

Using machine learning to forecast peak health care service demand in real-time during the 2022-23 winter season: a pilot in England, UK

Roger A Morbey¹ (<https://orcid.org/0000-0001-8543-477X>), Dan Todkill^{1,2} (<https://orcid.org/0000-0002-4325-4786>), Phil Moura³, Liam Tollinton⁴, Andre Charlett⁵ (<https://orcid.org/0000-0001-7154-0432>), Conall H. Watson⁶ (<https://orcid.org/0000-0001-7154-0432>), Alex J Elliot¹ (<https://orcid.org/0000-0002-6414-3065>).

¹ Real-time Syndromic Surveillance Team, Field Services, Health Protection Operations, UK Health Security Agency, Birmingham B2 4BH, United Kingdom.

² Health Sciences, Warwick Medical School, University of Warwick, Coventry, United Kingdom

³ Department of Health and Social Care, London, United Kingdom

⁴ Health Analytics and Automation, Data Analytics and Surveillance, UK Health Security Agency, London, United Kingdom

⁵ Statistics, Modelling and Economics Division, UK Health Security Agency, London, United Kingdom

⁶ Immunisation and Vaccine Preventable Diseases Division, UK Health Security Agency, London, United Kingdom

Corresponding author: Roger Morbey, Real-time syndromic Surveillance Team, UK Health Security Agency, 23 Stephenson Street, Birmingham B2 4BH, United Kingdom.

Roger.Morbey@ukhsa.gov.uk

Keywords: syndromic surveillance; epidemiology; respiratory; respiratory syncytial virus; public health

Abstract

During winter months, there is increased pressure on health care systems in temperate climates due to seasonal increases in respiratory illnesses. Providing real-time short-term forecasts of the demand for health care services helps managers plan their services.

During the Winter of 2022-23 we piloted a new forecasting pipeline, using existing surveillance indicators which are sensitive to increases in respiratory syncytial virus (RSV). Indicators including telehealth cough calls and ED bronchiolitis attendances, both in children under 5 years. We utilised machine learning techniques to train and select models that would best forecast the timing and intensity of peaks up to 28 days ahead. Forecast uncertainty was modelled using a novel gamlss approach which enabled prediction intervals to vary according to the level of the forecast activity.

The winter of 2022-23 was atypical because the demand for healthcare services in children was exceptionally high, due to RSV circulating in the community and increased concerns around invasive Group A streptococcal (GAS) infections. However, our short-term forecasts proved to be adaptive forecasting a new higher peak once the increasing demand due to GAS started. Thus, we have demonstrated the utility of our approach, adding forecasts to existing surveillance systems.

Introduction

Healthcare services are prone to experiencing periods of high burden and demand for services ('pressures') during winter months each year. These pressures can lead to severe problems in delivering critical health services. During winter months, healthcare pressures are exacerbated by factors that can increase demand, including cold weather, respiratory pathogens, gastrointestinal pathogens and subsequent workforce absences.¹ In particular, the role of influenza and respiratory syncytial virus (RSV) in driving winter pressures has been extensively documented.

RSV is a major cause of bronchiolitis and bronchitis amongst young children² and although mainly produces mild symptoms, RSV infection can lead to severe illness in the immunocompromised³ and is a major cause of death in infants globally.⁴

During periods of heightened influenza and RSV activity, increases in demand can occur across a range of healthcare services from community physicians (general practitioners; GPs) through to specialist secondary care facilities. In England, RSV accounts for approximately 30,000 paediatric admissions in children aged <5 annually.⁵

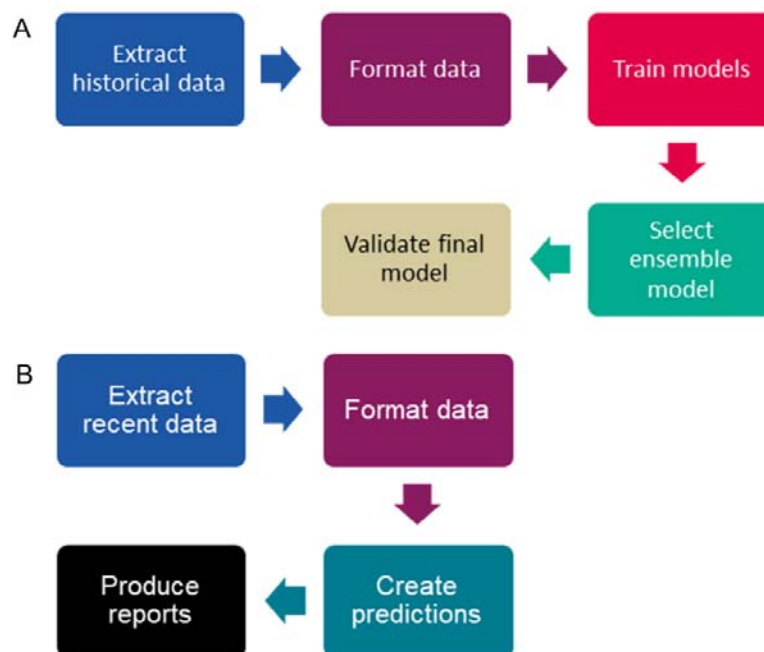
Identifying the key drivers underlying winter pressures is critical to understanding, managing and responding to the periods of high demand. Surveillance is a cornerstone of public health, monitoring changes in community-based activity of certain pathogens, diseases and conditions. Surveillance can provide a 'view' of key metrics that can be used to understand the drivers of pressures. Routinely collected surveillance data provide intelligence on those factors known to cause pressures e.g. monitoring increases in influenza cases. Surveillance data can also provide the opportunity to anticipate these pressures through predictions or forecasting.

Recently, advancements in Machine Learning (ML) have made it possible to develop more powerful and accurate forecasting models, utilising larger and more complex datasets. However, the key to developing accurate and timely models is the availability of suitable surveillance data that informs on healthcare service usage. Here, we use real-time syndromic surveillance data that are routinely collected as part of the UK Health Security Agency (UKHSA) public health surveillance programme to create short term forecasts for peak health care demand during periods of rising seasonal respiratory activity. We calculated forecast reliability to describe uncertainty around forecasts and piloted forecasts during the 2022-23 winter season and compared forecasts to actual activity.

Methods

We created two automated machine-learning pipelines in R, firstly to select and train forecast models, secondly to create daily forecasts (Figure 1). We describe here the methods used following the flow of the pipelines.

Figure 1: Pipelines used for A) creating models and B) producing daily forecasts



Data selection

The UKHSA coordinates a programme of real-time syndromic surveillance that supports and augments other UKHSA health surveillance programmes.⁶ For this pilot study, we used two syndromic indicators that are routinely part of the ongoing UKHSA daily syndromic surveillance service; NHS 111 telehealth calls for ‘cough’ and emergency department (ED) attendances for ‘acute bronchiolitis’. Both syndromic indicators were restricted to children aged five years or under because they are known to be sensitive to seasonal outbreaks of RSV.⁷⁻⁹ Using established indicators that are well understood aids interpretation and enables comparison with previous years.

Data cleaning and formatting

Firstly, data were smoothed to remove day of the week effects caused by weekends and public holidays.¹⁰ Secondly data were normalised, so that variables were in the range zero to one. Finally, derived variables were created that were used to create the forecast models (although noting that not all the derived variables were used in all the models tested; Table 1).¹¹

Table 1: derived variables used to create forecast models

Derived variable	Descriptor
Trend slope	difference between today and yesterday
Rate of change	difference between today’s slope and yesterday
Seasonality	using either month of year or Fourier transformations
Secular trends	linear and quadratic
Quadratic term	square of daily activity
Three-point-moving-average	daily activity, trend slope and rate of change of slope

Training models

Rather than restrict forecasts to a single methodology, for both indicators we tested a wide range of alternate models, using the data to select the best method for each indicator ¹¹.

Firstly, we choose seven alternate supervised machine learning methods; linear regression, generalised linear models with elastic net regularization (with and without internal optimisation of parameter lambda), k-Nearest-Neighbour regression, random forest regression, support vector machine for regression, and eXtreme Gradient Boosting regression. For each of these seven different methods we included the following options:

1. No seasonality or including month of year or with Fourier transformation
2. No secular trend, a linear trend or a quadratic trend
3. With or without a quadratic term for current activity
4. With or without using a three-day moving average for activity, slope and change of slope to avoid undue influence of single-day spikes in activity.

The combination of seven methods and the options above gave 252 alternate model specifications to be tested. Furthermore, for each model specification 28 models were created and trained to forecast from 1 to 28 days ahead respectively. Models were trained using historical data prior to October 2022, the supervised learning using actual data from either 1 to 28-days ahead of daily forecasts as the ‘labels’ for the target forecast. Historical data were split randomly into training and test data sets, 80% of the historical data being used for training.

We used anonymised health service data that is routinely used by UKHSA for public health surveillance of respiratory illnesses, including RSV. This study was part of ongoing work to improve the capabilities of UKHSA surveillance systems. As such, no specific approvals were required to use the anonymised data included this study.

Selection of ensemble model

For each of the 252 alternate model specifications an ensemble forecast was created which forecast when and how high activity was going to peak over the next 28 days. The ‘forecast peak’ for each model specification was defined as the highest value forecast by the individual 1 to 28-day ahead forecast models. Using the test data set, the forecast peaks were compared with the actual highest value or ‘peak’ that occurred in the 28 days following the forecast. An ‘intensity error’ was calculated as the difference between the height of the forecast peak and the actual peak. Similarly, a ‘timing error’ was calculated which was the difference in days between the day when activity peaked and the date when activity was forecast to peak. The intensity and timing errors were combined to give a single ‘forecast peak error’. The forecast peak error includes weighting to emphasise the importance of accurately forecasting peaks when activity is high.¹¹ The model specification which resulted in the smallest mean forecast peak error was selected for daily forecasts.

Model validation

Once the best model specification has been selected based on the training data, forecasts were retrained using all the available historical data. The historical intensity and timing errors were calculated for forecast peaks and used to estimate forecast uncertainty. To allow for variation in the standard deviation of errors as activity approaches a peak, a gamlss model was used to estimate standard deviation variation against the level of current activity. Thus, we can create uncertainty intervals which vary as activity approaches a seasonal peak.

Creating daily reports

To produce daily forecasts, recent data is extracted and formatted using the same data processes as for training models. The validated model for each indicator is used to produce 1 to 28-day ahead forecasts based on the latest data available. These forecasts are used to create

daily reports which predict when activity will peak in the next 28 days and at what level, including the uncertainty intervals.

Results

Model selection

The model specifications with the lowest forecast peak errors for both indicators used a random forest learning method, with seasonality modelled by Fourier transformations. However, the other specification options differed between the two indicators. The lowest forecast peak errors for the NHS 111 cough calls data included a quadratic term for activity, a quadratic secular trend and averaging over three consecutive days' data points. Whilst for ED acute bronchiolitis attendances the lowest errors involved a linear trend and no quadratic term for activity or averaging over consecutive days. Supplementary table 1 shows the forecast peak errors for each model specification. In general, errors were lower for NHS 111 calls than for ED bronchiolitis attendances, with 52 NHS 111 model specifications performing better than the best ED specification. Overall, including seasonality improved peak forecasts, with Fourier transformations performing better than seasonality using months. The learning method with the lowest mean errors was random forest, followed by k-nearest neighbour regression (Table 2).

Table 2: Mean forecast peak errors by method and syndromic surveillance system.

Learning method	ED ¹	NHS 111
	attendances	calls ²
random forest	0.0267	0.0195
k-nearest-neighbour	0.0359	0.0239
support vector machine	0.0399	0.0287
glm with elastic net regularization including optimised lambda	0.0483	0.0355
linear regression	0.0490	0.0378
generalised linear model (glm) with elastic net regularization	0.0487	0.0418
extreme gradient boosting	^0.1483	0.1307

^Only 6 out of 36 model specifications converged. ¹Emergency department; ²National Health Service 111

Model validation

The gamlss models show that the variation in intensity errors increase as actual counts increase (Table 3). By contrast, the variation in timing errors decrease as actual counts increase.

Table 3: Gamlss model coefficients for variation in error standard deviation vs actual counts

Error type	System	Intercept	Coefficient
Intensity	NHS 111	24.5	1.0006
	EDSSS	18.9	1.0055
Timing	NHS 111	5.9	0.9994
	EDSSS	12.0	0.9998

The gamlss model coefficients were used to create confidence intervals around the timing and intensity errors, which varied depending on the number of actual counts at the time of the forecast (Figures 2 and 3). Intensity errors are not symmetric because forecast peaks cannot be negative, thus when actual counts are low a forecast peak can over-estimate by more than it can under-estimate.

Pilot season 2022-23

During October 2022, ED acute bronchiolitis attendances in children aged under 5 years reporting to EDSSS increased until a peak of 220.0 attendances on 31 October. Subsequently, there was a decrease until 5 November before attendances increased again reaching a seasonal high of 311.4 attendances on 29 November. Similarly, NHS 111 calls for cough in children under 5 years rose to a peak of 991.6 calls on 22 October, decreased until 2 November and then started rising. However, whilst the increase in NHS 111 calls slowed prior to 30 November it was then followed a sharp increase in calls, reaching a seasonal high of 1,842.9 calls on 6 December 2022.

The seasonal peak in ED attendances at the end of November coincided with the usual timing of peak RSV activity seen in previous years (as monitored by laboratory reporting).¹² The additional increase in NHS 111 calls after 30 November 2022 was unprecedented, being 39.3% higher than the previous highest winter peak, 1,323.4 on 7 December 2019. Consequently, the level of activity was outside the range of anything seen in the training data.

Figure 2: Intensity errors during training period for A) emergency department acute bronchiolitis attendances and B) NHS 111 cough calls. Lines show 50% and 95% confidence intervals.

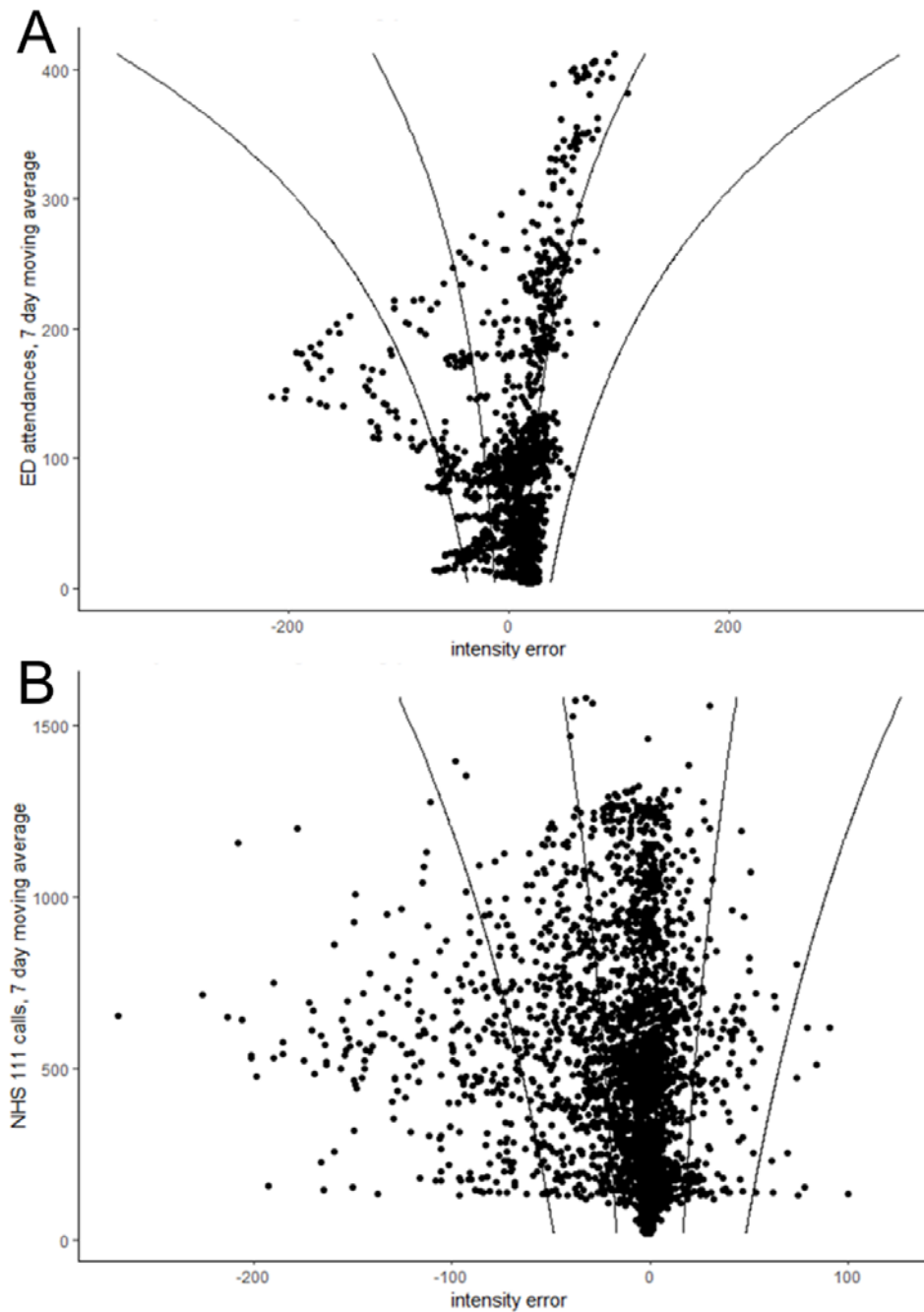
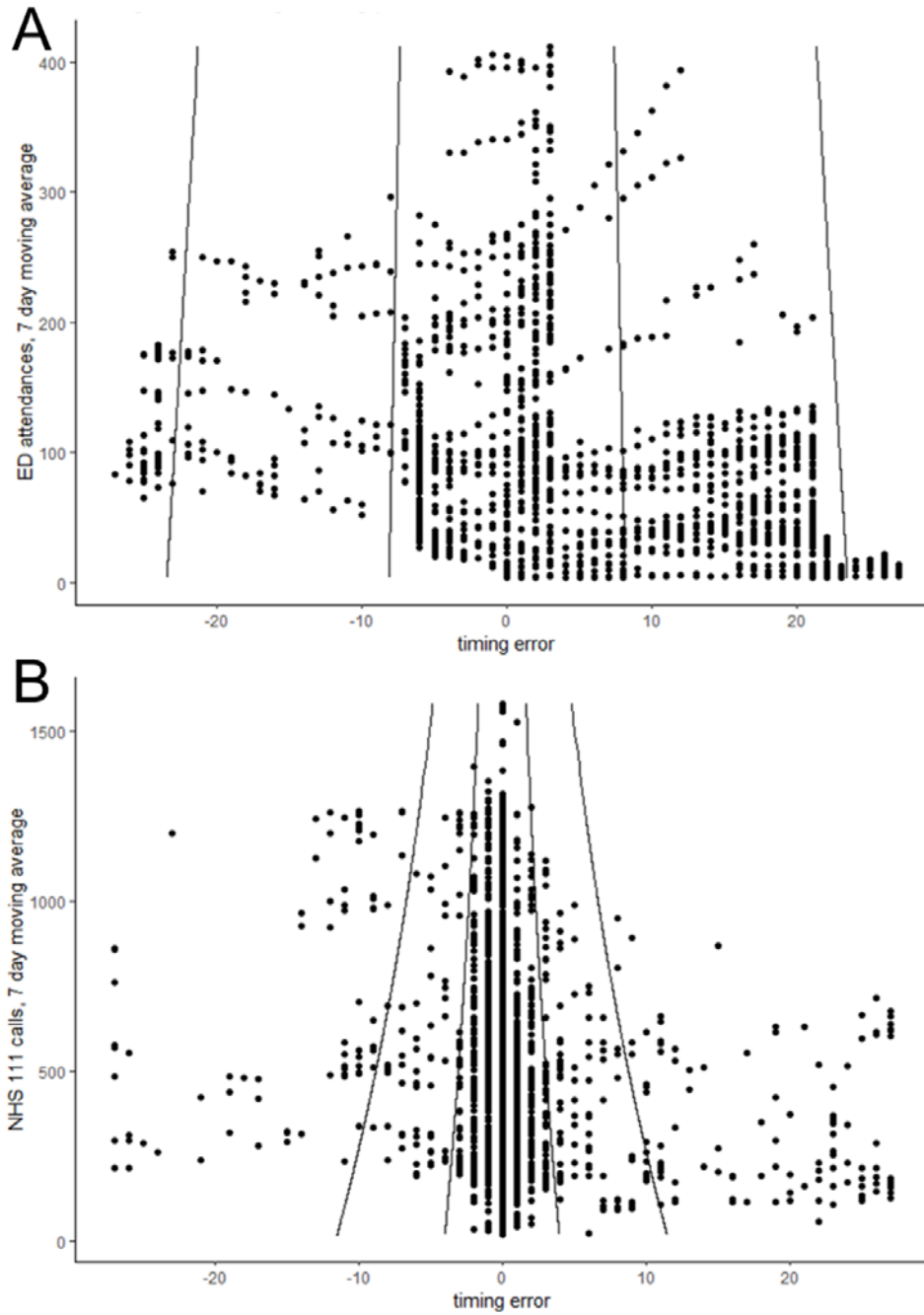
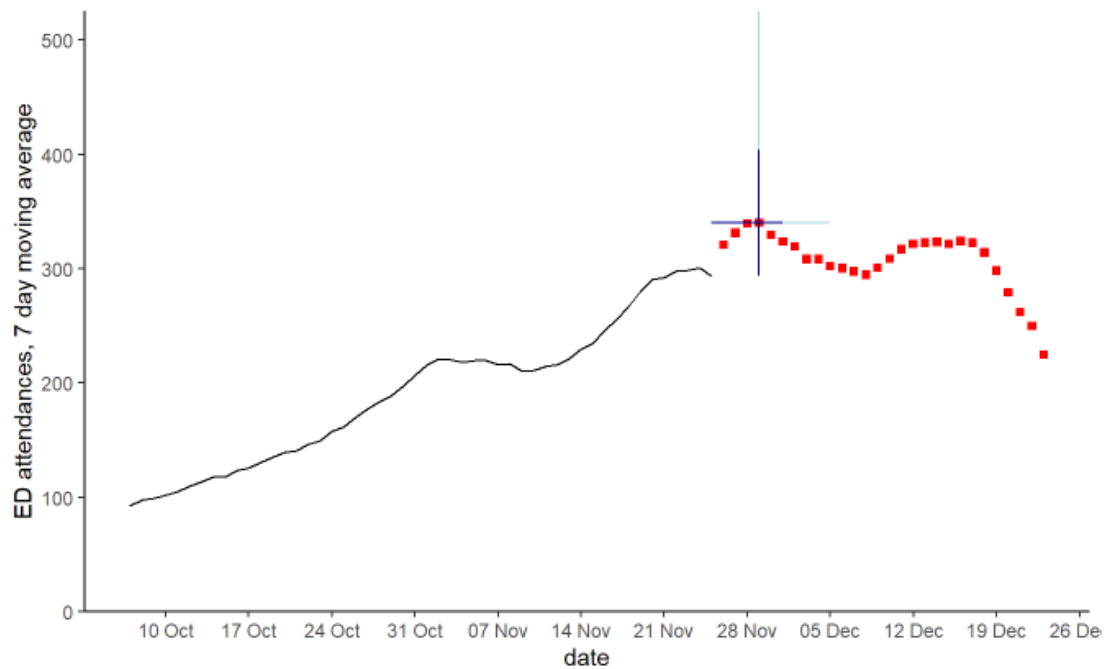


Figure 3: Timing errors during training period for A) emergency department acute bronchiolitis attendances and B) NHS 111 cough calls. Lines show 50% and 95% confidence intervals.



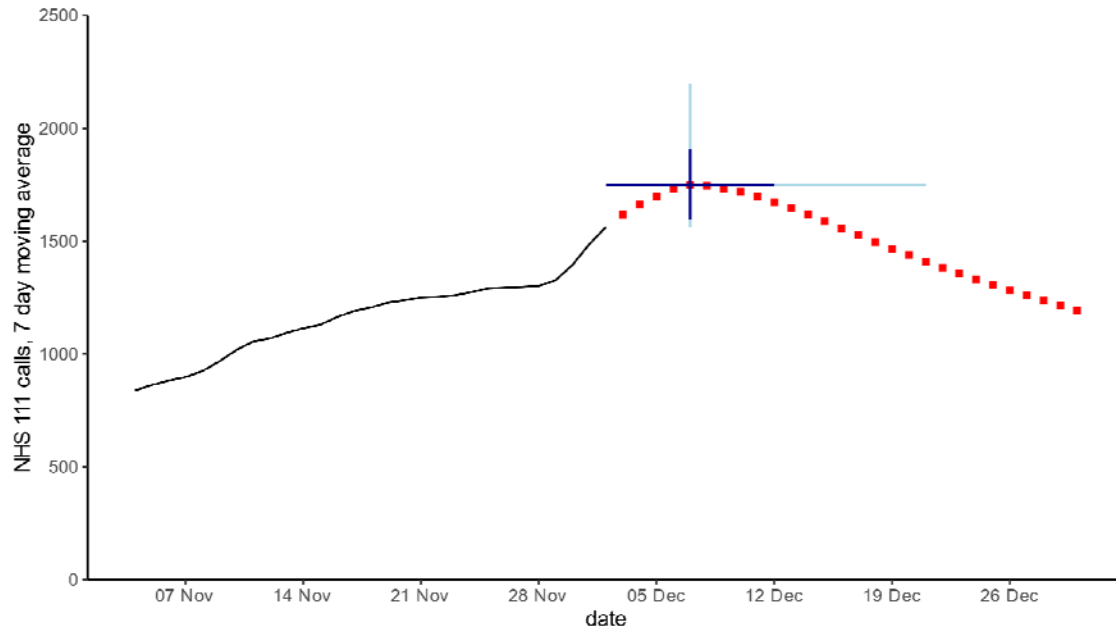
The pilot forecast made on 26 November 2022, forecast that ED attendances for acute bronchiolitis would peak at 340.1 attendances on 29 November (Figure 4). The same day forecast for NHS 111 calls predicted that they had already peaked. The timing for the ED forecast was correct but the level of the peak was an over-estimate of 29.3 (9.4%) attendances.

Figure 4: 28 day forecast for peak emergency department acute bronchiolitis attendances in children aged <5 years. Red squares are 28 day forecast, blue lines show 50% (dark blue) and 95% (light blue) data intervals around the peak forecast.



The NHS 111 forecasts failed to predict the unprecedented rise in NHS 111 cough calls in children in December 2022 until the rise had started. However, a forecast using a linear regression learning method proved to be adaptive, forecasting a later and higher seasonal peak once activity began to rise sharply at the start of December 2022 (Figure 5).

Figure 5: 28 day forecast for peak NHS 111 cough calls in children aged <5 years. Red squares are 28 day forecast, blue lines show 50% (dark blue) and 95% (light blue) data intervals around the peak forecast.



Discussion

Key findings

Machine learning pipelines can be used to train, select models and create daily forecast reports that predict the peak in demand for RSV activity over the following 28 days. During 2022 our pilot forecasts were able correctly predict the peak in ED acute bronchiolitis attendances in children under 5 years. During November 2022, our forecasts for NHS 111 cough calls in children aged under 5 years, predicted a similar peak was going to occur as in previous pre-pandemic years. However, as cough calls began to increase sharply at the start of December, our forecasts also began to change, predicting a later peak in December that was higher than previous years.

What was known before

Prior to the COVID-19 pandemic, seasonal activity for RSV was predictable in its seasonality, peaking in England at the end of November to beginning of December.¹³⁻¹⁵ However, during 2020 and 2021, the seasonality of RSV was disrupted with no winter peak in 2020 and a deferred peak occurring in summer 2021.⁷ Traditional surveillance methods based on historical data and recurring seasonality¹⁶ continued to predict syndromic indicators would rise in winter 2020 due to RSV. Whilst short-term forecasts based on recent trends are more adaptive, these too would not perform well during atypical seasons, unless seasonality was excluded from model variables.¹¹

Interpretation of findings

The unexpected dramatic rise in NHS 111 cough calls in December 2022 coincided with media reporting about unusual increases in invasive Group A streptococcal (GAS) infections in children.^{17 18} Extensive news coverage reported children with GAS becoming severely ill very quickly with symptoms including sore throat, fever and cough. Over the days following the release of the GAS news coverage, the NHS 111 telehealth service saw an unprecedented increase in calls relating to children. However, ED attendances were less affected by changes in patient presenting behaviour, and consequently ED syndromic indicators did not have an additional large peak. The differences between syndromic systems illustrate some of the strengths and weaknesses of syndromic surveillance. Syndromic indicators cannot identify specific causal pathogens, thus NHS 111 cough calls although sensitive to RSV are not specific enough to exclude other causal factors. Thus, NHS 111 cough calls were not a reliable indicator for assessing the total burden of health care demand attributable to RSV. However, if policy and decision-makers need to understand the current pressures on health services from all causes then syndromic indicators are more sensitive than pathogen-specific surveillance such as laboratory reporting. Importantly, a syndromic surveillance service that

comprises a range of data sources across the spectrum of health services is better able to distinguish between pressures due to changes in underlying disease incidence and those due to changing patient behaviour.

Limitations

Inevitably, forecasts trained on historical data will perform better when current data is within the range and seasonality seen previously. Previously, we have shown that a model that is not trained to expect a recurring seasonal pattern performs better during atypical seasons ¹¹.

However, the unexpected peak in December 2022 NHS 111 calls was not out of season and so the accuracy of forecasts was not due to inclusion of seasonality variables. In this case, we found that some of the regression methods which performed best using our forecast peak error measure, e.g. random forest, generated forecasts that assumed activity had already peaked. By contrast, using simple linear regression generated forecasts that correctly predicted activity was going to continue to increase in line with current trends. Therefore, the simpler method outperformed the method automatically selected by our machine learning algorithm.

The unprecedented increase in NHS 111 cough calls in children during December 2022 revealed a limitation in the use of this indicator for forecasting peak pressures due to RSV. The exceptional additional winter pressures were not due to RSV but due to reaction to iGAS media reports. Thus, without additional intelligence, a report intended to show pressures due to RSV could have misinterpreted as showing that RSV activity was exceptionally high.

Public health implications

Short-term forecasts can provide additional information compared to existing surveillance baselines based on previous years because they are more adaptive to recent changes in trends. Our automated pipeline for creating short-term forecasts of seasonal peaks is useful in

identifying the timing and intensity during typical seasons. Furthermore, the process of using machine learning methods to produce a reproducible automated pipeline means new indicators can easily and quickly be added to syndromic reports. However, automated reporting of trends and forecasts for syndromic data should always be accompanied with expert interpretations which can warn of emerging events. For instance, where a real-time change in patient behaviour means one or more indicators is no longer comparable with previous years. Improved automation and real-time interpretation is important as we may need to create forecasts quickly when notified that there is an increase in disease incidence. Also, the same pipeline can be used to assess other causal pathogens, including influenza and SARS-Cov2. Similarly, it may be possible to model non-infectious diseases such as allergic rhinitis (hay fever) where the historical data includes recurring seasonal peaks.

Recommendations and future work

Importantly, when current data is outside the range of training data, or seasonality does not match the training data forecasts should be interpreted in caution. We'd recommend that any forecasts used for routine surveillance include tests for data that is outside the range of testing data. Also, we suggest that when increases start to occur out-of-season forecast models are selected that do not include seasonality variables.

In future, it may be possible to provide better forecasts during atypical seasons by weighting the training data to give more emphasis to the rare events. Also, synthetic data could be incorporated in training data to allow for plausible events that have not yet occurred in the training data, e.g. out-of-season outbreaks, or more virulent pathogens.

Author contributions: RAM conceived the study and undertook all data collection and analysis. PM and LT provided technical, data science and script coding support. AC and CW provided topic specific expertise and advice. RAM, DT and AJE drafted the first manuscript and all authors contributed to subsequent versions and approved the final version for submission.

Conflict of interest: Nothing to declare.

Funding: This study was funded by the National Institute for Health Research (NIHR). This work also benefits from the infrastructure and partnerships assembled by Health Data Research UK, including through the Data and Connectivity National Core Study, funded by UK Research and Innovation [grant reference MC_PC_20058].

Ethical statement: All data used in this study were anonymised. The UKHSA has access to a range of data sources under Regulation 3 (Health Protection) of the Health Service (Control of Patient Information) Regulations 2002. This study therefore did not require specific ethics approval.

Data Availability Statement: Applications for requests to access relevant anonymised data included in this study should be submitted to the UKHSA Office for Data Release. Available at: <https://www.gov.uk/government/publications/accessing-ukhsa-protected-data>.

Acknowledgements: RAM, DT and AJE are affiliated with the NIHR Health Protection Research Unit (HPRU) in Emergency Preparedness and Response at King's College London. AJE is affiliated with the NIHR HPRU in Gastrointestinal Infections at University of Liverpool. DT is supported by the NIHR Applied Research Collaboration (ARC) West Midlands. The views expressed are those of the author(s) and not necessarily those of the NIHR, the UK Health Security Agency or the Department of Health and Social Care.

References

- 1 Scobie S. Snowed under: understanding the effects of winter on the NHS. 2018.
Available: <https://www.nuffieldtrust.org.uk/resource/snowed-under-understanding-the-effects-of-winter-on-the-nhs>.
- 2 Li Y, Wang X, Blau DM, Caballero MT, Feikin DR, Gill CJ, et al. Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in children younger than 5 years in 2019: a systematic analysis. *Lancet* 2022;399:2047-64.
- 3 Obando-Pacheco P, Justicia-Grande AJ, Rivero-Calle I, Rodriguez-Tenreiro C, Sly P, Ramilo O, et al. Respiratory syncytial virus seasonality: a global overview. *J Infect Dis* 2018;217:1356-64.
- 4 Zhang S, Akmar LZ, Bailey F, Rath BA, Alchikh M, Schweiger B, et al. Cost of respiratory syncytial virus-associated acute lower respiratory infection management in young children at the regional and global level: a systematic review and meta-analysis. *J Infect Dis* 2020;222:S680-S7.
- 5 Reeves RM, Hardelid P, Gilbert R, Warburton F, Ellis J, Pebody RG. Estimating the burden of respiratory syncytial virus (RSV) on respiratory hospital admissions in children less than five years of age in England, 2007-2012. *Influenza Other Respir Viruses* 2017;11:122-9.
- 6 UK Health Security Agency. Syndromic surveillance: systems and analyses. 2023.
Available: <https://www.gov.uk/government/collections/syndromic-surveillance-systems-and-analyses>.
- 7 Bardsley M, Morbey RA, Hughes HE, Beck CR, Watson CH, Zhao H, et al.
Epidemiology of respiratory syncytial virus in children younger than 5 years in

- England during the COVID-19 pandemic, measured by laboratory, clinical, and syndromic surveillance: a retrospective observational study. *Lancet Infect Dis* 2022.
- 8 Hughes HE, Morbey R, Hughes TC, Locker TE, Pebody R, Green HK, et al. Emergency department syndromic surveillance providing early warning of seasonal respiratory activity in England. *Epidemiol Infect* 2016;144:1052-64.
- 9 Morbey RA, Harcourt S, Pebody R, Zambon M, Hutchison J, Rutter J, et al. The burden of seasonal respiratory infections on a national telehealth service in England. *Epidemiol Infect* 2017;145:1922-32.
- 10 Buckingham-Jeffery E, Morbey R, House T, Elliot AJ, Harcourt S, Smith GE. Correcting for day of the week and public holiday effects: improving a national daily syndromic surveillance service for detecting public health threats. *BMC Public Health* 2017;17:477.
- 11 Morbey R, Todkill D, DeAngelis D, Charlett A, Elliot A. DiD IT?: a differences-in-differences investigation tool to quantify the impact of local incidents on public health using real-time syndromic surveillance health data. *Epidemiol Infect* 2023;151:e56.
- 12 UK Health Security Agency. Respiratory infections: laboratory reports 2022. 2022. Available: <https://www.gov.uk/government/publications/respiratory-infections-laboratory-reports-2022>.
- 13 Goddard NL, Cooke MC, Gupta RK, Nguyen-Van-Tam JS. Timing of monoclonal antibody for seasonal RSV prophylaxis in the United Kingdom. *Epidemiol Infect* 2007;135:159-62.
- 14 Fleming DM, Taylor RJ, Lustig RL, Schuck-Paim C, Haguinet F, Webb DJ, et al. Modelling estimates of the burden of respiratory syncytial virus infection in adults and the elderly in the United Kingdom. *BMC Infect Dis* 2015;15:443.

- 15 Hardelid P, Pebody R, Andrews N. Mortality caused by influenza and respiratory syncytial virus by age group in England and Wales 1999-2010. *Influenza Other Respir Viruses* 2013;7:35-45.
- 16 Morbey RA, Elliot AJ, Charlett A, Verlander NQ, Andrews N, Smith GE. The application of a novel 'rising activity, multi-level mixed effects, indicator emphasis' (RAMMIE) method for syndromic surveillance in England. *Bioinformatics* 2015;31:3660-5.
- 17 Guy R, Henderson KL, Coelho J, Hughes H, Mason EL, Gerver SM, et al. Increase in invasive group A streptococcal infection notifications, England, 2022. *Eurosurveillance* 2023;28:pii=2200942.
- 18 UK Health Security Agency. Group A streptococcal infections: activity during the 2022 to 2023 season. 2023. Available: <https://www.gov.uk/government/publications/group-a-streptococcal-infections-activity-during-the-2022-to-2023-season/group-a-streptococcal-infections-12th-update-on-seasonal-activity-in-england>.