A simulation–based multiobjective optimization approach for health care services management

Stefano Lucidi, Massimo Maurici, Luca Paulon, Francesco Rinaldi and Massimo Roma

***Abstract*—Hospitals are huge and complex systems. However, for many years the management was commonly focused on improving the quality of the medical care while less attention was usually devoted to operations management. In recent years, the need of containing the costs while increasing the competitiveness along with new policies of National Health Services hospital financing forced hospitals to necessarily improve their operational efficiency.**

In this work we focus on a management problem usually arising in health care. In particular, we deal with optimal resources allocation of a ward of a big hospital. To this aim we propose a simulation–based optimization approach based on a discrete event simulation model reproducing the hospital services and combined with a derivative–free multiobjective optimization method. The results obtained on the obstetrics ward of an Italian hospital are reported, showing the effectiveness of the new approach proposed.

***Note to Practitioners*—In the last years, reducing health care costs while providing high quality health care services became a critical issue. Hence the necessity to make available to health care practitioners a decision support system for determining an optimal resources allocation.**

In this paper, we develop a simulation-based optimization framework that combines a simulation model reproducing the main processes of a specific hospital ward with a multiobjective optimization algorithm in order to find an approximate optimal resources allocation.

The proposed approach can be used in practice by decision makers in order to adjust the allocation of resources in a given ward. The results obtained on a real obstetrics ward of an Italian hospital show that the proposed approach is viable in practice and allow practitioners to adopt the best strategy according to specific indicators related to clinical risk, quality of the care provided, economical benefits both for patients, hospitals and for the National Health Service.

***Index Terms*—Health care operations management, Logistics of hospital services, Discrete event simulation, Simulation–based optimization, Derivative–free multiobjective optimization methods.**

Stefano Lucidi and Massimo Roma are with the Dipartimento di Ingegneria Informatica, Automatica e Gestionale "A. Ruberti", SAPIENZA Università di Roma, Italy (e-mail: lucidi,roma@dis.uniroma1.it).

Massimo Maurici and Luca Paulon are with the Dipartimento di Biomedicina e Prevenzione and Laboratory on Simulation and Optimization of Health Services, Università di Roma "Tor Vergata". Italy (e-mail: maurici@med.uniroma2.it, luca.paulon@gmail.com).

Francesco Rinaldi is with Dipartimento di Matematica, Università di Padova. Italy (e-mail: rinaldi@math.unipd.it).

Manuscript received

I. Introduction

In the last years, controlling health care costs while providing the best possible health outcomes became a more and more critical issue (see e.g. [1], [2], [3]). Moreover, recently in many National Health Services (NHS), health care services providers financing has changed from a budget-oriented system to a fee-for-service system. As a consequence, an optimal resources allocation is now strongly needed.

Hence the central role of the so called *health care operations management* that, according to [4], stands for "the quantitative management of the supporting business systems and processes that transform resources into health care services". In this context, again quoting from [4], *logistics* is "the efficient coordination and control of the flow of all the operations—including patients, personnel, and other resources".

In particular, the efficient management logistics of a hospital ward along with the design and performance evaluation of any hospital department is greatly important (see e.g. [5]). The choice of the resources (number of beds, doctors, nurses, and so on) to be employed, the patient flows, the supply chain management, the inventory management, the operational planning and scheduling, the staffing level and other similar items strongly affect the management costs and the income, as well as the quality of the services. The health care services of a hospital essentially represent specialized procedures for diagnosing or treating a disease of a given patient. Reducing the overall costs for delivering such services is currently at the forefront of any health care operations management.

On the basis of these observations, in this paper we consider the optimal resources allocation of the emergency room and obstetrics ward of a big hospital. The services under study are the caesarean section without complications or comorbidities and the vaginal childbirth without complications or comorbidities. In this case, the sources of the costs are several and mainly due to staff salaries and management of medical equipments and consumable goods. The incomes derive from the refunds through the NHS of the services delivered.

In the allocation of the resources of a hospital ward several constraints must be taken into account. They are either structural constraints or deriving from clinical and regulatory needs. For an obstetrics ward a crucial role is played by the rate of caesarean sections with respect to the overall childbirths. Indeed, due to the higher risk for mother or child in the case of caesarean delivery [6], this rate should be low. Since 1985, the World Health Organization (WHO) recommends a caesarean sections rate not higher than 15% (of the overall childbirths), but in many OECD countries[1] this value is often widely exceeded [7]. In recent decades the rate of caesarean sections has been even increasing in some countries usually because of economic reasons related to a lower profit associated to the natural childbirth. For instance, Italian National NHS standard would require a threshold value of 25%, but in some regions of Italy the value is over 40%.

Therefore, the current goals of an obstetrics ward should be maximizing the overall net profit and minimizing the caesarean sections rate. Thus, two contrasting objectives must be taken into account in the operations management of the ward.

Discrete Event Simulation (DES) methods have been widely used over the last decade for modelling health care systems and analyzing their performance (see e.g. [8], [9], [10]). The use of simulation models is motivated by the need of considering patient flow dynamics and all uncertainties related to activities of health care providers which cannot be described by means of analytical models. Therefore, an health care system is represented by a *stochastic model* whose output is a random vector sampled by computer simulation. Moreover, simulation methods enable to examine the responses obtained for a number of different input combinations (*scenarios*). Very often the number of scenarios considered in a DES approach is very small due to high computational burden. However, in practical problems usually the “best scenario” is sought. To this aim, recently such DES methods have been combined with optimization techniques (see e.g. [11], [12]). Hence, the term *simulation optimization* (or *simulation–based optimization*) commonly used to refer to this combination. However, quoting from [11], “combining the two techniques is a more recent development and software effectively integrating the two is relatively limited; thus, simulation optimization remains an exciting and fertile area of research”.

Indeed, for many years most of the optimization routines available in commercial simulation packages were based only on evolutionary algorithms and metaheuristics. More recently, many deterministic optimization algorithms have been employed in the simulation optimization context (see [13] for a recent survey). However, very often, real problems involve multiple objective functions, i.e. many conflicting objects must be optimized, but as far as we are aware, almost all the optimization algorithms embedded within simulation packages are only for single objective problems, or reduce to this case by aggregating the different objective functions into a single one. The latter procedure could be a serious drawback within a support decision system, since the solution will consist of a single point and no choice is left to the decision maker. Instead, when the problem is multiobjective, the solution provided in the form of a set of (nondominated) points allows the decision maker to choose among different strategies, according to specific demands or preferences.

To deal with the aforementioned optimal resources allocation problem of an obstetrics ward, in this paper we propose to represent the behavior of the given ward by means of a DES model and optimize its performance by using a novel Derivative-Free Multiobjective Optimization method. In [14] the management of an obstetrics ward was already tackled by the same authors of this paper, but a single objective model was considered, being the objective to be maximized only the net profit. In that model, only the growth of the rate of caesarean sections was controlled by adding a constraint (an upper bound) to that rate according to the WHO recommendations.

The paper is organized as follows. In Section II, a literature review is reported. Section III describes the methodology used in our work, namely the service delivery description, the model formulation, the DES model, the derivative–free multiobjective algorithm and the implementation. Section IV includes the case study, namely the resources allocation problem for the obstetrics ward of one of the most important Italian hospital for childbirth located in Rome. In Section V some concluding remarks and future study directions are reported, along with some policy recommendations.

II. Related Works

In the recent years, multiobjective simulation optimization techniques have been used in many different contexts: industrial engineering, systems management, design technology, production and inventory planning and many others. Some examples are described in the papers [15], [16], [17], [18], [19], [20], However, very few papers have been published proposing the use of multiobjective optimization in connection with a DES model without reducing the multiobjective problem to a single objective problem. Namely, the so called “a priori” articulation of preferences approach is applied, i.e. the multiobjective optimization is transformed into a single objective one by aggregating the different objective functions. However, as well known, this procedure presents a serious drawback since in this case the solution is very sensitive to the preferences used [21].

In health care, multiobjective simulation optimization method have been also used in some cases study. As example, in [22], Baesler and Sepulveda developed a methodology integrating simulation and Genetic Algorithms to solve a problem with four objectives arising in health care treatment. In [23] the authors study the inpatient flow process of a large acute–care hospital by means of multiobjective DES optimization.

However, also in the health care framework, the multiobjective problem is usually transformed into a single objective one. At this regards, see also the discussion reported in Section II of [24]. The latter very recent paper [24] by Song, Qiu and Liu represents the closest to our approach. Indeed, they studied the optimal patient flow distribution both “intra-hospital” and “inter-health care facilities” by integrating a DES model and an multiobjective optimization algorithm. Their aim is to improve the overall system performance by finding an approximate Pareto set representing the patient flow distribution.

III. Methodology

This work is based on a simulation–based optimization methodology. First we construct a DES model reproducing the real patients flows through the Emergency Room to the

[1]Countries members of the *Organisation for Economic Co-operation and Development*

obstetric ward. Then a careful validation of the model is performed to guarantee its good accuracy. The simulation model is then used to estimate some relevant performance indexes related to the processes of interest. Since the problem is stated as a bi-objectives optimization problem, a derivative–free multiobjective algorithm is connected to the simulation model by using a suitable interface. The simulation optimization procedure is then executed starting form the current operating condition of the ward. Finally, results are analyzed and compared with those obtained by an "a priori" preferences based approach which aggregates the two objective functions into a single one.

A. Service Delivery Description

The service delivery under study is related to the caesarean section without complications or comormidities and the vaginal childbirth without complications or comormidities which are the most common health services provided by hospitals.

The service delivery can be described as follows: pregnant women go through the Emergency Room. Also pregnant women for which a caesarean section was scheduled in advance arrive to the ER for registration and verification. At the beginning, nurses perform a first triage and assign a priority. In case of a scheduled caesarean section the patient flows directly to the ward, and wait for the availability of an operating room. Otherwise another triage (a specialistic one) is performed by obstetricians along with a continuous fetal monitoring. Moreover, a gynaecologist visits the patient, confirms or changes the assigned priority, decides if the hospitalization is required and if a caesarean section is needed or not. As concerns the subsequent activities, the patients undergo different treatments on the basis of the assigned priority. Patients which do not need hospitalization are discharged. The patients flow keeps on as described in the sequel.

- Those patients for which the highest priority is confirmed (or newly assigned) need to quickly flow to delivery room in case of vaginal childbirth or to the operating room in case of caesarean section. Therefore the availability of a bed or a stretcher in the ward is checked, and the patient is driven to the required room, eventually waiting for its availability. After the delivery, the patient remains for a while in the room under observation, then if a bed is available she is driven to the ward, otherwise she settles herself on a stretcher. If neither a bed or a stretcher is available, due to the emergency, the delivery takes place anyhow, but both woman and newborn are not hospitalized and after a period under observation they are transferred to another hospital. In the sequel, this occurrence will be named "extra" childbirth or "extra" delivery.
- The patients with a low assigned priority (i.e. which do not need an immediate delivery), undergo some visits and clinical exams in order to decide if hospitalization is needed. If it is not required, the patient is discharged, otherwise the availability of a bed in the ward or a stretcher is checked. In case of no availability, the patient is transferred to another hospital. Otherwise, the patient is driven to the ward and prepared for the delivery. After the childbirth the women goes back to her bed, if it has been previously assigned, otherwise (i.e. only a stretcher was assigned to the patient) a check is carried out to verify if in the meantime a bed has been released. If no accommodation is available, the patient will settle herself again on a stretcher.

The length of the hospitalization depends on the delivery: it usually lasts less in the case of vaginal childbirth (e.g., 2 days) than in case of caesarean section (e.g., 3 days). Finally, the discharge of mother and newborn from the hospital can occur only in a specific time slot when a gynecologist in charge of this task is available.

In Figure 1, the main patient flow (the pregnant women for which a caesarean section is not scheduled in advance) and the related service processes are reported. Note that two other more simple patients flows (not reported in Figure 1) are also included in the model: pregnant women for which a caesarean section is scheduled in advance and women that need hospitalization in the ward with diagnosis different from childbirth. Even if the patients belonging to the latter flows are not part of the services under study, if hospitalized in the ward, they use ward resources and hence must be considered.

In this organization, the sources of the costs are several and mainly due to staff salaries and management of medical equipments, consumable goods and utilization of the operating rooms. The income derives from the refunds through the NHS of the services delivered. At each choice of the resources corresponds a different "case-mix", i.e. a different number of patients to treat for each of the two kind of childbirth. The allocation of the resources is subject to several constraints. They are structural constraints or derive from clinical and regulatory needs.

The hospital top managers require the maximization of the net profit determined by the overall childbirths and the minimization of the caesarean rate. These are two goals which are contrasting since the profit for a caesarean section is greatly higher than the one for natural childbirth.

B. Model Formulation

The *variables* represent the resources which can be controlled by the hospital manager. Namely there are 7 counters z_i of allocated resources and one service demand indicator t_1:

z_1 :	number of stretchers
z_2 :	number of gynecologists
z_3 :	number of gynecologists who discharge a patient from the hospital
z_4 :	number of nurses
z_5 :	number of midwives
z_6 :	number of hospital beds
z_7 :	number of operating rooms
t_1 :	mean value of the patient interarrival time (in hours).

Note that, even if t_1 is not a resource, its value can be controlled due to the possibility, in some cases, to reduce or rise admissions of patients by adopting appropriate strategies. We denote by $z = (z_1, z_2, z_3, z_4, z_5, z_6, z_7) \in \mathbb{Z}^7$ the vector of the integer variables and by $t = t_1 \in \mathbb{R}$ the real variable.

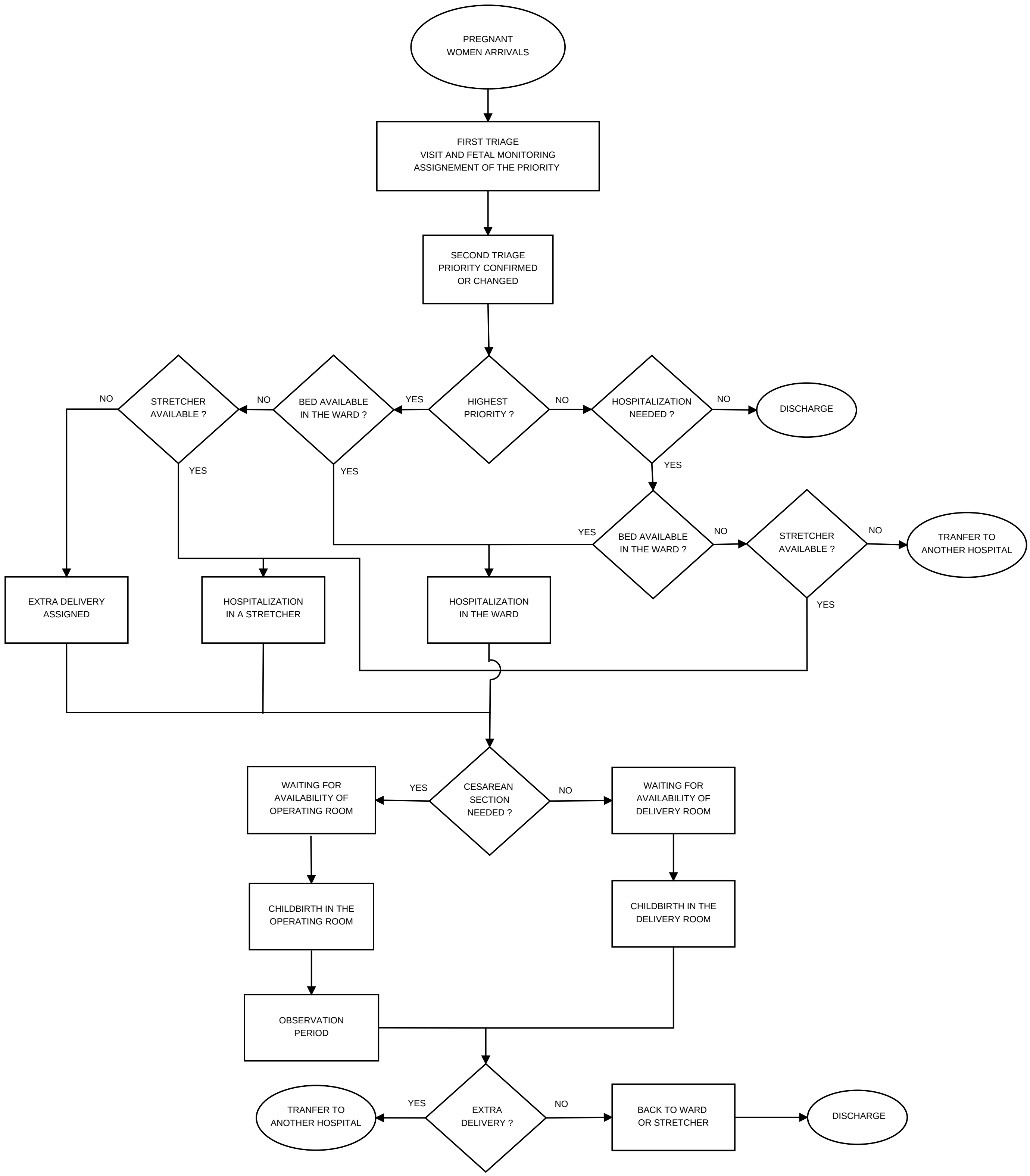

Fig. 1. Patient flow for pregnant women for which a caesarean section is not scheduled in advance

Moreover, the patient case-mix of the service provider is given by the following 6 counters y_j (expressed as number of patients per year):

y_1 : number of caesarean sections
y_2 : number of vaginal childbirths
y_3 : number of "extra" caesarean sections
y_4 : number of "extra" vaginal childbirths
y_5 : number of hospitalized woman not for childbirths
y_6 : number of woman transferred to another hospital before the childbirth.

Actually, they are estimates of the expected values of the output of the service delivery model which depends on z and t. In practice, the values $y_j = y_j(z,t),\ j = 1,\dots,6$, are obtained as an average over the output of a certain number of independent replications of the simulation. We denote by

$$y(z,t) = (y_1(z,t), y_2(z,t), y_3(z,t), y_4(z,t), y_5(z,t), y_6(z,t))$$

this response vector.

The *objective functions* are two: the first one represents the net profit to be maximized and can be stated as follows:

$$\begin{aligned} f_1(z,t) &= P_{cs}(y_1(z,t) - y_3(z,t)) + P_{vc}(y_2(z,t) - y_4(z,t)) \\ &-C_1 \max\{0, z_1 - z_1^0\} - C_2 \max\{0, z_2 - z_2^0\} \\ &-C_3 \max\{0, z_3 - z_3^0\} - C_4 \max\{0, z_4 - z_4^0\} \\ &-C_5 \max\{0, z_5 - z_5^0\} - C_6 \max\{0, z_6 - z_6^0\} \\ &-C_7 \max\{0, z_7 - z_7^0\} - C_8 z_1 - C_9 z_6. \end{aligned}$$

The first two terms correspond to the profit due to caesarean sections (cs) and vaginal childbirths (vc), being P_{cs} and P_{vc} the corresponding unit profit. The terms of the form $C_i \max\{0, z_i - z_i^0\}$ correspond to set up costs and the last two terms correspond to some additional costs for stretchers and beds utilization.

The second objective function represents the rate of caesarean sections (with respect to the overall childbirths) to be minimized and it can be stated as follows:

$$f_2(z,t) = \frac{y_1(z,t) - y_3(z,t)}{y_1(z,t) - y_3(z,t) + y_2(z,t) - y_4(z,t)}.$$

The *constraints* are general constraints and box constraints on the variables. They are derived from some guidelines of the NHS or from local clinical and logistic requirements:

- a lower bound on the number of caesarean sections to guarantee a minimum number of expected caesarean sections per year
$$y_1(z,t) \geq Y^1_{min};$$
- a lower bound on the overall number of childbirths per year required by some guidelines in order to guarantee a good efficiency of the ward
$$y_1(z,t) + y_2(z,t) \geq Y^{12}_{min};$$
- a lower bound on the overall patient occupation rate in order to avoid the underutilization of the ward; this rate is defined as the ratio between the effective overall length of the patients stay and the (theoretical) length of stay available
$$\frac{1}{365(z_1+z_6)}\Big(L_{vc}(y_2(z,t) - y_4(z,t)) + L_{cs}(y_1(z,t) - y_3(z,t)) + L_{others} y_5(z,t)\Big) \geq O_{rate};$$
- an upper bound on the number of transferred women before delivery imposed to keep low the risks of transfers
$$y_6(z,t) \leq T_{rate}(y_1(z,t) + y_2(z,t)).$$

The box constraints, namely lower and upper bound on the variables z_i , $i = 1\dots,7$, are mainly due to budget and logistic restrictions, while for t_1 derive from specific clinical and managerial policy on patients admission. They are the following:

$$\begin{aligned} Z_1^l \leq z_1 \leq Z_1^u \\ Z_2^l \leq z_2 \leq Z_2^u \\ Z_3^l \leq z_3 \leq Z_3^u \\ Z_4^l \leq z_4 \leq Z_4^u \\ Z_5^l \leq z_5 \leq Z_5^u \\ Z_6^l \leq z_6 \leq Z_6^u \\ Z_7^l \leq z_7 \leq Z_7^u \\ T_1^l \leq t_1 \leq T_1^u. \end{aligned}$$

Thus, the resulting problem is a bi-objective *mixed integer nonlinearly constrained problem with box constraints* on the variables z and t, namely a problems of the following general form

$$\begin{aligned} \min \quad & F(z,t) = (f_1(z,t), \dots, f_l(z,t))^\top \\ & g_1(z,t) \leq 0 \\ & \vdots \\ & g_m(z,t) \leq 0 \\ & 0 \leq l_z \leq z \leq u_z \\ & 0 \leq l_t \leq t \leq u_t \end{aligned} \tag{1}$$

where the objective functions f_h, $h = 1,\dots l$ and the general constraints g_i, $i = 1,\dots,m$ are real valued functions, $f_h, g_i : \mathbb{Z}^p \times \mathbb{R}^q \times \mathbb{R}^r \longrightarrow \mathbb{R}$. The distinguishing feature of this problem with respect the one considered in [14] is the multiobjective formulation, i.e. the presence of two of objective functions.

C. Discrete-Event Simulation model

The simulation model of the Hospital ER and obstetrics ward is implemented by using *Arena 14.7* simulation software [25], [26], a general–purpose simulation environment and one of the most popular DES software. In order to construct an accurate simulation model, a database containing all the data related to hospitalizations (e.g. hospital childbirth records, hospital discharge forms, all cost and income items) of a given period is needed. By simple database queries, it is possible to obtain clinical and economical information for each childbirth. Our particular focus is on: operational times of any activity of the entire service delivery; interarrival times of pregnant women to the Emergency Room; arrival times of pregnant women for which a caesarean section was scheduled in advance; percentage of the different priorities assigned to patients at the obstetric triage; information on all the possible movements of patients.

On the basis of these information it is possible to perform an accurate input analysis for determining the service-time probability distributions (with the related parameters) of all the processes used in the model along with the corresponding resources seized.

D. Derivative-free multiobjective optimization algorithm

In this section, we describe the algorithm used to deal with the mixed integer nonlinear multiobjective optimization problem (1) within the simulation-based optimization framework. Since both the objective functions and constraints values come from a simulation tool, there is no way to obtain first order information for the problem. Therefore, derivative based methods cannot be used in this context. Furthermore, due to the presence of noise coming from the simulation runs, finite-differences derivative cannot be applied (since wrong estimates of the first order derivatives would be obtained). Hence, a Derivative-Free Optimization (DFO) approach needs to be considered in this case (see [27] for an overview on DFO methods).

The proposed approach is basically obtained by the combination of the DFMO method proposed in [28], which is an efficient DFO algorithm for constrained multiobjective continuous problems, with a rounding step that guarantees satisfaction of the integrality constraints. The main features of the algorithm are the following:

- An *exact penalty approach*, which is needed in order to handle the constraints $g_1, \dots, g_m$. Those constraints are simply removed from the model, and a penalty measuring their violation is included in the objective functions. Hence, the new problem to be solved is

$$\begin{aligned} \min \quad & Z\left(z,t;\epsilon\right) = \left(Z_1\left(z,t;\epsilon\right), \dots, Z_l\left(z,t;\epsilon\right)\right)^\top \\ & 0 \le l_z \le z \le u_z \\ & 0 \le l_t \le t \le u_t \end{aligned}$$

  where, for all $h = 1, \dots, l$,

$$Z_h(z,t;\epsilon) = f_h(z,t) + \frac{1}{\epsilon}\sum_{i=1}^{m} \max\left\{0, g_i(z,t)\right\},$$

  is the so called penalty function, and $\epsilon > 0$ is the penalty parameter (used for weighting the penalty term).
- The use of a *list* of “candidate” Pareto points that evolves as the algorithm goes on. In practice, at each iteration, the list is updated by including new suitably generated nondominated points and by filtering those ones that become dominated.
- A *line-search approach* for obtaining the new nondominated points.
  At each iteration, first a search direction is generated. Then, starting from each point in the list, a line-search is performed along that direction. More specifically, a point is suitably selected along the given direction and, in case it satisfies a specific condition of “sufficient” decrease (i.e. there exists at least one objective function that reduces enough), a “sufficiently” large movement is performed in order to generate some new nondominated points. This way of moving along the search direction is somehow needed in order to guarantee that the points are properly spread and get close enough to the real Pareto front.
- A *rounding step*, performed in order to guarantee that variables z satisfy integrality constraints. In the algorithm, integrality is relaxed and all variables are considered continuous. Hence, before passing the point to the simulation software, the algorithm needs to properly round z variables up:

$$z_j = \lfloor z_j + 0.5 \rfloor, \quad \text{for all} \quad j = 1, \dots, p.$$

More specifically, the k-th iteration of the algorithm can be summarized in the scheme reported below, where $\left[z,t\right]_r$ denotes the projection (followed by a proper rounding) of the point (z,t) on the box-feasible set of the previous multiobjective problem, and γ is a positive constant:

Algorithm 1 Scheme of the algorithm (iteration k)

1 *given* the list L^k of “candidate” Pareto points;

2 *choose* a direction $d^k = (d_z^k, d_t^k)$;

3 *compute* all $\left[\tilde z + \tilde\alpha d_z^k, \tilde t + \tilde\alpha d_t^k\right]_r$ where $(\tilde z, \tilde t) \in L^k$ and $\tilde\alpha$ is an initial stepsize associated to $(\tilde z, \tilde t)$;

4 *if* there exists $(\hat z, \hat t) \in L^k$ such that

$$Z_h(\left[\tilde z + \tilde\alpha d_z^k, \tilde t + \tilde\alpha d_t^k\right]_r; \epsilon) > Z_h(\hat z, \hat t; \epsilon) - \gamma\alpha^2$$

for every $h = 1, \dots, l$,

then $\left[\tilde z + \tilde\alpha d_z^k, \tilde t + \tilde\alpha d_t^k\right]_r$ is rejected and $\tilde\alpha$ is halved;

else

- $\tilde\alpha$ is doubled until $(\hat z, \hat t) \in L^k$ exists such that

$$Z_h(\left[\tilde z + 2^{\hat r}\tilde\alpha d_z^k, \tilde t + 2^{\hat r}\tilde\alpha d_t^k\right]_r; \epsilon) > Z_h(\hat z, \hat t; \epsilon) - \gamma(2^{\hat r}\tilde\alpha)^2$$

  for every $h = 1, \dots, l$;
- $\tilde\alpha$ is updated to the value $2^{\hat r - 1}\tilde\alpha$:
- L^{k+1} is constituted by all non-dominated points contained in the set

$$L^k \cup \left\{[\tilde z + 2^i\tilde\alpha d_z^k, \tilde t + 2^i\tilde\alpha d_t^k]_r : i = 0, \dots, \hat r - 1\right\};$$

endif

In the continuous case, if the sequence $\{d^k\}$ of the search directions used in the algorithm satisfies a suitable assumption, the previous algorithm has interesting *theoretical properties*. Indeed, in [28] it is proved that every accumulation point of a sequence of points belonging to the candidates list satisfy necessary optimality conditions to be a Pareto point.

E. Implementation

As we already mentioned, the simulation model of the Hospital ER and obstetrics ward is implemented by using *Arena 14.7* simulation software. Afterwards, in order to connect this model with an implementation of DFMO algorithm described in Section III-D, an interface between the Fortran90 code of the optimization algorithm and *Arena* simulation software is constructed. The Visual Basic for Applications (VBA) tool included in *Arena* is used to this aim.

The procedure implemented is the following: the DFMO algorithm selects the values for the decision variables (z, t). These values are transferred to the *Arena* model and a prefixed number of independent simulation runs are performed to estimate the response vector y. The DFMO algorithm uses these responses to select new values for the decision variables to transfer to *Arena*. The loop is carried on until a stopping criterion is satisfied.

IV. Case–study

The case study considers the optimal resources allocation of the emergency room and obstetrics ward of the Fatebenefratelli San Giovanni Calibita (FBF-SGC) Hospital in Rome. It is one of the most important hospitals of the Italian NHS in terms of number of childbirth cases. The study was carried out within a project named "Business Simulation for Healthcare" (*BuS-4H*) by a research group composed by doctors, managers, engineers, statisticians and other experts in health care. A database containing all the data concerning the hospitalizations for a two years timeline was expressly constructed for this project. This allowed us to easily obtain the data needed to build an accurate simulation model.

A. Input Analysis

In the sequel, we report the details of the main stochastic processes in the simulation model. Namely, we specify the probability distributions and the resources involved.

As regards the *arrival processes* to the system, we distinguish three kind of arrivals: pregnant women going through the ER, pregnant women for which a caesarean section was scheduled in advance (also going through the ER), women which flow to the ward for diagnosis different from childbirth. The probability distribution of the interarrival times (in hours) are reported in Table I.

TABLE I
ARRIVALS PROCESSES

Pregnant women (to the ER)	Pregnant women with scheduled caesarean section (to the ER)	Women with different diagnosis (to the ward)
EXP(2.4)	*fixed schedule*	Gamma(7.9,1.25)

In case of caesarean sections scheduled in advance, the arrivals scheme is based on a timeline of seven days and fixed, namely one, two or three arrivals from 8:00 a.m to 10:00 a.m. for each day.

As concerns the processes of the service delivery, in Table II we report the probability distribution of the *service times*, along with the *resources* required. The time for the delivery room includes a period of observation just after the delivery, while an additional time of 2 hours of observation at the surgical unit must be considered just after a caesarean section. All the queues discipline are based on the priority assigned in the triage. The only exception regards the queue discipline of the discharging process which is first come, first served, taking into account that gynecologists who discharge a patient are available only between 8:00 and 12:00 a.m.

TABLE II
SERVICE DELIVERY PROCESSES

	service times	*resources*
First triage	Triangular(3, 5, 10) (minutes)	1 nurse
Specialistc triage (fetal monitoring) and visit	Normal(30, 2) (minutes)	1 midwife 1 gynaecologist
Delivery in the operating room (caesarean section)	Uniform(70, 90) (minutes)	1 midwife 1 gynaecologist 1 operating room
Delivery in the delivery room (vaginal childbirth)	Uniform(8, 10) (hours)	1 midwife 1 gynaecologist 1 delivery room
Discharging	Constant 5 minutes	1 gynecologist who discharge a patient

As regards the stay at the ward, it depends on the type of delivery (vaginal childbirth and caesarean section). Moreover, a short stay before delivery and a stay after delivery is usually expected. Finally, women hospitalized in the ward for a diagnosis different from childbirth require a different stay. Table III reports the probability distributions of the *stay times* (in hours). Moreover, on the basis of the data available, we infer the following probabilities of assigning the priority in the first triage and the conditional probabilities to confirm or change this priority in the second triage (see Table IV). In case of a caesarean section scheduled in advance, the lowest priority is conventionally assigned. As expected, in most cases, the priority assigned at the first triage is confirmed in the second one. Finally, as concerns the decision on the hospitalization, 85% of pregant women arriving at the ER (excluding the caesarean sections scheduled in advance) are hospitalized.

TABLE III
STAY TIMES AT THE WARD

	before delivery	*after delivery*
Caesarean section	Uniform(0.17,0.25)	48+Lognormal(87.8, 162)
Vaginal childbirth	Uniform(1,1.7)	20+Lognormal(2.68, 1.21)
Different diagnosis	Gamma(200, 0.501)	

TABLE IV
PRIORITIES ASSIGNEMENT

	First triage		
	Priority 1	Priority 2	Priority 3
	0.4	0.3	0.3
	Second triage		
Priority 1	0.9	0.1	0.05
Priority 2	0.05	0.8	0.2
Priority 3	0.05	0.1	0.75

B. Model verification/validation and design of experiments

A careful verification of the simulation model has been carried out by using standard techniques (e.g. self-inspection, structured walkthrough, interactive debugger). An interaction on regular basis with the hospital management was really helpful, too. Moreover, thanks to the availability of the operating information of the ER and obstetrics ward of the hospital, a careful validation of the model was possible, by comparing the responses of the simulation model with the real observations in correspondence of some relevant indexes (e.g., the case-mix).

As regards the design of experiments, the length of a simulation run was set to one year, the number of replications was 10 and the warm-up period was 42 days.

C. Current state

The current operating condition of the FBF-SGC, i.e. the values currently used in the hospital for the resources are denoted by (z^0, t^0) and reported in Table V.

TABLE V
RESOURCES FOR THE CURRENT OPERATING CONDITION

z_1^0	z_2^0	z_3^0	z_4^0	z_5^0	z_6^0	z_7^0	t_1^0
10	5	1	1	6	42	1	2.400

The patient case-mix (estimate of the expected values) corresponding to the current operating condition (denoted by y^0) obtained from simulation is reported in Table VI.

TABLE VI
PATIENT CASE-MIX FOR THE CURRENT OPERATING CONDITION

y_1^0	y_2^0	y_3^0	y_4^0	y_5^0	y_6^0
883.40	2514.70	12.80	220.60	1080.00	551.70

The cost parameters (which appears in the first objective function f_1) are specified in Table VII.

The resulting net profit and the rate of caesarean sections corresponding to the current operating condition are $f_1(z^0, t^0) = 400876.00$ euros and $f_2(z^0, t^0) = 0.27$, respectively.

The values of the parameters of the general constraints are reported in Table VIII.

Finally, Table IX reports lower and upper bounds of the box constraints. Moreover, since in the obstetrics ward of FBF-SGC Hospital three beds are in each room, it is required that $z_6 = 3\ell$, $\ell \in \mathbb{Z}$.

TABLE VII
COSTS PARAMETERS (IN EUROS)

P_{cs}	382.00
P_{vc}	309.00
C_1	4500.00
C_2	10352.00
C_3	10352.00
C_4	9589.00
C_5	9589.00
C_6	5000.00
C_7	50000.00
C_8	2737.00
C_9	14600.00

TABLE VIII
GENERAL CONSTRAINTS PARAMETERS

Y^1_{min}	Y^{12}_{min}	L_{vc}	L_{cs}	L_{others}	O_{rate}	T_{rate}
500	3500	3.3	5.0	5.0	0.75	0.25

D. Optimization experiments

In the experiments, the current operating condition, namely the point (z^0, t^0), is taken as starting point even if it is infeasible. This is possible since the optimization algorithm used is based on an exact penalty approach. We refer to [28] for all the details concerning the algorithm DFMO and its implementation.

The use of the DFMO algorithm enables us to obtain a set of Pareto points whose objective functions values are reported in Table X.

These results clearly point out that, as expected, starting from the current operating condition and due to the tight constraints provided (especially the small width of the box constraints) the decrease of the caesarean section rate with respect to the current one is moderate. Indeed, from the value 0.27 corresponding to the current operating condition the least value obtained is approximately 0.23. Anyhow, a decrease by 4% is assessed relevant from the hospital management.

As regards the net profit, a significant increase can be obtained with respect to the current one, namely from 400876.00 euros to at least 499632.00 euros. Of course, to a higher value of the profit corresponds a higher caesarean sections rate. In Table XI, the values of the resources (z, t) for the two "best" points are reported, namely the one corresponding to the best value of the net profit and the one corresponding to the best value of the caesarean sections. In between, the

TABLE IX
LOWER AND UPPER BOUND CONSTRAINTS PARAMETERS

	l	u
z_1	8	15
z_2	2	7
z_3	1	3
z_4	1	5
z_5	2	9
z_6	33	45
z_7	1	3
t_1	1.000	4.000

TABLE X
PARETO POINTS OBJECTIVE FUNCTIONS VALUES

	net profit (euros) f_1	*c.s. rate* f_2
1	499632.00	0.22966
2	510629.30	0.22973
3	523259.60	0.22986
4	527757.40	0.22987
5	530838.20	0.23085
6	538476.40	0.23147
7	546969.00	0.23246
8	553506.00	0.23359
9	560811.10	0.23383
10	565000.90	0.23483
11	572146.00	0.23846

TABLE XI
RESOURCES VALUES CORRESPONDING TO THE TWO "BEST" POINTS

	z_1^*	z_2^*	z_3^*	z_4^*	z_5^*	z_6^*	z_7^*	t_1^*
point 1	15	7	1	1	6	45	1	1.805
point 11	15	5	1	1	6	42	1	1.753

remaining points are other nondominated points representing "intermediate" solutions. It is worthwhile to highlight that to provide a set of (nondominated) points (instead of a single point) as solution enables the hospital managers to select the strategy to be adopted according to their preferences (or some special needs).

By comparing Table XI with Table V, it can be easily observed that, the improvements in terms of net profit and/or in terms of rate of caesarean sections are obtained even if few changes are required with respect to the present setting. This is very appreciated by the hospital managers, since they can adopt new strategies without dramatically changing the current conditions.

The DFMO algorithm used in our simulation optimization framework belongs to the so called class of *methods with "a posteriori" articulation of preferences*, i.e. methods which try to reconstruct the whole Pareto front for the multiobjective problem under analysis. As far as the authors are aware, this is a novel feature in the solution of a simulation–based multiobjective optimization problem. Indeed, the optimization procedures embedded in simulation packages are usually able to only tackle single objective problems (see e.g. *OptQuest for Arena* [29]). In other cases, *methods with "a priori" articulation of preferences* are adopted to handle the multiobjective problems. This means, as we already said, that objective functions are combined into a single one by means of an aggregation criterion (see e.g. [30]) and the original problem is transformed into a single objective one. As consequence, this class of algorithms provides a unique solution point.

In order to compare the results obtained by using the DFMO algorithm and those obtained by transforming the original multiobjective into a single objective one, we transformed the bi-objective problem from our case study into a single objective problem by means of weighted sum of the two objective functions f_1 and f_2. Namely, we defined several combinations of the form

$$\eta_1 \frac{f_1(z,t)}{f_1(z^0,t^0)} + \eta_2 \frac{f_2(z,t)}{f_2(z^0,t^0)}, \tag{2}$$

where $\eta_1 \geq 0$, $\eta_2 \geq 0$, $\eta_1 + \eta_2 = 1$.

We tried several combinations obtained by selecting different weights η_1 and η_2. For each combination we applied the DFL (single objective) algorithm proposed in [31], i.e. the same used in [14], obtaining one solution point for each combination. For sake of brevity we do not report the detailed results of this experiment, but we only display in Figure 2 (top) these points (red squares) along with the points obtained by DFMO and reported in Table X (blue circles). Note that the values on the y-axis are reported with the minus sign, since the multiobjective problem is reformulated in terms of the minimization of the two objective functions. It can be easily seen that all the points obtained by different minimizations of the transformed single objective problem are dominated by the points obtained by DFMO algorithm which represent an approximate Pareto front. Note that both the strategies aim at finding approximations of local Pareto optimal points. However, the obtained results seem to indicate that the use of a list of candidate Pareto points allows the proposed DFMO algorithm to have better global properties.

In order to assess the robustness of the proposed approach, we performed further experiments by considering two different scenarios. In particular, we focused on processes where uncertainty is a more critical issue, in the sense that changes in the probability distributions related to these processes may significantly change the performance of the overall system. They are the service delivery processes (see Table II). The two situations we considered are reported in Table XII. The first scenario corresponds to an improvement on the services provided with respect to the real case study (obtained by considering a decrease of the service delivery times), while the second one is related to a worsening (obtained by considering an increase of the service delivery times). Even though those two scenario can be considered reasonable (according to the experts analysis), each one represents a critical problem from the multiobjective optimization point of view. We will analyze in depth this fact hereinafter.

TABLE XII
TWO SCENARIOS OF THE SERVICE DELIVERY PROCESSES

	Scenario 1	*Scenario 2*
First triage	Triangular(2, 3, 7) (minutes)	Triangular(6, 7, 10) (minutes)
Specialistc triage (fetal monitoring) and visit	Normal(20, 2) (minutes)	Normal(40, 2) (minutes)
Delivery in the operating room (caesarean section)	Uniform(50, 70) (minutes)	Uniform(80, 100) (minutes)
Delivery in the delivery room (vaginal childbirth)	Uniform(6, 8) (hours)	Uniform(10, 12) (hours)
Discharging	Constant 5 minutes	Constant 5 minutes

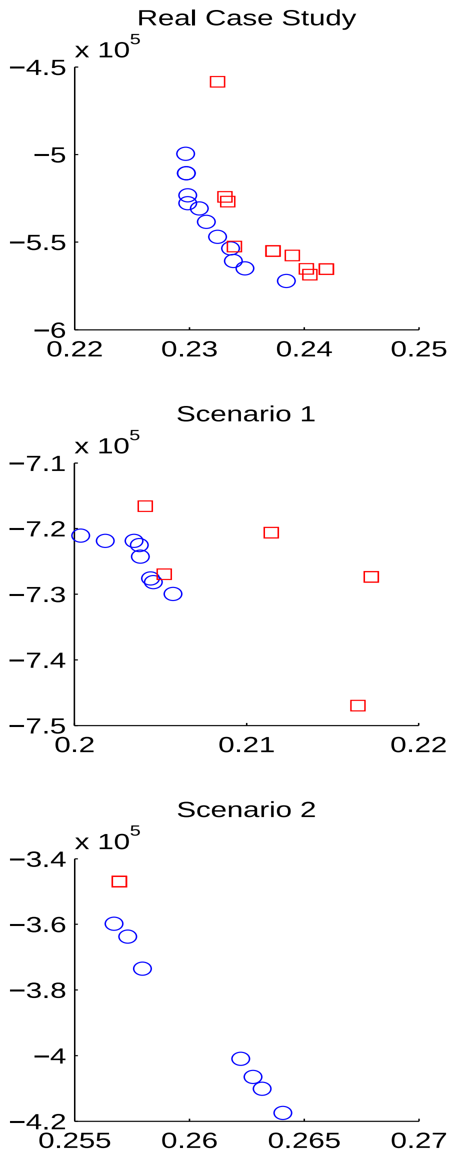

Fig. 2. Approximate Pareto front for the original two–objectives problem (blue circles), and points obtained by different minimizations of the transformed single–objective problem (red squares) - Real Case Study (top), Scenario 1 (center), Scenario 2 (bottom)

In Figure 2 (center), we report the points obtained for the *Scenario 1*. In this case, thanks to the reduction of the service delivery times, we get that the interarrival times may be reduced (so that more pregnant women may arrive to the ER) still maintaining the general constraints inactive. This is due to the fact that, to a decrease of the service delivery times, somehow corresponds a reduction of the conflict existing between the two objectives. Hence, when using our approach we get that the points in the final list are clustered in some way (it is almost like we get a single point). When considering the points obtained aggregating the objective functions, we can notice that there is no cluster effect and all those points are dominated by the points generated by our algorithm. Such a bad behavior might be due to the fact that the single objective algorithm easily gets stuck in local solutions.

In Figure 2 (bottom), we report the points obtained for the *Scenario 2*. In this case, the increase of the service delivery times gets the resources management crucial, thus making the problem harder to be solved. Indeed, this implies that the two objectives become more conflicting, and it is also easier to get stuck in local solutions. By observing the figure, we notice that our algorithm is able to generate a Pareto front, but the number of points obtained is smaller than the number of points in the Pareto front obtained for the original case study. Anyway, aggregating the objective functions gets worse results, since a single point is obtained that is dominated by the Pareto front generated by our algorithm. Hence, we can conclude that our approach is fairly robust, also when compared with the approach based on the aggregation of functions. Indeed, it gives good results for the original case study and it also "reacts" properly when considering changes of the parameters of the probability distributions leading to the two critical scenarios considered.

In Figure 3, we finally report a comparison with *OptQuest for Arena* [29] on the original case study. *OptQuest* is the optimization tool included in the Arena package and it is one of the most commonly used optimization algorithms in the simulation based optimization context. We highlight that, since *OptQuest* only performs single objective optimization, we need to aggregate the two objective functions. To this aim we use the same approach described before, i.e. the weighted sum defined in (2). Regarding the parameters used in *OptQuest*, they were all set to their default values. The tolerance used in the stopping criterion is the same for both the algorithms. As we can easily see by observing the figure, the Pareto front obtained by our method is better than the one obtained by *OptQuest*. Indeed, we get a larger number of points with a better distribution.

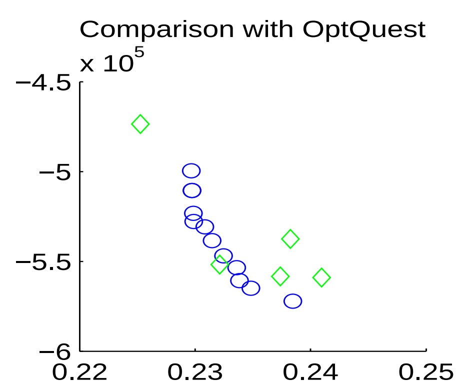

Fig. 3. Comparison between our approach (blue circles) and OptQuest (green diamonds) on the real case study

As final remark, we highlight that, from the computational point of view, the use of DFMO algorithm is less expensive with respect the approaches which aggregate the two objective functions. This is due to the fact that the latter approaches require a complete run of the simulation–optimization process for each generated point. Therefore a significant computational saving is also obtained by using our approach.

V. Conclusion and Future Research

This paper proposes a novel approach for health care services management. In particular, the use of a simulation optimization approach is described for the optimal resources allocation of a ward of a big hospital.

From the methodological point of view, the main contribution of this work is the use of a simulation optimization framework, which integrates a DES model and an optimization algorithm, allowing to study the problem in hand as a multiobjective optimization problem. Then, the DFMO algorithm used enables to obtain an approximate Pareto set of points.

From the practical point of view, this work represents an attempt to provide a quantitative framework for deciding the resource allocation in a hospital ward. This is an innovative contribution since the choice of such resources is usually left to managers which rarely make use of a decision support system. Moreover, the solution of the multiobjective formulation of the problem is provided as a set of points and this helps decision makers to propose the best strategy according to specific indicators related to clinical risk, quality of the care provided, economical benefits both for patients, hospitals and for the NHS.

The application of the approach proposed in this paper to a specific case study, namely the FBF-SGC Hospital in Rome, showed its reliability and allowed significant improvements of the system performance and its efficiency.

As regards future research, two different directions may be followed. On the multiobjective optimization side, it would be crucial developing suitable algorithms for mixed integer problems that guarantee better theoretical and computational properties. On the simulation side, it would be important to use more complex and detailed models that give a better description of the real phenomenon under analysis.

Acknowledgment

The authors wish to thank all the participants to the project "*Business Simulation for Healthcare*" (*BuS-4H*) which allowed the realization of this study and FILAS s.p.a. for granting the project. Moreover, the authors are indebted with the FBF-SGC Hospital and the FBF Foundation for providing the case study. Finally, the authors are grateful to the Department of Biomedicine and Prevention at the University of Rome "Tor Vergata" in supporting the startup of the *Laboratory on Simulation and Optimization of Health Services (SOS-NHS)*. The authors wish to thank the Associate Editor and anonymous Referees for the helpful comments and suggestions which led to improve the paper.

References

[1] R. M. Kaplan and Y. M. Babad, "Balancing influence between actors in healthcare decision making," *Bmc Health Services Research*, vol. 11, p. 14, 2011.

[2] J. Y. Kim, P. Farmer, and M. E. Porter, "Redefining global health-care delivery," *Lancet*, vol. 382, no. 9897, pp. 1060–1069, 2013.

[3] J. Langabeer and J. Helton, *Health Care Operations Management: A Systems perspective*. Burlington, MA: Jones and Bartlett Learning, 2015.

[4] J. Langabeer, *Health Care Operations Management: A Quantitative Approach to Business and Logistics*. Sudbury, MA: Jones and Bartlett Publisher, 2008.

[5] M. P. Fanti, A. M. Mangini, M. Dodoli, and W. Ukovich, "A three–level strategy for the design and performance evaluation of hospital departments," *IEEE Transactions on Systems, Man, and Cybernetics Systems*, vol. 43, no. 4, pp. 742–756, 2013.

[6] B. Moore, "Appropriate-technology for birth," *Lancet*, vol. 2, no. 8458, pp. 787–787, 1985.

[7] OECD, "Caesarean sections per 100 live births, 2009 and change between 2000 and 2009," Report, 2011, available at http://dx.doi.org/10.1787/888932524887.

[8] S. Brailsford, P. Harper, B. Patel, and M. Pitt, "An analysis of the academic literature on simulation and modelling in health care," *Journal of Simulation*, vol. 3, pp. 130–140, 2009.

[9] M. Günal and M. Pitt, "Discrete event simulation for performance modelling in health care: a review of the literature," *Journal of Simulation*, vol. 4, pp. 42–51, 2010.

[10] B. Mielczarek and J. Uzialko-Mydlikowska, "Application of computer simulation modeling in the health care sector: a survey," *Simulation-Transactions of the Society for Modeling and Simulation International*, vol. 88, no. 2, pp. 197–216, 2012.

[11] M. C. Fu, Ed., *Handbook of Simulation Optimization*. New York: Springer, 2015.

[12] M. C. Fu, G. Bayraksan, S. G. Henderson, B. L. Nelson, W. B. Powell, I. O. Ryzhov, and B. Thengvall, "Simulation optimization: a panel on the state of the art in research an practice," in *Proceedings of the 2014 Winter Simulation Conference*. IEEE, 2014, pp. 3696–3706.

[13] S. Amaran, N. Sahinidis, B. Sharda, and S. Bury, "Simulation optimization: a review of algorithms and applications," *4OR*, vol. 12, pp. 301–333, 2014.

[14] S. Lucidi, M. Maurici, L. Paulon, F. Rinaldi, and M. Roma, "A derivative-free approach for a simulation-based optimization problem in healthcare," *Optimization Letters*, vol. 10, pp. 219–235, 2016.

[15] H. Ding, L. Benyoucef, and X. Xie, "Stochastic multi-objective production distribution network design using simulation-based optimization," *International Journal of Production Research*, vol. 47, no. 2, pp. 479–505, 2009.

[16] L. H. Lee, S. Teng, E. P. Chew, K. W. Lye, P. Lendermann, I. A. Karimi, Y. Chen, and C. H. Koh, "Application of multi-objective simulation-optimization techniques to inventory management problems," in *Proceedings of the 2005 Winter Simulation Conference*. IEEE, 2005, pp. 1684–1691.

[17] A. Sallem, M. Fakhfakh, E. Tlelo-Cuautle, and M. Loulou, "Multi-objective simulation-based optimization for the optimal design of analog circuits," in *2011 International Conference on Microelectronics (ICM)*. IEEE, 2011, pp. 1–4.

[18] A. Sallem, B. Benhala, M. Kotti, M. Fakhfakh, A. Ahaitouf, and M. Loulou, "Simulation-based multi-objective optimization of current conveyors: Performance evaluations," in *7th International Conference on Design Technology of Integrated Systems in Nanoscale Era (DTIS)*. IEEE, 2012, pp. 1–5.

[19] J. A. Joines, M. A. Gupta, D. and. Gokce, R. E. King, and K. M. G., "Supply chain multi-objective simulation optimization," in *Proceedings of the 2002 Winter Simulation Conference*. IEEE, 2002, pp. 1308–1314.

[20] A. Ammar, H. Pierreval, and S. Elkosantini, "A multiobjective simulation optimization approach to define teams of workers in stochastic production systems," in *2015 International Conference on Industrial Engineering and Systems Management (IESM)*. IEEE, 2015, pp. 977–986.

[21] K. Miettinen, *Nonlinear Multiobjective Optimization*. Kluwer Academic Publishers, 1999.

[22] F. Baesler and J. Sepulveda, "Multi-objective simulation optimization: A case study in healthcare management," *International Journal of Industrial Engineering: Theory, Application and Practice*, vol. 13, no. 2, pp. 156–165, 2010.

[23] Y. Wang, L. H. Lee, E. P. Chew, S. S. W. Lam, S. K. Low, M. E. H. Ong, and H. Li, "Multi-objective optimization for a hospital inpatient flow process via discrete event simulation," in *Proceedings of the 2015 Winter Simulation Conference*. IEEE, 2015, pp. 3622–3631.

[24] J. Song, Y. Qiu, and Z. Liu, "Integrating optimal simulation budget allocation and genetic algorithm to find the approximate Pareto patient flow distribution," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 1, pp. 149–159, 2016.

[25] *Arena User's guide*, Rockwell Automation, 2010, Allen–Bradley, Rockwell Software.

[26] W. Kelton, R. Sadowski, and D. Sturrock, *Simulation with Arena*, Fourth ed. McGraw–Hill, 2007.

[27] A. Conn, K. Scheinberg, and L. Vicente, *Introduction to derivative–free Optimization*, ser. MOS/SIAM Series on Optimization. Philadelphia, MA: SIAM, 2009.

[28] G. Liuzzi, S. Lucidi, and F. Rinaldi, "A derivative-free approach to constrained multiobjective nonsmooth optimization," 2015, available at http://www.optimization-online.org/DB_HTML/2015/07/5037.html.

[29] *OptQuest for Arena User's guide*, Rockwell Automation, 2012, Allen–Bradley, Rockwell Software.

[30] C. Audet, G. Savard, and W. Zghal, "Multiobjective optimization through a series of single-objective formulations," *SIAM Journal on Optimization*, vol. 19, no. 1, pp. 188–210, 2008.

[31] G. Liuzzi, S. Lucidi, and F. Rinaldi, "Derivative-free methods for mixed–integer constrained optimization problems," *Journal of Optimization Theory and Applications*, vol. 164, pp. 933–965, 2015.