Socio-Economic Planning Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

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

Ambulance allocation optimization model for the overcrowding problem in US emergency departments: A case study in Florida

Jorge A. Acuna *, José L. Zayas-Castro, Hadi Charkhgard

Industrial and Management Systems Engineering, University of South Florida, 4202 E. Fowler Avenue, Tampa, FL, 33620, USA

ARTICLE INFO

Keywords:
Overcrowding
Ambulance allocation
Game theory
Bi-objective optimization
Min–max technique
Mixed integer programming

ABSTRACT

In the last decade, emergency department (ED) overcrowding has become a national crisis for the US healthcare system. Increasing mortality rates, decreasing quality of care, financial losses due to walkouts, and ambulance diversion are some of the consequences of ED overcrowding. Given the increasing demand in terms of ambulance utilization, being able to assign service requests to EDs efficiently, becomes a key function of the emergency medical services. This paper presents new ambulance allocation optimization models to reduce patients' total time to treatment, waiting times; therefore, ED overcrowding. Disparities and fairness are considered in the development of the mixed integer programming models. Under a set of assumptions, we apply our strategies to allocate 75 ambulance emergencies to 11 EDs in a specific county in Florida. Heterogeneous types of patients, demand characteristics, and geographical/facility information are considered in the models. Based on numerical experiments and the situation in Florida, we show that the optimization techniques can be utilized for large problems and result in up to 31% improvement of the current decentralized model. Further analysis reveals the negative or positive impact that the strategies have on each patient, giving new insights for future policy modifications. Bi-objective, single objective, and game theory optimization models are implemented in this study.

1. Introduction

Emergency departments (EDs) have traditionally been the places where patients go to receive care in life-threatening situations, or when acute events occur [1]. These health facilities are focused on the care and management of patients who need to be treated within a short period of time. Due to the diverse diseases and conditions treated in an ED, admissions are driven by priority-based policies. Patients may arrive at EDs via special emergency vehicles or on their own (walk-in). These arrivals are not planned or easily predictable, and their stochastic nature impacts the workload and has negative consequences in terms of quality of care and patient waiting time. Therefore, the appropriate assignment of priorities is needed to ensure that each patient receives correct and timely treatment [2,3]. As the United States (US) health care system has evolved under the pressure of political, economic, and clinical matters, the EDs system is being forced to expand their role despite being recognized as the most expensive care [4]. Nowadays, the EDs represent a baseline for the safety net of the health care system regardless of the economic or social status of patients [5]. From the 145 million of patients attending US EDs annually, over 12 million are uninsured [6]. Additionally, EDs have taken on the responsibilities of public health surveillance, disaster preparedness, caring for indigent people, and primary care due to barriers to access at this level of care [7,8]. In 2016 the number of annual ED visits per 1000 habitants reached 441 emergencies [9]. Despite the increase in demand for and utilization of emergency services, the number of EDs in the US, along with the total number of hospitals and hospital beds, have decreased significantly in the last two decades [10,11]. As a consequence, the ability of emergency teams and physicians to deliver adequate care is affected [12]. The disproportion between supply and demand, caused by the factors stated above, has led to the national crisis of overcrowding [5,13,14].

Although there is no agreement on an exact definition of ED overcrowding, it often refers to an overburden of patients in treatment areas that exceeds the ED resources capacity, frequently requiring that medical care be provided in ED corridors and other temporary examination areas [15]. A national US survey showed that more than 90% of large hospitals have their EDs operating at or over capacity [13]. Overcrowding has been associated with long waiting times, especially for patients who are not critically ill, thus decreasing the quality of care; increasing mortality; increasing patient walkouts; increasing ambulance offload delays; and increasing ambulance diversion among other externalities [16–18].

* Corresponding author.
E-mail addresses: jorge@mail.usf.edu (J.A. Acuna), josezaya@usf.edu (J.L. Zayas-Castro), hcharkhgard@usf.edu (H. Charkhgard).

https://doi.org/10.1016/j.seps.2019.100747
Received 21 December 2018; Received in revised form 26 August 2019; Accepted 30 September 2019
Available online 1 October 2019

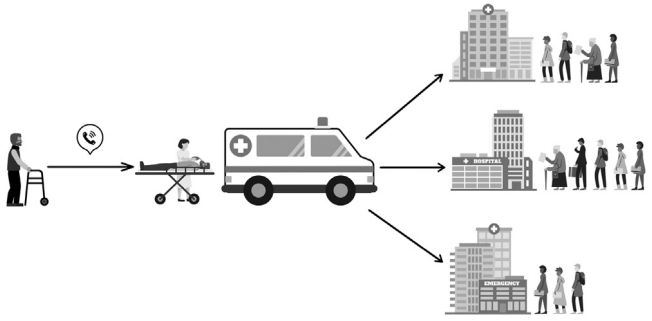


Fig. 1. Standard system process.

In addition to the impact perceived by the patients, hospitals are facing financial losses due to walkouts and ambulance diversion as consequences of overcrowding in EDs. Researchers have found that there is a revenue loss of USD 3.9 million per year for a 500-bed hospital [19]. According to the National Hospital Ambulatory Medical Care Survey, between 2003 and 2015 the number of emergency patients who arrived at EDs via ambulances increased from approximately 16 million to 21 million per year in the US [20,21]. In this context, the allocation of service requests to EDs plays an essential role in overcrowding, not only in terms of life-threatening scenarios but also from a financial perspective.

1.1. Problem description

In the US, emergency medical services (EMS) refers to the treatment and transport of injured or sick patients to hospitals. Despite the tremendous importance of the system, the lack of strong representation at the federal level and the increasing number of local agencies and system models make a fully encompassing definition impossible [22,23]. There are more than 21,000 licensed local EMS agencies in the US that can be classified into three main groups according to the tasks they perform: (1) Agencies that focus primarily on 911-based emergencies with or without transport; (2) EMS that provide scheduled medical transport; and (3) EMS that provide emergent inter-facility transport [24]. In this study, we focus on the first category with transport, particularly the utilization of ambulances. Fig. 1 shows the most common process patients go through. The steps include calling 911 to request help, receiving initial treatment by the ambulance personnel, and choosing the destination ED. Across the US, different protocols are followed in determining the destination ED. The different states and counties customize their policies. For example, Hillsborough County in Florida gives the responsibility of determining the most appropriate facility to the senior caregiver at the scene. He or she has to decide the final destination in accordance with the state-approved Hillsborough County Trauma Agency Plan. Basically, a patient is transported to the chronological/developmental, age-appropriate, treatment facility [25]. It is necessary to add other elements to these customized protocols, which can lead to the inefficient allocation of service requests to EDs. For example, Medicare only covers ambulance service to the nearest appropriate medical facility that can provide the care [26], and some private insurances cover medical transportation only if the ambulance company is a partner organization or the provider is part of the network. Additionally, the large number of ambulance providers by area (e.g., county) is a barrier to coordination with EDs.

Several studies have been conducted in the framework of EMS planning and facility allocation. In this paper, we divided the literature into two main pillars. The first considers publications related to ambulance assignment to call request, dispatch policies, and allocation of EMS to increase coverage. This pillar has been widely studied in the last decade; from these research efforts, many publications have emerged. [27–35]. We are interested in the second pillar, which focuses on the allocation of medical emergency service requests to ED, the decision that pairs ambulance emergencies and hospitals' ED. The following are publications that provide interesting results and understanding of the different techniques available for the allocation of services request to ED or similar problems.

In [36], through queuing theory, the interface between regional EMS provider and EDs serving ambulances and walk-ins was modeled. Markov chain was utilized to analyze the ambulance offload delays due to overcrowding in EDs. A simulation was implemented to validate the model assumptions. In [37], a stylized queuing model with blocking was used to analyze the effect of routing decisions on EMS and to help in creating proper patient allocation policies for multiple hospitals in a region. [38] used mixed integer programming (MIP) to optimize the allocation of ambulance request to EDs, considering geographical information in a region of Italy. A complete offline picture of an optimal assignment was used to evaluate the price of anarchy, where centralized allocation is useful as a reference for the state of the art of the decentralized approach and future reorganization ideas. In [39], through a queuing game between two EDs that minimizes the waiting time, it was found that the decentralized decisions related to diversion generated a lack of pooling benefits. The existence of a defensive equilibrium was also discovered. Additionally, the benefits of a centralized planner that maximizes the social optimum in terms of diversions were estimated. In [40], approximate dynamic programming optimization was utilized to reallocate ambulances after first-aid requests to the waiting location in a time-efficient manner. In [41], a design of pragmatic retrospective cohort analysis of all the planned and unplanned ED visits was studied. The planned visits followed an ambulance service secondary telephone triage, and the impact on ED admissions was measured. The study shows that planned visits were more likely to be admitted to hospital and to find a suitable ED, compared to emergencies that were not planned ED visits. [42] studied proactive destination selection considering real-time data of the regional EMS system. The authors showed that proactive destination selection could improve regional capacity and helps to reduce ambulance diversions. [43] implements a "nurse navigator program" to identify the ideal destination ED between two nearby hospitals. The improvement of load-balancing considered real-time information from EDs that is utilized to inform the EMS. The article showed that a proactive mechanism of nurse navigator with real-time data could decrease EMS turnaround time.

Our contribution: In the present work, we investigate the overcrowding problem from the viewpoint of the allocation of emergency requests to EDs in the US system. Three optimization strategies are proposed, taking into account both the EDs' workloads and service allocation. We consider remote triage management, as suggested by [38]. That is, anticipating patient priority can help in handling emergency requests to EDs, thus assuring the best possible service level.

Unlike previous approaches, we recognize the fact that minimizing total or average time across all emergencies, may favor certain patients over others. This is a common problem in optimization models focused on efficiency-based objectives [44]. By incorporating terms of fairness and disparities, we aim to provide patient-centered solutions that recognize the necessity of improving efficiency in the context for which they are designed.

When working on healthcare systems, there is an imperative need to acknowledge the issues previously described. Prioritizing the common welfare of the community over individual patients implies potential loss of lives. This situation may be disregarded in other fields, where the thing being optimized is inert, or the externalities only have an economic impact.

The first strategy that we propose focuses on system efficiency and is modeled using MIP with a single objective function. Our second strategy focuses on the reduction of disparities, and the Min–max technique is used to develop a bi-objective MIP problem. Finally, the

Table 1
Model indices.

Symbol	Definition
i	Emergency in the system, where $i \in I$.
j	Emergency department in the system, where $j \in J$.
t	Period of time, where $t \in T$.
p	Pathology, where $p \in P$.

integration of game theory and MIP enabled us to generate a grand coalition between our objectives/imaginary players to improve the efficiency of the system and provide fair payoffs to the patients. A non-symmetric bargaining game is utilized in this last strategy.

While each strategy improves the current decentralized system, we analyze and discuss the differences among them. Total waiting times, fairness, and efficiency are some of the critical points under analysis. These strategies represent potential guidelines to modify policies, increase service capacity, and reduce overcrowding in EDs.

1.2. Structure

The following section describes the notation and definitions for the mathematical representation of the system. Section 3 covers the formulation of the problem and presents the required modifications of each strategy under analysis. In Section 4, the formulation is tested through a numerical experiment. Section 5 presents a case study in Florida, where the strategies are applied. Finally, Section 6 comprises the conclusion and future research directions.

2. Notation and definitions

We modeled the time horizon of the system using a discrete ordered set $T = 1, \dots, t$. Each element of the set represents a period of time when emergencies may occur, and decisions need to make. Let I be the set of emergencies that need to be assigned to EDs in the time horizon T. G denotes the set of areas in which the region under analysis is divided. Let U represent a subset of areas G where the hospitals' EDs are located. The set of hospitals' EDs in the region is defined by J. Finally, let P and C represent the set of pathologies and priority codes, respectively.

Similar to the work done in [38], each emergency i is modeled by a quadruple (g_i, ψ_i, c_i, p_i), where $g_i \in G$ represent the location of emergency i, $\psi_i \in T$ is the time period in which the emergency i occurs, $c_i \in C$ is the priority code assigned by the ambulance personnel to i, and $p_i \in P$ is the pathology associated with i.

Every ED j is also modeled by a quadruple $(u_j, w_j^t, \sigma_{pj}, q_{pj})$, where $u_j \in U$ represents the location of ED j. w_j^t is the waiting time in ED j at period t, which depends on the number of ambulance emergencies assigned to j, as well as walk-ins and patients already waiting at j. The capability of the ED j to treat the pathologies is defined by a binary parameter σ_{pj}, where $p \in P$. Finally, the quality of treatment that ED j can provide for a given pathology p is denoted by q_{pj}.

Tables 1–3 summarize all indices, parameters, and variables required for the mathematical formulations.

3. Mathematical formulation and strategies

This section introduces the model formulation of the problem and describes the modifications required for each strategy.

Table 2
Model parameters.

Symbol	Definition
c_i	The priority code assigned to the emergency i.
d_{ij}	The travel time of emergency i to arrive at ED j.
q_{pj}	The quality of care for pathology p offered by ED j.
φ_j	The initial number of patients waiting at ED j.
l_j^t	The number of walk-in patients at ED j in time period t.
α_j^t	The proportion of patients still waiting to be treated in j and period t.
σ_{pj}	Binary parameter, 1 if ED j can treat pathology p, 0 otherwise.
δ_i^t	Binary parameter, 1 if emergency i occurs in period t, 0 otherwise.
ψ_i	Time period in which emergency i occurs.
w_{max}	The maximum waiting time before starting ambulance diversion.
β	The minimum quality of care expected for every patient.
a_j	Slope coefficient of the waiting time function in j.
b_j	Intercept coefficient of the waiting time function in j.
r_i	Binary parameter, 1 if i is traveling more than 15 min, 0 otherwise.
θ_i	Time to treatment for i without the creation of coalition (status quo).
np	Number of players in the bargaining game.

Table 3
Model variables.

Symbol	Definition
x_{ij}	1 if patient i is assigned to ED j, 0 otherwise.
w_j^t	Waiting time at ED j in time period t.
n_j^t	Total number of patients being assigned to ED j in time period t.
z_j^t	1 if ED j in time period t is not doing diversion, 0 otherwise.
v_{ij}	Total time that emergency i waits to receive treatment if assigned to ED j.
λ_{c_i}	Maximum time that i with priority code c_i has to wait for treatment.
y_i	Expected time to treatment of i.
$\hat{y}_{if}$	Expected time to treatment of i and his/her f copies.

3.1. Mixed-integer programming formulation

To describe the MIP formulation, which is the core of our strategies, we introduce the basic constraints.

- Constraint (1) ensures that each emergency is assigned to one ED among all periods.

$$\sum_{j \in J} x_{ij} = 1 \qquad \forall i \in I \tag{1}$$

- Constraint (2) captures the total number of emergencies assigned to a particular ED j in a time period t.

$$\sum_{i \in I} x_{ij} \delta_i^t = n_j^t \qquad \forall j \in J \quad \forall t \in T \tag{2}$$

- Constraint (3) initializes the variable n_j^t in period $t = 0$.

$$n_j^0 = \varphi_j \qquad \forall j \in J \tag{3}$$

- Constraint (4) represents the function of waiting time at each ED j for a given period t. Different studies have been performed to understand the relationship between ED waiting time and its occupancy levels [45–47]. In [38], the authors showed, through the analysis of historical data, that the average waiting time can

be expressed as a convex piecewise linear function of occupancy levels. Based on this contribution and for simplicity, we assumed the waiting time of each ED as a linear function of their workload in a given period t. This function considers the arrival of emergencies through ambulances and walk-ins in the current period, furthermore, incorporates the patients from previous periods still waiting for treatment at ED j. For more information regarding convex piecewise linear functions, we refer the readers to [48].

$$w_j^t = a_j[\alpha_j^t(n_j^{t-1} + l_j^{t-1}) + n_j^t + l_j^t] + b_j \quad \forall j \in J \quad \forall t \in T \tag{4}$$

- In (5), we ensure that if ED j cannot treat pathology p, emergency i is not assigned to j.

$$x_{ij} \leq \sigma_{p_i j} \qquad \forall i \in I \quad \forall j \in J \tag{5}$$

- (6) defines the threshold of maximum waiting time before ambulance diversion across all EDs. M represents a large value (e.g., the value of the worst scenario for any ED in terms of waiting time) and z_j^t is a binary variable equal to one if $w_j^t \leq w_{max}$, and zero otherwise.

$$w_j^t \leq w_{max} + M(1 - z_j^t) \quad \forall j \in J \quad \forall t \in T \tag{6}$$

- Constraint (7) generates the diversion of emergencies if ED j exceeds the maximum waiting time threshold.

$$x_{ij} \leq z_j^t \quad \forall i \in I \quad \forall j \in J \quad t = \psi_i \tag{7}$$

- (8)–(11) define the range and type of variables.

$$x_{ij} \in \{0,1\} \quad \forall i \in I \quad \forall j \in J \tag{8}$$

$$n_j^t \in \mathbb{Z}^+ \quad \forall j \in J \quad \forall t \in T \tag{9}$$

$$w_j^t \in \mathbb{R}^+ \quad \forall j \in J \quad \forall t \in T \tag{10}$$

$$z_j^t \in \{0,1\} \quad \forall j \in J \quad \forall t \in T \tag{11}$$

3.2. Strategy 1: System efficiency

Our first strategy aims to reduce the total time that emergencies have to wait to be treated and to minimize the transfer and waiting time at EDs for all emergencies. To give preference to the patients in the worst condition, we use an emergency priority code as a weight in each summation of the objective function. This approach, based on maximizing overall system efficiency, is similar to the work done in [38–40].

The following problem provided the general formulation of strategy 1.

Problem 1.

$$\min \sum_{i \in I} \sum_{j \in J} \sum_{t \in T} c_i(d_{ij} + w_j^t)x_{ij}\delta_i^t \tag{12}$$

$$\max \sum_{i \in I} \sum_{j \in J} q_{p_i j} x_{ij} \tag{13}$$

subject to (1)–(11)

In Problem 1, the first objective function (12) minimizes the summation of times, where for each patient the travel time and the waiting time at the ED are multiplied by the corresponding priority code. At the same time, the second objective function (13) maximizes the summation of the quality of care that patients receive. This last objective cannot be easily added to the previous function because the parameter to transform quality into time is unknown and difficult to estimate. However, the second objective function can be turned into a constraint to facilitate calculations.

From Problem 1, we can notice that the first objective function (12) is non-linear. Therefore, we next describe the linearization; four new constraints are added, and the objective function is modified.

- Constraint (14) defines a new variable that is smaller or equal to the sum of the travel and waiting times.

$$v_{ij} \leq d_{ij} + w_j^t \quad \forall i \in I \quad \forall j \in J \quad t = \psi_i \tag{14}$$

- Constraint (15) forces the new variable for total time to be zero if the binary variable x_{ij} is equal to zero. Basically, the time to treatment in ED j for emergency i cannot exist if the patient is not assigned to that ED. When x_{ij} is equal to one, the value of M has to be large enough to ensure that the variable v_{ij} is smaller or equal to the sum of the travel and waiting times.

$$v_{ij} \leq M x_{ij} \quad \forall i \in I \quad \forall j \in J \tag{15}$$

- If variable x_{ij} is equal to one, then constraint (16) combined with (15) makes the new time variable v_{ij} equal to the sum of the travel and waiting times.

$$v_{ij} \geq d_{ij} + w_j^t - M(1 - x_{ij}) \quad \forall i \in I \quad \forall j \in J \quad t = \psi_i \tag{16}$$

- Constraint (17) defines the range and the type of variable to v_{ij}^t.

$$v_{ij} \in \mathbb{R}^+ \quad \forall i \in I \quad \forall j \in J \tag{17}$$

In Appendix A, we further discussed how the linearization procedure applied to constraints (14)–(17) works, including a reference that provides additional insights about this topic.

The next step converts the objective function (13) into a constraint. This transformation helps to solve larger instances of the problems and in a shorter period of time. This is a common advantage of models with a single objective function versus bi-objective models [49]. However, proper knowledge about the minimum level of quality expected for every patient is required; an incorrect value may make the problem infeasible.

- Constraint (18) sets a minimum level of quality of care for every patient. Parameter β is used as the lower bound.

$$\sum_{j \in J} q_{p_i j} x_{ij} \geq \beta \quad \forall i \in I \tag{18}$$

With the previous constraints and modifications added to the model, the new problem for strategy 1 is:

Problem 2.

$$\min \sum_{i \in I} \sum_{j \in J} c_i v_{ij} \tag{19}$$

subject to (1)–(11) and (14)–(18)

In Problem 2, the objective function (12) is replaced by objective function (19). The main advantage is the change from a non-linear to a linear function that can be easily managed with commercial solvers. Additionally, the objective function (13) is converted into a constraint to speed up the processing time of the solver.

Consequently, using Problem 2, we can solve the first strategy to improve the allocation of service requests to EDs. It is important to notice that constraints (14) and (15) can be ignored due to the minimization characteristic of the problem.

As mentioned in the previous section, the approach described for strategy one is based on optimizing an efficiency-based objective. The development of the next two strategies is founded on the necessity of finding fair approaches to improve the system.

3.3. Strategy 2: Min–max on disparities

In our second strategy, we implemented the min–max technique to reduce disparities [49]. Given that the literature on healthcare shows the existence of vast inequality in access to care between rural and urban areas [50], we focus our approach on patients traveling more than 15 min to EDs. Previous models in the literature do not consider that patients with the same severity of illness, but distinct travel times should be prioritized differently in EDs.

A new objective function has to be added to the model in strategy one to generate a non-dominated frontier between the system efficiency and the reduction of disparities. In a bi-objective problem, the non-dominated frontier is defined as the set of feasible solutions in the criterion space that cannot be improved in the value of one objective without worsening the other [51].

Problem 3.

$$\min \sum_{i\in I}\sum_{j\in J} c_i v_{ij} \tag{19}$$

$$\min \sum_{i\in I} \lambda_{c_i} r_i \tag{20}$$

subject to (1)–(11), (14)–(18), and

$$v_{ij} r_i \le \lambda_{c_i} \quad \forall i \in I \quad \forall j \in J \tag{21}$$

Problem 3 presents the formulation required in strategy 2. The objective function (20) minimizes the sum of maximum times by priority class. Only emergencies with travel time longer than 15 min are considered in this objective. Using independent time variables by priority class helps to develop different upper bounds by type of patient. Constraint (21) sets the time limit for emergencies with travel time longer than 15 min. Finally, the model for the min–max on disparities considers objective functions (19), (20), and constraints from (1)–(11), (14)–(18) and (21) as presented in Problem 3.

Strategy two generates a non-dominated frontier of solutions, where each solution has components of efficiency and fairness. The fact that a non-dominated frontier is created in the criterion space (space of objective function values) is directly related to the trade-off existing between the two components. In other words, the improvement of one component can be obtained only at the expense of the deterioration of the other component [49].

3.4. Strategy 3: Game theory and cooperative model

In our last strategy, we implement a non-symmetric version of the two-player cooperative games from [52,53], also known as the non-symmetric bargaining games. A bargaining problem is a cooperative game where the players create a grand coalition instead of competing with each other to get better payoffs [50].

Two important axioms relating to bargaining games are individual rationality and Pareto optimality. The first one establishes that no player will accept a payoff lower than the one under disagreement. In our problem, this axiom represents the patients who, based on their total current time, are not willing to accept a longer total time. The second axiom introduces the trade-off among the players when a solution has been obtained. Thus, the solution guarantees that the total time for one patient cannot be decreased without negatively affecting the total times of other patients. It is noted that in practice patients do not negotiate; however, the model pretends that they do to safeguard the interest of each patient while improving overall efficiency.

In the allocation of ambulance request to EDs, the patients or emergencies represent the players, and the time to treatment is the payoff they received. To model the bargaining game, non-symmetric powers are needed to represent the priority codes of the emergencies. Therefore, the priority of each emergency is the capacity of the player to influence the negotiation to obtain a better time or payoff.

The objective function (22) maximizes the differences between the status quo or disagreement and the expected times obtained through the new coalition for more than two players. The priority code c_i gives different negotiation power to the players depending on the emergency condition.

$$\max \prod_{i\in I} (\theta_i - y_i)^{c_i} \tag{22}$$

We next present the basic formulation of strategy 3.

Problem 4.

$$\max \prod_{i\in I}\prod_{f=1}^{c_i} (\theta_i - \hat{y}_{if}) \tag{23}$$

subject to (1)–(11), (14)–(18), and

$$y_i = \sum_{j\in J} v_{ij} \quad \forall i \in I \tag{24}$$

$$y_i = \hat{y}_{if} \quad \forall i \in I \quad \forall f = 1, \dots, c_i \tag{25}$$

$$y_i \le \theta_i \quad \forall i \in I \tag{26}$$

In Problem 4, the objective function (23) is the transformation of the objective function (22) into the standard Nash optimization problem. c_i copies are created for each patient, increasing the number of players artificially. Constraint (24) is added to match the value of the expected time to treatment y_i with the value of the variable v_{ij}. Meanwhile, constraint (25) guarantees that the copies of the expected time to treatment have the right value. Finally, in (26), the axiom of individual rationality is established; the patients only accept payoff greater or equal to the status of disagreement.

As can be observed, the objective function (23) has to be linearized. Because the variables are not binaries, a special reformulation is necessary. The second-order cone problem (SOCP) can be used to reformulate this model and has the advantage that commercial solvers can solve it. Basically, a new non-negative variable γ and a geometric constraint are added to the model. To avoid computational issues with the geometric constraint, a final reformulation is presented in Problem 5. This last model replaces the geometric constraint with a set of non-negative variables and constraints.

Next, we show the mathematical reformulation of SOCP. The objective function (27) and the constraints (28)–(32) are introduced to our model. Let κ be the smallest integer value such that $2^\kappa \ge np$. The set of non-negative variables Γ and τ are generated for the mathematical transformation of Problem 4; accordingly, they do not have an explicit meaning for our model.

Problem 5.

$$\max \gamma \tag{27}$$

subject to (1)–(11), (14)–(18), (24)–(26), and

$$0 \le \gamma \le \Gamma \tag{28}$$

$$0 \le \Gamma \le \sqrt{\tau_1^{\kappa-1}\tau_2^{\kappa-1}} \tag{29}$$

$$0 \le \tau_j^l \le \sqrt{\tau_{2j-1}^{l-1}\tau_{2j}^{l-1}} \quad \forall i = 1, \dots, 2^{\kappa-l} \quad \forall l = 1, \dots, \kappa-1 \tag{30}$$

$$0 \le \tau_j^0 = \theta_j - \hat{y}_j \quad \forall j = 1, \dots, np \tag{31}$$

$$0 \le \tau_j^0 = \Gamma \quad \forall j = np+1, \dots, 2^\kappa \tag{32}$$

Finally, the strategy of the cooperative model has to consider objective function (27) and constraints (1)–(11), (14)–(18), (24)–(26), and (28)–(32) to solve the allocation of ambulance request to EDs. As established by the second axiom, the solution obtained guarantees the trade-off among players. A unique solution can be found in the non-symmetric bargaining game.

4. Numerical experiment

This section introduces the numerical experiment of our strategies. Different random instances are run to verify the size of problems that can be solved by our formulations. Strategy one is selected to run the analysis, given that its formulation is utilized by strategies two and three. Additionally, it is the fastest model to solve due to the single objective function. The model is implemented in Julia language and uses Gurobi 7.5.2 as the MIP solver. The experiments are conducted on a Dell OptiPlex 5040 with an Intel(R) Core(TM) i7-6700U @ CPU 3.40 GHz, 16 GB RAM, with a 64-bit operating system, Windows 10 Pro.

We performed all instances based on simulated data for both patients' conditions and EDs' characteristics. The random values are obtained based on the case study presented later. The maximum waiting time is fixed at 120 min, the minimum quality of care at 70%, and the value of big M at 200. See Appendix B for more information regarding the data utilized in this section. In our problem, big M should take the value of the worst scenario for any ED in terms of waiting time. The value of big M has to be carefully calculated to improve the performance.

Six classes are defined to run the computational study. The first index of a class represents the number of emergencies, and the second represents the number of EDs. Table 4 shows the results of 30 random instances and their respective averages by class. As can be observed, classes 50–100, 80–15, 100–15, and 100–20 are solved to optimality with 0% gap, while classes 150–20 and 200–30 show gap values greater than 0%. The time required to solve the first four classes is less than a minute, which we associated with a good capacity of the formulation to solve medium and large size problems to optimality.

Because instances with 150 or more patients and 20 or more EDs are not solved in a reasonable period of time, we have limited the solver elapsed time to three minutes. The same instances are shown in Table 5, but in this case, the solver is limited to one hour. These two settings show the performance deterioration of the formulation as classes increase in size. The solver takes three minutes to obtain the 10.22% and 11.14% average optimality gap, respectively. The gap is reduced to 8.60% for class 150–20 and 9.19% for class 200–30 after one hour running, which is equivalent to less than one minute improvement in the average mean time to treatment. Therefore, the time effort cannot be justified through a real improvement of the solution, and the deterioration of the solver becomes evident.

When looking to the number of variables presented in Table 4, we see the complexity of the problem under analysis, and the thousands of possible interactions that need to be considered when handling the allocation of emergency requests to EDs.

The following section introduces the analysis for a real scenario, where the three strategies are implemented.

5. Case study

This section presents a case study based on the context of Hillsborough County (Florida) in the United States. Hillsborough County has a population of 1.4 million of habitants and covers an area of 1,266 square miles in the state of Florida. The case uses the Healthcare Cost and Utilization Project (HCUP) databases for the state of Florida of the year 2014. We combined the State Emergency Department Databases (SEDD) [54], with the State Inpatient Databases (SID) [55]. The first database covers all ED visits per state that did not result in an admission, while the second, includes 97% of all the discharges in the community hospitals across the United States. Based on the Five Year Trauma Plan Update of Hillsborough County [56], there are two trauma and nine non-trauma centers that meet the criteria of receiving centers for emergency stabilization. These eleven emergency departments were selected from the Florida State HCUP database to conduct the analysis.

Considering the national percentage of arrivals to EDs through ambulances (14%) [57] and the HCUP database, we estimated that 150 ambulance emergencies and 900 walk-ins are received daily in the EDs of Hillsborough County. Reports [6,58] were utilized to collect data related to the priority code assigned to each emergency. We used the information provided in [59] as the source for the quality of care by type of pathology and the hospitals' EDs capability to provide treatment. We considered seven official subdivisions of the county (Brandon, Keystone, Palm River, Plant City, Ruskin, Tampa, and Wimauma-Riverview) [60] where emergencies may occur and the six diagnoses with the higher frequency in the HCUP databases (chest pain, complication of pregnancy, respiratory infections, headache, abdominal pain, and urinary tract infections). Combining the data available in Hillsborough County GeoHub [61] and the software ArcGIS Desktop 10.5, a geographical representation of the county was made. Fig. 2(a) shows the location of the EDs and county subdivisions with different layers. Fig. 2(b) presents in gray the high-density population areas by zip code. Based on the locations of the EDs and county subdivisions, the travel time was estimated in a range of plausible values. Considering the population levels [62], the number of emergencies per area was assigned. Using R software 3.6.1 and the HCUP database, we fitted a linear regression for each ED to estimate the parameter a_j of the waiting time function (average adjusted R-squared of 0.937). At zero occupancy the waiting time in the EDs is zero; therefore, the intercept b_j of each function is zero. The 2018 Florida Statues [63] gives hospitals the opportunity to develop emergency room diversion programs but do not specify standards. The State of Florida, in a reassessment of EMS document from 2013, points out the necessity of matching system resources with patient needs in developing diversion policies [64]. Given that the national average waiting time in EDs is around 58 min [65], we established a maximum waiting time of 150 min before diversion. Basically, no patient should wait more than 2.5 times the national average. This setting is used for our three strategies and the current decentralized model. An instance with 75 ambulance emergencies, 450 walk-ins, and 11 EDs, which approximates half a day of service demand and resources in Hillsborough County, was applied to our models. For simplicity, we considered a 12-hour time frame and divided it evenly in 12 periods. The literature shows that different levels of priorities have been studied in EDs [66,67]. We considered a five-level triage system as described by the National Hospital Ambulatory Medical Care Survey (non-urgent, semi-urgent, urgent, emergent, and immediate) [6]. Notice that the models developed can consider different number of areas, periods, and pathologies in accordance to the context under analysis. The minimum quality of care was fixed at 70% based on the national average percentage of patients receiving recommended acute care [68].

The bi-objective model for the min–max on disparities is solved using the perpendicular search method [69,70]. The perpendicular search method utilizes a search direction that is always perpendicular to the parameter axis. The last implies that one parameter at a time changes, while all other parameters keep the same value. After one step, the next parameter is employed, and the process continues [71].

5.1. Results

This section provides graphical views of the results. The emergencies are sorted based on the total time to treatment on the decentralized model in increasing order. Fig. 3 presents the results for total time to treatment of each emergency in the current/decentralized model and strategy one (efficiency model). The horizontal axis shows the emergencies, and the vertical axis shows the total time to treatment

Table 4
Numerical results, part I.

Class	Objective value	Mean time to treatment (min)	Mean travel time (min)	Mean waiting time (min)	Time running (s)	Primal–Dual gap (%)	No. of variables
50–10	1717	52	14	38	0.04	0	1200
	1546	49	15	34	0.09	0	1200
	1948	64	13	51	0.11	0	1200
	1542	56	13	43	0.05	0	1200
	2329	79	12	67	0.40	0	1200
Average	1816	59	14	47	0.138	0	1200
80–15	2380	50	11	38	1.33	0	2705
	2542	51	14	37	2.77	0	2505
	2537	51	13	38	1.06	0	2505
	2706	52	13	39	1.52	0	2505
	2622	54	13	41	0.31	0	2505
Average	2557	52	13	39	1.39	0	2505
100–15	3511	56	13	43	1.71	0	3325
	3103	50	12	37	3.42	0	3325
	3612	55	13	41	3.87	0	3325
	3135	54	12	42	1.51	0	3325
	3373	50	13	36	5.35	0	3325
Average	3347	53	13	40	3.17	0	3325
100–20	3121	49	13	36	2.88	0	4400
	2786	42	11	31	19.03	0	4400
	3014	48	13	35	6.47	0	4400
	3057	48	12	36	34.33	0	4400
	2607	45	14	31	2.87	0	4400
Average	2917	46	12	34	13.11	0	4400
150–20	5184	55	13	42	180	4.55	6450
	4268	50	12	37	180	11.87	6450
	4682	49	12	36	180	12.19	6450
	5340	56	12	43	180	10.21	6450
	4734	50	13	37	180	12.30	6450
Average	4842	52	12	39	180	10.22	6450
200–30	6004	47	11	36	180	6.00	12650
	5120	43	12	31	180	14.16	12650
	7290	51	11	39	180	8.03	12650
	5727	43	11	32	180	10.65	12650
	5874	47	12	35	180	16.86	12650
Average	6003	46	11	35	180	11.14	12650

Table 5
Numerical results, part II.

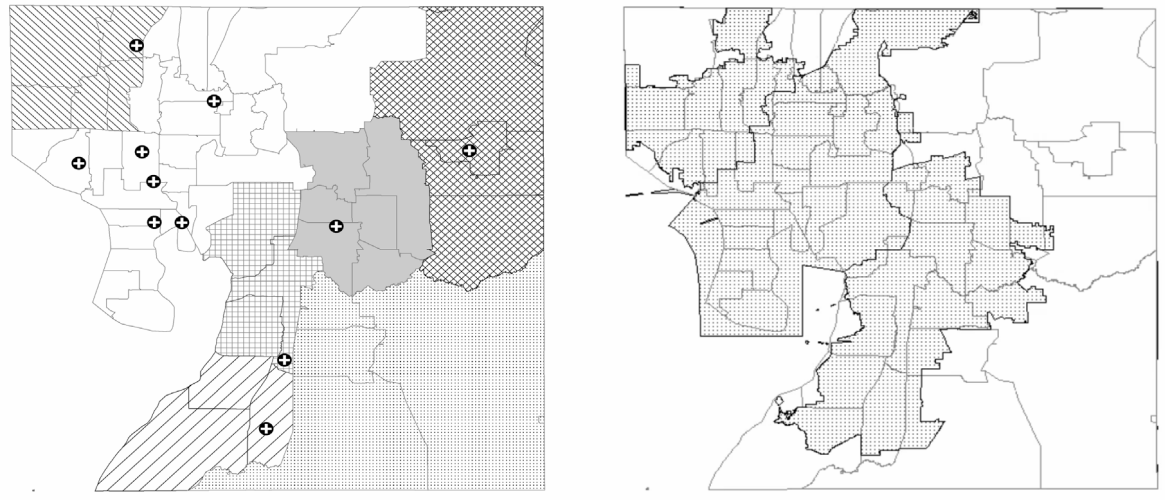
Class	Objective value	Mean time to treatment (min)	Mean travel time (min)	Mean waiting time (min)	Time running (s)	Primal–Dual gap (%)	No. of variables
150–20	5183	55	13	42	3600	4.38	6450
	4255	50	12	38	3600	10.04	6450
	4652	48	12	37	3600	10.36	6450
	5337	56	13	43	3600	8.65	6450
	4725	50	13	37	3600	9.55	6450
Average	4830	52	13	39	3600	8.60	6450
200–30	5997	47	11	36	3600	5.00	12650
	5063	43	11	31	3600	11.86	12650
	7280	51	12	39	3600	6.62	12650
	5694	43	11	32	3600	9.35	12650
	5783	47	12	35	3600	13.14	12650
Average	5963	46	11	35	3600	9.19	12650

in minutes. In a decentralized model, the emergencies are assigned to the closest ED that can treat the patients properly. The results obtained using the model mentioned above are fundamental to contrast with the proposed strategies. As can be observed, in the efficiency strategy, the values of total time to treatment of most patients are reduced significantly compared to the decentralized model. However, another small group of emergencies received a negative impact on its times to treatment, particularly, emergency 1, 19, 20, 26, and 27. This last group may represent the favoring of the common good over the welfare of some patients. We analyze the possible implications of this strategy later.

In Fig. 4, the non-dominated frontier for the bi-objective problem is shown. For the min–max strategy, the trade-off between efficiency (horizontal axis) and fairness (vertical axis) is created. With this technique, the decision-maker can choose a particular solution according to the necessities of the system. Objectives one and two of Fig. 4 are functions of time; therefore, the smaller the value of the objective is, the better the efficiency and fairness of the system are, respectively.

Fig. 5 shows the results when comparing strategy two and the current model. A similar analysis to Fig. 3 can be done in this comparison, where a group is affected positively by the min–max strategy and another smaller group negatively. The values of the min–max are obtained when the point (6944, 4244) is chosen from the non-dominated frontier. These objective values provide the solution with the highest level of fairness and the lowest level of efficiency that the min–max strategy can generate. Therefore, the solution of point (6944, 4244) shows the best contrast when compared with the efficiency strategy solution. These changes are analyzed in detail in Figs. 7 and 8. See Appendix C for more information related to solutions to other points in the non-dominated frontier.

Fig. 6 shows the results of our last strategy versus the current decentralized model. One of the main advantages of the non-symmetric



(a) Emergency departments (b) Population density by zip code

Fig. 2. Hillsborough County.

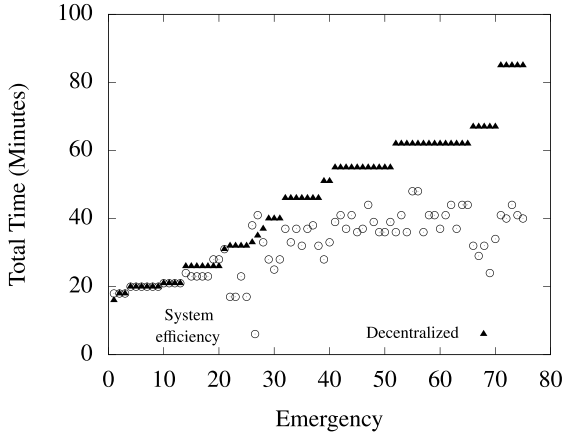


Fig. 3. Total times, decentralized vs. system efficiency.

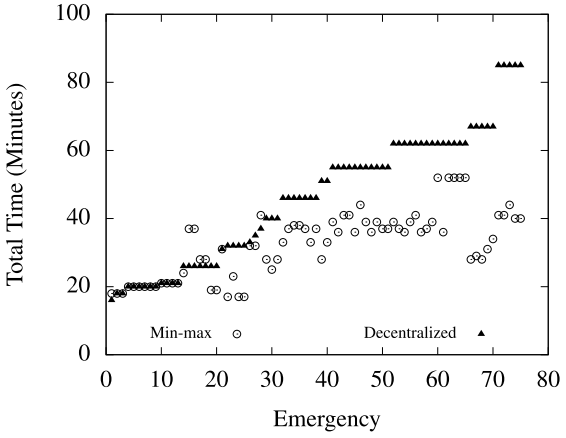


Fig. 5. Total times, decentralized vs. min–max.

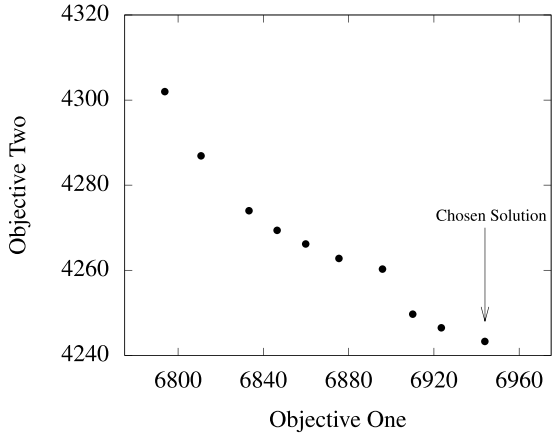


Fig. 4. Min–max non-dominated frontier.

bargaining strategy is that it results in a fair improvement for patients and not just for the system. In this case, all the observations of strategy three have their times located at the same or under the decentralized conditions, ensuring that only positive impacts in emergencies are allowed. Figs. 7 and 8 present a comparison between the efficiency and min–max strategy when the point (6944, 4244) is chosen from the non-dominated frontier. Fig. 7 shows the emergencies for which the travel time exceeded 15-minute threshold. Those observations in which the travel time is the longest for the first strategy are reduced by the min–max. As mentioned in previous sections, the min–max imposes an upper bound of travel time by priority code class. Therefore, as much we decrease the disparities of the system, more observations with travel time greater than 15 min have their total time to treatment reduced.

From Fig. 8, we can observe the negative impact that patients with a travel time less than or equal to 15 min experienced due to the fairness strategy. Given that the trade-off existing between fairness and efficiency represented in the non-dominated frontier is not a one-to-one relationship, to decrease one unit of time for a patient in an unfair situation, more than one unit of time has to be given up for another patient. Therefore, in general, the distance between the triangles and the circles is longer in Fig. 8 than in Fig. 7. The gap may be reduced depending on the solution chosen by the decision-maker.

Table 6 summarizes some statistics of the decentralized model and the proposed strategies. Average travel, waiting, and total time are presented. The improvement row shows in minutes and percentage the difference in terms of total time to treatment between the current model and each of the strategies. Therefore, these numbers represent earnings

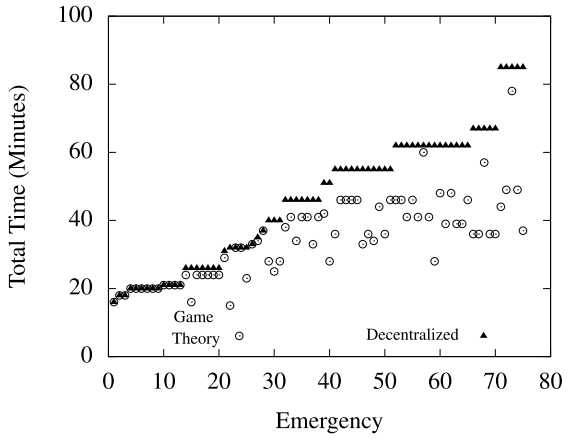


Fig. 6. Total times, decentralized vs. game theory.

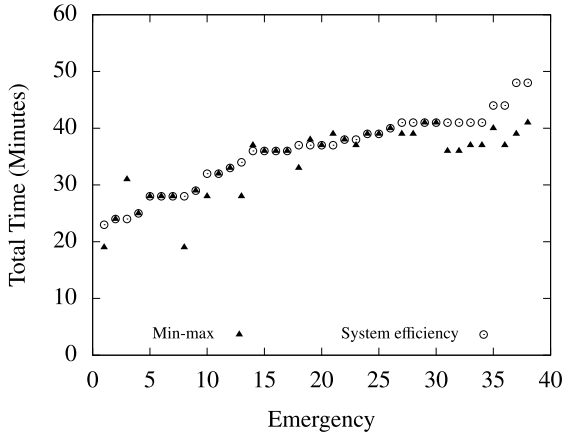


Fig. 7. Emergencies with travel time longer than 15 minutes.

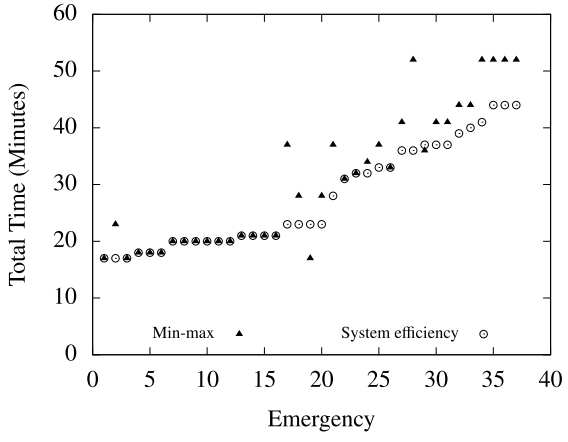


Fig. 8. Emergencies with travel time equal to or less than 15 minutes.

for the system in terms of time. The results of average total time and standard deviation indicate that the efficiency and min–max strategy, provide a substantial improvement to the current system. Nevertheless, the fact of patients being affected negatively in favor of the common good needs further analysis. Additionally, the results of the min–max strategy depend on the chosen solution in the non-dominated frontier. In the case of the game strategy, the overall system improvement is not as good as in the previous two models but guarantees non-negative results for each emergency. However, how the status quo is defined impacts the results. In our case, the status quo was considered as the decentralized model. Based on the conditions of the current model, we use Fig. 9 to analyze the impact of the different strategies on each patient. Every time that a color red, green, or blue line takes negative values, a patient is in an unfavorable situation in relation to the original status. Consequently, having positive values means an improvement in total patient times. Some negative peaks can be observed in Fig. 9; all of them belong to the min–max and system efficiency strategy. Of the total number of patients with priority codes 3 (urgent), 6% are being affected negatively by strategy one and 9% by strategy two. For priority code one and two (semi-urgent and non-urgent), 10% are being affected negatively by strategy one and 16% by strategy two. Priority level 4 and 5 (emergent and immediate) are not affected. In models one and two, the total number of patients in unfavorable situations is equal to 7% and 8% of the total number of emergencies, respectively. In contrast, the game theory strategy never takes negative values in the impact axis, which means this strategy guarantees only positive changes to the patients.

6. Conclusion and future research directions

The allocation of emergency service requests to EDs has a significant effect on the whole system of healthcare delivery, from both financial and life-saving perspectives. To the best of our knowledge, this is the first study that incorporates MIP optimization models for the allocation of ambulances to reduce waiting times for treatment in EDs and thus decrease overcrowding in the US healthcare system. The incorporation of remote triage plays a fundamental role in this task. The patient priority codes combined with the utilization levels of EDs, provide the framework for centralized decision systems. Fully centralized allocation strategies by region can be used to improve the current policies. Nevertheless, the fact of existing disparities must be considered.

We have formulated three different strategies based on MIP, applying single-objective optimization, and bi-objective optimization. The first strategy is an efficiency-based model. The results presented in Table 6 show this model provided a significant improvement to the overall system. However, further analysis revealed the negative effect that the first strategy may have on some patients, despite the priority class assigned to them. The second strategy is a bi-objective model where we minimize the maximum time by priority class when patients travel more than 15 min. This consideration is added as a second objective function to the first strategy. This technique allows the decision-maker to control better the distribution of resources in circumstances of unfairness. As a result, there are multiple combinations of efficiency and fairness. Nevertheless, the min–max strategy has some disadvantages. Similar to the first strategy, some patients experienced a negative impact compared to the current conditions, particularly those traveling 15 min or less. Finally, our game theory strategy provides a solution that presents 26.42% system improvement. Additionally, it guarantees a positive effect for every patient with respect to the current conditions.

Given the results obtained, the present study provides a crucial quantitative advance in comparing centralized decision systems with decentralized ones. More importantly, the proposed strategies not only offer optimal assignment with fairness elements but can also be used as analysis tools for any EMS system. Under desired conditions, infeasible solutions can be interpreted as limitations of the system in terms of handling all emergency requests. The adjustment of parameters should result in insights about the changes required by the network, for example, changing the waiting time function parameter a_j of a given ED to estimated the lack of capacity.

Finally, based on the results presented in the previous section, the utilization of either strategy results in improved time to treatment, decreased waiting times, and less overcrowding. However, choosing the right one is vital to ensure the wellness of all patients. The results show that an efficiency-based model cannot be used by itself to improve

Table 6
Results summary.

Metrics	Decentralized	Efficiency	Min–max	Game theory
Average travel time (mins)	10.06	15.21	16	16.47
Average ED waiting time (mins)	36.41	16.67	16.48	17.72
Average total time (mins)	46.47	31.88	32.48	34.19
Total system improvement (mins, %)	–	1094 , 31.39	1049, 30.10	921, 26.42
Standard deviation	19.28	8.80	9.62	10.85

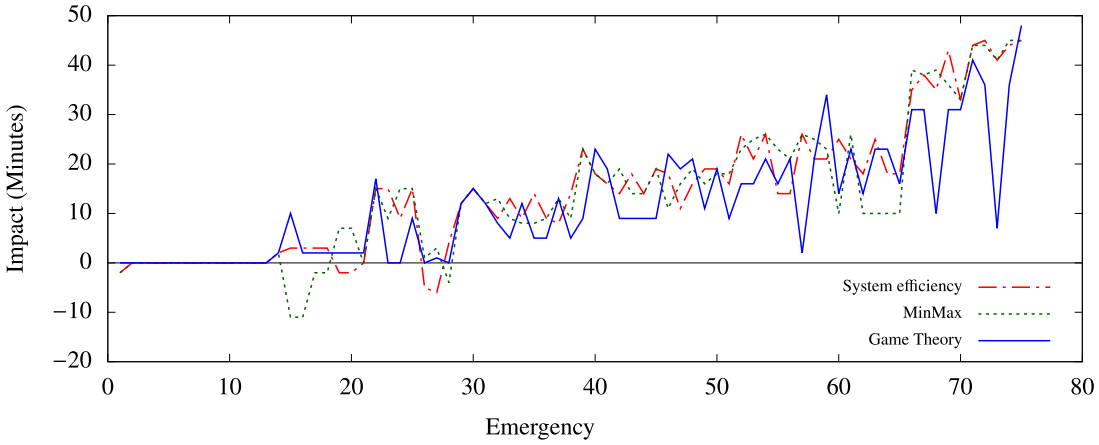


Fig. 9. Impact perceived by the patients in minutes.

systems in which patients at risk are involved. Efficiency strategies that incorporate fairness characteristics, as in the case of the min–max model, are potential solutions when scenarios similar to those described here are being studied. It is our opinion that the game theory strategy is the best way to guarantee the absence of negative impacts on all patients. Another positive aspect of strategies two and three is the flexibility offered by bi-objective models and non-symmetric bargaining games to adjust the results based on the system requirements.

In the future, we would like to generate a hybrid strategy where a new objective function is incorporated into the non-symmetric bargaining problem. This objective function could be related to rural areas, unattended illnesses, and other types of disparities. Besides, the development of new metrics to compare the strategies from an economic point of view can provide new insights to choose the best methodology. The models we propose and the results obtained disclose some understanding of the impact that allocation decisions in healthcare have on patients and the overall system. Future research that relaxes some of the assumptions and limitations of this study might prove useful. Uncertainties related to traffic or weather conditions may affect the final destination of an emergency. Also, the integration of a complete system view can help to disclose relationships between destination ED and dispatch policies for ambulances.

We consider implementation the next step, where the impact of strategies can be measured in a real-time data environment, providing insights on a daily basis and contrasted with the theoretical assumptions.

Appendix A. Model linearization

Linearization is a commonly used practice in the modeling of nonlinear functions. Linearization will come with the cost of introducing some new decision variables and constraints. First, let us consider a simplified case where both variables are binary. Suppose the model has the following product:

$$y \times z \tag{A.1}$$

where y and z are binary variables. Since $y \times z$ can take only 0 or 1 values, we can replace any instance of $y \times z$ in the model with a new binary variable x. We then add the following three constraints to the model to ensure that x will take proper values.

$$x \leq z \tag{A.2}$$

The previous inequality ensures that x will be equal to zero if z is zero.

$$x \leq y \tag{A.3}$$

Inequality (A.3) ensures that x will be equal to zero if y is zero.

$$x \geq z + y - 1 \tag{A.4}$$

The last inequality makes sure that if z and y are equal to one, then x is set to one. In this way, x will always represent the result of the product despite the values that y and z can take.

Now, let us consider a case where we try to linearize the product between a binary and a continuous variable, y and c respectively. Since $y \times c$ can take only 0 or c values, we can replace any instance of $y \times c$ in the model with a new continuous variable x. If c is bounded below by zero and above by a Big M (Large number), then we can add the following three constraints to ensure that x will take proper values.

$$x \leq M \times y \tag{A.5}$$

Inequality (A.5) ensures that if the binary variable is equal to zero, then the product of the continuous and binary variable has to be equal to zero. Conversely, if y is equal to one, the inequality will make x always smaller or equal than the upper bound of c.

$$x \leq c \tag{A.6}$$

The previous inequality considers that y is a binary variable; therefore, the value of the product can be at most equal to the value of the continuous variable c.

$$x \geq c - M(1 - y) \tag{A.7}$$

The inequality (A.7) will make sure that x is always greater or equal than c. Finally, considering the previous inequalities, we can be sure that x represents the product of y and c. Additional analysis can be done based on the structure of each model. For example, if the model

is trying to minimize the value of x, then inequalities (A.5) and (A.6) representing upper bounds are not needed. For more scenarios and similar cases we recommend to check the following Ref. [48].

Appendix B. Supplementary data

The data utilized in section four was randomly generated through processes that can be hard to replicate. Consequently, the data files and the instructions for their utilization are available at https://github.com/jorgeacunam.

Appendix C. Min–max strategy, additional solutions

Table 7 presents the results for each point in the non-dominated frontier of Fig. 4. The last column shows the number of patients receiving an adverse impact based on the chosen objective values.

Table 7
Min–max strategy results.

Objective values	Total time	Avg. time	SD	No. of negative impacts
6943.9, 4243.3	2436	32.48	9.62	6
6923.5, 4246.5	2425	32.33	9.57	6
6910.1, 4249.7	2416	32.21	9.52	5
6895.9, 4260.3	2423	32.31	9.25	7
6875.5, 4262.8	2409	32.12	9.23	6
6859.9, 4266.2	2407	32.09	9.11	7
6846.5, 4269.4	2405	32.07	8.99	5
6833.3, 4274.0	2395	31.93	8.95	5
6810.8, 4286.9	2393	31.91	8.82	5
6793.8, 4302.0	2391	31.88	8.80	5

References

[1] Nielsen TK. Implementation of the national emergency department overcrowding score to analyze crowding patterns and simulate an ambulance diversion protocol [Ph.D. thesis], Brandman University; 2017, https://search.proquest.com/docview/1983522890?pq-origsite=gscholar.

[2] Shiver JM, Eitel D. Optimizing emergency department throughput: Operations management solutions for health care decision makers. Productivity Press; 2009, https://www.taylorfrancis.com/books/e/9781420084979.

[3] Gilboy N, Tanabe P, Travers D, Rosenau AM, et al. Emergency severity index (ESI): A triage tool for emergency department care, version 4. Agency Healthc Res Qual 2012. https://www.ahrq.gov/sites/default/files/wysiwyg/professionals/systems/hospital/esi/esihandbk.pdf.

[4] Morganti KG, Bauhoff S, Blanchard JC, Abir M, Iyer N, Smith A, et al. The evolving role of emergency departments in the United States. RAND Health Q 2013;3(2). https://www.rand.org/pubs/research_reports/RR280.html.

[5] Richardson LD, Asplin BR, Lowe RA. Emergency department crowding as a health policy issue: Past development, future directions. Ann Emerg Med 2002;40(4):388–93. http://dx.doi.org/10.1067/mem.2002.128012.

[6] Rui P, Kang K, Ashman J. National hospital ambulatory medical care survey: 2016 emergency department summary. Tech. rep. Centers for Disease Control and Prevention; 2019, https://www.cdc.gov/nchs/data/nhamcs/web_tables/2016_ed_web_tables.pdf.

[7] Rust G, Ye J, Baltrus P, Daniels E, Adesunloye B, Fryer GE. Practical barriers to timely primary care access: Impact on adult use of emergency department services. Arch Intern Med 2008;168(15):1705–10. http://dx.doi.org/10.1001/archinte.168.15.1705.

[8] Moskop JC, Sklar DP, Geiderman JM, Schears RM, Bookman KJ. Emergency department crowding, part 1- concept, causes, and moral consequences. Ann Emerg Med 2009;53(5):605–11. http://dx.doi.org/10.1016/j.annemergmed.2008.09.019.

[9] TrendWatch chartbook 2018: Trends affecting hospitals and health systems. Tech. rep. American Hospital Association; 2018, https://www.aha.org/system/files/2018-07/2018-aha-chartbook.pdf.

[10] Pitts SR, Niska RW, Xu J, Burt CW, et al. National hospital ambulatory medical care survey: 2006 emergency department summary. Tech. rep. Centers for Disease Control and Prevention; 2008, p. 1–38, https://www.cdc.gov/nchs/data/nhsr/nhsr007.pdf.

[11] Health Forum. Hospital statistics 2018. American Hospital Association; 2018, https://www.aha.org/statistics/2016-12-27-aha-hospital-statistics-2018-edition.

[12] Eitel DR, Rudkin SE, Malvehy MA, Killeen JP, Pines JM. Improving service quality by understanding emergency department flow: A white paper and position statement prepared for the american academy of emergency medicine. J Emerg Med 2010;38(1):70–9. http://dx.doi.org/10.1016/j.jemermed.2008.03.038.

[13] Trzeciak S, Rivers E. Emergency department overcrowding in the United States: An emerging threat to patient safety and public health. Emerg Med J 2003;20(5):402–5. http://dx.doi.org/10.1136/emj.20.5.402.

[14] Salway R, Valenzuela R, Shoenberger J, Mallon W, Viccellio A. Emergency department (ED) overcrowding: Evidence-based answers to frequently asked questions. Rev Méd Clín Las Condes 2017;28(2):213–9. http://dx.doi.org/10.1016/j.rmclc.2017.04.008.

[15] Gordon JA, Billings J, Asplin BR, Rhodes KV. Safety net research in emergency medicine proceedings of the academic emergency medicine consensus conference on "The Unraveling Safety Net". Acad Emerg Med 2001;8(11):1024–9. http://dx.doi.org/10.1111/j.1553-2712.2001.tb01110.x.

[16] Kaushal A, Zhao Y, Peng Q, Strome T, Weldon E, Zhang M, et al. Evaluation of fast track strategies using agent-based simulation modeling to reduce waiting time in a hospital emergency department. Socio-Econ Plan Sci 2015;50:18–31. http://dx.doi.org/10.1016/j.seps.2015.02.002.

[17] Cowan RM, Trzeciak S. Clinical review: Emergency department overcrowding and the potential impact on the critically ill. Crit Care 2004;9(3):291. http://dx.doi.org/10.1186/cc2981.

[18] Richardson DB. Increase in patient mortality at 10 days associated with emergency department overcrowding. Med J Aust 2006;184(5):213–6, https://www.mja.com.au/system/files/issues/184_05_060306/ric10511_fm.pdf.

[19] Falvo T, Grove L, Stachura R, Zirkin W. The financial impact of ambulance diversions and patient elopements. Acad Emerg Med 2007;14(1):58–62. http://dx.doi.org/10.1197/j.aem.2006.06.056.

[20] Rui P, Kang K. National hospital ambulatory medical care survey: 2015 emergency department summary tables. Tech. rep. Centers for Disease Control and Prevention; 2015, https://www.cdc.gov/nchs/data/nhamcs/web_tables/2015_ed_web_tables.pdf.

[21] McCaig LF, Burt CW. National hospital ambulatory medical care survey: 2003 emergency department summary. Tech. rep. 358, Center for Disease Control and Prevention; 2005, https://www.cdc.gov/nchs/data/ad/ad358.pdf.

[22] http://hdl.handle.net/10945/5785 Mims L. Improving emergency medical services (EMS) in the United States through improved and centralized federal coordination [Master's thesis], Monterey, California. Naval Postgraduate School; 2011.

[23] Pozner CN, Zane R, Nelson SJ, Levine M. International EMS systems: The United States past, present, and future. Resuscitation 2004;60(3):239–44. http://dx.doi.org/10.1016/j.resuscitation.2003.11.004.

[24] Mears G, Armstrong B, Fernandez AR, Mann NC, McGinnis K, Mears CR, et al. 2011 national EMS assessment. Tech. rep. National Highway Traffic Safety Administration; 2012, https://www.ems.gov/pdf/2011/National_EMS_Assessment_Final_Draft_12202011.pdf.

[25] Uniform trauma transport protocol. Tech. rep. Hillsborough County Trauma Agency; 2016, https://www.hillsboroughcounty.org/library/hillsborough/media-center/documents/trauma/revised-change-13-of-hillsborough-county-s-uttp_submitted-to-fdoh-sept-2016.pdf.

[26] Department of Health and Human Services. Medicare coverage of ambulance services. Tech. rep. Centers for Medicare and Medicaid Services; 2017, https://www.medicare.gov/Pubs/pdf/11021-Medicare-Coverage-of-Ambulance-Services.pdf.

[27] Lim CS, Mamat R, Braunl T. Impact of ambulance dispatch policies on performance of emergency medical services. IEEE Trans Intell Transp Syst 2011;12(2):624–32. http://dx.doi.org/10.1109/TITS.2010.2101063.

[28] Rajagopalan HK, Saydam C. A minimum expected response model: Formulation, heuristic solution, and application. Socio-Econ Plan Sci 2009;43(4):253–62. http://dx.doi.org/10.1016/j.seps.2008.12.003.

[29] Farahani RZ, Asgari N, Heidari N, Hosseininia M, Goh M. Covering problems in facility location: A review. Comput Ind Eng 2012;62(1):368–407. http://dx.doi.org/10.1016/j.cie.2011.08.020.

[30] Kahraman C, Topcu Y. Operations research applications in health care management. Springer; 2018, https://www.springer.com/us/book/9783319654539.

[31] Nasrollahzadeh AA, Khademi A, Mayorga ME. Real-time ambulance dispatching and relocation. In: Manufacturing & service operations management. INFORMS; 2018, http://dx.doi.org/10.1287/msom.2017.0649.

[32] Van Barneveld T, Jagtenberg C, Bhulai S, Van der Mei R. Real-time ambulance relocation: Assessing real-time redeployment strategies for ambulance relocation. Socio-Econ Plan Sci 2018;62:129–42. http://dx.doi.org/10.1016/j.seps.2017.11.001.

[33] Boutilier JJ, Chan TC. Ambulance emergency response optimization in developing countries. 2018, ArXiv Preprint ArXiv:1801.05402, http://arxiv.org/abs/1801.05402.

[34] Iannoni AP, Morabito R, Saydam C. Optimizing large-scale emergency medical system operations on highways using the hypercube queuing model. Socio-Econ Plan Sci 2011;45(3):105–17. http://dx.doi.org/10.1016/j.seps.2010.11.001.

[35] Sorensen P, Church R. Integrating expected coverage and local reliability for emergency medical services location problems. Socio-Econ Plan Sci 2010;44(1):8–18. http://dx.doi.org/10.1016/j.seps.2009.04.002.

[36] Almehdawe E, Jewkes B, He Q-M. A Markovian queueing model for ambulance offload delays. European J Oper Res 2013;226(3):602–14. http://dx.doi.org/10.1016/j.ejor.2012.11.030.

[37] Almehdawe E, Jewkes B, He Q-M. Analysis and optimization of an ambulance offload delay and allocation problem. Omega 2016;65:148–58. http://dx.doi.org/10.1016/j.omega.2016.01.006.

[38] Leo G, Lodi A, Tubertini P, Di Martino M. Emergency department management in lazio, Italy. Omega 2016;58:128–38. http://dx.doi.org/10.1016/j.omega.2015.05.007.

[39] Deo S, Gurvich I. Centralized vs. decentralized ambulance diversion: A network perspective. Manage Sci 2011;57(7):1300–19. http://dx.doi.org/10.1287/mnsc.1110.1342.

[40] Schmid V. Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. European J Oper Res 2012;219(3):611–21. http://dx.doi.org/10.1016/j.ejor.2011.10.043.

[41] Eastwood K, Smith K, Morgans A, Stoelwinder J. Appropriateness of cases presenting in the emergency department following ambulance service secondary telephone triage: A retrospective cohort study. BMJ Open 2017;7(10). e016845. http://dx.doi.org/10.1136/bmjopen-2017-016845.

[42] McLeod B, Zaver F, Avery C, Martin DP, Wang D, Jessen K, et al. Matching capacity to demand: A regional dashboard reduces ambulance avoidance and improves accessibility of receiving hospitals. Acad Emerg Med 2010;17(12):1383–9. http://dx.doi.org/10.1111/j.1553-2712.2010.00928.x.

[43] Felice J, Coughlin RF, Burns K, Chmura C, Bogucki S, Cone DC, et al. Effects of real-time EMS direction on optimizing EMS turnaround and load-balancing between neighboring hospital Campuses. Prehospital Emerg Care 2019. http://dx.doi.org/10.1080/10903127.2019.1587123.

[44] Ogryczak W, Luss H, Pióro M, Nace D, Tomaszewski A. Fair optimization and networks: A survey. J Appl Math 2014;2014. http://dx.doi.org/10.1155/2014/612018.

[45] Hoot NR, LeBlanc LJ, Jones I, Levin SR, Zhou C, Gadd CS, et al. Forecasting emergency department crowding: A discrete event simulation. Ann Emerg Med 2008;52(2):116–25. http://dx.doi.org/10.1016/j.annemergmed.2007.12.011.

[46] Kulstad EB, Hart KM, Waghchoure S. Occupancy rates and emergency department work index scores correlate with leaving without being seen. West J Emerg Med 2010;11(4):324, https://www.ncbi.nlm.nih.gov/pubmed/21079702.

[47] Tekwani KL, Kerem Y, Mistry CD, Sayger BM, Kulstad EB. Emergency department crowding is associated with reduced satisfaction scores in patients discharged from the emergency department. West J Emerg Med 2013;14(1):11. http://dx.doi.org/10.5811/westjem.2011.11.11456.

[48] Chen D-S, Batson RG, Dang Y. Applied integer programming: Modeling and solution. John Wiley & Sons; 2011, https://www.wiley.com/en-us/Applied+Integer+Programming.

[49] Ehrgott M. Multicriteria optimization. vol. 491. Springer Science & Business Media; 2005, https://www.springer.com/us/book/9783540213987.

[50] Gamm LD, Hutchison LL, Dabney BJ, Dorsey AM. Rural healthy people 2010: A companion document to healthy people 2010. Tech. rep. The Texas A&M University System Health Science Center; 2003, https://www.ruralhealthresearch.org/publications/311.

[51] Fattahi A, Turkay M. A one direction search method to find the exact nondominated frontier of biobjective mixed-binary linear programming problems. European J Oper Res 2018;266(2):415–25. http://dx.doi.org/10.1016/j.ejor.2017.09.026.

[52] Nash J. Two-person cooperative games. Econometrica 1953;128–40. http://dx.doi.org/10.2307/1906951.

[53] Charkhgard H, Savelsbergh M, Talebian M. A linear programming based algorithm to solve a class of optimization problems with a multi-linear objective function and affine constraints. Comput Oper Res 2018;89:17–30. http://dx.doi.org/10.1016/j.cor.2017.07.015.

[54] Healthcare Cost and Utilization Project (HCUP). HCUP state emergency department databases (SEDD). Rockville, Maryland: Healthcare Cost and Utilization Project, Agency for Healthcare Research and Quality; 2014, www.hcup-us.ahrq.gov/seddoverview.jsp.

[55] Healthcare Cost and Utilization Project (HCUP). HCUP State Inpatient Databases (SID). Rockville, Maryland: Healthcare Cost and Utilization Project, Agency for Healthcare Research and Quality; 2014, www.hcup-us.ahrq.gov/sidoverview.jsp.

[56] County Trauma Agency. 2015 five year trauma plan update. Tech. rep. Hillsborough County Florida; 2019, https://www.hillsboroughcounty.org/en/residents/public-safety/trauma-agency/2015-hillsborough-county-trauma-plan.

[57] Rui P, Kang K. National hospital ambulatory medical Care survey: 2014 emergency department summary tables. Tech. rep. National Center for Health Statistics; 2014, https://www.cdc.gov/nchs/data/nhamcs/web_tables/2014_ed_web_tables.pdf.

[58] AHCA Multimedia Design. Emergency department utilization report 2017. Tech. rep. Agency for Health Care Administration; 2017, http://www.floridahealthfinder.gov/researchers/studies-reports.aspx.

[59] Best hospitals in Tampa-St. Petersburg, Fla. Tech. rep. U.S. News and World Report; 2017, https://health.usnews.com/best-hospitals/area/tampa-st-petersburg-fl.

[60] 2010 census of population and housing, population and housing unit counts, CPH-2-11, florida, US. Tech. rep. U.S. Census Bureau; 2012, https://www.census.gov/library/publications/2012/dec/cph-2.html.

[61] Hillsborough County GIS. Hillsborough county GeoHub. Hillsborough County; 2010, http://gis2017-01-10t133755357z-hillsborough.opendata.arcgis.com.

[62] Statistical Atlas. Population of Hillsborough county, Florida (county). 2018, https://statisticalatlas.com/county/Florida/Hillsborough-County/Population.

[63] The Florida Legislature. The 2018 Florida statutes: Access to emergency services and care. Florida Legislature's website; 2018, http://www.leg.state.fl.us/statutes/index.cfm?App_mode=Display_Statute&URL=0300-0399/0395/Sections/0395.1041.html.

[64] Delbridge TR, Flaherty S, Kuykendall DR, Pratt D, Whitney JR. A reassessment of emergency medical services. Tech. rep. State of Florida; 2013, http://www.floridahealth.gov/licensing-and-regulation/trauma-system/_documents/nationalhighwaytrafficsafetyadmin-ems.pdf.

[65] Hing E, Bhuiya F. Wait time for treatment in hospital emergency departments, 2009. Tech. rep. No. 102. Center for Disease Control and Prevention; 2012, https://www.cdc.gov/nchs/data/databriefs/db102.pdf.

[66] Travers DA, Waller AE, Bowling JM, Flowers D, Tintinalli J. Five-level triage system more effective than three-level in tertiary emergency department. J Emerg Nurs 2002;28(5):395–400. http://dx.doi.org/10.1067/men.2002.127184.

[67] Parenti N, Ferrara L, Reggiani MLB, Sangiorgi D, Lenzi T. Reliability and validity of two four-level emergency triage systems. Eur J Emerg Med 2009;16(3):115–20. http://dx.doi.org/10.1097/MEJ.0b013e328310b594.

[68] Schuster MA, McGlynn EA, Brook RH. How good is the quality of health care in the United States? Milbank Q 2005;83(4):843–95. http://dx.doi.org/10.1111/j.1468-0009.2005.00403.x.

[69] Chalmet L, Lemonidis L, Elzinga D. An algorithm for the bi-criterion integer programming problem. European J Oper Res 1986;25(2):292–300. http://dx.doi.org/10.1016/0377-2217(86)90093-7.

[70] Boland N, Charkhgard H, Savelsbergh M. A criterion space search algorithm for biobjective integer programming: The balanced box method. INFORMS J Comput 2015;27(4):735–54. http://dx.doi.org/10.1287/ijoc.2015.0657.

[71] Bazaraa MS, Sherali HD, Shetty CM. Nonlinear programming: Theory and algorithms. John Wiley & Sons; 2013, https://www.wiley.com/en-us/Nonlinear+Programming.

Jorge A. Acuna, MS: He obtained his bachelor degree from the Universidad de la Frontera (Chile). In 2013 he earned a scholarship and complete part of his studies at the Alpen-Adria-Universität Klagenfurt (Austria). He is perusing his doctoral degree in the Industrial and Management Systems Engineering department at the University of South Florida, where he obtained his MS degree in 2018. His research interests center around healthcare systems engineering, specifically waiting list management, overcrowding of emergency departments, and the design of admissions policies to different units.

José L. Zayas-Castro, PhD: is professor of Industrial and Management Systems Engineering at the University of South Florida. Over the past two decades, he has been working with colleagues across engineering, health sciences, and healthcare providers in understanding, analyzing, modeling and improving the delivery of care as well as reducing unnecessary costs. Additionally, he has developed instructional initiatives that cut across engineering and the health sciences.

Hadi Charkhgard, PhD: is Assistant Professor of Industrial and Management Systems Engineering and he is the director of the multi-objective optimization laboratory at the University of South Florida. Prior to this position, he was a postdoctoral research fellow at the Georgia Institute of Technology. He has a track record of creating innovative techniques for solving optimization problems that are published in highly-ranked journals in Operations Research.