2: Illustration of the calibration procedure by bounding the model predictions using the upper limit (upper black curve, see Section 2.2) and the curve of definite NIs (lower black curve). The predicted time evolution of the NI rate is shown in orange, for probability thresholds between 0.24 and 0.30. The discharge date was used to calibrate the model given the nature of the extraction procedure from HES data; this leads to an expected time lag between the results and the dynamics of the pandemic.



Supplementary material Figure S3: Time evolution of the extracted number of patients diagnosed with COVID-19, as function of date of discharge.



Supplementary material Figure S4. Receiver Operating Characteristics (ROC) curve obtained when applying the trained random forest classifier to the test set. The Area Under the ROC curve (AUROC) was found to be 0.89. The dotted diagonal line corresponds to a stochastic model with no learning; a perfectly accurate model would be represented by a single point on the top left corner (0, 1). The blue point corresponds to a probability threshold fixed at 0.25. The predictions shown in orange in **Supplementary material Figure S2** correspond to a threshold between 0.24 and 0.30, as constrained by the designed lower and upper limits.

References

1. Griffin KM, Karas MG, Ivascu NS, Lief L. Hospital preparedness for COVID-19: a practical guide from a critical care perspective. *American journal of respiratory and critical care medicine* 2020; **201**(11): 1337-44.
2. Patterson B, Marks M, Martinez-Garcia G, et al. A novel cohorting and isolation strategy for suspected COVID-19 cases during a pandemic. *Journal of Hospital Infection* 2020; **105**(4): 632-7.
3. Asad H, Johnston C, Blyth I, et al. health care workers and patients as Trojan horses: a COVID19 ward outbreak. *Infection Prevention in Practice* 2020; **2**(3): 100073.
4. Abbas M, Nunes TR, Martischang R, et al. Nosocomial transmission and outbreaks of coronavirus disease 2019: the need to protect both patients and healthcare workers. *Antimicrobial Resistance & Infection Control* 2021; **10**(1): 1-13.
5. Read JM, Green CA, Harrison EM, et al. Hospital-acquired SARS-CoV-2 infection in the UK's first COVID-19 pandemic wave. *The Lancet* 2021.
6. Ponsford MJ, Jefferies R, Davies C, et al. The burden of nosocomial covid-19: results from the Wales multi-centre retrospective observational study of 2518 hospitalised adults. *medRxiv* 2021.
7. Graham NS, Junghans C, Downes R, et al. SARS-CoV-2 infection, clinical features and outcome of COVID-19 in United Kingdom nursing homes. *Journal of Infection* 2020; **81**(3): 411-9.
8. Carter B, Collins J, Barlow-Pay F, et al. Nosocomial COVID-19 infection: examining the risk of mortality. The COPE-Nosocomial Study (COVID in Older PEople). *Journal of Hospital Infection* 2020; **106**(2): 376-84.
9. Oswald M. Anonymisation Standard for Publishing Health and Social Care Data Specification (Process Standard). *Leeds, UK, Information Standards Board for Health and Social Care* 2013.
10. European Centre for Disease Prevention and Control. Surveillance definitions for covid-19. 2021. https://www.ecdc.europa.eu/en/covid-19/surveillance/surveillance-definitions.
11. Sundararajan V, Henderson T, Perry C, Muggivan A, Quan H, Ghali WA. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *Journal of clinical epidemiology* 2004; **57**(12): 1288-94.
12. Gilbert T, Neuburger J, Kraindler J, et al. Development and validation of a Hospital Frailty Risk Score focusing on older people in acute care settings using electronic hospital records: an observational study. *The Lancet* 2018; **391**(10132): 1775-82.
13. Soong JT, Kaubryte J, Liew D, et al. Dr Foster global frailty score: an international retrospective observational study developing and validating a risk prediction model for hospitalised older persons from administrative data sets. *BMJ open* 2019; **9**(6): e026759.
14. Ministry of Housing and Communities and Local Government. English indices of deprivation. 2019. https://www.gov.uk/government/collections/english-indices-of-deprivation (accessed 2021-08-25.
15. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 2011; **12**: 2825-30.
16. Breiman L. Random forests. *Machine learning* 2001; **45**(1): 5-32.
17. developers s-l. sklearn.ensemble.RandomForestClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed 2021-08-05.
18. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics* 2007; **8**(1): 1-21.
19. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC bioinformatics* 2008; **9**(1): 1-11.
20. Navaratnam AV, Gray WK, Day J, Wendon J, Briggs TW. Patient factors and temporal trends associated with COVID-19 in-hospital mortality in England: an observational study using administrative data. *The Lancet Respiratory Medicine* 2021; **9**(4): 397-406.
21. Gray WK, Navaratnam AV, Day J, Wendon J, Briggs TW. Changes in COVID-19 in-hospital mortality in hospitalised adults in England over the first seven months of the pandemic: An observational study using administrative data. *The Lancet Regional Health-Europe* 2021; **5**: 100104.

22. Faes C, Abrams S, Van Beckhoven D, Meyfroidt G, Vlieghe E, Hens N. Time between symptom onset, hospitalisation and recovery or death: statistical analysis of Belgian COVID-19 patients. *International journal of environmental research and public health* 2020; **17**(20): 7560.
23. Office for National Statistics. Coronavirus (COVID-19) Infection Survey technical article: waves and lags of COVID-19 in England, 2021.
24. Taylor J, Rangaiah J, Narasimhan S, et al. Nosocomial COVID-19: experience from a large acute NHS Trust in South-West London. *Journal of Hospital Infection* 2020; **106**(3): 621-5.
25. Rickman HM, Rampling T, Shaw K, et al. Nosocomial transmission of coronavirus disease 2019: a retrospective study of 66 hospital-acquired cases in a London teaching hospital. *Clinical infectious diseases* 2021; **72**(4): 690-3.
26. Lauer SA, Grantz KH, Bi Q, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine* 2020; **172**(9): 577-82.
27. World Health Organization. Clinical management of severe acute respiratory infection (SARI) when COVID-19 disease is suspected: interim guidance, 13 March 2020: World Health Organization, 2020.
28. Arons MM, Hatfield KM, Reddy SC, et al. Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. *New England journal of medicine* 2020; **382**(22): 2081-90.
29. Public Health England. Covid-19: investigation and initial clinical management of possible cases. 2020. https://www.gov.uk/government/publications/wuhan-novel-coronavirus-initial-investigation-of-possible-cases (accessed 2021-07-12.
30. Public Health England. Covid-19: guidance for households with possible coronavirus infection. 2020. https://www.gov.uk/government/publications/covid-19-stay-at-home-guidance (accessed 2021-07-12.
31. Public Health England. Covid-19: guidance for healthcare providers who have diagnosed a case within their facility. 2020. https://www.gov.uk/government/publications/covid-19-guidance-for-healthcare-providers-who-have-diagnosed-a-case-within-their-facility (accessed 2021-07-12.
32. Zheng C, Hafezi-Bakhtiari N, Cooper V, et al. Characteristics and transmission dynamics of COVID-19 in healthcare workers at a London teaching hospital. *Journal of Hospital Infection* 2020; **106**(2): 325-9.
33. Keeley AJ, Evans C, Colton H, et al. Roll-out of SARS-CoV-2 testing for healthcare workers at a large NHS Foundation Trust in the United Kingdom, March 2020. *Eurosurveillance* 2020; **25**(14): 2000433.
34. Rimmer A. Covid-19: NHS staff express scepticism over promised twice weekly testing. British Medical Journal Publishing Group; 2020.
35. Shorten RJ, Haslam S, Hurley MA, et al. Seroprevalence of SARS-CoV-2 infection in healthcare workers in a large teaching hospital in the North West of England: a period prevalence survey. *BMJ open* 2021; **11**(3): e045384.
36. Tomlinson J, Khan S, Page G. Incorporating SPACES recommendations to the COVID-19 ward care approach at the Royal Bournemouth Hospital. *Clinical Medicine* 2020; **20**(6): e234.
37. Fink D, Khan P, Goldman N, et al. Development and internal validation of a diagnostic prediction model for COVID-19 at time of admission to hospital. *QJM: monthly journal of the Association of Physicians* 2020.
38. Yen M-Y, Lin Y-E, Lee C-H, et al. Taiwan's traffic control bundle and the elimination of nosocomial severe acute respiratory syndrome among healthcare workers. *Journal of Hospital Infection* 2011; **77**(4): 332-7.
39. Sultan S, Altayar O, Siddique SM, et al. AGA institute rapid review of the gastrointestinal and liver manifestations of COVID-19, meta-analysis of international data, and recommendations for the consultative management of patients with COVID-19. *Gastroenterology* 2020; **159**(1): 320-34. e27.
40. Gray WK, Navaratnam AV, Day J, et al. Variability in COVID-19 in-hospital mortality rates between national health service trusts and regions in England: A national observational study for the Getting It Right First Time Programme. *EClinicalMedicine* 2021; **35**: 100859.

41. Zazzara MB, Penfold RS, Roberts AL, et al. Probable delirium is a presenting symptom of COVID-19 in frail, older adults: a cohort study of 322 hospitalised and 535 community-based older adults. *Age and ageing* 2021; **50**(1): 40-8.
42. NHS Digital. Covid-19 national clinical coding standards (01 april 2021). 2021. https://hscic.kahootz.com/gf2.ti/f/762498/96552069.1/PDF/-/COVID19_NCCS_01_April_2021.pdf (accessed 2021-07-13.
43. Hulsen T, de Vlieg J, Alkema W. BioVenn–a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC genomics* 2008; **9**(1): 1-6.
44. NHS England and NHS Improvement. Healthcare associated covid-19 infections – further action. 2020. https://www.england.nhs.uk/coronavirus/wp-content/uploads/sites/52/2020/06/Healthcare-associated-COVID-19-infections--further-action-24-June-2020.pdf (accessed 2021-07-30.
45. Infection Prevention and Control. Training resources: COVID-19 infection prevention and control (IPC) recommendations for healthcare settings. 2020.
46. Health and Social Care Information Centre. Methodology to create provider and CIP spells from HES APC data. 2014. http://content.digital.nhs.uk/media/11859/Provider-Spells-Methodology/pdf/Spells_Methodology.pdf2021-07-13).