

Control problems and management policies in health systems: application to intensive care units

Fermín Mallor · Cristina Azcárate · Julio Barado

Published online: 16 December 2014
© Springer Science+Business Media New York 2014

Abstract The stochastic nature of both patient arrivals and lengths of stay leads inevitably to periodic bed shortages in healthcare units. Physicians are challenged to fit demand to service capacity. If all beds are occupied eligible patients are usually referred to another ward or hospital and scheduled surgeries may be cancelled. Lack of beds may also have consequences for patients, who may be discharged in advance when the number of occupied beds is so high as to compromise the medical care of new incoming patients. In this paper we deal with the problem of obtaining efficient bed-management policies. We introduce a queuing control problem in which neither the arrival rates nor the number of servers can be modified. Bed occupancy control is addressed by modifying the service time rates, to make them dependent on the state of the system. The objective functions are two quality-of-service components: to minimize patient rejections and to minimize the length of stay shortening. The first objective has a clear mathematical formulation: minimize the probability of rejecting a patient. The second objective admits several formulations. Four different expressions, all leading to nonlinear optimization problems, are proposed. The solutions of these optimization problems define different control policies. We obtain the analytical solutions by adopting Markov-type assumptions and comparing them in terms of the two quality-of-service components. We extend these results to the general case using optimi-

F. Mallor (✉) · C. Azcárate
Department of Statistics and Operational Research, Public University of Navarre, Pamplona, Spain
e-mail: mallor@unavarra.es

C. Azcárate
e-mail: cazcarate@unavarra.es

J. Barado
Hospital of Navarre, Pamplona, Spain
e-mail: X017905@cfnavarra.es

zation with simulation, and propose a way to simulate general length of stay distributions enabling the inclusion of state-dependent service rates.

Keywords Queuing theory · Healthcare · Optimization · Control problem · Simulation · ICU

Mathematics Subject Classification 60K25 · 80M50 · 93E20 · 93C65

1 Introduction

Intensive care units (ICU) are specialized sections of a hospital that provide care for critically ill patients requiring immediate attention. The type of equipment and clinicians needed to staff intensive care units makes them very costly to run and it is a challenge for hospital managers to fit demand to service capacity.

An ICU can be mathematically modelled as a queuing system, where the clients are the patients, the servers are the beds and there is no waiting room. Queuing theory has been widely used in the health care context, (see Lakshmi and Sivakumar (2013) for a review of its applications in health care management problems). Some studies have used queuing models to address management problems in various hospital departments [emergency department (ED), ICU, wards, etc.]. As an example, the reader is referred to Green (2002), a study estimating bed unavailability in obstetrics and ICUs units; de Bruin et al. (2007), which included ED, ICU and cardiac care wards; Cochran and Roche (2008), an analysis of hospital bed capacity at four care levels (ICU, telemetry, medical/surgical and obstetrics); and de Bruin et al. (2010), a sizing analysis of 24 wards in a medical center, including general and critical care. Nevertheless, there are few papers that focus on ICUs: Shmueli et al. (2003) used queuing theory to find optimal admission policies; McManus et al. (2004) analyzed bed availability, utilization rates and rejection rates using queuing theory; and Griffiths et al. (2013) proposed a queuing model to improve bed management distinguishing between emergency and elective surgery patients.

Queuing theory is a powerful analytical tool for building simple models with relatively few little data, while including randomness. Simulation is another common approach in queuing modelling, which enables a more detailed representation of the complexity in health systems. Reviews and discussion papers dealing with the application of simulation modelling in health care can be found in Brailsford et al. (2009), Günal and Pidd (2010), Eldabi et al. (2007) and Katsaliaki and Mustafee (2011). Many studies use simulation to analyze hospital capacity and bed allocation, but, again, only a few deal specifically with ICUs. Worth noting are Kim et al. (1999, 2000), in which ICU admission and discharge processes are analyzed through simulation and several bed allocation rules are evaluated; Litvak et al. (2008), Ridge et al. (1998), Costa et al. (2003) and Zhu et al. (2012), in which the ICU capacity problem is studied; Masterson et al. (2004) which presents a case study of an ICU at a military medical centre analyzing staffing, sizing and operational policies; and Kolker (2009), in which an ICU simulation model is used to establish a quantitative link between the daily load levelling of elective surgeries and ICU diversion.

Queue system performance (particularly in an ICU) is greatly influenced by the client (patient) arrival pattern and the service time (length of stay in the ICU). The ICU arrival pattern usually consists of a mix of emergency and elective patients. Emergency patient arrivals are typically modeled as Poisson processes (see, for example, Griffiths et al. (2005), de Bruin et al. (2010)), including patient arrivals from different sources (emergency surgery, ward, emergency room, another hospital, etc.). Elective arrivals, however, are scheduled and therefore very unlikely to follow a Poisson process. While some authors use other arrival distributions for scheduled patients: empirical distributions (Mallor and Azcárate (2014)), non-stationary Poisson processes (Griffiths et al. (2005)), and deterministic distributions (Zhu et al. (2012)), other consider it reasonable to assume Poisson arrivals [see for example Litvak et al. (2008) or Kim et al. (2000)]. Inter-arrival exponential distribution is taken in Green (2002), Shmueli et al. (2003), McManus et al. (2004), Griffiths et al. (2006) and Troy and Rosenberg (2009).

Patient length of stay (LoS) has also been extensively studied (Zimmerman et al. 2006). Some authors consider exponential distributions to approximate patient LoS in ICU analytical approaches. Examples can be found in the model proposed by McManus et al. (2004) and that presented by Shmueli et al. (2003). Kim et al. (1999) assume exponentially distributed service times for all the patient groups included in their simulation model. Nevertheless, several studies have proved that LoS distributions are usually heavily skewed to the right (see for example, Rauner et al. (2003) and Vasilakis and Marshall (2005)). These empirical findings support the choice of weighted-tail distributions for LoS modeling purposes: Weibull distributions are used in Ridge et al. (1998); Lognormal distributions in Costa et al. (2003) and in Masterson et al. (2004); Pearson VI in Griffiths et al. (2005); phase-type distributions in Griffiths et al. (2006) and Bowers (2013). Mallor and Azcárate (2014) proposed non-normal regression models including variables with the power to explain some of the LoS variability, such as the Apache index. An overview of LoS and patient flow modeling techniques can be found in Marshall et al. (2005).

The stochastic nature of both patient arrivals and LoS periodically lead to inevitable bed shortages. If all beds are occupied, eligible patients are usually referred to another ward or hospital and scheduled surgeries can be cancelled. Bed shortages may also have consequences for ICU patients, who may be eligible for early discharge when the number of occupied beds is so high as to compromise the medical care of new incoming patients. Therefore, both patient admissions and discharges are triaged (Shmueli et al. (2003), Sinuff et al. (2004), Capuzzo et al. (2010) and Anderson et al. (2011), Chan et al. (2012, 2014).

Ridge et al. (1998) pointed out that the “early discharge” of less critical patients to other wards is a solution commonly adopted to cope with ICU bed shortages. Costa et al. (2003) observed the dynamical changes in ICU management when units become full and physicians attempt to limit admissions or to discharge patients who are closer to recovery. Although these studies suggest early discharge as a bed management tool, they do not include it in their mathematical models. The impact of workload on service time in health care has been considered by Kc and Terwiesch (2009), who show that hospital resources are sensitive to load levels and that service workers can adapt to system needs by increasing the service rate. Mallor and Azcárate (2014) demonstrated in a real setting that patient LoS is not independent of

the ICU workload and may be influenced by the ICU bed occupancy level. As a consequence, they signalled the need to include these discharge policies to obtain a valid simulation model.

The bed capacity planning problem has drawn much attention in ICU management studies, where various economic and quality-of-service (QoS) criteria have been considered. They include minimization of total operating costs, full demand coverage, reasonable bed utilization rates, a lower rejection rate, minimization of the number of surgery cancellations, maximization of the expected incremental number of lives saved. We consider the following two QoS criteria. The first is of a social nature: i.e., the percentage of population that can benefit from the ICU care when needed. The second is of an individual nature: i.e., the degree of recovery reached by a patient on discharge from the ICU. Thus, high quality of service in an ICU means both a low percentage of rejected patients and sufficient LoS.

This paper deals with the problem of obtaining efficient bed-management policies to control the ICU bed occupancy level. Queuing theory addresses the problem of resource allocation under uncertainty (Gross and Harris 2008) to provide queue designs and control policies that optimize some measure of interest. Here, we consider a different queuing control problem in which neither the arrival rates nor the number of servers can be modified. The bed occupancy control problem is addressed by adjusting the service time rates to make them dependent of the state of the system. Best values for these service times must then be determined to achieve the two QoS targets already mentioned: i.e., minimal patient rejection (due to full ICU occupancy) and minimal shortening of the LoS. The first objective has a clear mathematical formulation: minimize the probability of rejecting a patient. The second objective admits several formulations, which will be discussed later. Four different expressions, all leading to nonlinear optimization problems, are proposed. The analytical solutions to these problems-which define the desired management policies-are obtained and compared in terms of the two QoS components.

The paper is organized as follows. Section 2 presents a Markovian analysis of the ICU and provides some theoretical results for the ICU queuing model. We focus on the mathematical modelling of the ICU control problem, considering the two QoS components already mentioned. We propose four different nonlinear multi-objective problems for the Markovian case and give the efficient management policies based on the analytical solutions to the different control problems. Section 3 presents the extension of the control problems and the management policies to the general $G/G/c/c$ model and proposes a hazard function method to simulate general LoS distributions with service time dependent on the number of occupied beds. The optimal management policies are obtained by combining optimization with simulation. Finally, Sect. 4 contains the conclusions and final remarks regarding our work in progress and some suggestions for future research.

2 Markovian analysis for an ICU

2.1 Queuing modelling and analysis

As mentioned in the introduction, an ICU can be mathematically modelled as a queuing system, where the clients are the patients, the servers are the beds and there

is no waiting room. Then a general model that fits this system is a G/G/c/c queue model with several types of customers, each characterised by an arrival pattern and a service time. The complexity of this model generally demands the use of simulation to obtain a full and detailed probabilistic analysis. We first study a simplified version of the ICU queuing model by adopting Markovian assumptions on arrivals and service times. These assumptions could be considered not totally unrealistic in an ICU serving only emergency patients. Specifically, we assume a M/M/c/c queue model where arrival rate is a constant λ and the service rates $\boldsymbol{\mu} = (\mu_1, \dots, \mu_i, \dots, \mu_c)$ can be dependent on the state of the system i , that is, dependent on the number i of occupied beds.

Let $X(t)$ be the variable denoting the number of occupied beds at time t and $\mathbf{P} = (p_0, p_1, \dots, p_c)$ the stationary probability distribution of process $\{X(t)\}$ when the service is provided according to service rates $\boldsymbol{\mu} = (\mu_1, \dots, \mu_i, \dots, \mu_c)$.

Proposition *Given a constant arrival ratio λ and a probability distribution $\mathbf{P} > 0$, there is a unique set of service rates $\boldsymbol{\mu} = (\mu_1, \dots, \mu_c)$ such that the stationary probability distribution of process $\{X(t)\}$ is \mathbf{P} .*

Proof Straightforward from steady state equilibrium equations:

$$\mu_i = \lambda \frac{p_{i-1}}{ip_i} \quad \forall i = 1, \dots, c \quad \text{and} \quad \sum_{i=0}^c p_i = 1$$

□

This proposition shows that there exist service-rate values leading to any probability distribution of the number of occupied beds.

Let us consider a queuing model with input and service rates λ and μ , respectively, and denote the steady state probabilities by $p_i^* \ i = 0, \dots, c$. Suppose that a new target distribution for the state probabilities $p_i, \ i = 0, \dots, c$, is desired for the queuing model and that this new distribution has to be achieved by determining new, state-dependent service rates.

Then, by the previous proposition, the new service rates $\mu_i \ i = 1, \dots, c$ verify the expression:

$$\frac{\mu_i}{\mu} = \frac{p_{i-1}/p_i}{p_{i-1}^*/p_i^*} = \frac{p_{i-1}/p_{i-1}^*}{p_i/p_i^*} = \frac{\varphi_{i-1}}{\varphi_i} \tag{1}$$

where φ_i denotes the relative change with respect to the initial probabilities:

$$\varphi_i = p_i/p_i^*$$

This result relates the necessary changes in the service rates with the magnitude of change in the probabilities and shows that this change is a continuous function:

$$\forall \varepsilon > 0, \exists \delta > 0 / |\mu_i - \mu| < \varepsilon, \forall p_i^* / |p_i^* - p_i| < \delta$$

Let us now consider the LoS of a patient in an ICU when the service rates are dependent of the number of occupied beds. The transitions of the patient among the

different states of the queuing system can be modeled as a continuous-time Markov process with an absorbing state representing the patient’s discharge from the health facility.

Thus, a patient remains at state i during an exponentially distributed time with rate $(\lambda + i\mu_i)$ and then jumps to

- state $i + 1$ with probability $\lambda/(\lambda + i\mu_i)$
- state $i - 1$ with probability $(i - 1)\mu_i/(\lambda + i\mu_i)$
- the absorbing state with probability $\mu_i/(\lambda + i\mu_i)$

The LoS of a patient can be seen as the absorption time of that continuous-time Markov process in which the initial distribution for the states is the bed occupancy stationary distribution conditioned to a state smaller than c . We denote this initial distribution by vector α . Therefore, the LoS, denoted by variable Z , follows a Phase-type distribution characterized by the following transition rate matrix:

$$\begin{pmatrix} -\lambda - \mu_1 & \lambda & \dots & 0 & 0 & 0 \dots & 0 & 0 & \mu_1 \\ \vdots & \dots & (i - 1)\mu_i & -\lambda - i\mu_i & \lambda \dots & \vdots & 0 & \mu_i \\ 0 & 0 & \dots & \dots & \dots & 0 & 0 & 0 \end{pmatrix}$$

This matrix can be expressed as $\begin{pmatrix} \mathbf{T} & \mathbf{T}_0 \\ \mathbf{0} & 0 \end{pmatrix}$

Then it is well known that the distribution function of the LoS, Z , verifies the following expression:

$$P\{Z \leq x\} = F_Z(x) = 1 - \alpha \exp\{\mathbf{T}x\} \mathbf{e}$$

where \mathbf{e} is the vector with all components equal to 1.

The conditional stationary distribution α is expressed in terms of the steady state probabilities p_i as

$$\alpha = \left(p_0/1 - p_c, p_1/1 - p_c, \dots, p_{c-1}/1 - p_c \right)$$

which leads to the following nonlinear expression in terms of $\mu_i, i = 1, \dots, c$:

$$\alpha = (\gamma_1/\theta, \gamma_2/\theta, \dots, \gamma_c/\theta)$$

where:

$$\theta = \lambda^{c-1} + \sum_{i=1}^{c-1} \lambda^{c-i-1} \frac{(c - 1)!}{(c - i - 1)!} \prod_{j=c-i}^{c-1} \mu_j$$

and

$$\gamma_i = \lambda^{i-1} \frac{(c-1)!}{(i-1)!} \prod_{j=i}^{c-1} \mu_j \quad i = 1, \dots, c$$

The first two moments of the random variable Z representing the LoS are given by:

$$E[Z] = -(\boldsymbol{\alpha} \mathbf{T}^{-1} \mathbf{e})$$

$$E[Z^2] = 2(\boldsymbol{\alpha} \mathbf{T}^{-2} \mathbf{e})$$

Let be $\boldsymbol{\alpha}_i$ a c -dimensional vector with 1 in position $i + 1$ and the rest equal to 0. The substitution of vector $\boldsymbol{\alpha}$ by $\boldsymbol{\alpha}_i$ in the above expressions provides the distribution function of variable $Z_{(i)}$ representing the LoS conditioned on the patient having entered the system at state i , $F_{Z_{(i)}}(x) = 1 - \boldsymbol{\alpha}_i \exp\{\mathbf{T}x\} \mathbf{e}$, and the first two moments:

$$E[Z_{(i)}] = -(\boldsymbol{\alpha}_i \mathbf{T}^{-1} \mathbf{e})$$

$$E[Z_{(i)}^2] = 2(\boldsymbol{\alpha}_i \mathbf{T}^{-2} \mathbf{e})$$

These expressions are used in next section in the mathematical definition of the queue control problems.

2.2 Mathematical modeling of control problems

Queuing theory has studied the problem of resource allocation under uncertainty (Gross and Harris 2008) to provide queue designs and control policies that optimize some measure of interest, such as a customer's expected waiting time. One example of this type of control problem is the determination of a policy for switching workers in a facility with front room and back room operations in order to cope with changing customer demand (Terekhov and Beck 2008, and references therein). Another is how to control the queue by developing optimal admission policies taking customer behavior characteristics into account, which has led to the idea of controlling arrivals by pricing (see Stidham 2002).

We consider a type of control problem in which the objective is to maximize QoS (by reducing as far as possible the probability of patient rejection and to minimize the shortening of patient's LoS). Nevertheless, two ICU characteristics need to be taken into account: fixed resources and unadjustable patient arrival rates. Thus, in our queuing control problem, neither the arrival rate nor the amount of resources can be modified. Therefore, bed occupancy is controlled by modifying service time rates to make them state-dependent.

In this section, we study this control problem in a M/M/c/c queue model with constant arrival rate λ and "normal" service rate μ which can be adjusted to individual service rate values μ_i , when i beds are occupied, $i = 1, \dots, c$. Then, the aim of the control problem is to determine new values for service rates

$\mu_i, i = 1, \dots, c$, that minimize the probability of rejecting a new patient with minimum adjustment of the LoS of an admitted patient.

The first objective has a clear mathematical formulation: *Minimize* p_c . By the PASTA property (Poisson Arrivals See Time Averages), p_c is the probability of a patient finding a full ICU on arrival. The second objective admits different formulations. We propose two approaches, one based on the variation in services rates and another on the variation in the expected LoS. Expressions (2) and (3) are two alternative mathematical formulations for the “variation in services rates” objective functions. Expressions (4) and (5) are proposed for “variation in the expected LoS” objectives.

The first approach considers the magnitude by which the original service rate μ is adjusted in order to obtain the new service rates $\mu_i, i = 1, \dots, c$:

$$\min \max_i |\mu_i - \mu| \quad (2)$$

or

$$\min \sum_{i=1}^c |\mu_i - \mu| \quad (3)$$

Expression (4) minimizes the difference between the expected LoS of a patient in the ICU “with no control policy” (constant service rate μ) and the expected LoS “with control policy” (service rate $\mu_i, i = 1, \dots, c$ depending on the bed occupancy level i):

$$\min \left| -(\boldsymbol{\alpha} \mathbf{T}^{-1} \mathbf{e}) - \frac{1}{\mu} \right| \quad (4)$$

Expression (5) modifies (4) by considering the patient’s expected LoS conditioned by the state i of the ICU at arrival time.

$$\min \max_i \left| -(\boldsymbol{\alpha}_i \mathbf{T}^{-1} \mathbf{e}) - \frac{1}{\mu} \right| \quad (5)$$

Each of these formulations for the second objective leads to a different multi-objective optimization problem:

$$\begin{aligned} & \min P_c \\ & \min \max_i |\mu_i - \mu| \\ & \text{subject to } \mu_1 \leq \mu_2 \leq \dots \leq \mu_c \end{aligned} \quad (\text{MOP1})$$

$$\begin{aligned} & \min P_c \\ & \min \sum_{i=1}^c |\mu_i - \mu| \\ & \text{subject to } \mu_1 \leq \mu_2 \leq \dots \leq \mu_c \end{aligned} \quad (\text{MOP2})$$

$$\begin{aligned} & \min P_c \\ & \min \left| -(\boldsymbol{\alpha}T^{-1}\mathbf{e}) - \frac{1}{\mu} \right| \\ & \text{subject to } \mu_1 \leq \mu_2 \leq \dots \leq \mu_c \end{aligned} \quad (\text{MOP3})$$

$$\begin{aligned} & \min P_c \\ & \min \max_i \left| -(\boldsymbol{\alpha}_i T^{-1}\mathbf{e}) - \frac{1}{\mu} \right| \\ & \text{subject to } \mu_1 \leq \mu_2 \leq \dots \leq \mu_c \end{aligned} \quad (\text{MOP4})$$

Observe that monotonicity constraints on the values of μ_i reflect the fact that service times become shorter as the ICU gets busier.

We estimate the Pareto frontier by using the ε -constraint method and considering different bounds for the probability of rejected patients (ε_j -values):

$$\begin{aligned} & \min \max_i |\mu_i - \mu| \\ & \text{subject to } \begin{cases} P_c \leq \varepsilon_j \\ \mu_1 \leq \mu_2 \leq \dots \leq \mu_c \end{cases} \end{aligned} \quad (\text{P1})$$

$$\begin{aligned} & \min \sum_{i=1}^c |\mu_i - \mu| \\ & \text{subject to } \begin{cases} P_c \leq \varepsilon_j \\ \mu_1 \leq \mu_2 \leq \dots \leq \mu_c \end{cases} \end{aligned} \quad (\text{P2})$$

$$\begin{aligned} & \min \left| -(\boldsymbol{\alpha}T^{-1}\mathbf{e}) - \frac{1}{\mu} \right| \\ & \text{subject to } \begin{cases} P_c \leq \varepsilon_j \\ \mu_1 \leq \mu_2 \leq \dots \leq \mu_c \end{cases} \end{aligned} \quad (\text{P3})$$

$$\begin{aligned} & \min \max_i \left| -(\boldsymbol{\alpha}_i T^{-1}\mathbf{e}) - \frac{1}{\mu} \right| \\ & \text{subject to } \begin{cases} P_c \leq \varepsilon_j \\ \mu_1 \leq \mu_2 \leq \dots \leq \mu_c \end{cases} \end{aligned} \quad (\text{P4})$$

All these models are nonlinear optimization problems with the following structure:

All optimization problems have c decision variables $(\mu_1, \mu_2, \dots, \mu_c)$. The nonlinearity of the problems comes from the absolute value functions and from the term P_c , which is given by the following nonlinear expression:

$$P_c = \frac{\lambda^c}{\lambda^c + \sum_{i=1}^c (\lambda^{c-i} i! \binom{c}{i} \prod_{j=c-i+1}^c \mu_j)} \tag{6}$$

Optimization problem (P3) also includes the following nonlinear function:

$$(\alpha T^{-1} e) = \frac{\beta_1 + \sum_{i=2}^c \beta_i \prod_{j=c-i+2}^c \mu_j}{\sum_{i=1}^c \beta_i \prod_{j=c-i+1}^c \mu_j} \tag{7}$$

where β_i are constant values depending on λ and c .

Expression $(\alpha_i T^{-1} e)$ in optimization problem (P4) is a complex, nonlinear function, which, for $i = c$, includes a monomial for each of the proper subsets of $\mu_1, \mu_2, \dots, \mu_c$, that is, $2^c - 2$ terms.

Furthermore, (P1) and (P4) have min max objective functions.

Observe that some of the nonlinear terms in the above problems admit reformulation. For example, minmax objective functions admit linear reformulation by considering an additional variable (δ), as illustrated in the following expression for the problem (P1) case:

$$\begin{aligned} & \min \delta \\ & \text{subject to } \begin{cases} |\mu_i - \mu| \leq \delta & i = 1, \dots, c \\ P_c \leq \varepsilon_j \\ \mu_1 \leq \mu_2 \leq \dots \leq \mu_c \end{cases} \end{aligned}$$

Let $dev_i, i = 1, \dots, c$ be the deviational variables associated with the deviation from 1 of ratios $\mu_i/\mu, i = 1, \dots, c$. Optimization problems (P1) and (P2) can also be reformulated [(P1b) and (P2b)] by considering both the steady state equilibrium equations and expression (1).

$$\min \max_j dev_j \tag{P1b}$$

or

$$\min \sum_{j=1}^c dev_j \tag{P2b}$$

$$\text{subject to } \begin{cases} \varphi_c P_c^* \leq \varepsilon_j \\ \frac{\varphi_{i-1}}{\varphi_i} - dev_i = 1, & i = 1, \dots, c \\ \sum_{i=0}^c \varphi_i P_i^* = 1 \\ \frac{\varphi_{i-1} \varphi_{i+1}}{\varphi_i^2} \leq 1, & \forall i = 1, \dots, c \end{cases}$$

This nonlinear formulation has $2c + 1$ variables ($\varphi_0, \dots, \varphi_c, dev_1, \dots, dev_c$) and avoids the calculation of the term P_c in expression (6).

2.3 Efficient management policies for the control problem

Management policies are obtained from the solution of the mathematical problems formulated in the previous section. Different optimization problems lead to different solutions and then to different management policies. Solutions differ not only in service rate values but also in structure, with management philosophy implications. In some cases, the only dramatical increase in the service rate occurs at full occupancy, while, in the other cases, the service rates are moderately and evenly increased in all states. We label these two types of policies *aggressive* and *equitable*, respectively. Another, intermediate structure is also found: the higher the occupancy level, the higher the service rate. We label this last policy as *cautious*.

To illustrate these results we study a small ICU with five beds and then a larger one with 20 beds, based on a real ICU analyzed in a previous study by the authors.

The 5-bed ICU is represented by the M/M/5 queuing model. Without loss of generality, in our study we consider the time unit as the average LoS, so that $\mu = 1$. Then, the different studied scenarios are defined by varying the values of λ to achieve certain desired occupancy ratios. In this way values of $\mu = 1$ and $\lambda = 4, 3$ and 2 lead to occupancy rates of 80, 60 and 40 %, respectively. The four optimization problems are solved in each scenario. Optimization problems (P3) and (P4) always provide aggressive policies, (P1) leads to equitable policies and (P2) leads to cautious policies. Results are presented in Tables 1, 2 and 3.

Table 1 shows the results for $\lambda = 4$ (server utilization = 80 %) and $P_5^* = 0.199067$. Recall that this means that 19.9 % of patients are rejected due to full ICU occupancy, in the absence of any control by physicians. To estimate the Pareto frontier by using the ε -constraint method, we consider the following ε_j -values, $\varepsilon_j = 19, 16, 13, 10, 7, 4$ and 1 %. Table 2 shows the results for $\lambda = 3$ (server utilization = 60 %) and $P_c^* = 0.110054$, considering the following ε_j -values, $\varepsilon_j = 10, 8, 6, 4, 2$ and 1 %. Table 3 shows the results for $\lambda = 2$ (server utilization = 40 %) and $P_c^* = 0.036697$, considering the following ε_j -values, $\varepsilon_j = 3, 2.5, 2, 1.5, 1$ and 0.5 %.

In these results, we see the three policy structures previously defined:

- Aggressive policy: $\mu_1 = \mu_2 = \dots = \mu_{c-1} = 1$ and $\mu_c \gg 1$ in (P3) [and also (P4)] results.
- Equitable policy: $\mu_1 = \mu_2 = \dots = \mu_c > 1$ in (P1) results.
- Cautious policy: $1 \leq \mu_1 \leq \mu_2 \leq \dots \leq \mu_c$ in (P2) results.

The same pattern appears in the results for the simplified version of a real ICU (the Hospital of Navarre, Spain) with 20 beds (M/M/20/20 queuing model).

In this case, we simplify the problem by distinguishing three levels for the bed occupancy state:

- low occupancy (<50 %): 1–10 occupied beds,
- moderate occupancy (50–75 %): 11–14 occupied beds,
- high occupancy (≥ 75 %): 15–20 occupied beds.

Table 1 Optimal management policies for a M/M/5/5, with $\lambda = 4$ and $\mu = 1$ (server utilization = 80 %)

	μ_1	μ_2	μ_3	μ_4	μ_5	Objective function
<i>Aggressive policies</i>						
19 %	1	1	1	1	1.0596	0.0175
16 %	1	1	1	1	1.3049	0.0726
13 %	1	1	1	1	1.6633	0.1239
10 %	1	1	1	1	2.2367	0.1718
7 %	1	1	1	1	3.3018	0.2166
4 %	1	1	1	1	5.9644	0.2586
1 %	1	1	1	1	24.6058	0.2981
<i>Equitable policies</i>						
19 %	1.0260	1.0260	1.0260	1.0260	1.0260	0.0260
16 %	1.1219	1.1219	1.1219	1.1219	1.1219	0.1219
13 %	1.2386	1.2386	1.2386	1.2386	1.2386	0.2386
10 %	1.3884	1.3884	1.3884	1.3884	1.3884	0.3884
7 %	1.5977	1.5977	1.5977	1.5977	1.5977	0.5977
4 %	1.9443	1.9443	1.9443	1.9443	1.9443	0.9443
1 %	2.9395	2.9395	2.9395	2.9395	2.9395	1.9395
<i>Cautious policies</i>						
19 %	1	1	1	1	1.0596	0.0596
16 %	1	1	1	1	1.3049	0.3049
13 %	1	1	1	1.1027	1.5534	0.6561
10 %	1	1	1	1.3507	1.8014	1.1521
7 %	1	1	1	1.7380	2.1887	1.9267
4 %	1	1	1.4516	2.2721	2.6332	3.3568
1 %	1	1.3959	2.9959	3.7080	3.9000	7.9997

We assume:

$$\mu_1 = \mu_2 = \dots = \mu_{10} \leq \mu_{11} = \mu_{12} = \dots = \mu_{14} \leq \mu_{15} \leq \mu_{16} \leq \mu_{17} \leq \mu_{18} \leq \mu_{19} \leq \mu_{20}$$

This assumption reduces the number of decision variables, thereby simplifying both the computational analysis and the presentation of results.

Table 4 includes the results for $\lambda = 16$ and $\mu = 1$ (server utilization = 80 %) and $P_c^* = 0.064411$. We consider the following ε_j -values, $\varepsilon_j = 6, 5, 4, 3, 2, 1$ %.

The question that clearly arises is which of the three types of ICU management policy is the best. Each is optimal in terms of one specific aspect of LoS shortening but none is optimal in all the LoS shortening objectives. Determining which is the best or the most appropriate management policy is therefore no easy task. The decision must be based on the performance assessment of each policy type in all the different LoS shortening-related measures and on the judgment of physicians.

Table 2 Optimal management policies for a M/M/5/5, with $\lambda = 3$ and $\mu = 1$ (server utilization = 60 %)

	μ_1	μ_2	μ_3	μ_4	μ_5	Objective function
<i>Aggressive policies</i>						
10 %	1	1	1	1	1.1130	0.0209
8 %	1	1	1	1	1.4221	0.0612
6 %	1	1	1	1	1.9374	0.0997
4 %	1	1	1	1	2.9679	0.1367
2 %	1	1	1	1	6.0595	0.1721
1 %	1	1	1	1	12.2420	0.1893
<i>Equitable policies</i>						
10 %	1.0413	1.0413	1.0413	1.0413	1.0413	0.0413
8 %	1.1388	1.1388	1.1388	1.1388	1.1388	0.1388
6 %	1.2680	1.2680	1.2680	1.2680	1.2680	0.2680
4 %	1.4582	1.4582	1.4582	1.4582	1.4582	0.4582
2 %	1.8103	1.8103	1.8103	1.8103	1.8103	0.8103
1 %	2.2046	2.2046	2.2046	2.2046	2.2046	1.2046
<i>Cautious policies</i>						
10 %	1	1	1	1	1.1130	0.1130
8 %	1	1	1	1.0788	1.3384	0.4172
6 %	1	1	1	1.3026	1.5622	0.8647
4 %	1	1	1.1001	1.6295	1.8732	1.6028
2 %	1	1	1.6175	2.1469	2.3318	3.0962
1 %	1	1.1015	2.2265	2.7318	2.8705	4.9304

Comparison of the three management strategies shows that *cautious* policies are the best for their associated objective ($\min \sum_{i=1}^c |\mu_i - \mu|$) and the second best for the other objectives. *Aggressive* policies are the best for the objectives associated with maximizing average LoS, the second best for minimizing the maximum deviation from the original service rate and the worst for the sum of service rate deviations. *Equitable* policies are the best for minimizing maximum deviation from the original service rate, the second best for the sum of service rate deviations, and the worst for the objectives associated with maximizing average LoS. See Fig. 1. From this analysis, *cautious* policies appear to display the best behavior overall. Furthermore, ICU physicians agree that *cautious* policies are the best choice in bed shortage situations. In fact, the representation of the bed-occupancy distribution under the different policies (Fig. 2) shows that the one corresponding to the *cautious* policy is the closest to histograms built from historical occupancy data collected in real ICUs [see, for example, Costa et al. (2003) and Mallor and Azcárate (2014)]. Common characteristics in both the cautious distribution and real data distributions are a sharp decline on the right side of the histogram and a skewed histogram with a peak around a set of desirable bed occupancy values (leading to high utilization of resources without compromising the entry of new patients).

Table 3 Optimal management policies for a M/M/5/5, with $\lambda = 2$ and $\mu = 1$ (server utilization = 40 %)

	μ_1	μ_2	μ_3	μ_4	μ_5	Objective function
<i>Aggressive policies</i>						
3 %	1	1	1	1	1.2317	0.1792
2.5 %	1	1	1	1	1.4857	0.3114
2 %	1	1	1	1	1.8667	0.4422
1.5 %	1	1	1	1	2.5016	0.0572
1 %	1	1	1	1	3.7714	0.0700
0.5 %	1	1	1	1	7.5810	0.8268
<i>Equitable policies</i>						
3 %	1.0665	1.0665	1.0665	1.0665	1.0665	0.0665
2.5 %	1.1285	1.1285	1.1285	1.1285	1.1285	0.1285
2 %	1.2069	1.2069	1.2069	1.2069	1.2069	0.2069
1.5 %	1.3122	1.3122	1.3122	1.3122	1.3122	0.3122
1 %	1.4697	1.4697	1.4697	1.4697	1.4697	0.4697
0.5 %	1.7667	1.7667	1.7667	1.7667	1.7667	0.7667
<i>Cautious policies</i>						
3 %	1	1	1	1.0615	1.1668	0.2283
2.5 %	1	1	1	1.1762	1.2814	0.4576
2 %	1	1	1.0438	1.3104	1.4122	0.7664
1.5 %	1	1	1.1916	1.4583	1.5497	1.1996
1 %	1	1	1.4224	1.6891	1.7680	1.8796
0.5 %	1	1.1785	1.8451	2.0860	2.1438	3.2534

3 Control problems and management policies in the general case G/G/c/c

Section 2 modeled the control problem of a health care service through a set of optimization problems which were solved by means of classical mathematical programming. This setting could be applied to the Markovian case. When other, non-exponential distributions are considered, the analytical expressions for the second objective in problems (P3) and (P4) do not hold and even the decision variables μ_i need to be reinterpreted. This section addresses these questions, first by assessing the combination of optimization with simulation as an appropriate optimization problem-solving tool and then by providing a way to simulate general LoS distributions allowing the inclusion of state-dependent service rates. We then present the management policies obtained from this new optimization and simulation framework when different Weibull-type LoS distributions are considered and compare them with those obtained in the Markovian case.

Table 4 Optimal management policies for the Markovian version of the ICU of the Hospital of Navarre, with 20 beds

	μ_1	μ_{11}	μ_{15}	μ_{16}	μ_{17}	μ_{18}	μ_{19}	μ_{20}
<i>Aggressive policies</i>								
6 %	1	1	1	1	1	1	1	1.0786
5 %	1	1	1	1	1	1	1	1.3086
4 %	1	1	1	1	1	1	1	1.6522
3 %	1	1	1	1	1	1	1	2.2260
2 %	1	1	1	1	1	1	1	3.3724
1 %	1	1	1	1	1	1	1	6.8166
<i>Equitable policies</i>								
6 %	1.0410	1.0410	1.0410	1.0410	1.0410	1.0410	1.0410	1.0410
5 %	1.0492	1.0492	1.0492	1.0492	1.0492	1.0492	1.0492	1.0492
4 %	1.0910	1.0910	1.0910	1.0910	1.0910	1.0910	1.0910	1.0910
3 %	1.1431	1.1431	1.1431	1.1431	1.1431	1.1431	1.1431	1.1431
2 %	1.2138	1.2138	1.2138	1.2138	1.2138	1.2138	1.2138	1.2138
1 %	1.3299	1.3299	1.3299	1.3299	1.3299	1.3299	1.3299	1.3299
<i>Cautious policies</i>								
6 %	1	1	1	1	1	1	1	1.0786
5 %	1	1	1	1	1	1	1.1029	1.1963
4 %	1	1	1	1	1	1.0837	1.2096	1.2972
3 %	1	1	1	1	1.0452	1.2102	1.3314	1.4081
2 %	1	1	1	1	1.2096	1.3746	1.4813	1.5419
1 %	1	1	1	1.2616	1.4742	1.6099	1.6848	1.7223

3.1 Finding optimal policies by combining optimization with simulation

Optimization problems (P1) to (P4) include the expected LoS, $E[Z]$, the expected conditional LoS, $E[Z_{(i)}]$, and the probability of rejecting a patient, P_{rej} , which have an explicit expression in terms of the decision variables in the Markovian case but not in the general case. The formulation of the optimization problems for the general case is:

$$\begin{aligned}
 & \min \max_i |\mu_i - \mu| \\
 & \text{subject to } \begin{cases} P_{rej} \leq \varepsilon_j \\ \mu_1 \leq \mu_2 \leq \dots \leq \mu_c \end{cases} \tag{GP1}
 \end{aligned}$$

$$\begin{aligned}
 & \min \sum_{i=1}^c |\mu_i - \mu| \\
 & \text{subject to } \begin{cases} P_{rej} \leq \varepsilon_j \\ \mu_1 \leq \mu_2 \leq \dots \leq \mu_c \end{cases} \tag{GP2}
 \end{aligned}$$

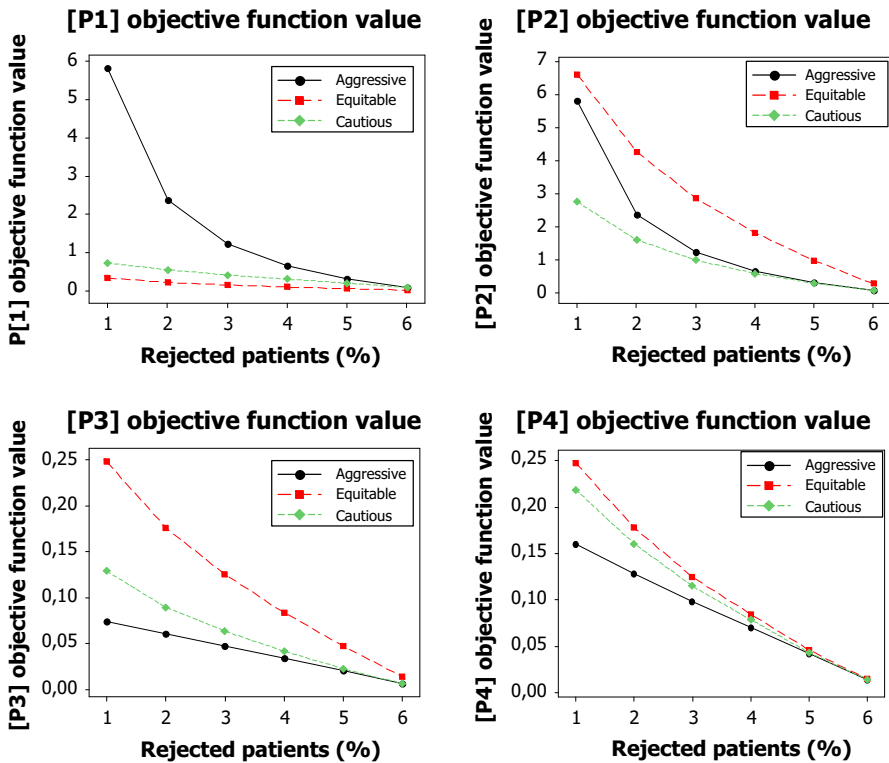


Fig. 1 Pareto frontiers for the optimal management policies

$$\begin{aligned}
 & \min \left| E[Z] - \frac{1}{\mu} \right| \\
 & \text{subject to } \begin{cases} P_{rej} \leq \varepsilon_j \\ \mu_1 \leq \mu_2 \leq \dots \leq \mu_c \end{cases}
 \end{aligned} \tag{GP3}$$

$$\begin{aligned}
 & \min \max_i \left| E[Z_{(i)}] - \frac{1}{\mu} \right| \\
 & \text{subject to } \begin{cases} P_{rej} \leq \varepsilon_j \\ \mu_1 \leq \mu_2 \leq \dots \leq \mu_c \end{cases}
 \end{aligned} \tag{GP4}$$

Now objective functions in (GP1) and (GP2) and the constraint on the probability of rejecting a patient $P_{rej} \leq \varepsilon_j$ have to be assessed by simulation. To solve this set of problems, we combine simulation with an optimization procedure, which determines a solution, that is, a value for the decision variables μ_i of the problem which define the configuration of the simulated system. The output of the simulation is used to evaluate the stochastic elements and assess the quality of the current solution. The optimization procedure, with this information and its search method, provides the next solution to be evaluated by simulation. This iterative process goes

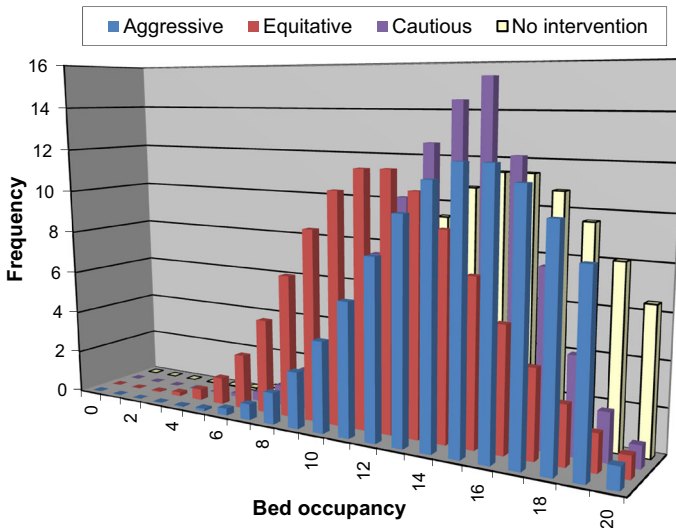


Fig. 2 Bed occupancy distribution plots for the different optimal management policies. Scenario with 1 % probability of rejecting a patient

on until the stopping conditions of the optimization method are met. The simulation-based optimization (SBO) methodology just described has been widely applied to solve stochastic optimization problems in different contexts, including healthcare. For example, in de Angelis et al. (2003) a simulation model is combined with nonlinear programming and neuronal networks to determine the optimal configuration of a transfusion center; in Azcárate et al. (2008) a multicriteria optimization simulation model is proposed to solve a hospital sizing problem; in Ahmed and Alkhamis (2009) one is used to obtain the optimal staff distribution for a hospital emergency unit; in Brailsford et al. (2007) a discrete-event simulation model is embedded in an ant colony optimization model for the optimal choice of screening policies for diabetic retinopathy; in Lin et al. (2013) simulation is combined with a genetic algorithm and data envelopment analysis to determine optimal resource levels in surgical services. The reader is referred to Fu et al. (2005) for a descriptive review of the main approaches for SBO.

The SBO methodology to solve the above control problems is validated by solving the Markovian case and comparing the results with the analytical ones. The SOB technique was implemented in ARENA simulation software and OptQuest optimization software. Figure 3 plots Pareto frontiers for optimization problems obtained by nonlinear programming techniques (black lines) and simulation and optimization techniques (red lines).

Figure 4 plots the relative error between analytical and simulated optimal solutions, expressed as

$$\frac{|\mu_{analytical_i} - \mu_{simul_i}|}{\mu_{analytical_i}}$$

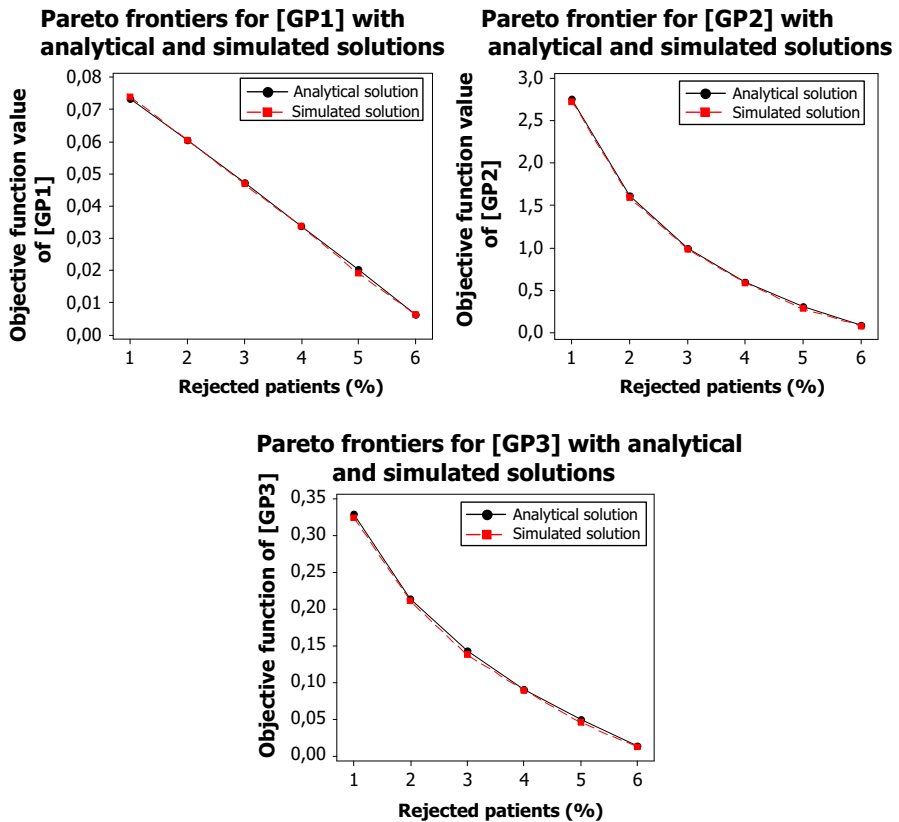


Fig. 3 Pareto frontiers for optimization problems obtained by nonlinear programming techniques (*black lines*) and simulation and optimization techniques (*red lines*)

It can be observed that, in general, simulation provides a good estimation of the analytical solutions. The relative error in almost all cases is less than 3 %, only a few cases for (GP2) show values between 4 and 6 %. Comparing the optimal objective functions, the differences between analytical and simulated results are less than 0.01. Thus, we can conclude that the SBO methodology is able to solve the proposed control problems and obtain the management policies.

3.2 Simulation from general distributions with varying service rates

The application of the SBO methodology outlined in previous section requires the simulation of queuing systems with varying service rates. In the Markovian case, the procedure is straightforward due to the lack of memory of the exponential distribution. When the system enters a new state i , which has to operate with a service rate of μ_i , the remaining LoS of each patient is updated by sampling a value from an exponentially-distributed random variable T_{μ_i} with mean $1/\mu_i$. Assuming that the LoS is initially distributed as an exponential random variable T_{μ} with mean

Relative error between optimal solutions (analytical vs simulated)

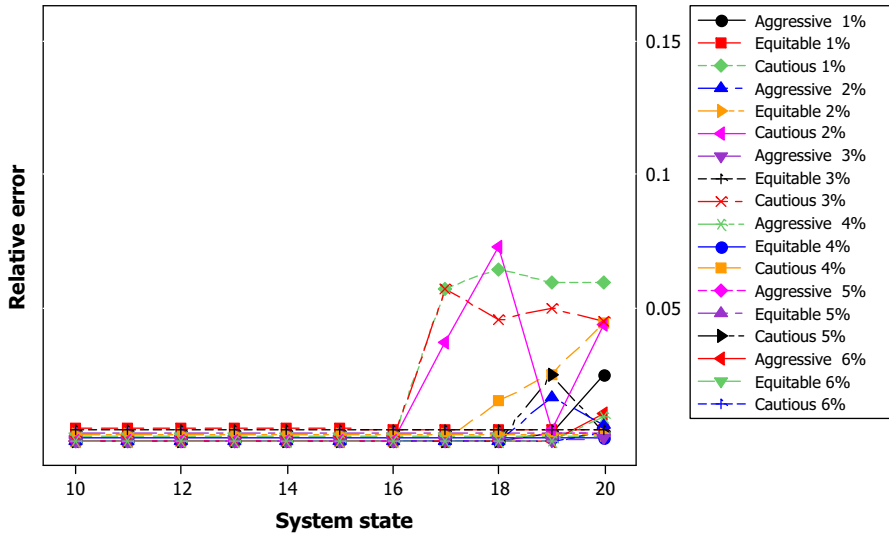


Fig. 4 Relative error between analytical and simulated optimal solutions

$1/\mu$, the variation to another service rate μ_i can be interpreted as a change in the time scale:

$$P\{T_{\mu_i} \leq t\} = 1 - e^{-\mu_i t} = 1 - e^{-\frac{\mu_i}{\mu} \mu t} = P\left\{T_{\mu} \leq \frac{\mu_i}{\mu} t\right\} = P\left\{\frac{\mu}{\mu_i} T_{\mu} \leq t\right\}$$

Then, $T_{\mu_i} = \frac{\mu}{\mu_i} T_{\mu}$ and $F_{T_{\mu_i}}(t) = F_{T_{\mu}}\left(\frac{\mu_i}{\mu} t\right)$.

When the system enters state i with an associated service rate μ_i the remaining LoS of that patient is adjusted by a scale factor of μ/μ_i .

This relationship is extended to the hazard function:

$$h_{T_{\mu_i}}(t) = \frac{F'_{T_{\mu_i}}(t)}{1 - F_{T_{\mu_i}}(t)} = \frac{F'_{T_{\mu}}\left(\frac{\mu_i}{\mu} t\right)}{1 - F_{T_{\mu}}\left(\frac{\mu_i}{\mu} t\right)} = \frac{\mu_i}{\mu} h_{T_{\mu}}\left(\frac{\mu_i}{\mu} t\right) \tag{8}$$

The hazard function at time t describes the probabilistic behaviour of a random variable at time t conditioned to a value greater than t :

$$h_{T_{\mu_i}}(t) = \lim_{\Delta t \rightarrow 0} \frac{P\{t < T_{\mu_i} \leq t + \Delta t / T_{\mu_i} > t\}}{\Delta t}$$

Then, the probability that a patient's LoS falls within the time interval $(t, t + \Delta t]$, given that at time t the patient is still in the health system, can be approximated by $\frac{h_{T_{\mu_i}}(t+\Delta t) + h_{T_{\mu_i}}(t)}{2} \Delta t$.

This expression is obtained from the two first terms of the Taylor series expansion [(1) in (9)] and numerical integration by trapezoidal method [(2) in (9)].

$$\begin{aligned} P\{T_{\mu_i} \in (t, t + \Delta t] / T_{\mu_i} > t\} &= \\ &= 1 - e^{-\int_t^{t+\Delta t} h_{T_{\mu_i}}(s) ds} \stackrel{(1)}{\approx} \int_t^{t+\Delta t} h_{T_{\mu_i}}(s) ds \stackrel{(2)}{\approx} \frac{h_{T_{\mu_i}}(t + \Delta t) + h_{T_{\mu_i}}(t)}{2} \Delta t \end{aligned} \quad (9)$$

This interpretation of the hazard function, together with expression (8), is used to simulate the LoS patients in the health system when both the general LoS distributions and varying service rate are considered.

As usual, the discrete event simulation model considers both patient arrival and end of service as state-changing events. However, the simulation is now driven by the patient LoS hazard functions as follows:

Step 0 Choose a small value for Δt .

Step 1 Simulate a time for the next patient arrival.

Step 2 If less than Δt time units left to a new patient arrival, then the system is updated, the clock is advanced to the arrival time of the new patient and step 1 is repeated. Let t_e be the patient's accelerated time in system. At the entry time of a patient, $t_e = 0$.

Step 3 If the system is in state i at time t (bed occupancy level is i) the hazard function $h_{T_{\mu_i}}(t_e)$ of the LoS of each patient is calculated, according to expression (8).

Step 4 For each patient the probability of leaving the health system in the next time interval of length Δt is calculated according to expression (9) and is used to decide whether the patient leaves or not the system.

Step 5 If, as a result of the simulation in step 4, none of the patients leaves the system, then the simulation clock is advanced by Δt time units and step 2 is repeated.

Step 6 If, as result of step 4, a patient leaves the system in the interval $(t, t + \Delta t)$, then the patient is discharged at a time uniformly distributed between t and $t + \Delta t$. The state of the system is updated and the simulation clock is advanced by Δt . Repeat step 2.

Each time the simulation clock is advanced, the accelerated time in system of a patient is updated according to expression:

$$t_e \leftarrow t_e + (t - t_e) \frac{\mu_i}{\mu}$$

These results allow us to use discrete-event simulation to simulate the $G/G/c/c$ queuing model with varying service times dependent on the number of occupied beds.

For example, in the case of the Weibull distribution with parameters α and β , a patient’s probability of leaving the health system during the next Δt time interval, if the current bed level occupancy is i , is calculated using the hazard function

$$h_{T_{\mu_i}}(t) = \frac{\mu_i}{\mu} \frac{\alpha}{\beta} \left(\frac{\mu_i}{\mu\beta} t \right)^{\alpha-1}.$$

Of course, the smaller step Δt , the greater the accuracy of the approximation, but also the greater computational effort because more steps are necessary to progress through time. Thus a compromise solution has to be adopted to balance accuracy and computational effort.

We have tested this method with the exponential distribution and obtained very good results taking 10 min-steps for average LoS = 1 day. Nevertheless, the accuracy of the approximation in (9) not only depends on the size of Δt , but also on the shape of the hazard function $h(t)$. When $h(t)$ is upper bounded, the error term in approximation (1) in (9) is proportional to $(\Delta t)^2$. This is the case for the usual distribution functions used to model LoS (loglogistic, lognormal and phase-type). When $h(t)$ increases as a linear function ($h(t) \sim ct$) then, the error term is proportional to $t\Delta t$. This is the case for Weibull distribution with shape parameter $a = 2$. Exponential increments in the hazard function produces large error terms, as for example in the Smallest Extreme Value distribution. However, this type of distribution has not been reported on the statistical modelling of LoS studies. The approximation error in trapezoidal method depends on $h''(t)$, the second derivative in the interval $(t, t + \Delta t)$:

$$|error| < k \frac{(\Delta t)^3}{48}, \text{ under the assumption } |h''(t)| < k.$$

We conduct a computational study to analyze the stability of the estimations of the rejection probability depending on the size of parameter Δt . Table 5 shows the results for different distribution functions and considering the optimal service rates for a rejection value of 4 % for the M/M/5/5 model, with $\lambda = 4$ and $\mu = 1$ (see Table 1). The distributions considered are: Exponential (with both Markovian and non-Markovian arrivals), Weibull with shape parameter $\alpha = 5$, Loglogistic with scale parameters $\sigma = 0.1$ and $\sigma = 0.6$, Smallest Extreme Value distribution with scale parameters $\sigma = 1.1$ and $\sigma = 1.5$. Six values for parameter Δt are analysed

Table 5 Estimations of the rejection probability with different sizes of parameter Δt

	Δt (min)					
	60	30	20	10	5	1
M/M/5/5	4.189	4.092	4.062	4.023	4.013	4.002
G/M/5/5	1.702	1.686	1.675	1.671	1.664	1.669
M/Weib ($\alpha = 5$)/5/5	4.109	4.046	4.028	4.014	4.010	4.006
M/Loglog ($\sigma = 0.1$)/5/5	4.008	3.996	3.997	3.999	3.999	3.999
M/Loglog ($\sigma = 0.6$)/5/5	3.680	3.810	3.858	3.924	3.953	3.985
M/SEV ($\sigma = 1.1$)/5/5	4.165	4.085	4.060	4.040	4.030	4.021
M/SEV ($\sigma = 1.5$)/5/5	8.602	8.485	8.444	8.404	8.394	8.377

(60, 30, 20, 10, 5 and 1 min), representing 4.2, 2.1, 1.39, 0.69, 0.35 and 0.069 % of the expected LoS, respectively. The simulations were extended to attain a confidence interval with precision ± 0.01 . A rapid convergence in the estimation and a small difference between the values obtained for $\Delta t = 10, 5, 1$ is observed for the Weibull and Loglogistic distributions. Nevertheless, according to above comments, the Smallest Extreme Value distribution presents a slower convergence.

3.3 Results for the general model

The methodology presented in Sects. 3.1 and 3.2 can be used to find optimal management policies in any general ICU model including elective patients and more realistic LoS distributions. As an example, we have chosen a 5-bed ICU (M/G/5/5 queuing model) and several distributions for modelling the patient LoS: Weibull, Loglogistic and Phase-type distributions. The Weibull distribution with shape parameter $\alpha = 1$ provides the exponential distribution, when $\alpha > 1$ ($\alpha < 1$) the hazard function is monotonically increasing (decreasing). As the shape parameter becomes closer to 1, the Weibull distribution gets closer to the exponential distribution. Loglogistic distribution, as the lognormal, is frequently used to model LoS. Its hazard function reaches a maximum and then decreases asymptotically to zero. The Phase-type distributions allow for a more realistic modelling of the LoS of a patient by associating each state of the distribution to a different health status. The influence of the distribution of the arrival process is also considered by mixing an arrival Poisson process for outpatients with a Deterministic arrival process for elective patients (G/G/5/5 queuing model).

Modelling the LoS by a phase-type distribution can prevent from discharge a patient not sufficiently recovered. In this study, we consider a phase-type distribution with three states. When a patient enters to the ICU, the Markov chain is in state 1, representing a very bad health condition. States 2 and 3 are used to represent better health status, in such a way that a patient can only be discharged from state 3. To model the clinical worsening of the patient's health (due to infections, for example) the transition from upper to lower states is allowed (Fig. 5).

Control problems (GP1) to (GP4) are solved by using optimization with simulation and the method presented in Sect. 3.2. Tables 6, 7, 8 show the results for several M/G/5/5 queuing models, considering different LoS distributions (Weibull with shape parameter $\alpha = 5$, Loglogistic with scale parameters $\sigma = 0.2$ and 3-phase-type distribution with $q_2 = 0.2$ and $q_3 = 0.15$). We study the phase-type distribution model allowing the shortening of a patient stay in both any phase-state (case 1) and only in state 3 (case 2), a more realistic situation. G/G/5/5 queuing

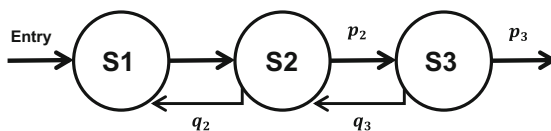


Fig. 5 Phase-type distribution diagram

Table 6 Aggressive policies: Optimal management policies for (GP3) and (GP4) control problems, for different G/G/5/5 models and percentage of rejected patients $\varepsilon_j = 16, 10,$ and 4%

	ε_j	μ_1	μ_2	μ_3	μ_4	μ_5
M/Weib ($\alpha = 5$)	16 %	1	1	1	1	1.31
M/Loglog ($\sigma = 0.1$)	16 %	1	1	1	1	1.3
M/Phase (case 1)	16 %	1	1	1	1	1.32
M/Phase (case 2)		1	1	1	1	1.85
G/M/		1	1	1	1	1
G/Weib ($\alpha = 5$)		1	1	1	1	1
M/Weib ($\alpha = 5$)	10 %	1	1	1	1.36	1.8
M/Loglog ($\sigma = 0.1$)	10 %	1	1	1	1.3	1.83
M/Phase (case 1)	10 %	1	1	1	1.39	1.77
M/Phase (case 2)		1	1	1.35	2.26	6.6
G/M/		1	1	1	1.15	1.26
G/Weib ($\alpha = 5$)		1	1	1	1.15	1.26
M/Weib ($\alpha = 5$)	4 %	1	1	1.49	2.37	2.5
M/Loglog ($\sigma = 0.1$)	4 %	1	1	1.4	2.47	2.47
M/Phase (case 1)		1	1	1.43	2.47	2.47
M/Phase (case 2)		Unfeasib	Unfeasib	Unfeasib	Unfeasib	Unfeasib
G/M/		1	1	1.1	1.81	1.81
G/Weib ($\alpha = 5$)		1	1	1.1	1.81	1.81

model is also analysed by mixing a Poisson process with rate arrival $\lambda = 2$ and 2 elective patients per day. This general arrival process is considered for both exponential and Weibull ($\alpha = 5$) LoS distributions. The parameters of the distributions has been adjusted to provide a distribution with mean $1 = E(G) = 1/\mu$. In all cases the arrival rate is constant, $\lambda = 4$. The Erlang Loss expression, provides the probability of a full ICU, $P_c^* = 0.199067$. The optimization problems are solved for percentage of rejected patients of $\varepsilon_j = 16, 10,$ and 4% .

An important conclusion is that the same structure of solutions is obtained for the general case than for the Markovian case. *Aggressive, equitable* and *cautious* policies are obtained as solutions to the optimization problems. Furthermore, the values that determine the service rates are very similar to those provided by the exponential distribution (see Table 1) for all the Markovian arrival models but the Phase-type distribution (case 2). Results for the non-Markovian arrival case also differ from the M/M/5/5 model.

4 Conclusions and final remarks

In this paper we have formulated a new control problem to represent the triage process that many physicians frequently face: the allocation of the available health

Table 7 Equitable policies: Optimal management policies for (GP1) control problem, for different G/G/5/5 models and percentage of rejected patients $\varepsilon_j = 16, 10,$ and 4%

	ε_j	μ_1	μ_2	μ_3	μ_4	μ_5
M/Weib ($\alpha = 5$)/	16 %	1.12	1.12	1.12	1.12	1.12
M/Loglog ($\sigma = 0.1$)/	16 %	1.12	1.12	1.12	1.12	1.12
M/Phase (case 1)/		1.13	1.13	1.13	1.13	1.13
M/Phase(case 2)/		1.43	1.43	1.43	1.43	1.43
G/M/		1	1	1	1	1
G/Weib ($\alpha = 5$)/		1	1	1	1	1
M/Weib ($\alpha = 5$)/	10 %	1.38	1.38	1.38	1.38	1.38
M/Loglog ($\sigma = 0.1$)/	10 %	1.39	1.39	1.39	1.39	1.39
M/Phase (case 1)/		1.39	1.39	1.39	1.39	1.39
M/Phase (case 2)/		4.24	4.24	4.24	4.24	4.24
G/M/		1.15	1.15	1.15	1.15	1.15
G/Weib ($\alpha = 5$)/		1.15	1.15	1.15	1.15	1.15
M/Weib ($\alpha = 5$)/	4 %	1.93	1.93	1.93	1.93	1.93
M/Loglog ($\sigma = 0.1$)/	4 %	1.95	1.95	1.95	1.95	1.95
M/Phase (case 1)/		1.95	1.95	1.95	1.95	1.95
M/Phase (case 2)/		Unfeasib	Unfeasib	Unfeasib	Unfeasib	Unfeasib
G/M/		1.52	1.52	1.52	1.52	1.52
G/Weib ($\alpha = 5$)/		1.52	1.52	1.52	1.52	1.52

resources in the best way. We focus on intensive care units, where rejection of incoming patients has to be balanced with the early discharge of current patients.

Control problems (P1) to (P4) are not convex optimization problems. By using a convex representation of the posynomials included in the expressions (6) and (7), the problems turn out to be the maximization of a convex function over a set. Such problems are usually hard to solve. Queuing theory and classical nonlinear optimization methods provide the solution to the control problem in the Markovian case while a simulation based optimization method has been developed to deal with the general case.

An interesting contribution is the interpretation of the three types of solutions obtained from the optimization problems when different expressions for the shortening of the LoS are considered. Physicians of the ICU understood them and identified the cautious type solution as representative of the way to deal with the pressure of a near-full ICU in the real setting.

Through a broad computational analysis we can conclude that the three management policy structures emerge as solution of the control problems independently of the LoS distribution, the size of the ICU, the occupancy ratio and the arrival pattern.

Experimental results show the capability of the proposed methodology to deal with optimization problems posed on real size ICUs without computational issues. Studies analyzing the size of hospital's ICUs conclude that 8–12 beds are considered as the optimum [see for example Vallentin and Ferdinandi (2011)] while

Table 8 Cautious policies: Optimal management policies for (GP2) control problem, for different G/G/5/5 models and percentage of rejected patients $\varepsilon_j = 16, 10,$ and 4%

	ε_j	μ_1	μ_2	μ_3	μ_4	μ_5
M/Weib ($\alpha = 5$)	16 %	1	1	1	1	1.31
M/Loglog ($\alpha = 0.1$)	16 %	1	1	1	1	1.3
M/Phase (case 1)		1	1	1	1	1.32
M/Phase (case 2)		1	1	1	1	1.85
G/M		1	1	1	1	1
G/Weib ($\alpha = 5$)		1	1	1	1	1
M/Weib ($\alpha = 5$)	10 %	1	1	1	1.36	1.8
M/Loglog ($\sigma = 0.1$)	10 %	1	1	1	1.3	1.83
M/Phase (case 1)		1	1	1	1.39	1.77
M/Phase (case 2)		1	1	1.35	2.26	6.6
G/M		1	1	1	1.15	1.26
G/Weib ($\alpha = 5$)		1	1	1	1.15	1.26
M/Weib ($\alpha = 5$)	4 %	1	1	1.49	2.37	2.5
M/Loglog ($\sigma = 0.1$)	4 %	1	1	1.4	2.47	2.47
M/Phase (case 1)		1	1	1.43	2.47	2.47
M/Phase (case 2)		Unfeasib	Unfeasib	Unfeasib	Unfeasib	Unfeasib
G/M		1	1	1.1	1.81	1.81
G/Weib ($\alpha = 5$)		1	1	1.1	1.81	1.81

descriptive analysis indicates that the average number of beds is 17, 20 and 12 in cardiac, medical/surgical and pediatric intensive care units, respectively [Leleu et al. (2012)].

The recovery status of the patient has been modeled by using the states of a phase-type distribution to prevent the early discharge of a patient which has not recovered sufficiently.

The model presented in this paper can be readily extended to include different types of patients. In practice patients are clustered in different groups according to the type of illness, origin, necessary treatment, etc. Each group is characterized by its own arrival pattern and stochastic LoS.

Nevertheless, in order to make this theoretical analysis useful for the effective control of the health care system, it is necessary to take further steps in the analysis of the solution: physicians need flexible and medically-meaningful operative rules for shortening the LoS of a patient to the degree that will result in the service rates dictated by the theoretical analysis. The discussion of how the theoretical solutions can be transformed into effective management rules to guide doctors' decisions constitutes our work in progress.

Acknowledgments This paper has been in part supported by Grant MTM2012-36025.

References

- Ahmed MA, Alkhamis TM (2009) Simulation optimization for an emergency department healthcare unit in Kuwait. *Eur J Oper Res* 198:936–942
- Anderson D, Price C, Golden B, Jank G, Wasil E (2011) Examining the discharge practices of surgeons at a large medical center. *Health Care Manag Sci* 14:338–347
- Azcárate C, Mallor F, Gafaro A (2008) Multiobjective optimization in health care management. A metaheuristic and simulation approach. *Algorithm Oper Res* 3:186–202
- Bowers J (2013) Balancing operating theatre and bed capacity in a cardiothoracic centre. *Health Care Manag Sci*. doi:10.1007/s10729-013-9221-7
- Brailsford SC, Gutjahr W, Rauner MS, Zeppelzauer W (2007) Optimal screen policies for diabetic retinopathy using a new combined discrete event simulation and ant colony optimization approach. *Comput Manag Sci* 4:59–83
- Brailsford SC, Harper PR, Patel B, Pidd M (2009) An analysis of the academic literature on simulation and modelling in health care. *J Simul* 3:130–140
- Capuzzo M, Moreno RP, Alvisi R (2010) Admission and discharge of critically ill patients. *Curr Opin Crit Care* 16:499–504
- Chan CW, Farias VF, Bambos N, Escobar G (2012) Optimizing intensive care unit discharge decisions with patient readmissions. *Oper Res* 60:1323–1341
- Chan CW, Yom-Tov G, Escobar G (2014) When to use speedup: an examination of service systems with return. *Oper Res* 62:462–482
- Cochran JK, Roche K (2008) A queuing-based decision support methodology to estimate hospital inpatient bed demand. *J Oper Res Soc* 59:1471–1482
- Costa AX, Ridley SA, Shahani AK, Harper PR, De Senna V, Nielsen MS (2003) Mathematical modelling and simulation for planning critical care capacity. *Anaesthesia* 58:320–327
- de Angelis V, Felici G, Impelluso P (2003) Integrating simulation and optimisation in health care center management. *Eur J Oper Res* 50:101–114
- de Bruin AM, van Rossum AC, Viseer MC, Koole GM (2007) Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Manag Sci* 10:125–137
- de Bruin AM, Bekker R, van Zanten L, Koole GM (2010) Dimensioning hospital wards using the Erlang loss model. *Ann Oper Res* 178:23–43
- Eldabi T, Paul RJ, Young T (2007) Simulation modelling in healthcare: reviewing legacies and investigating futures. *J Oper Res Soc* 58:262–270
- Fu MC, Glover FW, April J (2005) Simulation optimization: a review, new developments and applications. In: *Proceedings of the 2005 winter simulation conference*, pp 83–95
- Green LV (2002) How many hospital beds? *Inquiry* 39:400–412
- Griffiths JD, Price-Lloyd N, Smithies M, Williams JE (2005) Modelling the requirement for supplementary nurses in an intensive care unit. *J Oper Res Soc* 56:126–133
- Griffiths JD, Price-Lloyd N, Smithies M, Williams J (2006) A queueing model of activities in an intensive care unit. *IMA J Manag Math* 17:277–288
- Griffiths JD, Knight V, Komenda I (2013) Bed management in a critical care unit. *IMA J Manag Math* 24:137–153
- Gross D, Harris CM (2008) *Fundamentals of queueing theory*. Wiley, New York
- Güral MM, Pidd M (2010) Discrete event simulation for the performance modelling in health care: a review of the literature. *J Simul* 4:42–51
- Katsaliaki K, Mustafee N (2011) Applications of simulation within the healthcare context. *J Oper Res Soc* 62:1431–1451
- Kc D, Terwiesch C (2009) Impact of workload on service time and patient safety: and econometric analysis of hospital operations. *Manag Sci* 55:1486–1498
- Kim SC, Horowitz I, Young K, Buckley TA (1999) Analysis of capacity management of the intensive care unit in a hospital. *Eur J Oper Res* 115:36–46
- Kim SC, Horowitz I, Young K, Buckley TA (2000) Flexible bed allocation and performance in the intensive care unit. *J Oper Manag* 18:427–443
- Kolker A (2009) Process modeling of ICU patient flow: effect of daily load leveling of elective surgeries on ICU diversion. *J Med Syst* 33:27–40
- Lakshmi C, Sivakumar A (2013) Application of queueing theory in health care: A literature review. *Operations Research for Health Care* 2:25–39

- Leleu H, Moises J, Valdmanis V (2012) Optimal productive size of hospital's intensive care units. *Int J Prod Econ* 136:297–305
- Lin RC, Sir M, Pasupathy KS (2013) Multi-objective simulation optimization using data envelopment analysis and genetic algorithm: specific application to determining optimal resource levels in surgical services. *Omega* 41:881–892
- Litvak N, van Rijsbergen M, Boucherie RJ, van Houdenhoven M (2008) Managing the overflow of intensive care patients. *Eur J Oper Res* 185:998–1010
- Mallor F, Azcárate C (2014) Combining optimization with simulation to obtain credible models for intensive care units. *Ann Oper Res* 221:255–271
- Marshall A, Vasilakis C, El-Zardi E (2005) Length of stay-based patient flow models: recent developments and future directions. *Health Care Manag Sci* 8:213–220
- Masterson BJ, Mihara TG, Miller G, Randolph SC, Forkner E, Crouter AL (2004) Using models and data to support optimization of the military health system: a case study in an intensive care unit. *Health Care Manag Sci* 7:217–224
- McManus ML, Long MC, Cooper A, Litvak E (2004) Queueing theory accurately models the need for critical care resources. *Anesthesiology* 100:1271–1276
- Rauner MS, Zeiles A, Schaffhauser-Linzatti MM, Hornik K (2003) Modelling the effects of the Austrian inpatient reimbursement system on length-of-stay distributions. *OR Spectrum* 25:183–206
- Ridge JC, Jones SK, Nielsen MS, Shahani AK (1998) Capacity planning for intensive care units. *Eur J Oper Res* 105:346–355
- Shmueli A, Sprug CL, Kaplan E (2003) Optimizing admissions to an intensive care unit. *Health Care Manag Sci* 6:131–136
- Sinuff T, Kahn moui K, Cook DJ et al (2004) Rationing critical care beds: a systematic review. *Crit Care Med* 32:1588–1597
- Stidham S (2002) Analysis, design and control of queueing systems. *Oper Res* 50:197–216
- Terekhov D, Beck JC (2008) A constraint programming approach for solving a queueing control problem. *J Art Int Res* 32:123–167
- Troy PM, Rosenberg L (2009) Using simulation to determine the need for ICU beds for surgery patients. *Surgery* 146:608–617
- Vallentin A, Ferdinandi P (2011) Recommendations on basic requirements for intensive care units: structural and organizational aspects. *Intensive Care Med*. doi:10.1007/s00134-011-2300-7
- Vasilakis C, Marshall AH (2005) Modelling nationwide hospital length of stay: opening the black box. *J Oper Res Soc* 56:862–869
- Zhu Z, Hen BH, Teow KL (2012) Estimating ICU bed capacity using discrete event simulation. *International Journal of Health Care Quality Assurance* 25:134–144
- Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL (2006) Intensive care unit length of stay: benchmarking based on acute physiology and chronic health evaluation (APACHE) IV. *Crit Care Med* 34:2517–2529

Fermín Mallor studied mathematics at the University of Zaragoza, Spain. He received the Doctorate degree in mathematics from the Public University of Navarre, Spain, in 1994. He is currently a Professor in the Department of Statistics and Operations Research, Public University of Navarre. In addition to having taught for more than 20 years university courses in simulation, operations research, and statistics, he has successfully applied his knowledge in simulation and statistical modeling to the analysis of complex real-world problems arising in several industrial companies and institutions. He is particularly interested in applications in the fields of health and energy. His research interests include simulation modeling, optimization with simulation, functional data analysis, and reliability.

Cristina Azcárate studied mathematics at the University of Zaragoza, Spain. She received his doctorate in mathematics from the Public University of Navarre, in 1995. Currently she is an Associated Professor in statistics and operations research. She teaches optimization and simulation to civil engineers. Her research interests are simulation modeling and optimization with simulation, applied to healthcare and energy.

Julio Barado is a physician at the ICU of the Hospital of Navarre, Spain. Currently he is a Ph.D. student at the Department of statistics and operations research of the Public University of Navarre. His research interests lie in the simulation modeling of ICU and in the study of medical decision making processes.