British Journal of Applied Science & Technology
18(4): 1-8, 2016; Article no.BJAST.30987
ISSN: 2231-0843, NLM ID: 101664541

SCIENCEDOMAIN *international*
www.sciencedomain.org

A Study of Input Variable Selection to Artificial Neural Network for Predicting Hospital Inpatient Flows

Samira Rasouli[1], Hamed Tabesh[2] and Kobra Etminani[2*]

[1]Student Research Committee, Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
[2]Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.

Authors' contributions

This work was carried out in collaboration between all authors. Author SR designed the study, performed the statistical and time series analysis, wrote the protocol and wrote the first draft of the manuscript. Authors HT and KE managed the analyses of the study. Author KE managed the literature searches. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJAST/2016/30987
Editor(s):
(1) Qing-Wen Wang, Department of Mathematics, Shanghai University, P.R. China.
Reviewers:
(1) M. H. Burton, Rhodes University, South Africa.
(2) Rebecca Fein, Laboratory Informatics Institute, USA.
Complete Peer review History: http://www.sciencedomain.org/review-history/17574

Original Research Article

Received 13th December 2016
Accepted 3rd January 2017
Published 21st January 2017

ABSTRACT

Aims: Analysis of hospital processes has become a necessity to improve its performance over time through developing improved policies and decisions. In this context, the knowledge gained from an accurate prediction of patient flows would provide valuable information for strategic planning. This study aims at exploring and evaluating the predictability of time-series artificial neural network (ANN) approach for hospital inpatient flow forecasting using the partial autocorrelation function (PACF) of time-series data to identify the relevant time lags of the series as ANN inputs.
Methodology: We collected retrospective data of the number of monthly inpatient flows from 2004 to 2015 of four hospitals. We evaluated the application of the ANN model that uses extracted PACFs from time series data to determine appropriate inputs for ANN. This approach was compared with the neural network auto-regression which automatically selects relevant lags. The

**Corresponding author: E-mail: Etminanik@mums.ac.ir;*

performance of the ANN models was measured based on the mean absolute percentage error (MAPE) accuracy measure.
Results: We used the ANN models for predicting monthly inpatient flows for first three months of 2016. The post sample analysis revealed that the ANN model using selected input variables based on the PACF analysis offered improvements in monthly inpatient flows predictions than neural network auto-regression. Totally, for all four hospitals, the integrated model of PACF-ANN had a MAPE ranging from 2.91% to 6.67%, indicating an accurate prediction.
Conclusion: The ANN model with inputs extracted from the PACF analysis performs well for estimation of hospital inpatient flows. According to the unique characteristics of different hospitals, performance of the ANN model can vary from hospital to hospital. However, the proposed method of selecting input variables for the ANN model in this study may assist other hospitals and emergency departments for forecasting purposes.

Keywords: Forecasting; inpatient flows; hospital; time series; ANN; input variable.

1. INTRODUCTION

Study and analysis of the health care systems have become a necessity to improve its performance over time as it must meet a number of often conflicting objectives such as providing better and more efficient patient care while minimizing the cost of health care and resources [1,2]. Hospital management as an important component of health care systems may face with numerous challenging tasks while achieving these goals [2-4]. Basically, the flow of patients is a determinant factor in a hospital system that affects the performance of healthcare delivery processes. In addition, the decision problems of a hospital are directly related to and affected by the month to month changes in patient flows [5].

The short-term forecasting of the patient flows is the fundamental input of short-term decision making and planning on hospital and laboratory equipment, staff resources, food and laundry service demands, and so on. Furthermore, the long-term forecasts of patient flows are vital to long-term planning decisions about resources and capital budgeting which it's positive gains, in the long run, will lead to a sequence of capital expenditure [5]. In this context, the knowledge gained from an accurate prediction of patient flows would provide valuable information for resource allocation and strategic planning and also has the potential to minimize patient care delays, increase the capacity of the already existing system and improve the overall quality of care [6].

Forecasting patient flows have been studied extensively for bed occupancy and admissions in the context of emergency medicine [7-13]. To the best knowledge of the authors, use of artificial neural network (ANN) time series analysis in health care has been limited and there is a research gap to study the predictability of ANN time series modeling for patient flow forecasting.

ANNs are more flexible when compared with classical time series methods, presenting the capabilities of non-parametric, data-driven approximation with no a priori assumption about the statistical distribution of the data [14]. In spite of these theoretical capabilities, ANNs have not been able to confirm their potential in forecasting competitions against classical statistical methods, such as ARIMA, time series regression or exponential smoothing [8,15]. Since, modeling neural networks has posed multiple challenges in specifying adequate network architectures of input, hidden and output nodes, and training parameters depending on the underlying structure of the time series data. Fortunately, the literature provides guidance in selecting the number of hidden neurons of an ANN [16]. While, the identification and selection of relevant input variables and time lags without domain knowledge remain a critical issue in modeling neural networks for time series prediction [17,18]. Therefore, variable selection can be applied to discard irrelevant time lags, resulting in simpler models that are easier to interpret and that usually offer better performances [19,20].

This study aims at exploring the predictability of ANN for inpatient flow forecasting in hospitals using the partial autocorrelation function (PACF) of time-series data to identify the relevant lags of the series as ANN inputs. This approach is compared with the neural network auto-regression method over four hospital inpatient series.

2. MATERIALS AND METHODS

2.1 Study Setting, Design and Data Collection

This was a cross-sectional study, conducted using data collected from four hospitals affiliated with the Mashhad University of Medical Sciences (MUMS). The four hospitals were chosen because they vary in the demographics of the communities they serve, in size and setting. Hospital 1 is a 918-bed, general hospital with an average of approximately 4517inpatients per month (mean 4517.3, standard deviation [SD] ± 417); Hospital 2 is a 837-bed, general hospital with an average of approximately 5121 inpatients per month (mean 5120.6, SD ± 385); Hospital 3 is a 136-bed, woman's hospital with an average of approximately 1073 inpatients per month (mean 1073.25, SD ± 149); and Hospital 4 is a 325-bed, trauma centre with an average of approximately 1353 inpatients per month (mean 1353.25, SD ± 120).

At each hospital, inpatient admissions were recorded using a hospital information system which was archived in the MUMS statistical office database. Aggregated monthly flow of inpatients presenting for service at each of the four hospitals for the period January 1, 2004, to March 31, 2016, were included in this analysis.

2.2 Study Protocol

In order to evaluate the model, the data were split into two sets: one used to develop the model (training set) and the other to assess the model performance (test set). Data from January 1, 2004, to December 31, 2015, were modeled and the last three-month of inpatient flows were held out for the model validation process to empirically evaluate the post-sample forecast accuracy. Forecast accuracy was assessed at horizons ranging from 1 to 3 months in advance. Data from each hospital were analyzed as individual time series and evaluated separately. The time series plots of four hospitals are presented in (Fig. 1).

In this paper, all reported experiments were written in the R environment, an open-source tool for statistical computing, data analysis and graphics [21].

2.3 Error Estimation Method

In this study, the mean absolute percentage error (MAPE) was used for error estimation for each horizon (1–3 months in advance). It can be estimated by the following equation:

$$MAPE = \frac{100}{N}\sum_{j=1}^{N}\left|(target_j - output_j)/target_j\right|$$

Where N is the size of the test set, *target* $_j$ and *output* $_j$ are the actual and forecasted value, respectively. The MAPE expresses prediction errors as a percentage and is a scale-independent statistic which is easier and understandable for examining forecast accuracy. It is desirable that for a good forecast the obtained MAPE should be as small as possible.

2.4 Artificial Neural Networks

ANNs were mainly biologically motivated and designed to mimic the architecture of a human brain into a machine. The excellent feature of time series ANNs is their inherent capability of modeling complex nonlinear relationships between inputs and outputs, which makes this model more practical and accurate in modeling complex data patterns [14]. The multi-layer perceptron (MLP) is the most common and popular among ANNs which use a single hidden layer feed-forward network. (Fig. 2) exhibits an example of the MLP used for inpatient flows forecasting. The input nodes are the inpatient flow time series values at some particular lags while the output is the number of inpatient flows for the current time period. For example, if input nodes are the lagged values at time t-1, t-2 and t-12, the value at time t will be forecasted using the values at lags 1, 2 and 12.

2.5 Neural Network Auto-regression

Firstly, we used the neural network auto-regression algorithm function ''nnetar()'' from the 'forecast' package in R software created by Hyndman et al. [22]. The nnetar() function is a feed-forward neural network with a single hidden layer which the lagged inputs are selected automatically for forecasting univariate time series. The nnetar() function uses an optimal number of lags (p) according to the Akaike information criterion (AIC) for the autoregressive AR(p) model, therefore It is known as a neural network auto-regression. We applied this function using the first 12 years of data for each study site (2004–2016) as a training set for the model in order to predict the following three months (2016). Since there was variance in the results, we applied 50 runs to the selected model, and then the mean value of results was estimated.

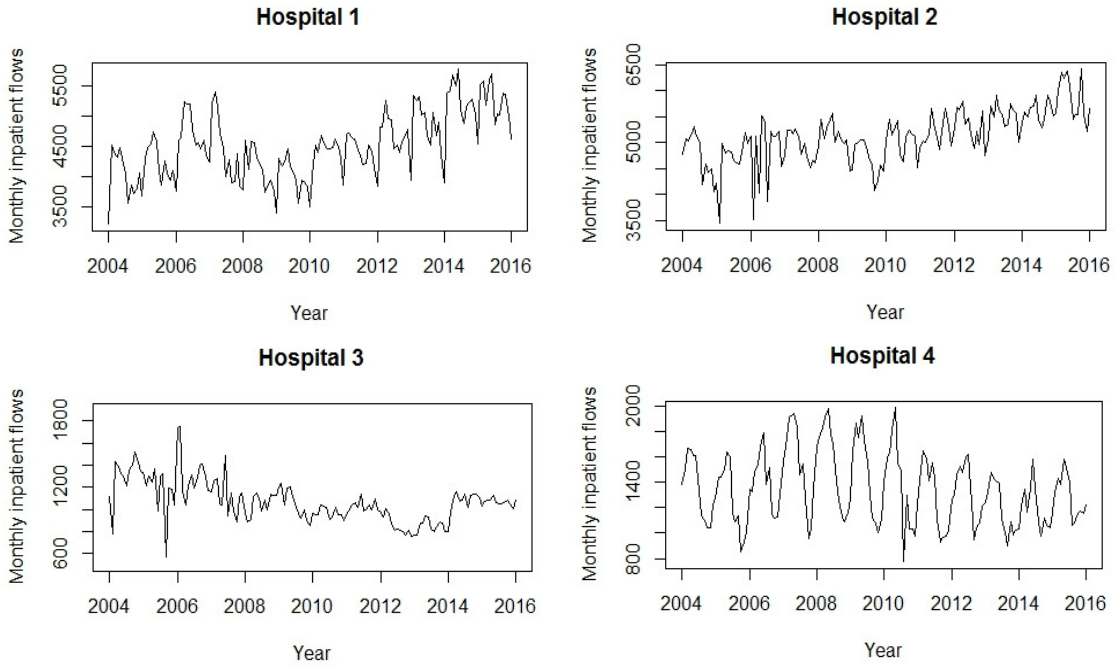

Fig. 1. Time plots for monthly inpatient flows for the period January 1, 2004, to December 31, 2015, at hospitals 1–4

2.6 Input Variable Selection for Multilayer Perceptron Model

Input variables usually used to create a set of training examples from the series. A small number of time lags will provide insufficient information and limit learning capabilities, whereas using a high number of them will increase the probability of having irrelevant inputs which may lead to over fitting. In this study, we held out the last three months of the training set for validation purpose that is known as the validation set. Thus, all of the available series were partitioned into three parts.

As explained by statistical Box and Jenkins methodology for AR process, the PACF plot can be used to determine to what extent and magnitude the past values of series are related to future values [23,24]. The PACF plot helps in identifying the maximum order of an AR process. As a result, we used an analysis of the AR from PACF-analysis to identify the relevant lags of the inpatient flows series. We have considered the PACF values up to lag 36 with 95% confidence level to determine the relevant time lags for each time series (2004-2016). Then statistically significant lags were selected. A better generalization, due to the reduced input time lags, is achieved if only relevant time lags are fed into the models [20]. In the proposed algorithm, in order to select the minimum number of time lags, adequate for representing the series, the training set was used to construct the MLP models based on different combinations of selected lags and the validation set was used to evaluate all models through MAPE accuracy measure. For example for lags {1,2,12} we have seven possible combinations including {{1}, {2}, {3}, {1, 2}, {1,12}, {2,12}, {1,2,12}}, note that the order is important and no higher time lags come before lower one. In some cases that the number of significant lags of PACF may be very high, up to first 6 significant lags were used as input variables to reduce the cost associated with the model training. Finally, a combination of lags that had the best in-sample performance for each time series was used for post-sample forecasting (Algorithm 1). We used the rminer library [25], which facilitates the use of ANN in R. We adopted the default R parameters for MLP. Due to the existed variances in the results, the mean value of results was estimated after 50 runs of model.

Algorithm 1. Steps of the Time lag selection algorithm.

Input: *S*: training set, *T*: test set, *W*: {Selected significant lags using PACF of *S*}, *n*: (2^|*W*|)-1)

1. $MAPE_{MAX} \leftarrow$ infinite; $p \leftarrow \{ p_1, p_2,\ldots p_n \}$: (All combinations of W)
2. Split S into training (N) and validation sets (V)
3. **For** $p \leftarrow \{ p_1, p_2,\ldots p_n \}$
4. **do** $D \leftarrow$ create training data using N and p as input time lags
5. $M \leftarrow$ fit MLP model with D
6. $MAPE_k \leftarrow$ MAPE estimate of M in V
7. **If** $MAPE_k < MAPE_{MAX}$
8. **then** $MAPE_{MAX} \leftarrow MAPE_k$; $p_b \leftarrow p$
9. **End**
10. **End**
11. $D \leftarrow$ create training data using S and p_b as input time lags
12. $M_{Final} \leftarrow$ fit MLP model with D
13. $MAPE_{Final} \leftarrow$ MAPE estimate of M_{Final} in T
14. **Return** $MAPE_{Final}, p_b$

According to the significant PACFs (
Fig.), we can see that lagged variables {1, 2, 11, 12, 13, and 14} were significant for hospital 1, lagged variables {1, 2, 5, 12, 13, 23, 26 and 28} were significant for hospital 2, lagged variables {1, 2, 3, 5, 16, and 22} were significant for hospital 3 and lagged variables {1, 2, 3, 4, 5, 7, 8, 9, 11, and 13} were significant for hospital 4.

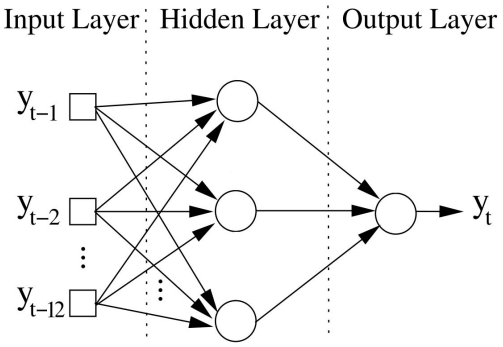

Fig. 2. Example of a multilayer perceptron

After implementing the proposed algorithm for each time series, for hospital 1, five lagged variables Y_{t-1}, Y_{t-2}, Y_{t-11}, Y_{t-12} and Y_{t-14} were selected as MLP inputs. In other words, inpatient flows in 1 month before (Y_{t-1}), 2 months before (Y_{t-2}), 11 months before (Y_{t-11}), 12 months before (Y_{t-12}) and 14 months before (Y_{t-14}) have been selected as MLP inputs where the inpatient flows in the current month (Y_t) is the output. For hospital 2, 3 and 4, lag variables {Y_{t-12}, Y_{t-23}}, {Y_{t-5}, Y_{t-16}}, {Y_{t-3}, Y_{t-7}, Y_{t-9}} were selected, respectively.

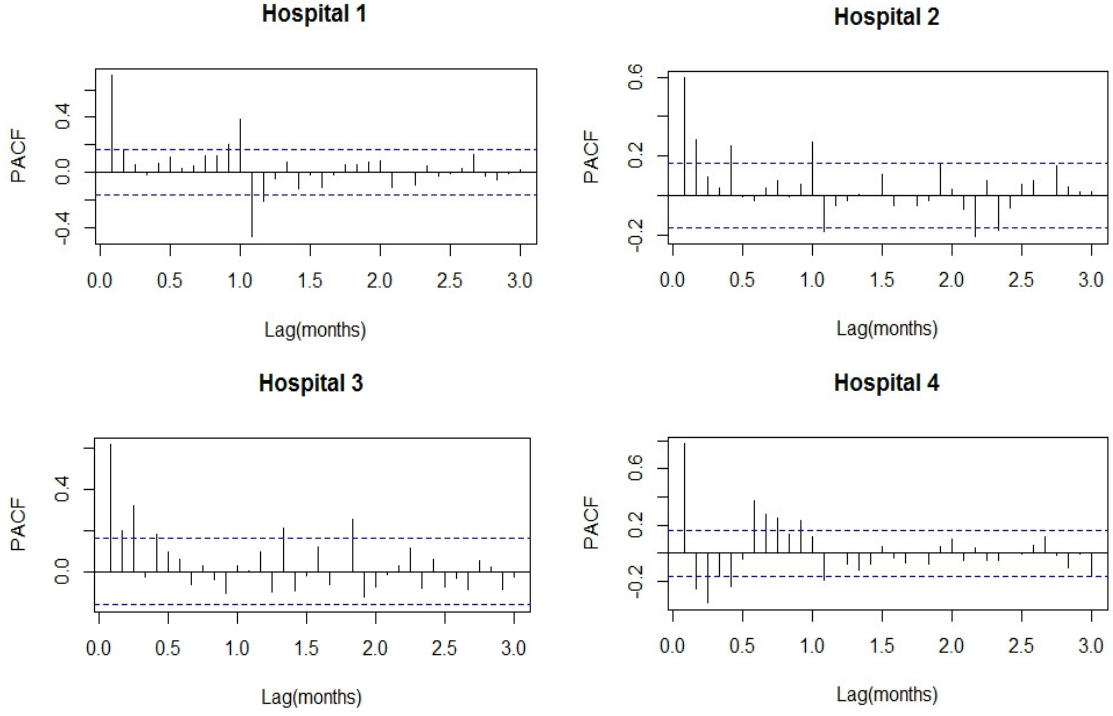

Fig. 3. The partial autocorrelation function of inpatient flows for the period January 1, 2004, to December 31, 2015, at hospitals 1–4

3. RESULTS AND DISCUSSION

In this study, the MLP model is used for estimation of monthly inpatient flows. In order to provide better insight about the PACF-MLP performances, at first we had conducted network auto-regression using nnetar() function which select lagged inputs automatically for forecasting univariate time series. The results of this model are presented in Table 1. After implementing the proposed time lag selection algorithm, we achieved improvements in monthly inpatient flows predictions than neural network auto-regression predictions (Table 2). We have calculated the in-sample and post-sample MAPE for each hospital. It was expected to have lower in-sample error than the post-sample. In spite of better in sample error of the neural network auto-regression than PACF-MLP model for hospital 2, 3 and 4, the PACF-MLP model outperformed neural network auto-regression due to lower post sample error.

The PACF-MLP represented satisfactory results in terms of MAPE for test data. This model has indicated a very consistent in-sample and post-sample error for Hospital 1 and 2, thus provided better predictions than the two other hospitals.

Table 1. Forecasting accuracy measurement of monthly inpatient flows with nnetar() function

Study site	nnetar() MAPE	
	In-sample error	Post-sample error
Hospital 1	4.53	4.75
Hospital 2	0.01	5.95
Hospital 3	0.35	15.28
Hospital 4	1.57	6.1

**MAPE = Mean absolute percentage error*

The primary outcome of interest for this study was the performance of MLP method in terms of post-sample forecast accuracy measure. We generated true post-sample forecasts, reflecting the differences between the numbers of predicted inpatient flows by the model and the actual observed values for first three months of 2016 (Fig. 4). The forecasts provided by the model are shown to compare very well with the actual observations.

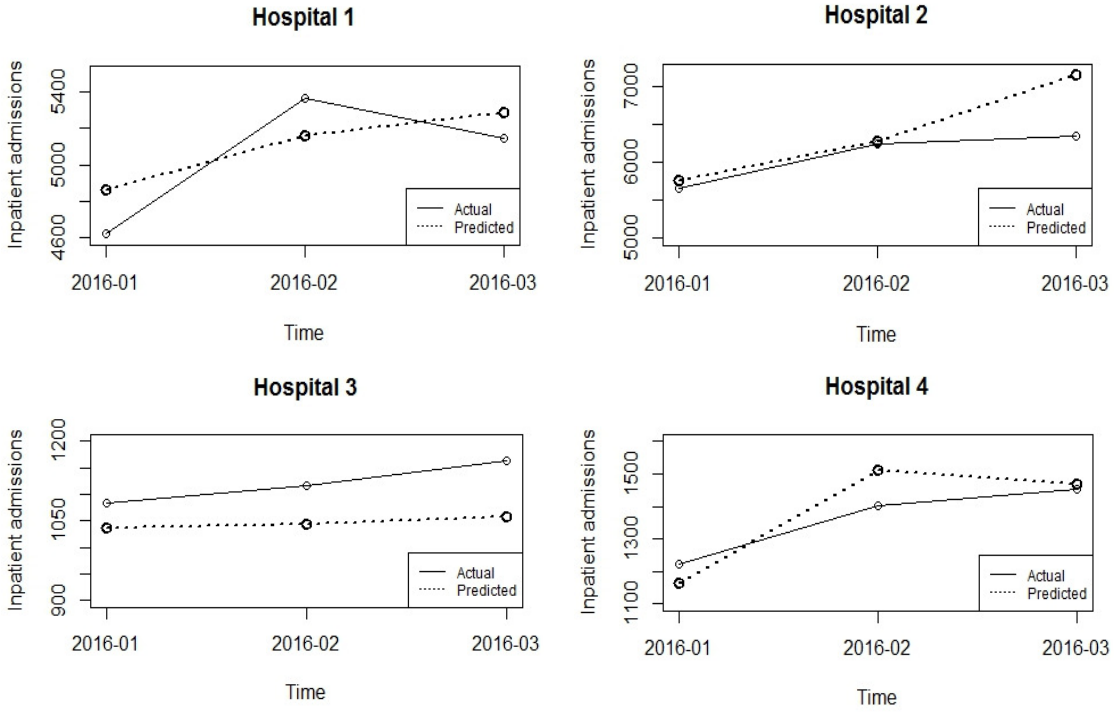

Fig. 4. A number of inpatients forecasted per month at hospitals 1–4

The ANN is a state-of-the-art approach to time series forecasting and has been applied to a wide variety of forecasting tasks [26]. Despite several studies that conducted time series techniques to analyze patient flows in the emergency departments, a few of them applied ANN methods for forecasting purposes [8,27]. Also, we found few studies about forecasting patient flows in hospitals. Our study fills a gap in the literature by considering ANN time series model for predicting hospital inpatient flows and demonstrates how to select the ANN time series input variables for short-term forecasting of such series.

Table 2. Forecasting accuracy measurement of monthly inpatient flows with PACF-MLP

Study site	PACF-MLP MAPE	
	In-sample error	Post-sample error
Hospital 1	2.67	2.91
Hospital 2	3.95	4.98
Hospital 3	1.17	6.67
Hospital 4	1.29	5.44

**MAPE = Mean absolute percentage error*
**MLP = Multilayer perceptron*
**PACF = Partial autocorrelation function*

Hospital administrators may use predictions of monthly inpatient flows for adjusting staffing levels and resource managing. However, there is limited research about these applications [28]. But in a similar study for predicting the number of patients sourced from the emergency department, the results of this study was used as a tool to help hospital management for purposes such as surgery scheduling problem and bed management. This study claims that this tool could have a considerable utility in health service planning and bed management [10].

As an important forecasting strategy, the input variable selection must update as more data become available. We would expect that better forecasts will be obtained if more recent data are available to the model [29]. In the future, we would apply multiple time series forecasting methods and compare their performances with the ANN model using proposed input variable selection.

4. CONCLUSION

Time series analysis is an important tool to support strategic decisions. We used the ANN model which is a more powerful and flexible algorithm. The performance of this algorithm depends on a correct setting of inputs (i.e. time lags used to build the training examples). As presented in this study, according to the unique characteristics of different hospitals, performance of the model can vary from hospital to hospital. However, the proposed method in this study for selecting the appropriate ANN input variables may assist other hospitals and emergency departments for forecasting purposes.

ETHICAL APPROVAL

This study only employed the summed numbers of inpatient admissions for each month, and no personal patient information was used. This study was conducted after obtaining the confirmation of the MUMS Ethics Committee for this project, No. 950080.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Ivatts S, Millard P. Health care modelling: Opening the 'black box'. British Journal of Health Care Management. 2002;8(7):251-255.
2. Bhattacharjee P, Ray PK. Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: A review and reflections. Computers & Industrial Engineering. 2014;78:299-312.
3. Brailsford S, Vissers J. OR in healthcare: A European perspective. European Journal of Operational Research. 2011;212(2): 223-234.
4. Abraham G, Byrnes GB, Bain CA. Short-term forecasting of emergency inpatient flow. Information Technology in Biomedicine, IEEE Transactions. 2009; 13(3):380-388.
5. Lin WT. Modeling and forecasting hospital patient movements: Univariate and multiple time series approaches. International Journal of Forecasting. 1989; 5(2):195-208.
6. System I.o.M.C.o.t.F.o.E.C.i.t.U.H. Hospital-based emergency care: At the breaking point. Washington, DC: National Academies Press; 2006.

7. Bergs J, Heerinckx P, Verelst S. Knowing what to expect, forecasting monthly emergency department visits: A time-series analysis. International Emergency Nursing. 2014;22(2):112-115.
8. Jones SS, et al. Forecasting daily patient volumes in the emergency department. Academic Emergency Medicine. 2008; 15(2):159-170.
9. Aboagye-Sarfo P, et al. A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in Western Australia. Journal of Biomedical Informatics. 2015;57:62-73.
10. Boyle J. PAPT—patient admissions prediction tool. Healthcare IT Manage. 2009;4:30-32.
11. Boyle J, et al. Predicting emergency department admissions. Emergency Medicine Journal. 2012;29(5):358-365.
12. Littig SJ, Isken MW. Short term hospital occupancy prediction. Health Care Management Science. 2007;10(1):47-66.
13. Zhu T, et al. Time series approaches for forecasting the number of hospital daily discharged inpatients.
14. Kamruzzaman J, Sarker RA, Begg R. Artificial neural networks: Applications in finance. Artificial Neural Networks in Finance and Manufacturing. 2006;1.
15. Makridakis S, Hibon M. The M3-Competition: Results, conclusions and implications. International Journal of Forecasting. 2000;16(4):451-476.
16. Sheela KG, Deepa S. Review on methods to fix number of hidden neurons in neural networks. Mathematical Problems in Engineering. 2013;2013.
17. Zhang G, Patuwo BE, Hu MY. Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting. 1998;14(1):35-62.
18. Curry B, Morgan PH. Model selection in neural networks: Some difficulties. European Journal of Operational Research. 2006;170(2):567-577.
19. Cortez P, Rocha M, Neves J. Time series forecasting by evolutionary neural networks. Artificial Neural Networks in Real-Life Applications. 2006;47-70.
20. He W, Wang Z, Jiang H. Model optimizing and feature selecting for support vector regression in time series forecasting. Neurocomputing. 2008;72(1):600-611.
21. Team RC. R: A language and environment for statistical computing. 2013.
22. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. OTexts; 2014. Available:https://www.otexts.org/fpp/9/3
23. Box GE, et al. Time series analysis: Forecasting and control. John Wiley & Sons. 2015;64-87.
24. Adhikari R, Agrawal R. An introductory study on time series modeling and forecasting. arXiv preprint arXiv:1302.6613; 2013.
25. Cortez P. Data mining with neural networks and support vector machines using the R/rminer tool. in Industrial Conference on Data Mining. Springer; 2010.
26. Zhang G. In GP Zhang & Pa. Hershey (Eds.), Business forecasting with artificial neural networks: An overview, neural networks in business forecasting. Idea Group Publishing; 2004.
27. Srikanth K, Arivazhagan D. Prediction model to enhance resource efficiently for hospitals.
28. Wargon M, et al. A systematic review of models for forecasting the number of emergency department visits. Emergency Medicine Journal. 2009;26(6):395-399.
29. Milner P. Forecasting the demand on accident and emergency departments in health districts in the Trent region. Statistics in Medicine. 1988;7(10):1061-1072.

© 2016 Rasouli et al.; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/17574