Operations Research Perspectives 8 (2021) 100208

Contents lists available at ScienceDirect

Operations Research Perspectives

journal homepage: www.elsevier.com/locate/orp

Demand analysis and capacity management for hospital emergencies using advanced forecasting models and stochastic simulation

Oscar Barros *, Richard Weber, Carlos Reveco

Department of Industrial Engineering, University of Chile, Santiago, Chile

ARTICLE INFO

Keywords:
Health care management
Emergency capacity management
Forecasting models
Process design
Simulation

ABSTRACT

Demand forecasting and capacity management are complicated tasks for emergency healthcare services due to the uncertainty, complex relationships, and high public exposure involved. Published research does not show integrated solutions to these tasks. Thus, the objective of this paper is to present results from three hospitals that show the feasibility of routinely applying integrated forecasting and capacity management with advanced operations research tools.

After testing several forecasting methods, neural networks and support vector regression provided the best results in terms of variance and accuracy. Based on this forecasting, a logic for managing hospital capacity was designed and implemented. This logic includes the comparison between the forecasted demand and the available medical resources and a stochastic simulation model to assess the performance of different configurations of facilities and resources. The logic also provides hospital managers with a decision tool for determining the number and distribution of medical resources on emergency services based on a cost/benefit analysis of resources and service improvement. Such results support the task of assigning doctors to different kinds of boxes, defining their work schedules, and considering additional doctors. The contribution of this paper consists of an integrated solution designed to implement the abovementioned logic. This solution combines forecasting, simulation for capacity management, process design, and IT support, facilitating the practical routine use of complex models. The integration explicitly considers a solution that also has adaptation capabilities to facilitate use under changing conditions.

The solution is also general and admits adaptation and extension to other services. Thus, we have already performed similar work for ambulatory and surgical services.

1. Introduction

Capacity management is a challenging task for emergency healthcare services due to demand uncertainty, necessary and complex health interventions, and patient risk [2]. A strategic issue for such management is to generate the best possible service, with the right capacity and determination of resources to assure a requisite service level. The value required for patients is to provide improved service in terms of waiting time and to provide the appropriate service, according to the risk involved. For such purposes, there is a need for predictive and resource analysis tools to ensure the proper level of service at the least cost [3]. These tools should be part of an integrated and redesigned process configuration, supported by Information Technologies, for emergency services, which reduces waiting time and improves medical service quality. Thus, this paper presents the development and implementation of integrated predictive and capacity analysis models and processes for emergency departments (EDs) in three hospitals in Chile. Given the support from hospital authorities and administration in one of these institutions, we could even implement our proposed solution.

In general, public hospitals in Chile have a capacity that is not enough to process demand for health services and assure patients a reasonable length of wait (LOW), especially for EDs. Thus, overcrowding is frequent, waiting time before attention usually exceeds an hour, and rejection of service also occurs. Hence, hospitals need to forecast demand with precision to adjust their capacity or take alternative courses of action, e.g., transferring patients to other facilities. To manage capacity, demand forecasts should predict not only on an aggregated level but also for different medical services, which renders the task technically more demanding. The forecasted demand for each service allows the determination of the required resources, such as doctors of different

* Corresponding author at: Republica 701, Santiago, Chile.
E-mail address: obarros@dii.uchile.cl (O. Barros).

https://doi.org/10.1016/j.orp.2021.100208
Received 13 August 2021; Received in revised form 9 November 2021; Accepted 9 November 2021
Available online 12 November 2021

specialties, reception areas, emergency room cubicle capacity, and operating room capacity. Comparing the resources needed to satisfy demand with available capacity allows decisions to adjust such capacity or prevent or transfer demand. Public hospitals in Chile, which process 75% of the country's demand for health, did not use any formal method to forecast demand and manage capacity when this work began. The procedures utilized were informal and based on the experience of the participants in the process; furthermore, such procedures were mainly oriented to solving the problem of excess demand.

Given the situation outlined, which motivated this paper, an agreement was reached with the Chilean Health Authority to perform an applied research program that would use state-of-the-art analytical tools, process design methodologies, and IT support to develop a general solution for demand forecasting, capacity management, and associated processes that could eventually be implemented in all Chilean public hospitals. Objectives were a smaller LOW and significantly improved service, including better use of resources for hospitals using forecasted demand to support their planning. The research started with three hospitals, for which an evaluation of the current situation of demand analysis and capacity management was made to determine the feasibility of introducing analytical tools and formal practices to improve the respective processes. The results presented in this paper comprise the product of collaborative work with the hospital staff. The hospitals' professionals reviewed each of the steps described here, which led to working processes for forecasting and managing demand by capacity adjustments.

Section 2 of this paper reviews the literature on the use of analytical methods in forecasting and hospital capacity management. Section 3 presents the methods proposed to perform demand forecasting and to convert the forecasts to the resources needed to satisfy demand, allowing capacity management. Section 4 gives the results from the application of the proposed methods, and Section 5 presents the conclusions and a discussion of the results.

2. Literature review

2.1. Demand forecasting

Demand forecasting is a well-studied subject, as presented in [4], that has generated important results in different areas, such as the retail industry [5] and inventory control at several enterprises, such as Dell [6]. Forecasts provide relevant information for making decisions about the resources needed to provide an adequate service to meet the potential demand and to avoid stock breakdowns or overstocking. A good example of this use is Walmart, which forecasts the demand for each of its sales points to feed models that determine the best actions over the supply chain to assure product availability at minimum cost [7].

In the case of hospital ED services, the capacity depends on available physical facilities, such as medical cubicles, operating rooms, and beds; also important are human resources, such as doctors, who perform diagnostic procedures and treatments on patients. The capacity should be planned to guarantee a given service level and improve the use of resources; for this planning, an accurate forecast of the number and type of patients who will arrive is necessary.

Publications about the formal demand forecasting of health services are reviewed subsequently. Many different proposals for forecasting exist, e.g., [4,5,8], and several studies compare such methods in terms of the accuracy of the results. One of these studies compares neural networks with traditional econometric methods and concludes that the former generally yields better results [9]. A good summary of cases using traditional methods is provided in [10], where the reported methods, such as autoregression, moving averages, and exponential smoothing, are simple. Other cases, which use more sophisticated methods, have focused on the prediction of the number of beds required to meet emergency demand [11–13]. A case that uses time series analysis for forecasting arrivals and occupancy levels in an ED is the case presented in [14]. A recent paper [15] presents a complete review of forecasting in an ED. The paper considers 102 articles and provides a good summary of the forecasted issues and the applied methods. The paper concludes that ED patient demand is the most popular issue with 43 papers, that the most employed method is times series and that hybrid approaches, including data mining, are the least employed methods, with just two papers for patient demand forecasting.

A similar result is reported in a literature review on the use of big data analytics in healthcare [16]. The authors reviewed a total of 804 articles that have been published between 2000 and 2016 and found that most papers use machine learning techniques to analyze clinical data, for health monitoring and prediction purposes.

One recent paper [17] on demand forecasting for an ED in a private hospital in Turkey uses linear regression, ARIMA, artificial neural networks, exponential smoothing, and hybrid methods such as ARIMA–ANN and ARIMA-LR. The forecasting performance of the methods was measured using the mean absolute percentage error, and the ARIMA–ANN hybrid model was shown to outperform the other methods in terms of forecasting accuracy.

Another important issue identified in the previously mentioned review [15] is Length of Stay (LOS) forecasting with 10 papers, which uses data mining in seven cases. This approach is an alternative to the simulation approach proposed in this paper for service level prediction. Of the papers reviewed, no paper used SVM in demand forecasting. However, a support vector regression (SVR) model was applied recently to the radiology department of a hospital in China [18] with low absolute percentage errors for the demand of emergency department patients.

Thus, the work reported in this paper intends to show that newer forecasting methods, such as SVR, have the potential to produce better results than traditional methods and neural networks and that it is possible to integrate forecasting and capacity analysis to determine ways that ED management can improve service. Such integration has not been explicitly considered in the forecasting literature, and only the possibility of using forecasts for capacity planning and staffing is suggested.

2.2. Capacity management

For capacity management, the usual practice has been to simulate the flow of patients through emergency facilities. In [10], this type of work is reviewed, and the cases evaluated mostly use discrete event simulation. The literature review in [19] confirms this result. Certain cases [10] include problems of capacity configuration and design, such as those considered in this work, e.g., using a fast track, resource modification, and staffing levels. Other simulation papers explicitly consider capacity design, perform capacity planning for an outpatient physical therapy service [20] and evaluate designs for outpatient ophthalmic services, defined using the technique of experimental design [21]. Some works that use the common approach of a static arrival probability distribution in ED simulations are those discussed in [22–25]. There are different simulation techniques, such as agent-based simulation (ABS), discrete event simulation (DES), system dynamics (SD), and Monte Carlo simulation. Discrete event simulation (DES) is selected in this work.

There is another line of a joint analysis of demand forecasts and capacity. One approach is to predict the number of clients who will demand a service to manage the capacity needed to provide a given level of service. For example, in [26], the proposal is a joint demand and capacity management for services in a restaurant, where the focus was optimizing revenue for a given dynamic demand. A similar study examined scheduling elective surgery under uncertainty [27] but did not consider a forecasted stochastic demand, which is the approach utilized in this paper.

A case study [1] presents a hybrid system combining forecasting and DES for endoscopy services. Official population projections are used together with historical demand data to forecast demand for the

respective diagnostic service. A more recent paper [28] integrates demand and capacity analysis in a case of "a hybrid application of discrete-event simulation and time-series forecasting across multiple centers in an urgent care network". This case is oriented to an operative, real-time solution to support patient flow decisions for several service units, in contrast to the structural modification to adapt capacity to demand, which is the focus of this paper.

A generalization of these ideas is "Hybrid Systems Modeling" [29], which considers the use of different OR/MS methods, such as forecasting, systems dynamics, agent-based systems, and game theory, with simulation [30]. In this modeling, the use of forecasting in simulation is defined as Case D, Hybrid System Model, of Hybrid Simulation, which is the focus of this paper.

While we are concerned about short-term forecasting and capacity management, alternative approaches use DES to study the effects of changes in the demographic structure which will significantly alter health needs for hospital treatment [31,32]. This is a long-range approach to plan future capacity, a problem different from the one this paper deals with.

Notice the tendency to use formal forecasting models, especially newer models and capacity management models, and the disintegration of such models. Additionally, no paper considers the processes necessary to routinely implement these models. Thus, an objective of this work is to show the need to integrate forecasting, capacity analysis, and process design to assure an adequate practical solution to support the capacity management of an ED.

A complementary approach is "to help patients in need of urgent care to make more informed decisions about available healthcare choices, thereby … reducing the wait time being experienced by the patients" [33].

3. Methods

This section evaluates forecast methods to predict demand and discusses capacity management using such predictions.

3.1. Demand forecasting

Four forecasting methods are evaluated: linear regression, weighted moving averages, neural networks, as suggested in [34,9], and support vector regression (SVR). The first two methods are well-known techniques used for forecasting that are described in the literature [4]. Neural networks and SVR are more recently employed techniques for forecasting that are summarized subsequently.

The particular type of network employed is the multilayer perceptron (MLP) [9]. Its basic units are neurons grouped into layers and connected using weighted links between two layers. Each neuron receives inputs from other neurons and generates a result that depends only on locally available information, which serves as an input to other neurons. The architecture of the network is shown in Fig. 1.

Each neuron operates according to the structure in Fig. 2, where the output y is determined as a function of the weighted inputs.

Function f in Fig. 2 is the activation function, which may take different forms; the most commonly employed function for continuous outputs is the logistic function, as shown in Eq. (1).

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

The network is trained with historical data, as described in Section 4.1. The basic idea is that previous data predict a given future month. In particular, the assumption is that the pattern is seasonal; therefore, values from previous years of the month to be predicted are inputs to the model. The structure of the network includes an output layer with one neuron that generates the desired forecast. The input layer contains the variables selected to explain the demand. In the hidden layer, the selection of neurons between input neurons and output neurons is such that it balances the following factors: a high number will tend to copy the data (overfitting), and a small number will not capture the pattern present in the data (underfitting).

The other method tested is support vector regression (SVR), as presented in [35–37], which is a variation of support vector machines [38] and based on the subsequent idea. SVR performs a linear regression in a high-dimensional feature space generated by a kernel function, as developed here, using the ε-insensitive loss function proposed in [39]. This function allows a tolerance degree to errors that are less than or equal to ε, as shown in Fig. 3. The subsequent presentation follows the applied structure and terminology [37].

SVR starts with a set of training data $\{(x_1, y_1), \ldots, (x_l, y_l)\}$, where each $x_i \in R^n$ denotes the input space of the sample and has a corresponding target value $y_i \in R$ for $i = 1, \ldots, l$; where l is the number of available data points to build the model. The SVR algorithm applies a

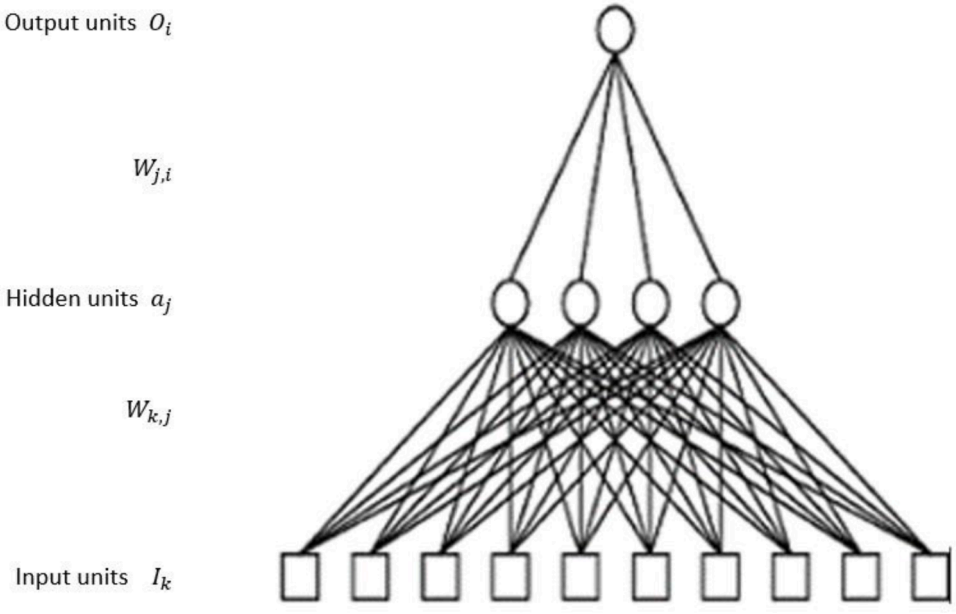

Fig. 1. Architecture of the Neural Network.

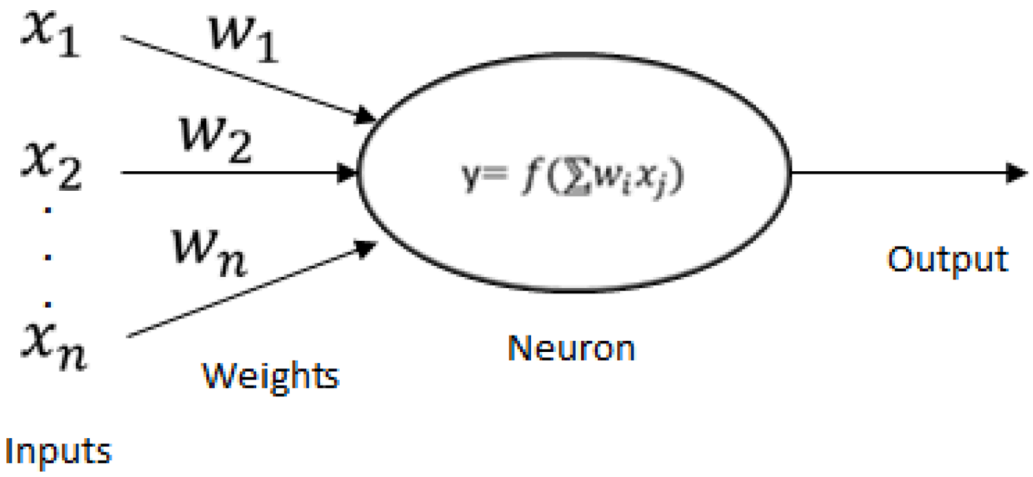

Fig. 2. Neuron Details.

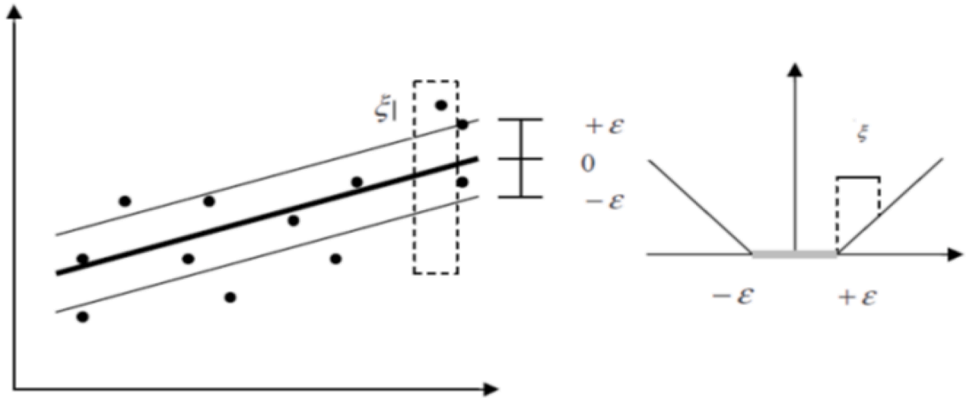

Fig. 3. SVR and ε-insensitive loss function to fit a tube with radius ε to the data and positive slack variables ξi [37].

function Φ that transforms the original data points from the initial input space (R^n) to a generally higher dimensional feature space ($F \subset R^m$). In this new space, a linear model f, which represents a nonlinear model in the original space, is constructed:

$$\Phi : R^n \rightarrow F \tag{2}$$

$$f(x) = \langle \omega, \Phi(x) \rangle + b \tag{3}$$

with $\omega \in R^m$ and $b \in R$ In Eq. (3), $\langle \cdot, \cdot \rangle$ denotes the dot product in R^m. When the identity function is employed, i.e., $\Phi(x) \rightarrow x$, no transformation is carried out, and linear SVR models are obtained.

When using the convex $\sum$-insensitive loss function, the goal is to obtain a function f that fits given training data with a deviation less than or equal to ε and is as flat as possible to reduce model complexity. Obtaining a flat function f means that a small weight vector ω is sought. One way to ensure this approach while maintaining a deviation less than or equal to ε is by minimizing the norm $\| \omega \|^2$ given a set of constraints [40], leading to the following convex optimization problem:

$$min \frac{1}{2} \| \omega \|^2 \tag{4}$$

$$s.t.$$
$$y_i - \langle \omega, \Phi(x) \rangle - b \leq \varepsilon$$
$$\langle \omega, \Phi(x) \rangle - y_i + b \leq \varepsilon$$

This problem could be infeasible. Therefore, slack variables ξ_I, ξ_i^* and $i = 1, \ldots, l$, are introduced to allow error levels greater than ε (refer to Fig. 3), arriving at the following convex formulation:

$$min \frac{1}{2} \| \varpi \|^2 + C \sum_{i=1}^{L} (\xi_i + \xi_i^*) \tag{5}$$

s.t.

$$y_i - \langle \omega, \Phi(x) \rangle - b \leq \varepsilon + \xi_i^*$$
$$\langle \omega, \Phi(x) \rangle - y_i + b \leq \varepsilon + \xi_i^*$$
$$\xi_i^*, \xi_i \geq 0$$

This problem is the primal problem of the SVR algorithm. The objective function considers two goals—the generalization ability and accuracy in the training set—and embodies the structural risk minimization principle [37]. Parameter $C > 0$ determines the trade-off between the generalization ability and the accuracy in the training data and the maximum value for which deviations larger than ε are tolerated. The ε-intensive loss function $|\xi|_\varepsilon$ is expressed in Eq. (6).

$$|\xi|_\varepsilon = \begin{cases} 0, & |\xi| < \varepsilon \\ |\xi| - \varepsilon, & |\xi| \geq \varepsilon \end{cases} \tag{6}$$

It is more convenient to represent the optimization problem (5) in its dual form [39]. For this purpose, a Lagrange function is constructed, and once applying saddle point conditions, the dual problem is converted to a quadratic optimization problem that is easier to solve [40] and that provides the estimation of f(x). The accuracy of the estimation depends on an appropriate set of parameters C and ε, among others [41]. Thus, the use of a grid search to obtain suitable parameters for SVR is

appropriate to test combinations of such parameters.

3.2. Capacity analysis and management

As stated in [12], having a demand forecast is not, by itself, a useful contribution to hospital management. Managers also need to know whether there is enough capacity to attend to forecasted demand with defined quality standards and how to rearrange or modify hospital resources to achieve this goal. The linkage between forecasting and resources provides managers with a quantitative basis for hospital capacity management, for which the proposal of this paper is subsequently presented.

With the demand and its behavior characterized, the next step is to determine whether the existing resources are sufficient to meet the forecasted demand. The measure selected to perform this comparison is medical hours per month. On the supply side, the total available medical hours per month is obtained by the simple multiplication of the number of doctors available by the number and length of shifts per doctor in that period. On the demand side, a simple method for calculating the medical hours required to meet the forecasted demand is considering a deterministic monthly prediction obtained from the model, divided uniformly within each month and distributed within each day according to the patients' arrival patterns.

In performing capacity management, two types of problems are considered. First, configuration management for determining how different designs of the hospital facilities may affect the quality of service, measured by the length of wait from the moment the patient arrives until service begins. This metric is justified because it depends on the design, since the time dedicated to medical services does not change because their methods are the same. The second problem is resource management, which decides how currently available resources should be assigned to increase the service level and which and where new resources are required to further improve the quality of service.

In the literature, several proposals for the configuration of emergency services exist [42]; they are summarized in Fig. 4. In the figure, option (b), which considers triage for patient pre-evaluation and a fast-track line for noncritical patients (C4), appears to be appropriate for the hospitals under consideration. The justification is that a line providing quick solutions with less qualified medical resources, e.g., medical students, is possible, which is good for hospitals that lack resources.

After deciding which configuration performs better on the emergency service, the next task is to determine the impact that a redistribution, reduction, or addition of medical resources would generate on the performance of the selected configuration. Based on the previous calculations, an estimate of the performance of the current distribution of resources, is static and deterministic. Thus, a simulation model that incorporates the stochastic behavior of the demand is necessary to provide a more dynamic and accurate evaluation of the emergency service performance with new resource distributions. The capacity analysis just outlined will be applied in Section 4.2.

4. Calculations and results

4.1. Forecasting

In forecasting, the first need is to obtain good historical data to develop the required models. Based on these data, demand forecasting proceeds as subsequently explained.

This work focuses on two public pediatric hospitals, HLCM and HEGC, and a general-purpose hospital, HSBA. On arrival at the emergency facilities of these hospitals, each patient registers, which generates a record of personal data, time of arrival, diagnosis, and a triage classification according to illness severity. The historical monthly data for the three hospitals obtained for this work are listed as follows: HLCM, nine years; HEGC, eight and a half years; and HSBA, ten years. Since these hospitals provided high-quality data regarding their emergency operation, historical demand is a good input to forecasting models. Monthly aggregated demand was the basis for constructing such predictive models. However, for data to be useful input information for forecasting models, further analyses and a series of transformations were necessary. In the analysis of the demand that arrives at the ED, outliers occur, as shown in Fig. 5. Visual inspection of aggregated demand, as shown in this figure for one of the hospitals, reveals a strong seasonal pattern. Low demand during the summer months (December, January, and February in the Southern Hemisphere) and a high influx of patients during the winter season (May, June, and July) occur. In general, a downward trend persists over the years.

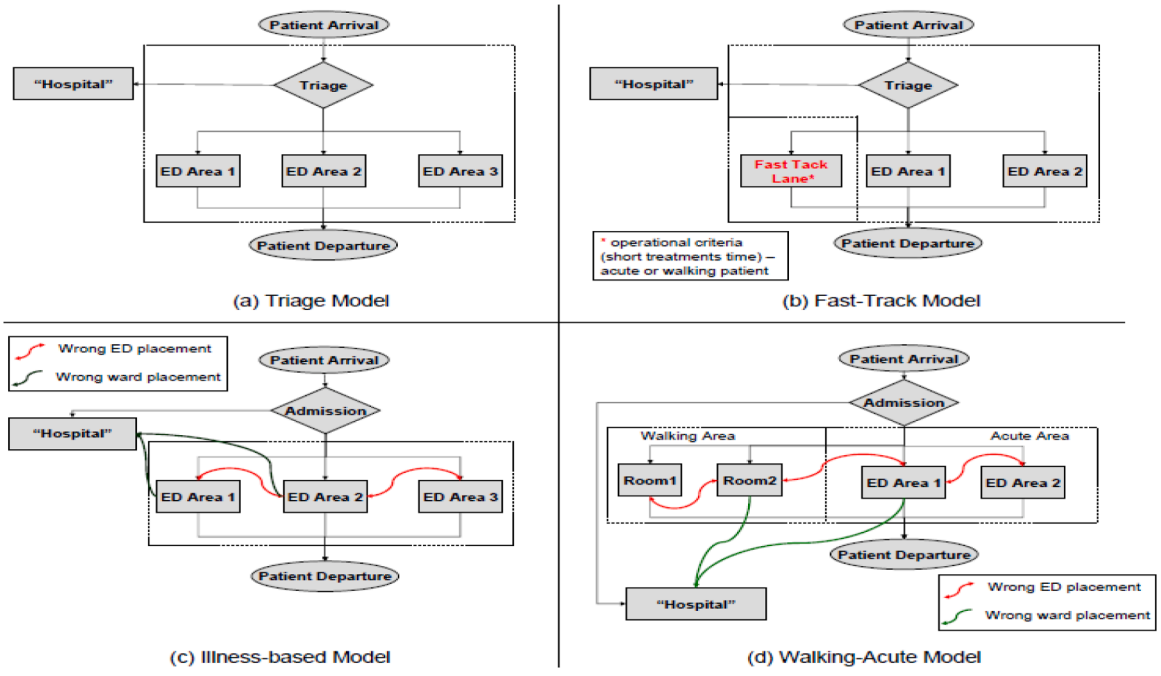

Fig. 4. Alternative configurations for emergency services.

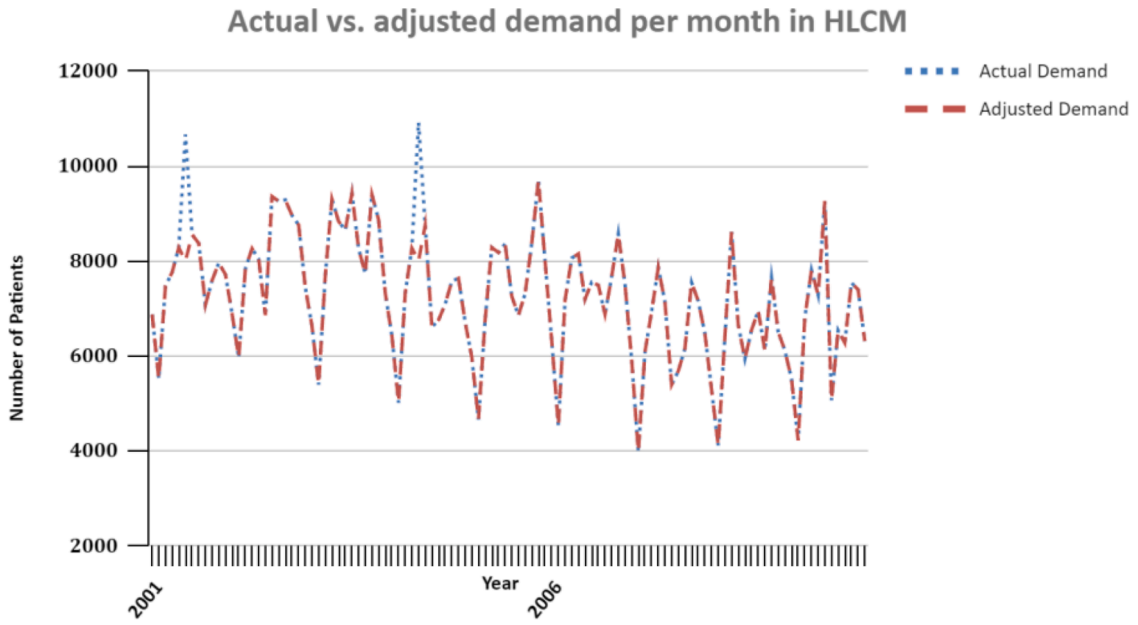

Fig. 5. Actual vs. adjusted demand per month in HLCM.

With data disaggregated by medical service type, there were important differences, as shown for medical and surgical demand in Figs. 6 and 7. The first difference is much more volatile over the years since it depends on factors such as temperature and flu-like illness rate, while the second difference is more stable over the years. Data also show that medical demand comprises 70% of emergency cases and that surgical demand corresponds to 30% of these cases.

Demands at HEGC and HSBA show behaviors that are very similar to the demand at HLCM.

In the development of the neural network forecasting models using the data just presented, the previous months' demand is the input. However, certain months are more relevant than others. A method for selecting relevant attributes is a genetic algorithm, as suggested in [43], but the results were not encouraging. In this work, a common pitfall of separating the dataset into just two groups, one group for training and one group for testing, was avoided, as discussed in [44], since this approach tries to minimize the error over the testing data, leading to overfitting of the resulting model. Thus, in this case, the applied method divides the data into three sets: 70% for training; 20% for testing, where the network is trained to minimize the test error; and the third set, in which 10% of the data is independently applied to validate the results. The use of an independent set provides a better evaluation of future results.

Regarding the determination of the architecture for the neural network model, several parameters were tested using the data, such as the number of epochs to use, the learning rate, and the number of hidden neurons. The best results obtained are for 10,000 training epochs, maintaining the model with minimum error in the training set and a learning rate of 0.2 with a momentum of 0.3. In addition, decaying was introduced, but this only helps to reach the solution faster with no significant changes in the results.

Based on the results given here, a neural network with 18 input neurons is employed. If N is the index of the forecasted month, the following neurons are obtained:

a) Three neurons that correspond to the values of the same month in previous years: N-12, N-24, and N-36.
b) Three neurons representing the tendency between two months given by the differences between N-12 and N-13, between N-24 and N-25, and between N-36 and N-37,
c) A set of 12 binary variables to represent the months of the year.

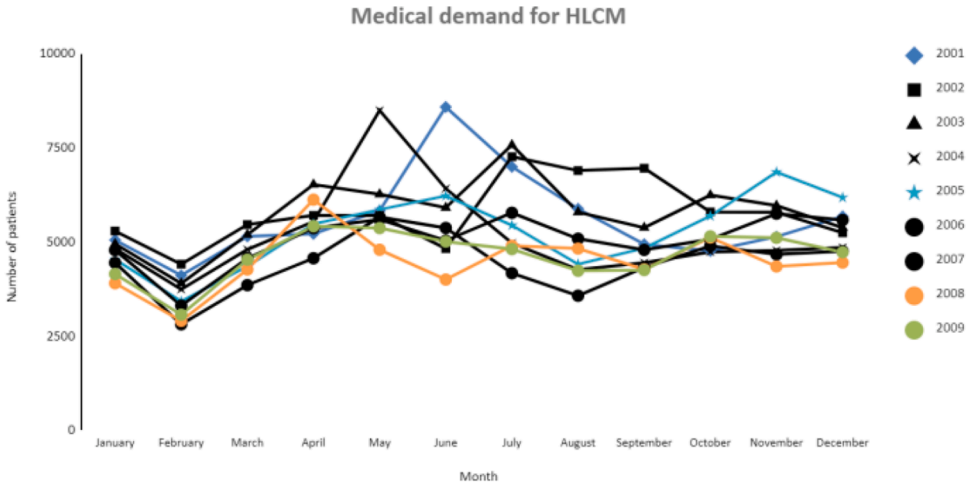

Fig. 6. Medical demand for HLCM.

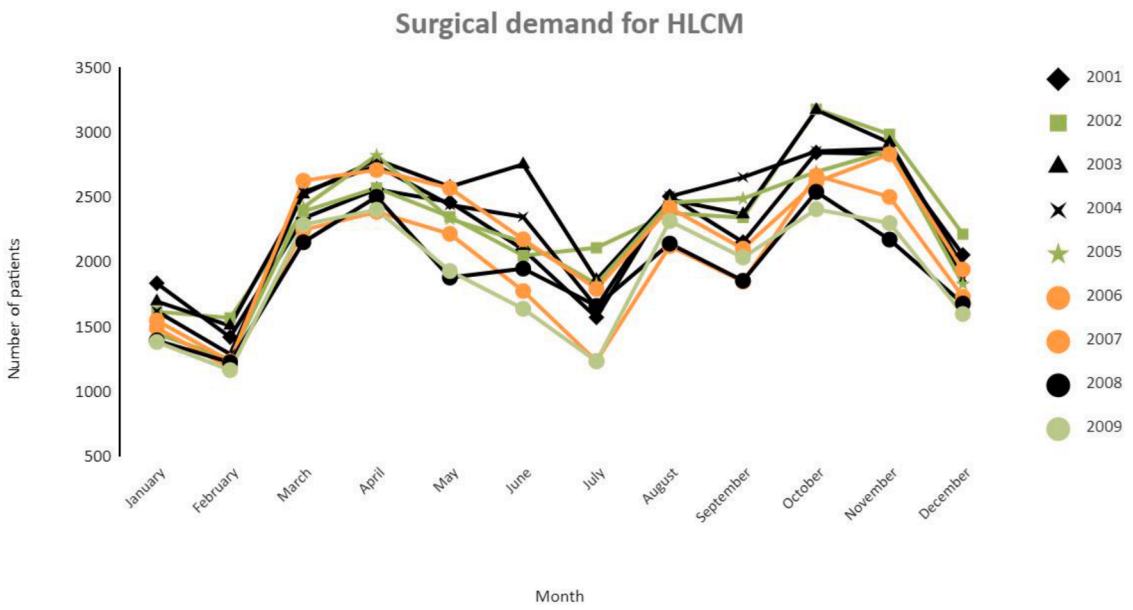

Fig. 7. Surgical demand for HLCM.

This process provided a model that can forecast up to a year in advance and considers the tendencies. Thus, the network has 18 input neurons plus an additional bias neuron that helps to separate cases and allows smaller neural networks than would be allowed without this bias.

The output layer simply contains one neuron that generates the forecasted demand in month N. The hidden layer contains 10 neurons, which provides the model an adequate degree of freedom, usually calculated by (Number of input neurons + Number of output neurons)/2. The resulting network is shown in Fig. 8.

The neural network model developed here and the other forecasting methods previously presented were run using RapidMiner 4.6.0, the Neural Network library from WEKA, and SVR from LIBSVM Library [45], but using RapidMiner as a graphic user interface. To determine the models' forecasting accuracy, the mean average percentage error (MAPE) and mean square error (MSE) are employed. The same data described at the beginning of this section feed all the models as previously described for the neural network method. The results obtained using these four methods for the validation sets of all hospitals are shown in Table 1.

As shown in Table 1, in five of the seven cases, the best results correspond to SVR when using MSE as a criterion to compare the performance of the different models. When using MAPE as a criterion for comparison, SVR appears to be the best option for demand forecasting in all cases.

Figs. 9 and 10 graphically display the results and show forecasts with SVR and actual demand for HLCM, with a 90% confidence interval for the forecast. The interval rests on the hypothesis that the forecast error has a normal distribution with a mean zero, confirmed using a Kolmogorov–Smirnov test.

Similar results obtained for other hospitals are shown in Figs. 11, 12, and 13 for HEGC and in Figs. 14 and 15 for HSBA.

Based on the previously presented results, the conclusion is that support vector regression is an appropriate method for predicting demand in hospitals but without disregarding the possibility of using simpler methods that may provide acceptable results under certain conditions. However, SVR will be employed for the simulations presented subsequently. All the aforementioned models generate forecasts in less than one minute on a standard PC.

4.2. Demand characterization and service times

Fig. 16 shows the patients' illness severity distribution for HLCM, which is the hospital selected as the case to present the capacity analysis, based on a triage with categories C1-C4 for patients, which varies over the different months of a year. Nevertheless, the severity of distribution per month remains relatively stable over the years. Therefore, in the following calculations, each month will have a deterministic distribution of patients for each category.

Given the emergency patients' forecast and the illness severity distribution, the expected number of patients per category is calculated. To determine the number of doctors required to attend such demand, the next step is to characterize the distribution of the attention time for each category. For this purpose, a representative sample of C1, C2, C3, and C4 patients was utilized. Each C1 patient who arrives at the emergency service proceeds to the reanimation room for resuscitation. When this situation occurs and depending on the complexity of the surgery or diagnosis, between one and three of the doctors are currently working in the attention cubicles to attend to the extreme-risk patients. After medical attention, nurses register the time required to stabilize and treat the C1 patient in a logbook, with the names of the doctors who performed the medical procedure. Using this information, a K-S test was performed to determine the distribution of C1 patients' attention times. The conclusion was that the attention times exhibit a lognormal distribution with a mean of 108 min and a standard deviation of 121 min. Of the number of doctors required to attend these patients, two doctors had a highly concentrated distribution; therefore, this value is used for the following calculations.

To characterize the attention time of C2 patients, the data applied did not provide enough information to determine the distribution of such time. However, the doctors reached a consensus that the average time to attend C2 patients was 60 min, with a standard deviation of 20 min. A normal distribution represents the behavior of this attention time. With a high level of confidence, the distribution of the attention time for C3 and C4 patients was lognormal, with means of 10 min and 7 min and standard deviations of 7 min and 3 min, respectively. With the exception of C1 patients, all patients receive attention from only one doctor.

The time distributions presented here provide a basis for estimating the time that doctors will spend attending to patients from each category who arrive at the emergency room. An analysis of patient arrivals at

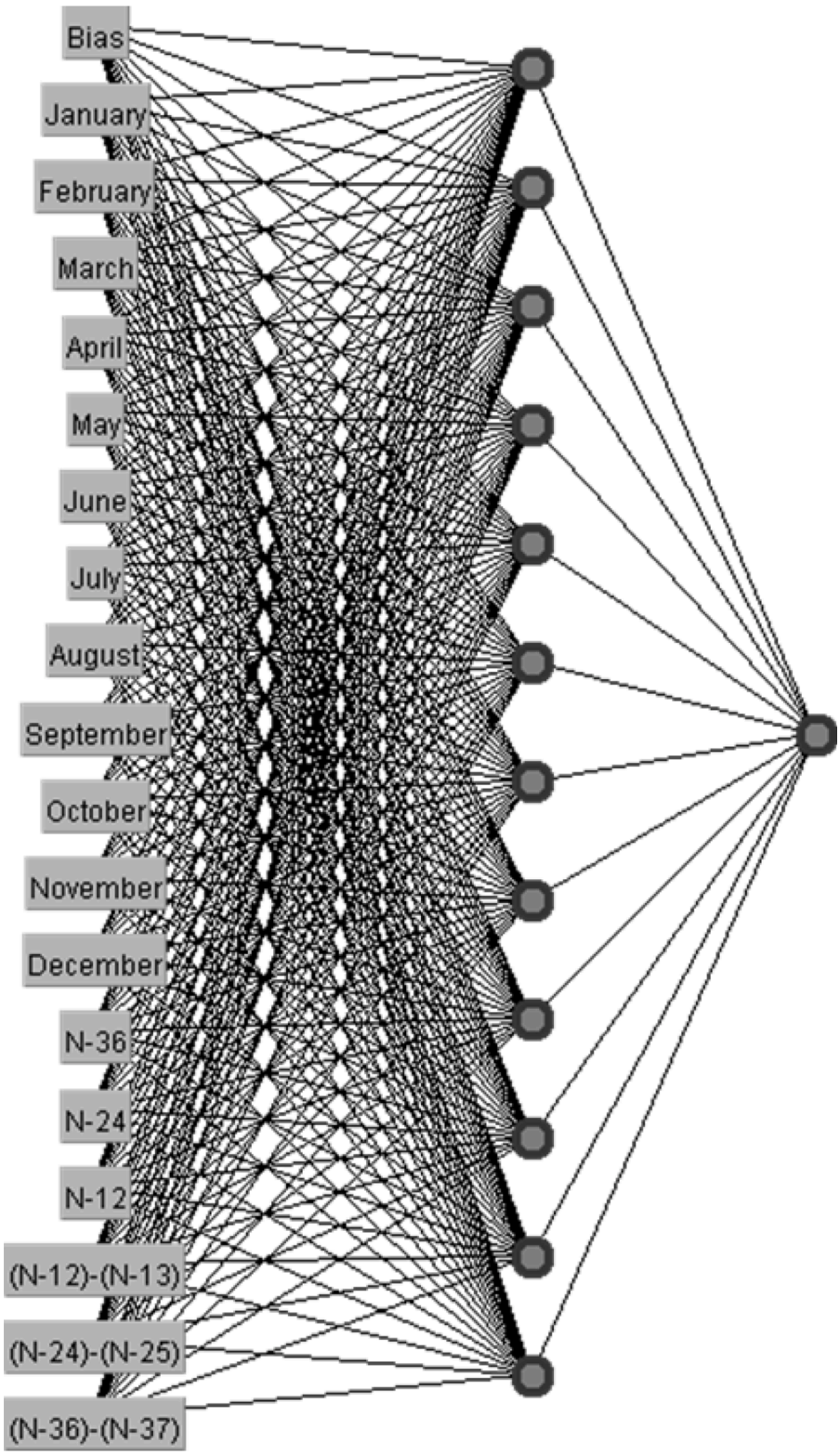

Fig. 8. Resulting Neural Network architecture.

different times of the day is shown in Fig. 17; the conclusion is that 59% of the patients arrive at the emergency service between 12 AM and 8 PM. A representative sample allowed us to determine that this distribution does not vary significantly among different days of the week or among the same days of different weeks. Therefore, this distribution used every day of the year.

4.3. Capacity management

Next, the use of the forecasted demand and its characterization on the management of ED capacity is presented for HLCM, using a representative case to show the methods valid for all hospitals.

Table 1
Forecast errors on validation sets (best results in bold).

	Linear Regression		Weighted Moving Average		Neural Network		SVR	
	MAPE	MSE	MAPE	MSE	MAPE	MSE	MAPE	MSE
HLCM Medical Demand	12.67%	150,686	7.53%	**144,729**	7.45%	161,689	**5.61%**	154,861
HLCM Surgery Demand	6.54%	27,097	7.36%	**20,137**	8.99%	22,947	**5.09%**	25,199
HEGC Medical Demand	15.91%	3114,376	16.5%	1978,332	7.7%	1043,753	**6.86%**	**606,324**
HEGC Surgery Demand	8.55%	14,302	8.96%	11,730	8.3%	12,155	**5.88%**	**8,120**
HEGC Orthopedic Surgery Demand	8.41%	35,940	8.60%	28,247	5.12%	29,851	**4.44%**	**25,460**
HSBA Medical Demand	8.27%	3125,071	11.83%	850,342	7.9%	1226,165	**6.97%**	**643,984**
HSBA Maternity Demand	10.54%	23,738	6.98%	12,408	10.6%	38,629	**3.24%**	**7,867**

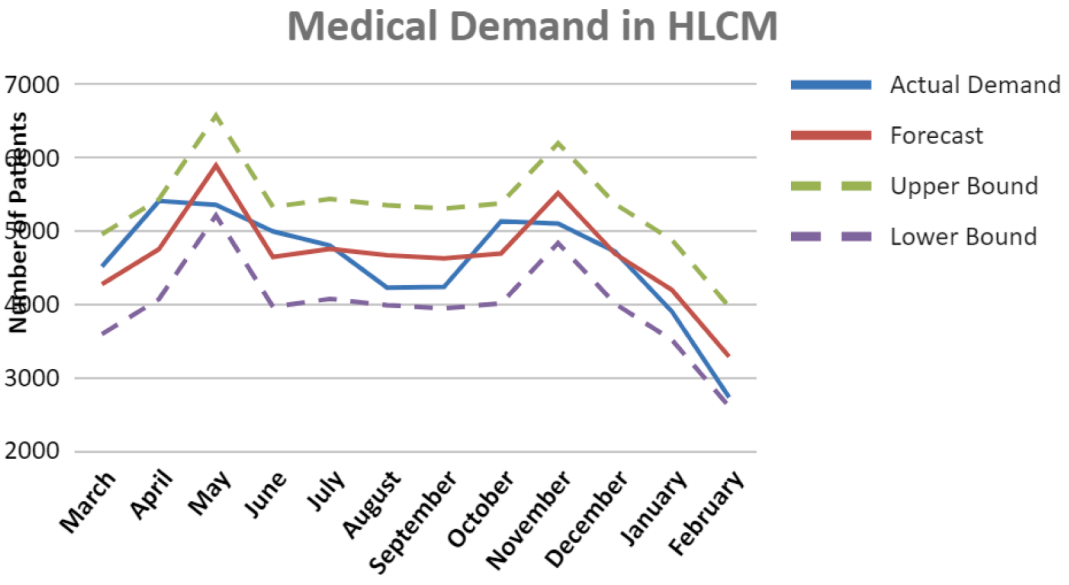

Fig. 9. Medical Demand in HLCM.

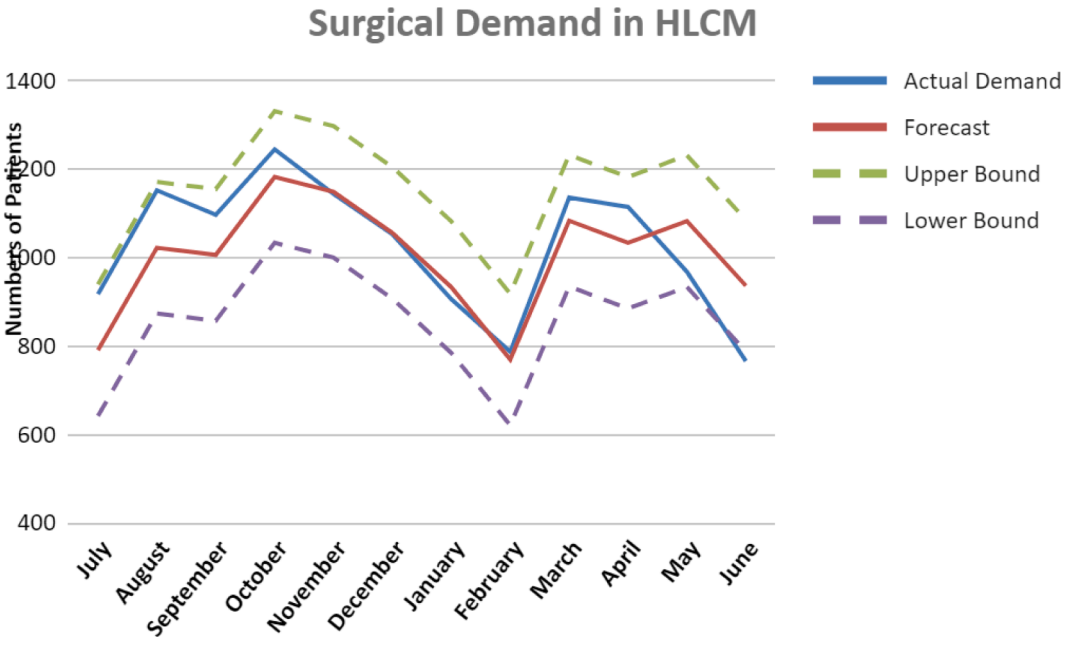

Fig. 10. Surgical Demand in HLCM.

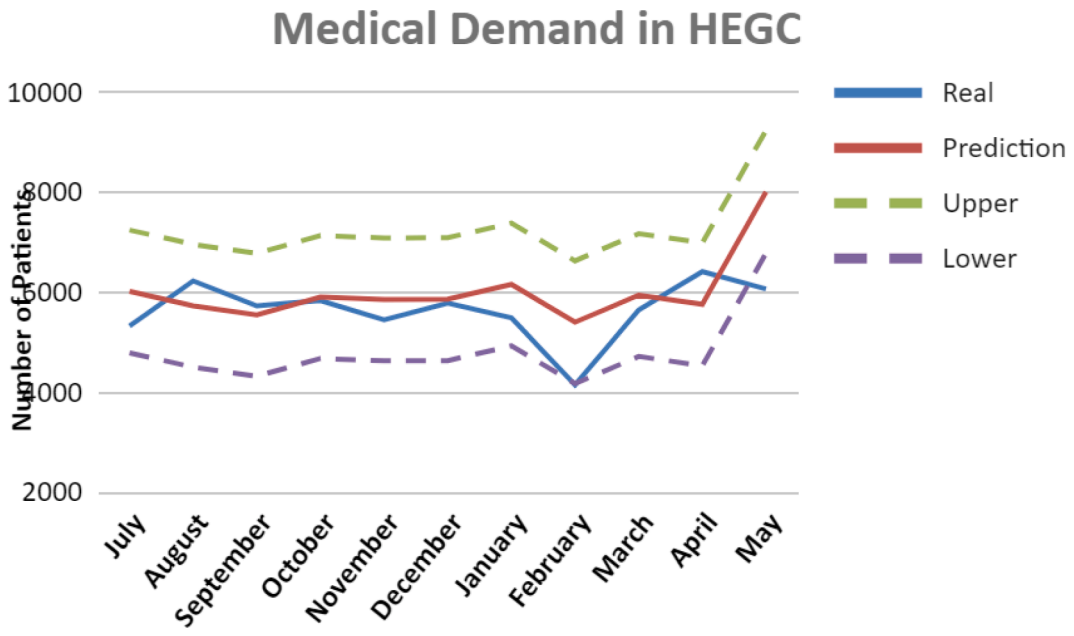

Fig. 11. Medical Demand in HEGC.

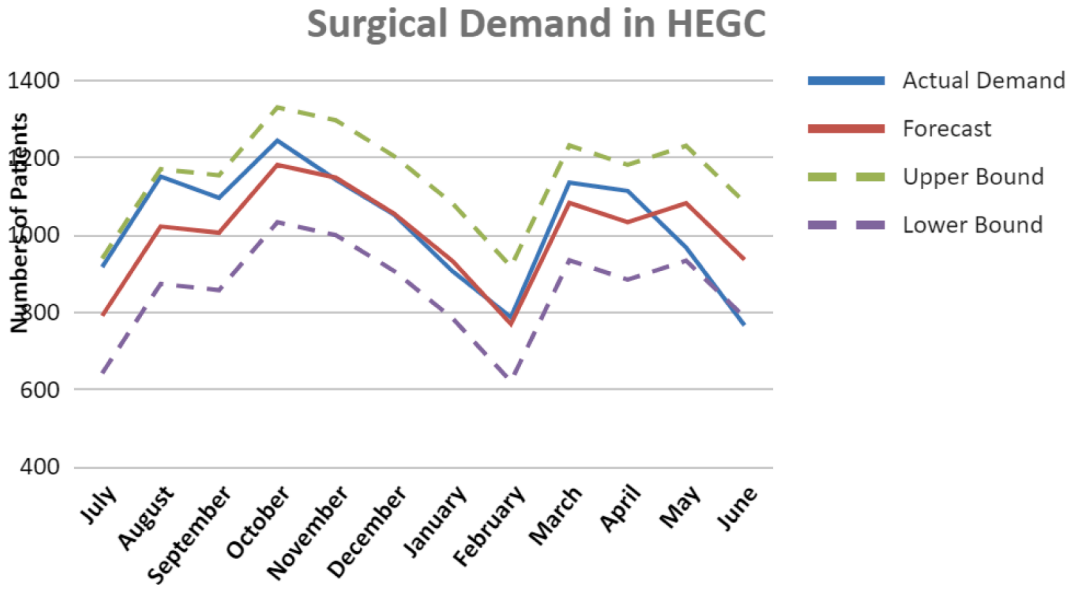

Fig. 12. Surgical Demand in HEGC.

4.3.1. Resource balance

On the demand side, as outlined in Section 3.2, the calculation of the medical hours required to meet the forecasted demand considers a deterministic monthly behavior obtained from the model, divided uniformly within each month and distributed within each day, as presented in Fig. 17. As explained previously, the assumption is that the severity distribution of these patients is deterministic for each month. With these considerations regarding the expected demand, the calculation of the forecasted number of patients per category is obtained by multiplying the total number of forecasted patients by the proportion of patients per category.

The final step is to convert the demand per category to medical hours required, distributed among each period of the day. To obtain a quick idea of the medical hours required to meet such demand, the forecasted number of patients per category, obtained with SVR, multiplied by the mean of the corresponding attention time distributions presented previously provides an initial approximation. Table 2 illustrates the expected behavior of demand during the day for the HLCM and the availability and rate of use of medical resources.

With these simple calculations, several key observations arise regarding the use of medical resources. For example, the period from 0:00 to 8:00 shows a high rate of idle resources, while between 10:00 and 21:00, there is high utilization of medical resources. Thus, some doctors from the night shift could be reassigned to work during peak hours but always remain prepared to meet a potential emergency by having one doctor on duty at home during the night. This finding gives ideas for alternatives to consider in the simulations of the following section.

4.3.2. Simulation model and capacity management

The simulation model that incorporates the stochastic behavior of the demand is for medical patients only since their waiting times and

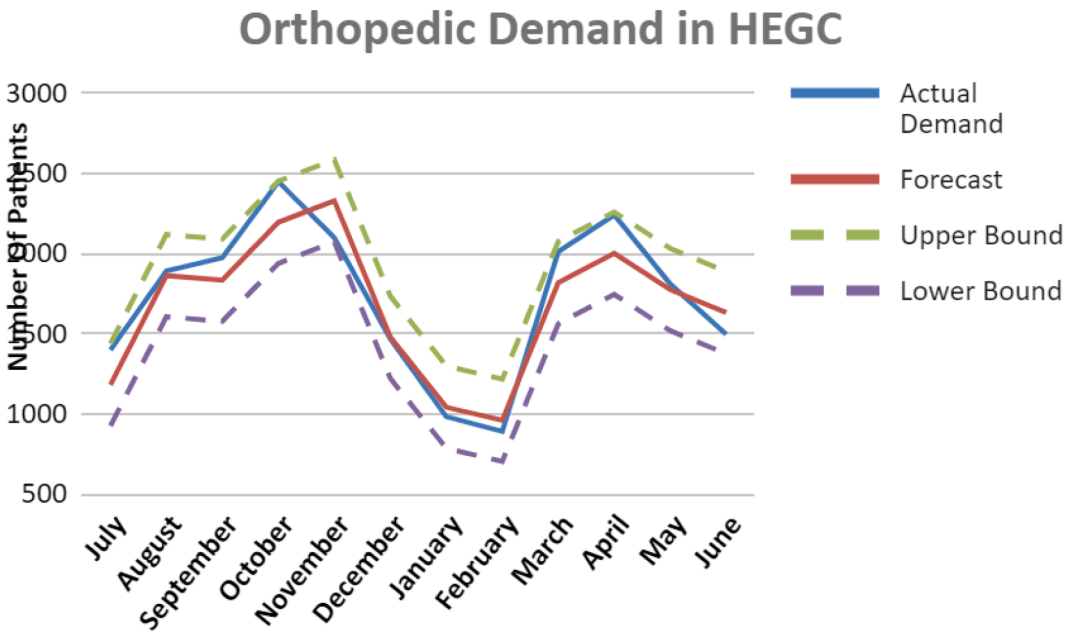

Fig. 13. Orthopedic Demand in HEGC.

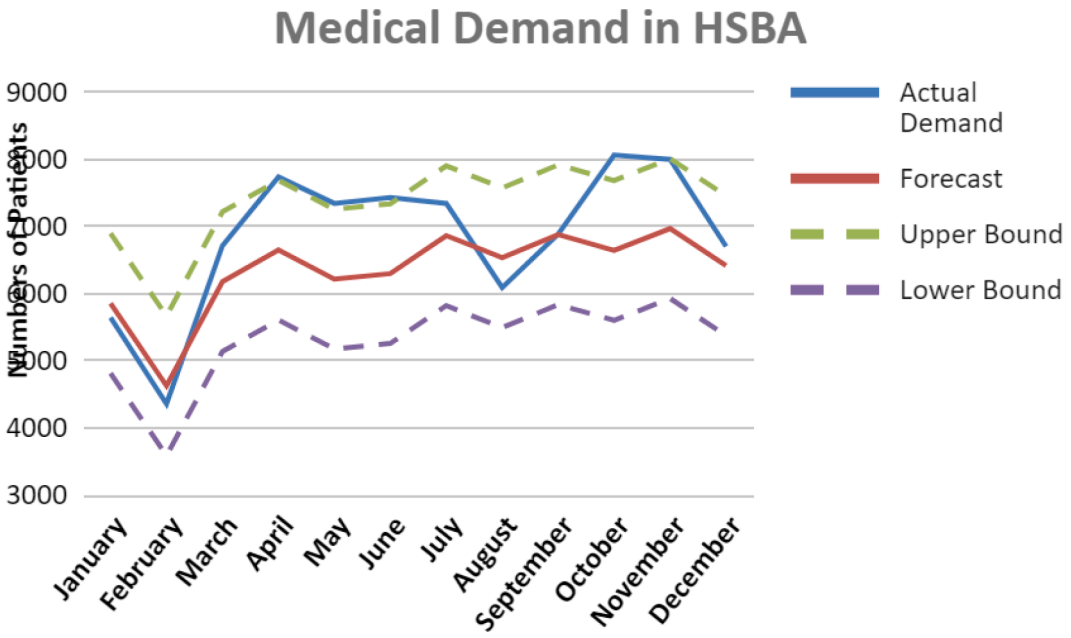

Fig. 14. Medical Demand in HSBA.

length of lines are significantly higher than those of other patients. We applied discrete event simulation (DES) using the software tool Arena™ version 10.0.

The SVR forecast detailed in Section 4.1 has an error with a normal distribution. Thus, several forecasts with times sampled from such a distribution of the error are the input for simulating the different demand scenarios for each month. Due to the stability of its daily behavior, the demand for each scenario has a uniform distribution across every day of the month. The daily demand disaggregates into hourly demand by using the distribution shown in Fig. 17. Consequently, several scenarios of monthly demand disaggregated per hour are available. Using the hourly forecasted demand from each of the scenarios generated, as previously described, allows generation of the average forecasted demand for each hour of the day. The assumption was that the hourly demand arrives according to a Poisson process; then, this average corresponds to the mean of the Poisson distribution per hour.

Patients categorized upon their arrival at the ED receive service according to the time distributions presented in Section 3.2. Because the stochastic behavior of the demand and medical attention is known, the simulation model construction and its role in the management of hospital capacity are presented as follows:

4.3.3. Configuration management

In the current configuration or base case for HLCM, only very ill patients (C1) receive priority service when arriving at the ED. They moved to the resuscitation service for stabilization and then referred to the operating room or the intensive care unit service. Patients who are not critically ill must provide their data upon their arrival and

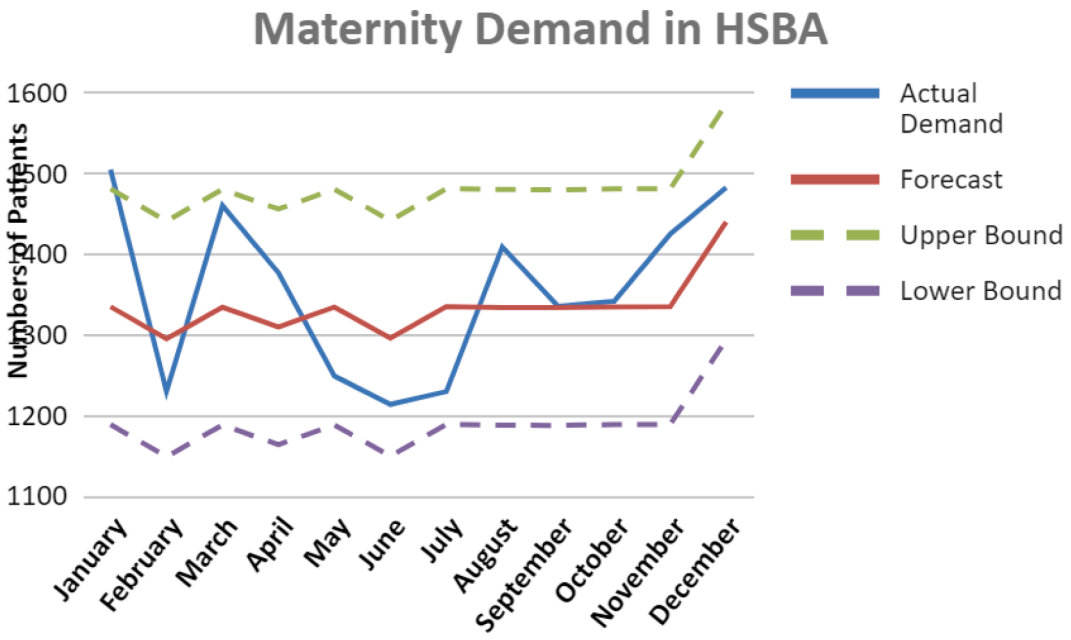

Fig. 15. Maternity Demand in HSBA.

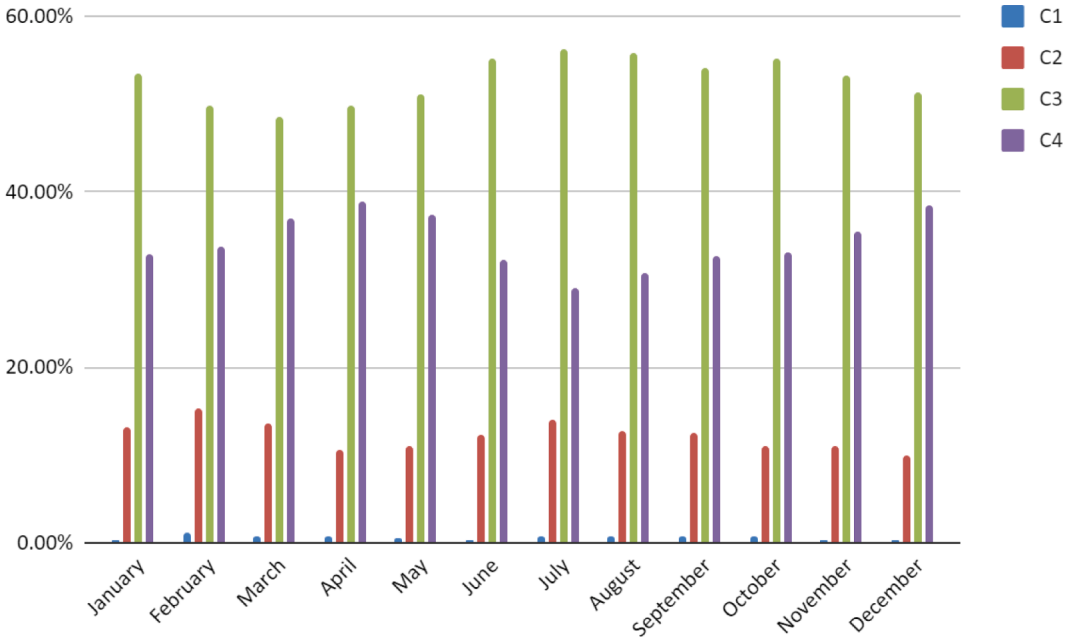

Fig. 16. Monthly categorization distribution.

subsequently wait for medical attention. The admission time had a uniform distribution between 5 min and 10 min. After medical attention, the patients can proceed to hospitalization for diagnostic testing or to immediate discharge. Hospitalization services do not belong to emergency services and hence do not use their medical resources. The time in diagnostic test services has a uniform distribution between 2 h and 4 h.

The fast track with triage configuration in Fig. 4 is an alternative to the current situation. The reason behind fast-track selection lies in the importance of liberating resources for the medical attention of patients with the most urgent needs (C2 and C3). Faster attention is attributed to the notion that half of the patients categorized as C4 in triage move to fast-track attention and do not use the medical resources of the emergency line.

The doctors' shift structure consists of two 12-hour shifts (day and night) with three doctors attending each shift. For the current situation of base case and fast track configurations, simulation models that include the previously presented demand and emergency service characterizations are employed. An example of the models is shown in Fig. 18.

The simulation model's role in the management of hospital capacity is subsequently given. To perform capacity configuration management, an estimation of the length of wait (LOW) for given configurations, which allows comparison of the performance of alternative designs, is necessary. This metric weights the demand per category by its respective average LOW. The results for this metric in the base case and fast track simulated configurations are shown in Table 3.

Based on the scenarios run in the simulation, a 95% confidence interval is obtained for the LOW of each configuration. The intervals

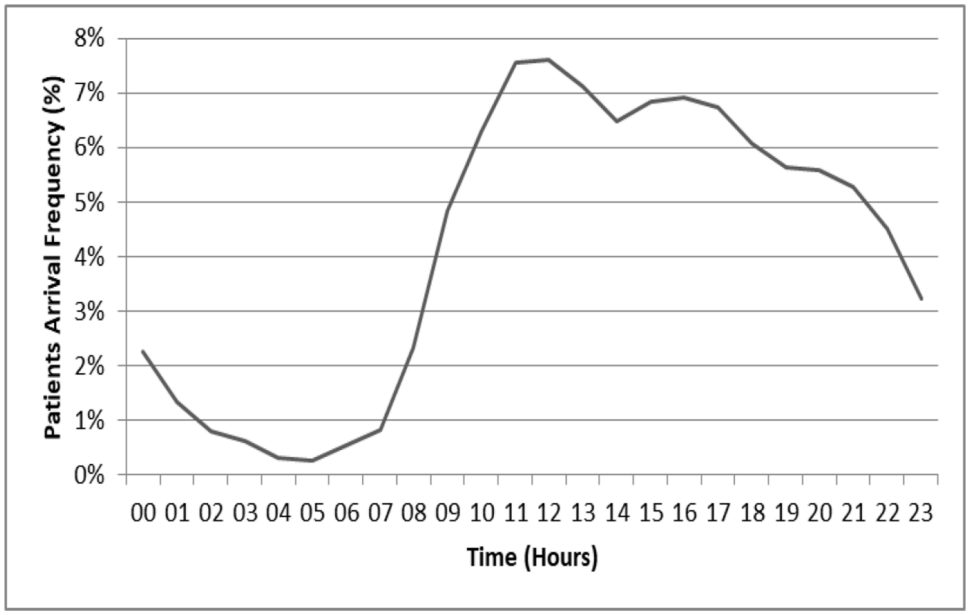

Fig. 17. Patient arrival distribution per hour.

Table 2
Forecasted medical resources occupation rate.

Hours	Available	Forecasted need	Excess use	Excess%
3:00 - 3:59	90	11	79	12%
4:00 - 4:59	90	6	84	7%
5:00 - 5:59	90	5	85	6%
6:00 - 6:59	90	10	80	11%
7:00 - 7:59	90	15	75	17%
8:00 - 8:59	90	42	48	47%
9:00 - 9:59	90	87	3	97%
10:00 - 10:59	90	114	−24	127%
11:00 - 11:59	90	136	−46	151%
12:00 - 12:59	90	138	−48	153%
13:00 - 13:59	90	129	−39	143%
14:00 - 14:59	90	117	−27	130%
15:00 - 15:59	90	123	−33	137%
16:00 - 16:59	90	125	−35	139%
17:00 - 17:59	90	122	−32	136%
18:00 - 18:59	90	110	−20	122%
19:00 - 19:59	90	102	−12	113%
20:00 - 20:59	90	101	−11	112%
21:00 - 21:59	90	95	−5	106%
22:00 - 22:59	90	81	9	90%
23:00 - 23:59	90	58	32	64%

obtained for the base case and fast track with triage were (61.8, 66.6) and (55.5, 59.1), respectively. The procedure proposed to test whether the LOW differs significantly between these two configurations is the procedure detailed in [46]. This comparison, based on the difference between their respective statistical distributions, is shown in Table 4. Since the confidence interval does not contain the value zero, there is confirmation that the difference shown in Table 3 is statistically significant.

Based on the results presented in Tables 3 and 4, we observe that the simulation model is a good representation of the current behavior of the system, since the average LOW, calculated with current statistics, is within the confidence interval of the simulated base case. The main result is that with current resources, the Fast Track with Triage configuration reduces the average LOW compared to the Base Case in 6.9 min, which corresponds to a 10.8% reduction in the current average waiting time.

Thus, HLCM management decided to implement the fast track with triage configuration; it is currently in use.

4.3.4. *Resource management*

The first resource management issue is that the current shift structure, including the number of doctors per shift, was not constructed based on the daily behavior of the demand, as shown in Table 2. To correct this situation, the simulation model considered several assignments and the number of doctors per shift with the fast track with triage configuration and assumed stochastic demand. If the HLCM current structure of two 12-hour shifts persists, an initial scenario would consider only redistributing the six doctors available in a different manner. Given the greater arrival of patients during the day, as observed in Fig. 17, a possible redistribution could include the reassignment of a doctor from the night shift to the day shift. Therefore, four doctors attended during the day shift, and only two doctors attended the night shift. The average LOW of this scenario is 45.1 min. Further resource management considerations may determine the addition or reduction of medical hours for attending to patients. Since these resources are expensive, the different scenarios simulated changed the existing capacity in half doctor intervals. The extra half doctor was included through the creation of a new shift of six hours, from 12:00 to 18:00, which is the period during which most patients arrive at the service. Thus, the number of 6.5 doctors available means that four doctors attend the day shift (8:00–20:00), one doctor attends the half shift (12:00–18:00), and two doctors attend the night shift (20:00–8:00). The average LOW obtained with this configuration is 40.5 min.

The simulation uses five to seven doctors within 24 h and distributed as previously explained. The idea behind analyzing the reduction in the current number of doctors is to assess whether the performance of the system substantially changes when these resources are lacking, either by management decisions or by absenteeism. The average LOW for different scenarios of numbers and assignments of doctors is shown in Fig. 19, including the 95% confidence interval for each of the points.

As expected, the addition of medical resources improves the service quality, measured as the average LOW. The interesting result is that the average LOW decreases dramatically when the number of doctors increases from 5 to 5.5 and decreases more gradually more resources are available. To show that they are statistically significant, the same procedure proposed in [46] is applied to analyze the differences in the LOW among all the scenarios in Fig. 19. Table 5 shows the confidence intervals when comparing the LOW of these scenarios. As observed, increasing doctors from 5 to 5.5 provides a significant improvement in the performance of the system, while the change between 5.5 and 6

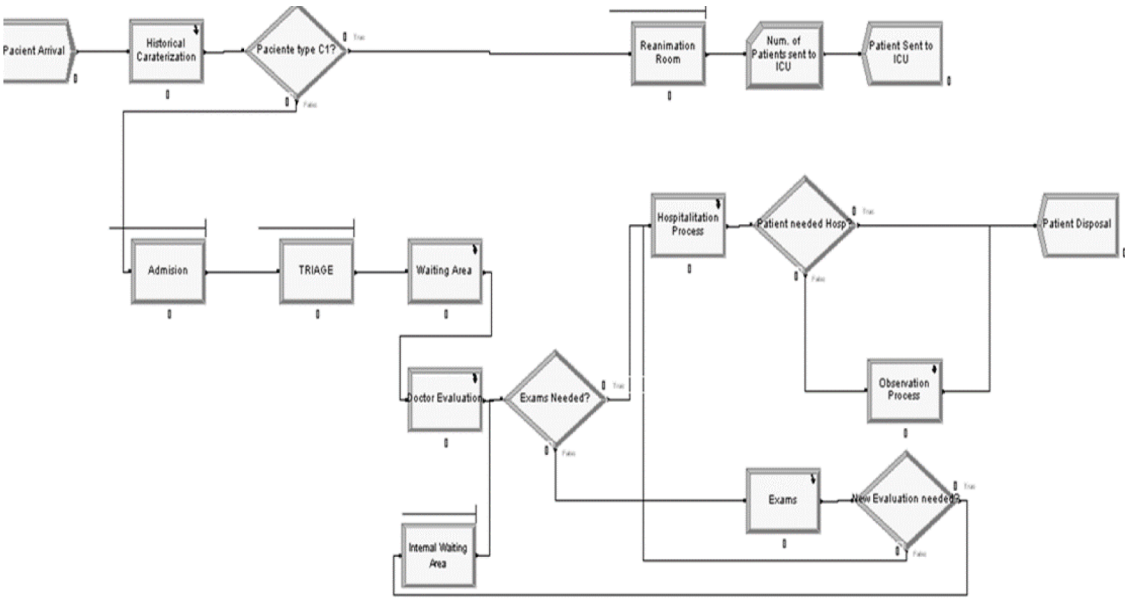

Fig. 18. The simulation model for the Base Case capacity analysis.

Table 3
Simulated LOW for HLCM emergency service configurations.

Configuration	Avg. (min)	Std. Dev. (min)
Base Case	64.2	1.2
Fast-Track with Triage	57.3	0.9

Table 4
Base Case and Fast Track with Triage configurations comparison.

Comparing Configurations	Avg. (min)	Std. Dev. (min)	Lower Bound 95% (min)	Upper Bound 95% (min)
Base Case / Fast Track with Triage	6.9	1.5	3.9	9.9

doctors is not significant. Nevertheless, increasing the number of doctors from 5.5 to 6.5 shows statistical significance. As shown in Fig. 19 by simple visual inspection, increasing the number of doctors from 5.5 to 6 and from 6.5 to 7 does not seem to provide important improvements. This finding is supported by statistical analyses, as shown in Table 5, where the respective confidence intervals do not show statistical significance.

4.4. Processes design

To use forecasting and simulation models to routinely manage capacity, there is the need for formally designed processes. Such processes define the routines that would be periodically executed to run models, determine the necessary actions based on their results, and act on the decision variables that the models consider, such as ED medical doctors' staff and schedules. There must also be a process that evaluates the results of the models' use and that detects the needs to adapt them to new situations, e.g., a change in the demand structure. Thus, for each process, IT support requirements, including required software packages, newly collected information, its integration with current databases, and required reports, should be defined.

Table 5
Confidence intervals for compared scenarios.

Comparing Scenarios	5.5	6	6.5	7
5	[30.4; 42.6]	[31.6; 43.9]	[36.3; 48.3]	[36.5; 48.7]
5.5	–	[−0.4; 3]	[4.5; 7.2]	[4.4; 7.8]
6		–	[3.3; 5.9]	[3.1; 6.5]
6.5			–	[−1.1; 1.6]

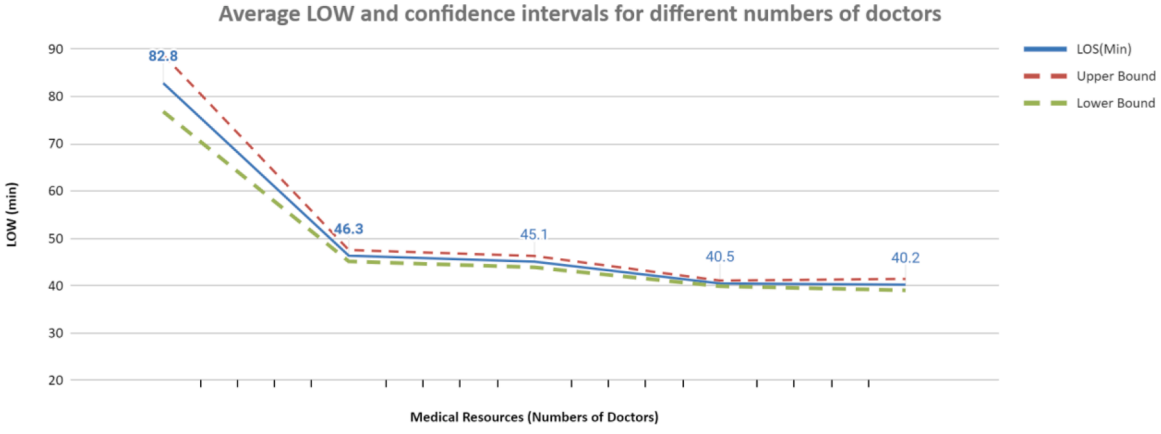

Fig. 19. Average LOW and confidence intervals for different numbers of doctors.

The required design definition performed in this case followed a general process architecture pattern for hospital management as a guide, as detailed in [47]. The designed processes were formalized using BPMN [47] notation, which explicitly establishes the process flow and its interaction with software packages and databases that contain information for model operation explicitly.

5. Conclusion and discussion

According to the results presented in this paper for three hospitals, it is possible to perform demand forecasting with great confidence for hospital emergency services. After testing several forecasting methods, SVR provided the best results in terms of variance and accuracy. Based on this forecasting, a logic for managing capacity was developed for one hospital. Such logic uses the comparison between the forecasted demand and the available medical resources and a simulation model to assess the performance of different configurations of facilities and resources. These analyses provide hospital managers with a decision tool for determining the number and distribution of medical resources on emergency services based on a cost/benefit analysis of resources and service improvement. The abovementioned results support the task of assigning doctors to different kinds of boxes, defining their work schedules, and considering additional doctors.

The forecasting method and capacity management logic proposed in this paper have been validated and accepted by hospital managers and staff and are currently in use in the Hospital Luis Calvo Mackenna (HLCM), which is a major pediatric hospital in Santiago, Chile. For this use, there was a need for formal processes that embed the forecasting model and resource management logic, including a support computing system. The results of the implemented processes have been encouraging, and the National Health Authorities are considering extending the whole design concept to other public hospitals in Chile.

Note that the design of the processes, with embedded analytics and IT support, is not a one-time effort. Its design includes the periodical execution and adaptation of the processes under changing conditions, such as unexpected demand, for example, epidemic episodes and new campaigns, which require adapting capacity.

An interesting feature of the work reported in this paper is the integration of several methods that are presented independently in different publications. Such methods are forecasting, simulation, process design, and IT support. The methods' integration facilitates the practical use of quantitative models, since they may produce interesting results when their use is independent and on a one-time basis, but practical impact is not guaranteed. Integration explicitly addresses the design of a solution for routine use, which also has adaptation capabilities to facilitate use under changing conditions.

The solution is also general and admits adaptation and extension to other services. Thus, similar works on an ambulance service operation and surgical services is ongoing. Possible future research directions are related to updating the proposed forecasting models as new data are gathered. A methodology for dynamic feature selection and support vector regression was presented in [35].

Authors statement

We have carefully considered all the comments provided by the reviewers, which were very helpful for the revision of the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge support from the following authorities: Begoña Yarza Director at HEGC, Osvaldo Artaza Director at HLCM, and Inti Paredes Director at HSBA. The second author acknowledges support from ANID PIA/APOYO AFB180003.

References

[1] Harper A, Mustafee N, Feeney M. A hybrid approach using forecasting and discrete-event simulation for endoscopy services. In: 2017 Winter Simulation Conference (WSC); 2017. p. 1583–94. https://doi.org/10.1109/WSC.2017.8247899. 2017.
[2] Assad D, Spiegel Th. Improving emergency department resourse planning: a multiple case study. Health Systems 2020;9(1):2–30.
[3] d'Etienne JP, Zhou Y, Kan C, Shaikh S, Ho AF, Suley E, et al. Two-step predictive model for early detection of emergency department patients with prolonged stay and its management implications. Am J Emerg Med 2021;40:148–58. https://doi.org/10.1016/j.ajem.2020.01.050.
[4] Armstrong S J. Principles of forecasting. Norwell: Kluwer Academic Publishers; 2001.
[5] Aburto L, Weber R. Improved supply chain management based on hybrid demand forecast. Applied Soft Computing 2007;7(1):136–44.
[6] Kapuscinski R, Zhang RQ, Carbonneau P, Moore R, Reeves B. Inventory decisions in Dell's supply chain. Interfaces 2004;34:191–205.
[7] Davenport TH, Dalle Mule L, Lucker J. Know What Your Customers Want Before They Do. Harv Bus Rev 2011;89:84–92.
[8] Box GEH, Jenkins GM, Reinsel GC. Time series analysis, forecasting and control. 3rd Ed. Prentice-Hall: Englewood Cliffs; 1994. NJ.
[9] Adya M, Collopy F. How efective are neural nets at forecasting and prediction? A review and evaluation. Journal of Forecasting 1998;17:451–61.
[10] Wiler JL, Griffey RT, Olsen T. Review of modeling approaches for emergency department patient flow and crowding research. Acad Emerg Med 2011;18:1371–9.
[11] Farmer RDT, Emami J. Models for forecasting hospital bed requirements in the acute sector. J Epidemiol Community Health 1990;44:307–12.
[12] Jones AJ, Joy MP, Pearson J. Forecasting demand of emergency care. Health Care Manag Sci 2002;5:297–305.
[13] Schweigler LM, Desmond JS, McCarthy ML, Bukowski KJ, Ionides EL, Younger JG. Forecasting models of emergency department crowding. Acad Emerg Med 2009;16:301–8.
[14] Whitt W, Zhang X. Forecasting arrivals and occupancy levels in an emergency department. Operations Research for Health Care 2019;21:1–18.
[15] Gul M, Celik E. An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. Health Systems 2020;9(4):263–84.
[16] Galetsi P, Katsaliaki K. A review of the literature on big data analytics in healthcare. J Oper Res Soc 2020;71:1511–29. https://doi.org/10.1080/01605682.2019.1630328. 10.
[17] Yucesan M, Muhammet I. A multi-method patient arrival forecasting outline for hospital emergency departments. Int J Healthc Manage 2018;13(sup 1):283–95.
[18] West YZ, Luo L, Ruixiao FZ, Yang J, Feng Y, Guo H. Emergency patient flow forecasting in the radiology department. Health Informatics J 2020;26(4):2362–74.
[19] Brailsford SC, Harper PR, Patel B, Pitt M. An analysis of the academic literature on simulation and modelling in health care. J Simul 2009;3:130–40. https://doi.org/10.1057/jos.2009.10. 3.
[20] Rau PF, Tsai SM, Liang SM, Tan J, Syu S, Jheng Y, Ciou T, Jaw F. Using discrete-event simulation in strategic capacity planning for an outpatient physical therapy service. Health Care Manag Sci 2013;16(4):352–65.
[21] Pan C, Zhang D, Kon AW, Wai CS, Ang WB. Patient flow improvement for an ophthalmic specialist outpatient clinic with aid of discrete event simulation and design of experiment. Health Care Manag Sci 2015;8(12):137–55.
[22] García ML, Centeno MA, Rivera C, DeCario N. Reducing Time in an Emergency Room Via a Fast-Track. In: Winter Simulation Conference; 1995. p. 1048–53.
[23] Khurma N, Bacioiu GM. Simulation-Based Verification of Lean Improvement for Emergency Room Process. In: Winter Simulation Conference; 2008. p. 1490–9.
[24] Rojas LM, Garavito LA. Analysing the Diana Turbay CAMI emergency and hospitalization processes using an Arena 10.0 simulation model for optimal human resource distribution. Revista Ingeniería e Investigación 2008;28(1):146–53.
[25] Samaha S, Armel WS, Stark DW. The Use of Simulation to Reduce the Length of Stay in an Emergency Department. Winter Simulation Conference 2003:1907–11.
[26] Hwang J, Gao L, Jang W. Joint Demand and Capacity Management in a Restaurant System. Eur J Oper Res 2009;207(1):465–72.
[27] Min D, Yih Y. Scheduling Elective Surgery under Uncertainty and Downstream Capacity Constraints. Eur J Oper Res 2010;206(3):642–52.
[28] Harper A, Mustafee N. Proactive service recovery in emergency departments: a hybrid modelling approach using forecasting and real-time simulation. igsim-pads '19. USA: Chicago, IL; 2019.
[29] Mustafee N., Powell J.H. From Hybrid Simulation to Hybrid Systems Modelling. Proceedings of the 2018 Winter Simulation Conference, Rabe M, Juan AA, N. Mustafee, N. Skoogh, Jain AS, Johansson B, eds.
[30] Gu Y, Kunc M. Using hybrid modelling to simulate and analyse strategies. J Model Manage 2020;15(2):459–90.
[31] Mielczarek B. Combining Simulation Techniques to Understand Demographic Dynamics and Forecast Hospital Demands. In: 2019 Winter Simulation Conference

(WSC); 2019. p. 1114–25. https://doi.org/10.1109/WSC40007.2019.9004855. 2019.
[32] Mielczarek B. A simulation approach to evaluate the effect of demographic changes on projected number of patients across disease categories. J Comput Sci 2021;53. https://doi.org/10.1016/j.jocs.2021.101393. available online.
[33] Mustafee N, Powell J. Providing Real-Time Information for Urgent Care. Impact 2021;1:25–9. https://doi.org/10.1080/2058802X.2020.1857601. 2021.
[34] McLaughlin D, Hays JM. Healthcare operations management. Washington, DC: AUPHA Press; 2008. p. 378–81.
[35] Guajardo J, Weber R, Miranda J. A Forecasting Methodology Using Support Vector Regression and Dynamic Feature Selection. J Inform Knowl Manage 2006;5(4): 329–35.
[36] Hofmann T, Schölkopf B, Smola AJ. Kernel Methods in Machine Learning. Ann Stat 2008;36:1171–220.
[37] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004; 14(3):199–222.
[38] Chen PH, Lin CJ, Schölkopf B. A tutorial on ν-support vector machines. Appl Stoch Models Bus Ind 2005;21:111–36.
[39] Vapnik V. The nature of statistical learning theory. Springer-Verlag; 1995.
[40] Vapnik V. Statistical Learning Theory. Wiley; 1998.
[41] Vladimir C, Ma YQ. Practical selection of SVM parameters and noise estimation of SVM regression. Neural Netw 2004;17(1):113–26.
[42] Marmor YN, Wasserkrug S, Zletyn S, Mesika Y, Greenshpan I, Carmeli B, Shtub A, Mandelbaum A. Toward Simulation-Based Real-Time Decision-Support Systems for Emergency Departments. Winter Simulation Conference 2009:2042–53.
[43] Shirxia Y, Xiang L, Li N, Shang-dong Y. Optimizing neural network forecast by immune algorithm. J Centr South Univ Tech 2007;13(5).
[44] Zhang GP. Avoiding pitfalls in neural network research. IEEE Trans Syst Man Cybernet—Part C Appl Rev 2007;37:3–13.
[45] Chang C.C., Lin C. J. LIBSVM: A Library for Support Vector Machines [EB/OL], 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[46] Law AM, Kelton WD. Simulation modeling and analysis. McGraw Hill; 2001.
[47] Barros O. A process architecture pattern and its application to designing health services: emergency case. Business Process Management Journal 2019;26(2): 513–27.