PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	A time-series cohort study to forecast emergency department visits in the city of Milan and predict high demand: a 2-day warning system
AUTHORS	Murtas, Rossella; Tunesi, Sara; Russo, Antonio; Andreano, Anita

VERSION 1 – REVIEW

REVIEWER	Strehlow, Matthew Stanford, Emergency Medicine
REVIEW RETURNED	03-Oct-2021

GENERAL COMMENTS	This is a strong research question and well designed approach pending statistical review to confirm the statistical methods. Major Revisions: 1. There are certain sections that would benefit greatly by English language and grammar review/revision. 2. There are some assumptions around Emergency Department services that are potentially controversial. Additionally, some points made about ED visits and care are not referenced adequately or discussed in a manner consistent with current ED literature. Some examples of this I will list here. What are the causes of ED crowding? These are often reported as hospital capacity issues as a key component and possibly more critical than moderate fluctuations in the number of new ED visits. Additionally, what are appropriate ED visits. In ED literature, it is generally demonstrated that most ED visits are appropriate from the standpoint of the patient as a non-expert in the cause and treatment of their undifferentiated symptoms. There are a number of statements in the manuscript where inappropriate ED visits are referred to. These should be referenced thoroughly if stated. Finally, the impact of <5% increase over median ED visits is probably very limited if the hospital is not at full capacity. It would be interesting to determine at what % increase ED directors would consider changing staffing or setting up changes to other operational items. In my subjective experience, this probably lies in the 10-20% increase. This makes the lack of predictive value for the Red Zone in the study very important and this should be discussed. This is a very important issue and I appreciate the authors tackling this issue. This paper deserves publication assuming the stats have been reviewed and deemed appropriate. The other edits above are more about interpretation than the actual design and methods.

REVIEWER	Thompson, Helen

	Queensland University of Technology, Mathematical Sciences , Statistics
REVIEW RETURNED	17-Nov-2021

GENERAL COMMENTS	I note that this manuscript has previously been submitted to a different journal and that this version of the manuscript is an exact copy without any incorporation of previous reviewer feedback. As such I refer the authors to the comments provided previously by reviewers.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Dr. Matthew Strehlow, Stanford
Comments to the Author:
This is a strong research question and well designed approach pending statistical review to confirm the statistical methods.
Major Revisions:
There are certain sections that would benefit greatly by English language and grammar review/revision.
There are some assumptions around Emergency Department services that are potentially controversial. Additionally, some points made about ED visits and care are not referenced adequately or discussed in a manner consistent with current ED literature. Some examples of this I will list here.
What are the causes of ED crowding? These are often reported as hospital capacity issues as a key component and possibly more critical than moderate fluctuations in the number of new ED visits.
We included some suggestions on the potential reason for ED overcrowding.
Additionally, what are appropriate ED visits. In ED literature, it is generally demonstrated that most ED visits are appropriate from the standpoint of the patient as a non-expert in the cause and treatment of their undifferentiated symptoms. There are a number of statements in the manuscript where inappropriate ED visits are referred to. These should be referenced thoroughly if stated.
With “inappropriate ED visits” we meant visits with a low level of priority, classified as white triage in Italy, which are considered inappropriate as they could have instead been managed by other levels of care like general practitioners. Causes for incorrect use of ED has been studied both in Italy and internationally, major causes are perception of a need to receive immediate care, preference for ED instead of primary levels of care and dissatisfaction or lack of trust in primary care services (Lega 2008). Increasing primary care accessibility acts as a restraint on the inappropriate use of emergency departments reducing inappropriate admission by 10-15% (Bruni 2016). We expanded the “inappropriate ED visits” definition and included references.
Lega, Federico, and Alessandro Mengoni. "Why non-urgent patients choose emergency over primary care services? Empirical evidence and managerial implications." Health policy 88.2-3 (2008): 326-338.
Bruni, Matteo Lippi, Irene Mammi, and Cristina Ugolini. "Does the extension of primary care practice opening hours reduce the use of emergency services?." Journal of health economics 50 (2016): 144-155.

Finally, the impact of <5% increase over median ED visits is probably very limited if the hospital is not at full capacity. It would be interesting to determine at what % increase ED directors would consider changing staffing or setting up changes to other operational items. In my subjective experience, this probably lies in the 10-20% increase. This makes the lack of predictive value for the Red Zone in the study very important and this should be discussed.

We thanks the reviewer for this comment. We agree with the consideration that a 5% increase over median ED visits is very limited and, for this reason, we defined the Red zone if higher or equal to 10%. This warning system has been used by selected hospital in the city of Milan for almost 2 months in January and February 2020. Unlikely, we are not aware of the minimum % increase ED directors consider the optimal cutoff for changing staff roasters. However, we further discussed the lack of predictive value for the red zone. In addition we compared our high demand definition with similar definition and with an algorithm currently used by the Lombardy region to define spikes.

This is a very important issue and I appreciate the authors tackling this issue. This paper deserves publication assuming the stats have been reviewed and deemed appropriate. The other edits above are more about interpretation than the actual design and methods.

Reviewer: 2
Dr. Helen Thompson, Queensland University of Technology
Comments to the Author:
I note that this manuscript has previously been submitted to a different journal and that this version of the manuscript is an exact copy without any incorporation of previous reviewer feedback. As such I refer the authors to the comments provided previously by reviewers.

Dear reviewer, this is the first time we submitted to BMJ Open, we previously submitted the paper to another journal but the submission procedure failed and the second round of revisions arrived too late to be considered. We believe that you are referring to this second round of revisions, the first was completely included in the paper, but we are not sure what suggestions you are referring too (that were 3 other reviewers). However, we tried to include all previous reviewers' suggestion in the paper.

##
PREVIOUS REVIEWERS' COMMENTS AND REPLIES

Associate Editor Comments:
Associate Editor: Ramlakhan, Shammi
Comments to the Author:
The concept of using predictive models for ED forecasting is topical and of interest to urgent care clinicians and administrators, particularly if the predictions accurately change and fit with particular variables/settings. 2-day advanced prediction would be potentially helpful when other factors are also taken into consideration. However, there are several issues outlined by the Reviewers which require addressing. If you can address these most prominent concerns, you would be welcome to submit a new manuscript which we would review again.

Please have the manuscript read by a native English speaker with a medical background to improve the standard of English grammar.

Some justification of the model choice, adequacy and validity should be provided. Which predictor variables were most important/strongest?
In particular, ARIMA models can give perform variably depending on the chosen parameters/context. A (simpler) baseline model should ideally be provided for comparison.
Please be clear whether the model was actually deployed and used (actioned) by the hospitals, EDs or public health, or whether the performance reported in Table 4 is essentially an extension of the validation. The performance deteriorates with more variation from the validation data, which is expected in deployed models (more so with the type of model chosen).
Dear Editor, thank you very much for these precious comments. To evaluate model adequacy and validity, in this new version of the paper we provided, as supplementary material, plots of

autocorrelation and correlation among residulas according to the Ljung-Box test. ACF plots of residuals were overall in significance limits and the Ljung-Box test showed overall no significant correlation between residuals at different lags, except Hospital E which showed residual autocorrelation up to lag 366. In this work we choose to include in each, thus leading to a different for each hospital, only factors statistically significant according to a p-value < 0.05. Strength of associations are displayed in Table 2. In addition, as you suggested, we compared the ARIMA model with two simpler models: a regression model (M1) containing only meteorological, environmental and festivities covariates and a generalized linear model (M2) containing in addition the Fourier terms to control for seasonality. Models M1 and M2, compared by likelihood ratio test with the full ARIMA model used, always fitted the data significantly worst then ARIMA with MAPE for M1 and M2 above 13.5% and 9.8% respectively. Validation and calibration of the model were performed with data from 2014 to 2019 while the waring system operated in January 2020. Thus, model choices were decided according to 6-years of data (2014-2019) and results were used in the operating period of the WS. We better specify this distinction along the paper.

Thank you for the opportunity to review your work. We hope you will consider EMJ again in the future.

Reviewer(s') Comments to Author:
Reviewer: 1

Comments to the Author
Thank you for giving me the opportunity to review this article.
Overall, I think the study provides important findings to the audience in the related field. However, I have some concerns as noted below that I would like the authors to address:
1. In general, MAPE for a few days ahead forecast should be lower. In this study, the MAPE for one to two days ahead forecasts lies between 6.6% to 11.2%. Therefore, it is better if the dynamic regression model results (regression with ARIMA errors) can be compared to another model, at least to a simple regression model to see how the forecast error changes. Also, I would like to ask the authors to represent the forecasts as a plot of actual versus predicted values for the validation periods they consider. The plot will allow us to better understand how the forecasts fit into observed values.
We really thank the reviewer for this comment. We compared the ARIMA model with a simple regression model (M1) containing only meteorological, environmental and festivities covariates and with a generalized linear model (M2) containing meteorological, environmental, festivities covariates and Fourier terms to control for seasonality. Given that the ARIMA model is a complex version of M1 and M2, we calculated the likelihood ratio test comparing the full model (ARIMA) to M1 and M2. Results suggested that all ARIMA models fitted the data significantly better than M1 and M2 with MAPE above 13.5% and 9.8 respectively (Supplementary Table 2).

2. In table 1, for example in the age-group, the first group is <=14, and the second group is 14-65. A range without parenthesis generally indicates that the two numbers are included to the range, which means the second age-group also consider 14. This can be seen throughout the table with other attributes. Therefore, reproduce the table with corrected marginal values and percentages.
We thank the reviewer for this correction, we updated the table accordingly.

3. In the study design, it says that 'using current health care databases of the emergency department "admissions" aggregated at hospital level'. If the athours used aggregated admission data, how the admissions were stratified into presentations? or is it incorrect wording?

We modified the sentence as “This is a retrospective study conducted in the territory of the Milan Agency for Health Protection (AHP) using current health care databases of daily ED visits aggregated at hospital level.”

4. It is better to investigate the effect of different temperature measures, such as daily minimum, and maximum. Authors have already found a statistically significant effect with daily average temperature, however, I believe this effect can be better accounted for with daily maximum temperature.
As sensitivity analysis we also investigated the effect of minimum, maximum and apparent temperature on daily ED visits. The greatest effect on ED visits were attributed to mean temperature while indicators of performance and AIC where generally superior for mean temperature compared with minimum, maximum and apparent temperature. In Supplementary Table 2 we also calculated, only for outliers’ days, the relative error mean of observed vs predicted values in order to evaluate if extreme temperatures were best outliers’ predictors than mean temperature. Number of outliers replaced ranged from 2 for hospital A to 7 for hospital D, results suggested an overall better fit of outliers using minimum temperature (3 over 5 hospital with smaller relative errors).

5. I would also like to ask how the future values of the covariates such as temperature, NO2 were obtained. I couldn't find a clear statement describing the predicted values of the covariates for the forecasting horizons.
The forecasts were made incorporating in the model past meteorological and environmental information via an Application Programming Interface (API) where 2-day future forecast of meteorological and environmental information were provided by ARPA Lombardia.

Reviewer: 2

Comments to the Author
The authors should be commended for providing a data-driven approach to inform 2-day short-term ED resource planning for high-demand events. Whilst I have recommended rejection of the manuscript in its current form, I encourage the authors to consider a resubmission after addressing major concerns.
The authors should be more accurate with their terminology. I believe what is meant is ED presentations, rather than admissions.
We referred, along the paper, to ED visits.

Weekly and yearly effects were included but it is not clear if seasonal effects, or monthly effects were investigated. Previous studies, e.g., Duwalage et al. (2020), have demonstrated time-varying effects beyond week or year. I don’t believe that year would be a periodic effect and the authors should take care if describing the effect appropriately. It would be useful to provide plots of the data to view the seasonal effects.
In this study we included two seasonal adjustment using Fourier terms, one for daily/day of the week variation and one for year-round variation. We modified the sentence in “Day of the week and year-round seasonality were controlled for by including Fourier terms, a series of sine-cosine functions able to approximate periodicity.14,28 For each seasonal period (up to 7 for day of the week seasonality and up to 365 for year-round seasonality), the number of Fourier terms was chosen to minimise the AIC. Each seasonal component can be written, in the model equation, as
∑_(j=1)^n▒[α_j sin⁡〖(2πjt/m)+β_j cos⁡(2πjt/m) 〗]
where n is the number of Fourier terms chosen to minimise the AIC (up to 7 for day of the week seasonality and up to 365 for year-round seasonality) and m is the seasonal period (7 for day of the week and 365 for year-round seasonality).”

An MAPE of 8.1% is not so great compared to previous studies, again Duwalage et al. (2020) reports more accurate predictions and also provides reference to accuracy of other similar studies.

We modified the sentence as “The models showed a good overall performance with the MAPEs always smaller between 5.5% and 8.1%. Our results are slightly better than other studies: Marcilio and colleagues forecasted daily ED visits with Generalized Linear Models, finding MAPEs between 5.4% and 11.5%, according to different forecasting horizon and controlling for temperature effect. Jones and colleagues, using similar models, found MAPEs that varied between 8.5% and 15.5%. However, Duwalage et al. using a Generalized Additive Model found MAPEs consistently lower than 5% for 14 day forecasts, which significantly improved including temperature in the model.”

Influenza is the only diagnostic category used in this study, with little justification. The authors need to provide readers with some understanding as to why influeza was considered and not other diagnostic categories, particularly given the inclusion of air quality measures that they cite are linked to cardiac and respiratory diseases. Additionally, influenza has been shown to have a lagged effect on hospital bed demand (I apologise for a lack of references here, the studies are old and established), but no such lagging was investigated in the ARIMA model.
Syndromic surveillance (such as ILI rates which in Italy are provided weekly by the National Health Service Sentinel System) is a very important key factor in the ED presentations, especially in winter season, and may be able to provide early warning of hospital bed pressures caused by seasonal respiratory disease (Wargon 2017, Morbey 2020). In addition, in Murtas 2021 we evaluated the hypothesis of the early presence of the COVID-19 epidemic in Italy by analysing data on trends of access to EDs using a Poisson regression model adjusted for seasonality and influenza outbreaks. In this work we found that predicting ED visits by considering both seasonality and ILI rates, compared to a model with only seasonality, increased notably the fitting of the model.
The authors also do not explain what festivities they are referring to. I do not have knowledge of festivities in Milan so have no intuition as to why this might be important to modelling ED presentations. Please provide some more details on the hospitals, e.g., size, type of hospital: public, private, teaching, etc.
We modified the methods paragraph including the festivities considered and some details on the hospital type.
The authors only consider daily mean temperature, which may mask temperature effects linked to ED high-demand. For example, it would be relatively straightforward to investigate daily minimum, maximum or a derived measure of extreme temperature. The authors also use relative humidity but perhaps the weather experienced by the patients might have more meaning, e.g., apparent temperature.
We provided supplementary analysis investigating the effect of minimum, maximum and apparent temperature on daily ED visits. In Supplementary Table 2 we compared ARIMA results for different temperature specification: mean, minimum, maximum and apparent temperature. The greatest effect on ED visits were attributed to mean temperature while indicators of performance and AIC where generally superior for mean temperature compared with minimum, maximum and apparent temperature. In Supplementary Table 2 we also calculated, only for outliers’ days, the relative error mean of observed vs predicted values in order to evaluate if extreme temperatures were best outliers’ predictors than mean temperature. Number of outliers replaced ranged from 2 for hospital A to 7 for hospital D, results suggested an overall better fit of outliers using minimum temperature (3 over 5 hospital with smaller relative errors).
Please comment on the suitability of the location of the monitoring stations to capture the air quality at the hospitals considered. For example, the monitoring station site appears to be in a roadside location but possibly hospitals A, B and E might not be well represented by roadside air quality measurements. Studies on air-quality measurement have established that measures of air-quality differ from road-side versus urban monitoring stations, e.g., De Jesus et al. (2019).
In this work environmental information was extracted from the Regional Environmental Protection Agency as collected by 1 monitoring station, classified as from urban-traffic, positioned in the centre of the city of Milan. For this reason, pollution estimated from the monitoring station (classified as from urban-traffic) used in the analysis might be of a greater magnitude than that really observed in each

hospital. However, even if hospitals were mostly located in the border of the city of Milan, they are all located in urban areas characterized by similar air pollution pattern. In fact, all hospitals are located in major traffic roads characterized high level of pollution (maps of NO2 and PM10 concentration in the city of Milan are available as supplementary materials in Magnoni et al. 2021).
Magnoni, Pietro, Rossella Murtas, and Antonio Giampiero Russo. "Residential exposure to traffic-borne pollution as a risk factor for acute cardiocerebrovascular events: a population-based retrospective cohort study in a highly urbanized area." International journal of epidemiology 50.4 (2021): 1160-1171.

The missing data imputation discusses the use of averaging data from the other monitoring stations. The authors also need to provide the amount of missing data. Some validation of the robustness (unbiasedness) of the imputations should be provided.
We thank the reviewer for this correction. We found a mistake in the data imputation paragraph which we corrected as “Missing values on a specific day were imputed with the average of the measure in that specific year.”. We provided missing information in supplementary material.

Exploratory plots of the data by predictor would be useful. A plot of the fitted models and the predictions over the test period should also be provided. The former might belong in the appendix but the latter should be provided in the main text so that readers have a clear understanding of the accuracy of the model.
We provided plots of predicted vs observed values in supplementary material.

The authors should consider alternative definitions of high demand, especially given the lack of accuracy in their current method, which might be due to the statistically modelling, or to the high demand definition. In a short-term high-demand forecasting scenario, is it really suitable to base this definition of the past 31 days? If you are using 31 days, wouldn’t it also be appropriate to forecast out further than 2 days and provide readers with an assessment of the utility of this methodology for forecasts beyond 2 days? For example, forecast out to the period aligned to staff rostering. If the argument is that the fluctuations can spike in shorter periods, then this supports the argument that a 31 day median is likely not adequate, and possibly also that the ARIMA is not suitable. If the model is intended to capture spikes, then more suitable models should be considered.
The fitted models are not sufficiently validated.
To evaluate the proposed definition, we further calculated high demands as: the number of visits exceeding the median of the preceding 7, 14 and 21 days and the number of visits exceeding the mean of the preceding 7, 14, 21 and 31 days, defining green, yellow and red level of high demand as above. We choose 7, 14 and 21 lag days in order to adjust for weekly variation in the number of ED visits by design. We further calculated high demands as defined by the Lombardy Region: when the number of visits exceeded the 91-percentile of the previous year time-series. Low demand days were defined as those with a number of visits smaller than 25-percentile, medium demand days as those with a number of visits between 25-percentile and 75-percentile, high demand days if between 75-percentile and 90-percentile, and finally very high demands days if over 91-percentile. There were slightly improvement in percentage accuracy between the definition used and the other algorithms and there was not a favourite algorithm for all hospitals: hospital B had a maximum improvement of 4% using the mean of the preceding 31 days or the median of the procedings 21 days, hospitals A and C had an improvement of 2% using the mean of the preceding 31 days, hospital D had an improvement of 2% using the mean of the preceding 21 days, and finally hospital E had an improvement of 1% using the mean of the preceding 21 or 31 days. Using the high demand definition used by the Lombardy Region we did not find any improvement in the accuracy, with an overall percentage of matched classification between 50% and 64%. High demand was always worse predicted compared to the definition used in our ED warning system. However, results showed a good prediction of very high demand days with a sensitivity between 38% and 67%.

The researchers consider only one type of model, ARIMA. It is not uncommon for ARIMA models with the same predictors and different parameters to produce several competing ARIMA models with very similar model choice statistics, and for these models to lead to very different interpretations of the causal effects of the predictors. The authors justify that the parameters are not sensitive to this extent but cite an article related to a (non-time-series) linear model, so this is not an appropriate or correct justification. Further model validation should be included in an appendix or through an open source repository, such as GitHub. These should show the top 10 competing ARIMA models. The auto.arima function in R produced this by default. Relevant plots should be produced to support the modelling process and validity.

Related to the points above is that ARIMA models can be good for short-term forecasting, which is the aim of this paper, but not as an explanatory model, which is also what the authors attempt to use this model for. There are a lot of fourier terms. Please describe the fourier terms and show how well they capture the periodic effects, e.g., overlaying the fitted model with the observed data. You might even show a plot with and without the fourier terms to understand what they are modelling.

Is the model proposed overly complex? How does the proposed model compare to something simpler, like a regression model with fourier-terms, but not the ARIMA of the residuals. Or even something very simple that requires no modelling, such as using the previous day (or previous 2 days) presentations to predict for tomorrow. Or even a simple moving average model. Does the ARIMA provide sufficient improved predication accuracy? A comparison to a baseline simpler model would help readers understand if they really need to invest the effort to model and understand ARIMAs. Given the frequency with which this model needs to be run and refitted to produce updated daily reports, this modelling requires a lot of manual fitting, unless the parameters are unchanged from the test set. If this is the case, this should be discussed.

As you suggested, we compared the ARIMA model with two simpler models: a regression model (M1) containing only meteorological, environmental and festivities covariates and a generalized linear model (M2) containing in addition the Fourier terms to control for seasonality. Models M1 and M2, compared by likelihood ratio test with the full ARIMA model used, always fitted the data significantly worst then ARIMA with MAPE for M1 and M2 above 13.5% and 9.8% respectively. Fourier terms were better described in the method section (see comments above). According to Fourier terms' specification, we modified the sentence in "Day of the week and year-round seasonality were controlled for by including Fourier terms, a series of sine-cosine functions able to approximate periodicity.14,28 For each seasonal period (up to 7 for day of the week seasonality and up to 365 for year-round seasonality), the number of Fourier terms was chosen to minimise the AIC. Each seasonal component can be written, in the model equation, as

$$\sum_{j=1}^{n}[\alpha_j \sin\{(2\pi jt/m)+\beta_j \cos(2\pi jt/m)\}]$$

where n is the number of Fourier terms chosen to minimise the AIC (up to 7 for day of the week seasonality and up to 365 for year-round seasonality) and m is the seasonal period (7 for day of the week and 365 for year-round seasonality)."

In the regression part of the modelling, are all the predictors significant? That is, are there issues with multicollinearity in the regression part of the model, prior to ARIMA. If there is multicollinearity then the estimation from the regression can be unstable. This can lead to poor performance in the predictions.

The predictors included were all statistically significant. Multicollinearity was evaluated calculating Pearson pairwise correlation between variables and variance inflation criterion (VIF). The Pearson correlation between predictors varied from weak (absolute correlation<0.3) to moderate (absolute correlation between 0.3 and 0.7), with a maximum of -0.67 between temperature and ILI and 0.61 between NO2 and PM10. VIF was smaller than 5 for all variables, with a maximum of 2.8 for temperature and 1.9 for ILI. We therefore included all the variables in the models, selecting the final model according to the statistical significance of predictors and minimal AIC.

How does trimming the outliers impact the predication accuracy? Don't you want to include these spikes to help better predict spikes (likely the red category). I understand this was needed to satisfy ARIMA constraints, so this might indicate that the ARIMA is not suitable, or that some further processing is needed. For example, the regression with Fourier terms is supplemented with the ARIMA to incorporate the time-dependent nature of the data. Further modelling could also be done to model the spikes, e.g., a model whose purpose is to capture extreme events such as a copula model. Or you might consider machine learning models, or stochastic process models. Browning et al. (2021) is one example of a good-fitting COVID model, albeit this paper considers country-level data. Please give information (numbers and plots) on the amount and extent of trimming and the impact this has.
We did not remove completely all spikes from time series but only those which significantly did not satisfy ARIMA constraint. We included the number of outliers day's replaced (between 2 and 7 over the whole validation period) in supplementary Table 2. In addition, in Supplementary Table 2 we calculated as sensitivity analysis the relative error mean of observed vs predicted values, only for outlier days, in order to evaluate if extreme temperatures were better outlier predictors than mean temperature. Number of outliers replaced ranged from 2 for hospital A to 7 for hospital D, results suggested an overall better fit of outliers using minimum temperature (3 out 5 hospitals with smaller relative errors). However, we really thanks the reviewer for the suggestion about copula models or machine learning techniques to better model spikes which we aim to implement in the future.

I have little knowledge on notable festive events in Milan. There is no explanation as to why August 15th is important. For the two events considered (NY Eve and August 15), did you investigate if there are lagged effects, i.e., that the effect due to the festivity doesn't impact the ED to its greatest extent until "x" days before/after the event. And if that is the case, is it due to any effect of it being a public holiday, if it is one?
In the method paragraph we included a description of the Italian festivities used. We did not investigate their lagged effect as we observed a specific increase or decrease in the number of ED visits in those exact days.

What might be an explanation for the difference in ARIMA parameters between hospitals? Is there something about the hospitals that would mean the patterns in presentations would differ? Or is it that the top few models for each of the hospitals are all fairly similar in AIC and there is a common model that might be suitable across all hospitals with minimal loss in predication accuracy. If there were a common model, and the hospitals behaved similarly, this would reduce the cost in producing the daily forecasts.
Before applying the current method to each of the five hospital we evaluated the hypothesis of a common model suitable across all hospital by summing the daily ED visits of the five hospitals. Results were very scarce and thus we decided to evaluate each hospital individually. Differences between hospitals are clear observing the mean number of ED visits which ranged between 124 for hospital C and 247 for hospital E.

There is a section in the results discussing explanatory effects of the predictors that should be more convincingly justified or omitted, given that the ARIMA is a forecasting model not an explanatory model. If you were to compare the competing ARIMA models (where the ARIMA parameters differ) and the coefficients were similar, then this is more convincing.
We thank the reviewer for this comment. The reviewer is right asserting that we out aim was not on describing an explanatory but a forecasting model and we better specified it in the discussion. However, we believe that specification of covariate effects can better help the readers to understand the approach especially considering that the model used is not a pure ARIMA model but a regression model with ARIMA errors.

The forecasting accuracy by MAPE or via classification using the high demand definition is not very good, even in the validation set and during the operating period, and especially for the red high

demand days. The MAPEs are not so bad but the classification error should be addressed, perhaps considering alternative definitions as previously discussed. The loss in accuracy during COVID is appropriately explained. The reference to papers with similarly good overall performance in MAPE refers to studies (a particular study here) with a longer forecasting horizon. With such a short forecasting horizon in this study, one should expect the forecasting error to be notably better than previously published studies with longer forecasting horizons. This is because the accuracy of ARIMA forecasts deteriorates as the forecasting horizon increases.
As supplementary analysis to evaluate the proposed definition, we further calculated high demand as: the number of visits exceeding the median of the preceding 7, 14, and 21 days and the number of visits exceeding the mean of the preceding 7, 14, 21, and 31 days, defining green, yellow, and red levels of high demand as above. We chose 7, 14, and 21 lag days in order to adjust for weekly variation in the number of ED visits by design. We further calculated high demand as defined by the Lombardy Region31: when the number of visits exceeded the 91st percentile of the previous year time-series. Low demand days were defined as those with a number of visits smaller than 25th percentile, medium demand days as those with a number of visits between 25th percentile and 75th percentile, high demand days if between 75th percentile and 90th percentile, and finally very high demand days if over 91th percentile. There was slight improvement in percentage accuracy between the definition used and the other algorithms and there was no favourite algorithm for all hospitals: hospital B had a maximum improvement of 4% using the mean of the preceding 31 days or the median of the proceeding 21 days, hospitals A and C had an improvement of 2% using the mean of the preceding 31 days, hospital D had an improvement of 2% using the mean of the preceding 21 days, and finally hospital E had an improvement of 1% using the mean of the preceding 21 or 31 days. Using the high demand definition used by the Lombardy Region we did not find any improvement in accuracy, with an overall percentage of matched classification between 50% and 64%. High demand was always predicted less well compared to the definition used in our ED warning system. However, results showed good prediction of very high demand days with a sensitivity between 38% and 67%.

The authors should also consider either submitting a draft for review to an editorial service for grammatical improvement, or invite a co-author to join the paper whose role is to improve the grammar to publication standard.

Duwalage et al. (2020) Forecasting daily counts of patient presentations in Australian emergency departments using statistical models with time-varying predictors - Duwalage - 2020 - Emergency Medicine Australasia - Wiley Online Library
De Jesus et al. (2020) Long-term trends in PM2.5 mass and particle number concentrations in urban air: The impacts of mitigation measures and extreme events due to changing climates - ScienceDirect
Browning et al. (2021) https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0250015

Reviewer: 3

Comments to the Author
Globally the area of predictive analytics, specifically where it comes to the area of emergency and unscheduled clinical visits is both an area garnering much interest as well as value as hospitals are looking to both provided appropriate and timely services while being more budget conscious. While there is a good amount of literature on the topic, few have looked at the volume of variable considered in this manuscript. Similarly, the attempt to use these tools to better predict visits across numerous hospitals across a geographical area, as opposed to simply one health care system (ie many hospitals all under one operating Health System) appears to be a unique approach. Moreover, the fact that temporal periodicity is used to predict in the near future (ie 2 days) as opposed to the

immediate fates (ie hours) or distant/months to years down the road lends to potentially good usability of the concepts and potentially would increase interest.

There are unfortunately some key areas that need some refinement. While there are a number of small typing corrections needed the content itself is difficult to read. I am making the assumption that this is most likely due to English not being the authors primary language. I am certain that it is much better than most of the readers Italian, (including mine) it would probably be worth refinement and reworking with special attention to this, if the primary audience is expected to be English speaking. The concepts in the paper are complex, especially as they explain the methodology around linear regression and model building and unfortunately the language flow makes following those concepts more difficult as well.
with regards to the methodology, the authors and data scientists do a good job outlining how the use of linear regression was used to develop a near future predictive model. This has long been a standard model employed to analyze groups of data in an attempt to develop predictive modeling. With the advent of machine learning models that can take into account dozens or even hundreds of variable, with subsequently hundreds of thousands or even millions of discrete data points, I do worry that the impact of the work will be more limited despite it being very well done and absolutely impactful for today's audience.

In the discussion the authors start to point to real world applications of the model and how it may be used and implemented. Given it temporal nature, there is, in fact, good potential use of resource allocation as the model indicates alteration in visits for the specific hospitals. It is very impressive that the model is able to predict so well the "Red" days and the discussion the authors outline around why and which variable most well correlate, it well done. What is hinted at in the first part of the discussion is the implementation of the model/tool though this is not specifically addresses with regards to what actions were taken once the understanding of whether the day was predicted to be Green vs Red, etc. A future work could discuss this implementation as the use of the model in daily operations, I believe, would be valuable to many readers as well. Moreover, there is a couple of loose associations between visits and "overcrowding" though it is never explicitly delineated how one causes the other. This might be intuitive to many readers. However it is, in fact, possible to have very large volumes of arrival that do not create an overcrowding situation if hospital flow open and barriers to admission and discharges are low or non-existent.

There are a number of little grammatical errors that should be addressed, though I attribute many to the above language concerns.

Overall, this is a worthwhile subject with a nicely put together methodology. It will surely be work that will be taken by others in the field and built upon. If the concerns mentioned could be addressed, should certainly be considered for publication.

Statistical Reviewer Comments to Author:

VERSION 2 – REVIEW

REVIEWER	Strehlow, Matthew Stanford, Emergency Medicine
REVIEW RETURNED	05-Jan-2022

GENERAL COMMENTS	Overall: Excellent job responding to the comments from prior reviews on discussion and conclusions, statistical review withstanding.

	Minor Comments: 1. The impact of hospital bed capacity is not addressed. This may have a direct effect on ED crowding and utilization but to my knowledge is not well established. In future, iterations, inclusion of hospital bed capacity in the analysis would be an interesting addition. One, it likely impacts ED utilization in many areas of the world as people are shunted to the ED if the hospital is full and people cannot be directly admitted. Two, any response to predicted ED volume increases must include hospital based solutions such as rescheduling of elective surgeries/procedures, early discharges, and other system wide solutions. 2. The inclusion of the Lombardy Region data and into the discussion is excellent. ED based responses to increases or decreases in demand will occur at the extremes of demand assuming hospital bed capacity is not highly limited. As an ED administrator the highest variable cost is staffing. Targeting the extremes allows for potential changes in staffing which could have significant impacts on costs and human resource limitations. 3. Under strengths and limitations of the study, bullet 5 (the final bullet) adds only a little. This might be better placed in the limitations/discussion section. Thank you for your work.

REVIEWER	Duwalage, Kalpani Ishara
REVIEW RETURNED	08-Mar-2022

GENERAL COMMENTS	I am satisfied with the revised version of the manuscript and the way author has addressed the comments. I have one minor revision. Figure 2 in supplementary material 2 needs to be revised. The forecast vs observed (actual) graph needs to produced along with time. That means, x axis of the graph should be forecast interval, and Y axis is the observed and forecasted values. You can represent actual and forecast values using two different colours. This way, the figure allows you to directly compare how close the forecasts to actual values, for instance, you can check whether how well the model capture the peaks and lows in the actual data in different time periods. But, the current graph doesn't provide such comparison.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Dr. Matthew Strehlow, Stanford

Comments to the Author:

Overall:

Excellent job responding to the comments from prior reviews on discussion and conclusions, statistical review withstanding.

Minor Comments:

1. The impact of hospital bed capacity is not addressed. This may have a direct effect on ED crowding and utilization but to my knowledge is not well established. In future, iterations, inclusion of hospital bed capacity in the analysis would be an interesting addition. One, it likely impacts ED utilization in many areas of the world as people are shunted to the ED if the hospital is full and people cannot be directly admitted. Two, any response to predicted ED volume increases must include hospital based solutions such as rescheduling of elective surgeries/procedures, early discharges, and other system wide solutions.

Answer: In the discussion paragraph we included a sentence of the dual nature of our high-demands forecasting project as of a proposal for a more reasoned choice of the hospital to which request assistance and an help for the evaluation of the available beds and of the staff needed to accommodate these expected visits, considering these issues as fundamental ingredients that shoul be considered in the future.

2. The inclusion of the Lombardy Region data and into the discussion is excellent. ED based responses to increases or decreases in demand will occur at the extremes of demand assuming hospital bed capacity is not highly limited. As an ED administrator the highest variable cost is staffing. Targeting the extremes allows for potential changes in staffing which could have significant impacts on costs and human resource limitations.

3. Under strengths and limitations of the study, bullet 5 (the final bullet) adds only a little. This might be better placed in the limitations/discussion section.

Answer: We removed the last bullet point and included it in the discussion paragraph.

Thank you for your work.

Reviewer: 3

Kalpani Ishara Duwalage

Comments to the Author:

I am satisfied with the revised version of the manuscript and the way author has addressed the comments. I have one minor revision.

Figure 2 in supplementary material 2 needs to be revised. The forecast vs observed (actual) graph needs to produced along with time. That means, x axis of the graph should be forecast interval, and Y axis is the observed and forecasted values. You can represent actual and forecast values using two different colours. This way, the figure allows you to directly compare how close the forecasts to actual values, for instance, you can check whether how well the model capture the peaks and lows in the actual data in different time periods. But, the current graph doesn't provide such comparison.

Answer: We modified supplementary figure 2 as requested.

Reviewer: 1

Competing interests of Reviewer: I have no competing interests to declare.

Reviewer: 3

Competing interests of Reviewer: None