Emergency Department Management in Lazio, Italy

Emergency Department Management in Lazio, Italy

Gianmaria Leo*[1], Andrea Lodi†[1], Paolo Tubertini‡[1], and Mirko Di Martino§[2]

[1]DEI, University of Bologna, Bologna, Italy
[2]Department of Epidemiology, Lazio Regional Health Service, Roma, Italy

January 15, 2015; revised May 4, 2015

Abstract

The assignment of service requests to emergency departments is of paramount importance both from a life-threatening and an economical viewpoints. In the process of a more general project that aims at defining optimal allocation policies of patients to regional hospital network facilities (together with the potential reorganization of the facilities), the Department of Epidemiology of the Regional Health Service of Lazio, Italy, was interested in obtaining a completely offline picture of the effect of an optimal assignment of requests to emergency departments. This is in the spirit of evaluating the so-called *Price of Anarchy*, where the fully centralized (admittedly unrealistic) allocation is used as a reference for both the state-of-the-art completely decentralized approach and future reorganization ideas.

We have implemented and tested with real-world data of all service requests of 2012 a Mixed-Integer Programming model that computes such an optimal request allocation by minimizing travel and waiting times and penalizing workload unbalance among emergency departments in the region. Within the development process we have studied special cases and relaxations of the complete model showing interesting mathematical properties that are, in turn, useful from a practical viewpoint, for example, in obtaining a real-time version of the approach.

The present study is an important, quantitative step in the evaluation of centralized allocation strategies like remote triage that could have a remarkable impact in making the allocation process much more efficient and effective. More precisely, the developed methodology as well as the software tools are currently used by the DEP-Lazio for the reorganization of the regional networks of emergency healthcare.

1 Introduction

The Department of Epidemiology of the Regional Health Service of Lazio, Italy (DEP-Lazio in the following), a regional center for Health monitoring

*gianmaria.leo@gmail.com
†andrea.lodi@unibo.it
‡paolo.tubertini@unibo.it
§m.dimartino@deplazio.it

and management, is currently involved in a project that aims at defining optimal allocation policies of patients to regional hospital network facilities. The reorganization of health centres in order to deliver services in an effective way by taking into account economic sustainability is a topic of increasing importance for Regional Health Services in Italy. In recent years several inputs have been given, through financial laws, to reorganize hospitals infrastructure in order to increase efficiency. Reorganization policies can be considered, from a strategic point of view, as composed by two main decision elements: the definition of the subset of hospital facilities that should be active within the regional territory and the allocation of demand of services to active facilities. Because the reorganization of a Regional Health system in terms of facility location and service allocation is a task of great complexity Regional managers decided to start by focusing their attention on emergency departments (ED). Indeed, EDs are a crucial access point to hospital network facilities and as a consequence their management is a critical factor in order to improve system effectiveness and efficiency. In Italy it is possible to state that the role of EDs is even more important than in other European countries because, in addition to real emergency and urgency services, they have to face a set of demands that should instead be managed by Primary care units or by General Practitioners. This is due to historical reasons associated with the development of the system and, recently, to the increase of (often illegal) immigration. The 2013-2015 operational programs of the Lazio Region require the activation of new clinical care pathways for emergencies, with a special priority for life-threatening diseases, such as acute coronary syndrome, stroke and trauma. For these situations, a timely medical intervention, performed in facilities with the necessary equipment, can save the patient's life and significantly improve the prognosis. For example, with respect to patients suffering from ST Elevation Myocardial Infarction (STEMI), it is suggested to perform a percutaneous coronary intervention (PCI) in hospitals with high volume of activity, equipped with catheterization laboratory and highly qualified teams. Moreover, according to clinical guidelines, it is strongly recommended to perform PCI within 90 minutes from the onset of the early symptoms. Therefore, it is essential that STEMI patients can immediately be transferred to a specialized hospital. Unfortunately, preliminary analyses showed that the current emergency networks are not able to provide an appropriate and timely healthcare assistance to all residents, especially in areas far from the city of Rome, generally characterized by a lower socioeconomic status.

Emergency department characteristics. An emergency department can be defined as an health facility that is dedicated to the management of emergency and urgency treatments, that is to say to that spontaneous or traumatic pathological conditions that need to be treated within a short period of time. Emergency activities are, for their own nature, nonelective and patients can reach ED facilities both by their own (walk-ins) or with the support of an emergency vehicle. Due to the impossibility of planning patients arrival, EDs have to provide an initial treatment for a wide number of diseases some of which can be life-threatening. Because the set of patients that ask for treatments is heterogeneous from the pathological point of view, the admission of patients is driven by a priority-based policy. The stochastic nature of arrival times and of pathological conditions can have a strong impact on workload and as a con-

sequence on patient waiting times and quality of care. It is then fundamental that the priority assignment is properly managed in order to meet patients' needs according to their critical condition. The process of assigning priorities to patients is defined as *triage* and it is usually coded at a regional or national level. Triage is a set of procedures that ensure, in the best possible way, that patients with a more critical condition are admitted before the others. The priority level is usually represented by a color code (white, green, yellow and red) that defines the increasing need of care. For each patient the priority is usually defined just after the arrival by a dedicated operator. The definition of triage procedures is then fundamental to guarantee an immediate care for the patient, to identify the priority level and the medical area that may treat him and, ranking lower priority patients, to reduce waiting times. Triage activities can directly address the patient to the most appropriate hospital ward in case of complex treatments, for less serious ones the patient can be directly treated by emergency department physicians and discharged. It is then important for health managers to plan EDs so as to meet a set of objectives that can be in some cases conflicting. At first it is fundamental to guarantee quality of care that is composed by treatment timeliness, according to the patient health condition, and appropriateness, according to the patient pathological condition. On the other hand the cost sustained to provide services has to be reduced as much as possible by taking into account a minimum standard of care.

Paper Contribution. As already discussed, triage is currently the first activity that is performed when a patient reach the ED. This means that ED triage only determines the care pathway within the hospital structure. In other words, the possibility that a better quality of care and/or a shorter waiting time could have been reached if the patient would have been sent to another ED is not considered.

The objective of the present study is to develop an hybrid model that considers both ED workload and service allocation, evaluating what could be the impact of a remote triage management that, anticipating the patient classification, can address population requests to the first-aid structure, thus assuring the best possible service level. In particular, the final allocation policy for emergency department requests needs to maximize quality of care and service timeliness. In order to develop a regional allocation approach we must suppose that all requests can be filtered at a regional level. That is to say that walk-in or ambulance referral that have not been screened by the triage management center are not accepted. Clearly, this is only an hypothetical scenario that is, however, potentially useful to define a reference solution (as well as a reference methodology) in terms of service quality (to be defined below), so as to evaluate, in comparison, new and more sophisticated allocation policies. In other words, the current case-study establishes a benchmark solution with respect to which the cost of a completely decentralized and loosely planned allocation is computed. In this sense we somehow follow the so-called price-of-anarchy viewpoint [23] although the techniques applied here do not exploit game theory in the computation of an equilibrium. Instead, we have implemented and tested with real-world data of all service requests of 2012 a Mixed-Integer Programming model that computes such an optimal request allocation by minimizing travel and waiting times and penalizing workload unbalance among emergency departments in the region. Within the development process we have studied

special cases and relaxations of the complete model showing interesting mathematical properties that are, in turn, useful from a practical viewpoint. Finally, one of those special cases allowed us to devise a real-time version of the first-aid requests allocation approach, which can be used as a Decision Support System for the Triage Center daily operations.

The present study is an important, quantitative step in the evaluation of centralized allocation strategies like remote triage that could have a remarkable impact in making the allocation process much more efficient and effective. More precisely, the developed methodology as well as the software tools are currently used by the DEP-Lazio for the reorganization of the regional networks of emergency healthcare. Our findings will be shared with the Regional Directorate for health and social-health integration and the Regional Healthcare Emergency Unit, which operatively manages the first aid requests in Lazio. The joint analysis of the results by those who plan emergency healthcare programs and by those who operationally run them in the territory are expected to be helpful to develop and quantitatively evaluate strategies to: (a) improve health assistance for the population living in disadvantaged areas, (b) reduce waiting times in emergency departments and (c) balance workload among EDs of the Lazio region. More generally, considering that the technical equipment is known for each hospital, this type of optimization (possibly coupled with simulation) techniques can be effectively used to reorganize the emergency networks in accordance with the hierarchical levels of the hospitals equipment complexity. This is likely to result in optimization of the current "Hub and Spoke" model, based on the distinction of the emergency departments in basic EDs, first level EDs and second level EDs, depending on the provided intensity of care and on the dimension of the hospital catchment area. Basic or Primary care hospitals are characterized by a catchment area of 80,000 to 150,000 inhabitants and have a limited number of active medical specialization departments. First level hospitals have a catchment area of 150,000 to 300,000 inhabitants and Radiology and Ultrasound with X-ray CT as well as laboratory and blood transfusion services should be available 24 hours per day. Second level hospitals have a catchment area of 600,000 to 1,200,000 inhabitants and are equipped with all medical specialization departments. Those facilities are supplied with the most advanced technological devices in order to properly treat complex patients.

Paper Organization. In Section 1.1 we review the literature reporting ED and emergency medical services (EMS) management approaches. In Section 2 we discuss the details of the problem and we introduce the required notation and definitions. In Section 3 a Mixed Integer Linear Programming (MIP) approach is proposed and several properties and relaxations are discussed. In Section 4 we extensively discuss computational experiments performed by solving the MIP model on real-world instances provided by DEP-Lazio. In Section 5 we propose a real-time algorithm for first-aid requests allocation. Finally, in Section 6 we draw some conclusions and discuss the use of the proposed optimization approach in current and future settings.

1.1 Literature Review

Operations Research has been widely applied to study ED management issues, such as capacity planning and patient flows, by using both optimization and simulation techniques. Literature case-studies can be classified according to the set of decisions that are taken into consideration, including capacity planning, staff scheduling and general planning for future development of the facility. Pure capacity planning case-studies evaluate the impact of resource resizing on patients waiting times. It is then important to evaluate which is the degree of complexity of a comprehensive simulation model. As an example, in [11] the authors consider triage, prioritization and several staff level types as well as imaging studies, laboratory studies, physical examination, nursing activity, consultations, and bedside procedures. However, the model does not consider technical resources reducing the potential analysis of supply shortages. In [4] the authors show how capacity planning can provide an efficient patient flow by calculating the maximum occupancy level of beds. In [33] the authors define an analytical model to describe patient flows in emergency departments taking into consideration scarce resources such as medical doctors, nurses, beds and diagnostic machines. The model is used to evaluate the impact of resource resizing policies. In a similar way in [21] the resizing of different resources is compared in order to identify which is the one that mainly influences ED performances. Patients arrival pattern can be also simulated in order to level the peak of resource utilization, leading to a significantly better planning of staff and resources [29]. Similarly, in [30] arrival analysis allows a reduction of patient turnaround times. Finally in [7] optimal control policy is applied to define the number of resources that should be used in order to prevent ED overcrowding. ED capacity management can be also analyzed from a different perspective through the evaluation of how budget restrictions and workforce reduction can be faced while preserving operational performances [28]. In that case-study patient flow patterns are fixed and the main goal of the problem is to evaluate how staffing management can influence waiting times.

It is clear that ED performances cannot be improved only by means of resource resizing; advanced prioritization models as well as new organizational designs can turn out to be more effective than simple capacity planning. In [22] the authors evaluate the introduction of the so-called split-flow concept that is an emerging approach to manage ED processes by a split of the patient flow according to their acuity and enabling parallel processing. The model, applied to a real ED, aims at reducing patients waiting times and system congestion. In [9] a new prioritization model for patients is evaluated by taking in consideration patient acuity mix, arrival patterns and volumes and trying to minimize the walk-away for patients waiting for a long time. In [3] simulation also proves to be of great potential for the evaluation of future expansion of an ED by increasing the understanding of the processes involved. An integration of simulation and optimization techniques is presented in [35] to reduce patient queuing time. Modeling the complex behavior of an ED is a challenging task, due to interaction of human and physical resources. Medical staff, for example, is rarely dedicated to one patient or task. Instead, the staff treats several patients at a time while waiting for other processes. This diversity of process interaction can be described as multitasking, a common feature of ED operations even if rarely considered in planning models (see, e.g., [17]).

Till now we focused our review on emergency department planning by con-

sidering this organizational unit as a unique component that is externally influenced only by patient arrivals. It is clear that ED inflow is strongly related to the definition of catchment areas because a department, and Hospitals in general, usually cover the health needs of a subset of the local population. Through 'covering' we mean that a specific (regional) population cluster has as a reference point for health needs a specific hospital that is usually defined on a distance basis. It is then clear that if we widen the focus of analysis we can develop capacity plans for hospitals and EDs taking in consideration the fact that a reorganization can strongly influence the volume of activities and as a consequence system performances both in terms of patient outcomes and quality of service. As an example, in [8] the authors propose a modeling framework to analyze the supply and demand matching of public hospital beds addressing the planning issues of hospital locations and service allocations, which include new service distribution as well as existing service redistribution. In [5] an optimization model is formulated using integer programming and heuristics, the goal of the case-study being to maximize coverage of severely injured patients by locating trauma centers and aeromedical depots. Finally, in [18] the authors propose a discrete-event geographical location/allocation simulation model for evaluating various options for the provision of services including the location of the service centers, service capacities, geographical distribution of patients, and ease of access to the health services.

ED overcrowding can cause, as a domino effect, ambulance diversions and an inefficient utilization of emergency medical services (EMS) tying up resources and reducing response time [6]. As a consequence, a consistent branch of research integrates ED workload management with ambulance management so as to coordinate two services that are strongly interdependent. In [13] a multi-dimension Markov chain queuing model is developed to coordinate ambulance traffic in order to solve the ED crowding problem. The case study takes into consideration two hospital EDs and simulates both ambulance and walk-in arrivals. Similarly, in [2] a Markovian queuing model is used to study ED crowding and ambulance offload delays. An alternative approach to minimize patients waiting time is proposed in [12], where the ambulance diversion problem is analyzed by modeling a queuing game between two EDs. In that article the authors demonstrate the potential benefit of a centralized planner that maximizes the social optimum. A comparison of hospital selection policies in order to identify the one that mostly reduces ED crowding problems is reported in [25].

It is clear that ambulance management cannot be reduced to the diversion problem analyzed in the previous paragraph. Ambulance dispatching, after a first-aid request is collected, and its relocation to the next waiting location are real-time problems that emergency providers should efficiently solve. In [27] an approximate dynamic programming approach is proposed to solve the above-mentioned problem in a time-efficient manner.

2 Notation and Definitions

Given a positive integer τ, the time horizon of our analysis is modeled by discrete ordered set $T := \{1, \ldots, \tau\}$, whose elements represent time slots based on the shift work periods of ED medical operators. Let U be the set of districts in which the territory under the authority of Lazio Region is partitioned, see

Figure 1. Let $V \subseteq U$ be the subset of districts in which an emergency depart-

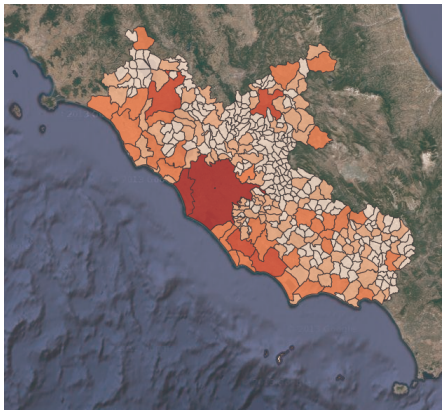

Figure 1: Municipal districts of Lazio Region.

ment is located. Let $d(u, v)$ be the expected time duration of a trip from $u \in U$ to $v \in V$. Throughout this paper we assume $d(u, v)$ is constant over T. Of course, this is not a completely harmless assumption because traffic conditions might in fact play an important role, although emergency vehicles are (far) less constrained by traffic. Let F be a set of first-aid medical treatments that can be supplied by healthcare centers.

Emergency departments. Let S be the set of emergency departments operating under the authority of Lazio Region. Each $s \in S$ is modeled by a quadruple composed by the following elements: (i) $v_s \in V$ is the district in which s is located; (ii) $F_s \subseteq F$ is the subset of specializations that s can offer; (iii) $w_s : \mathbb{R} \to \mathbb{R}$ is a function returning the expected time a patient has to wait at s before receiving first-aid service; (iv) $q_s \in \mathbb{R}^{|F_s|}$ reports the quality of service for each specific medical treatment in F_s, according to Lazio Region Evaluation Program for medical operations results.

Triage codes and pathologies of interest. Let C be the set of emergency department codes that can be assigned by triage diagnosis. Let P be a subset of pathologies that are known to be significant within emergency department management. Each $p \in P$ is characterized by a maximum estimated time $t^{\max}(p)$ that a person suffering p could wait without medical control. Let $f_p \in F$ a specific medical treatment for treating p.

Our analysis focuses on three pathologies, namely ST Elevation Myocardial Infarction, Acute Myocardial Infarction (AMI) and Femoral Fracture (FF), which have a remarkable impact on Lazio healthcare management system.

First-aid requests. Let R be the set of first-aid requests arising on the Lazio Region area, during time horizon T. Each $r \in R$ is modeled by a quadruple (u_r, t_r, c_r, p_r), where: $u_r \in U$ and $t_r \in T$ are the district and the time slot in which r arises, respectively, $c_r \in C$ is the expected triage code associated with r, namely its presumed emergency level, and p_r is the expected pathology, as diagnosed in terms of subjective symptoms.

First-aid requests assignment problem. We are given a set of emergency departments S and a set of first-aid requests R arising from a defined geographic area, during a fixed time horizon T. An assignment of first-aid requests to emergency departments is feasible if the following conditions are satisfied:

a) each request $r = (u_r, t_r, c_r, p_r) \in R$ is assigned to exactly one emergency department $s = (v_s, w_s, F_s, q_s) \in S$;

b) s supplies suitable medical treatment for p_r;

c) the expected duration of the trip from u_r to v_s does not exceed the maximum estimated time for avoiding life-threatening, i.e., $t^{\max}(p_r)$.

The goal is looking for feasible assignments that allow to maximize the overall benefit, in terms of efficiency and effectiveness of supplied emergency department services. In addition, condition (b) enforces the idea that the feasibility of requests assignment should be strongly correlated to health care delivery appropriateness [24].

3 A Mixed Integer Programming approach

In this section we introduce a basic MIP model for the problem (Section 3.1) and we then discuss some interesting and useful mathematical properties (Section 3.2) and relaxations (Section 3.3).

3.1 The Basic MIP model

In oder to define the backbone of the basic MIP model we need assignment variables and constraints. More precisely, we introduce a binary variable x_{rs} with $r \in R$, $s \in S$ for each assignment, such that $x_{rs} = 1$ if and only if request r is assigned to emergency department s. Thus, the following constraints guarantee a feasible assignment.

$$\sum_{\substack{s \in S: \\ f_{p_r} \in F_s}} x_{rs} = 1 \qquad \forall r \in R \tag{1}$$

$$x_{rs} \leq 1 - \min\left\{1, \left\lfloor \frac{d(u_r, v_s)}{t^{\max}(p_r)} \right\rfloor\right\} \qquad \forall r \in R, \forall s \in S \tag{2}$$

$$x_{rs} \geq 0 \qquad \forall r \in R, \forall s \in S \tag{3}$$

Let observe that (1) forces each request r to be assigned to exactly one emergency department that is able to supply the required medical treatment; thus, both conditions (a) and (b) are satisfied. Moreover, (2) forbids any assignment that does not respect condition (c).

Evaluating efficiency. As mentioned, the first required step is to appropriately define service quality indicators. Our model allows to evaluate the efficiency of each assignment in terms of time components, and we do distinguish two in particular.

1. **Travel time.** An initial version of the model can evaluate how to assign requests to emergency departments in order to minimize the overall time needed to reach the first-aid facility. The travel time between the place

where the call is made and the hospital is an element of paramount importance because, if the patient has compromised vital functions (consciousness, respiration, heart rate, shock) and is in life-threatening conditions, then the time needed to reach the closest hospital can strongly impact on the probability of surviving.

2. **Waiting time.** As a second factor the workload of the emergency department, quantifiable as "waiting time", has to be evaluated. Using data from the Health Emergency Information System, it is possible to empirically estimate the workload of the hospital for each day of the week and time of the day. For each ED we used 2012 data to sample waiting times in correlation with the volume of first-aid requests per time period. In this way we use the waiting time function as a proxy of the ED capacity. As a consequence, the choice of the structure may be evaluated considering penalty coefficients "proportional" to the estimated waiting time.

The first cost contribution is easily given by $d_{u_r v_s}$, whereas the second is given by function w_s, which depends on the number n of patients waiting for medical treatment (at emergency departments). In particular, w_s allows to estimate the needed waiting time for processing all first-aid requests assigned to s with the aim of penalizing emergency department overload situations. In our analysis, we have obtained ED waiting functions from a statistical study of DEP-Lazio. For each ED and for each time slot, DEP-Lazio provided us a set of points explaining the stochastic relation between the median waiting time and the number of patient waiting for healthcare services. For any dataset that was statistically relevant, we have retrieved ED waiting functions by suitably interpolating the provided data points. Of course, the function obtained is only approximately convex but this approximation has been validated by the statistical office of DEP-Lazio. Finally, we model w_s as a (convex) piecewise linear function as follows.

Definition 1. *Given k_s nonnegative integers $0 < n_1 < \cdots < n_{k_s}$ for each $s \in S$, let*

$$w_s(n) := \begin{cases} a_h^s n + b_h^s & n_h \le n < n_{h+1} \; \forall h = 1, \ldots, k_s - 1 \\ a_{k_s}^s n + b_{k_s}^s & n \ge n_{k_s} \end{cases} \tag{4}$$

such that the following conditions hold:

$$a_h^s < a_{h+1}^s \qquad \forall h = 1, \ldots, k_s - 1 \tag{5}$$

$$b_1^s = 0 \tag{6}$$

$$b_{h+1}^s = b_h^s + \left(a_h^s - a_{h+1}^s\right) n_h. \qquad \forall h = 1, \ldots, k_s - 1 \tag{7}$$

In our study, parameters a_h^s have been estimated by analyzing real waiting time data provided by DEP-Lazio. Moreover, because w_s is a piecewise convex function, it is easy to check that the following property holds:

$$w_s(n) = \max_{h=1,\ldots,k_s} \left\{a_h^s n + b_h^s\right\}, \tag{8}$$

i.e., for any value of (the number of patients waiting for medical treatment) n the slope of the linear segment such that $n_i \le n < n_{i+1}$ is the leading one in (8). This is depicted in Figure 2.

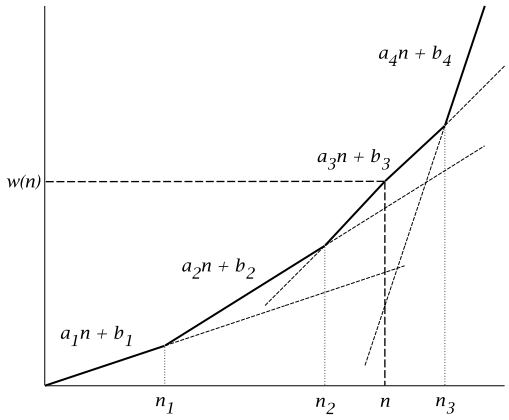

Figure 2: Example of piecewise linear convex function.

For each time slot $t \in T$, let $\bar{z}_t$ be the average waiting time of all emergency departments of Lazio. In order to balance the overall regional emergency department workload in each t, we introduce a fixed cost λ_t for each emergency department whose waiting time w_s exceeds the constant $\bar{z}_t$. In our computational experience, we discuss the impact of different choices of λ_t. Note that the workload threshold $\bar{z}_t$ has been provided by DEP-Lazio based on a stochastic analysis, which considers ED median waiting time and the regional healthcare service target.

Evaluating effectiveness. We evaluate the effectiveness of each assignment by considering the quality of healthcare service for pathologies of interest. Each hospital can be classified according to a penalty coefficient based on the quality of care provided, as estimated by the indicators of outcome and process of the Regional Program for the Evaluation of Outcomes [14, 26]. If, at the time of the request, the patient's symptoms are not clearly defined, a summary measure of hospital quality of care (taking into account some of the most relevant indicators and proceeding to their synthesis) is applied. Otherwise, if a patient has more defined symptoms, the penalty coefficient may be applied using specific indicators according to the pathological area.

For each emergency department $s \in S$, vector q_s gives the quality for each medical treatment supplied by s. In particular, the quality of care service supplied by s for treating p is denoted by q_{sp} and it is computed according to two indicators, namely the ratio of medical treatments for p over the total number of medical services supplied by s and the ratio of successful clinical interventions for p. In our model, we relate q_s components to time dimension by introducing a suitable parameter γ, which expresses the amount of time a patient is prepared to wait for achieving a one-percentage point improved service.

MIP formulation. We are now ready to define the basic MIP formulation. Let n_t^s be a nonnegative integer variable representing the total number of first-aid requests assigned to emergency department s during time slot t. Let z_t^s the workload of s during t, estimated by waiting time function w. Let y_t^s be

a binary variable such that $y_t^s = 1$ if the total workload of s during t exceeds the fixed threshold $\bar{z}_t$. Moreover, let $\bar{n}_s$ be a nonnegative integer constant corresponding to the expected total number of patients who have been waiting or receiving medical treatments in s at starting time of first time slot of T. Finally, let $\alpha_t \in [0,1]$ be a real-valued constant that reports the expected ratio of patients who have required first-aid services during $t-1$, but are still waiting or receiving medical treatments during t.
Thus, let **MIP** be the following mixed integer (linear) formulation of the *first-aid requests assignment problem*:

$$\min \sum_{r \in R} \sum_{s \in S} d(u_r, v_s) x_{rs} + \sum_{t \in T} \sum_{s \in S} (z_t^s + \lambda_t y_t^s) - \gamma \sum_{s \in S} \sum_{p \in P} q_{sp} \sum_{\substack{r \in R: \\ p_r = p}} x_{rs} \tag{9}$$

s.t.

$$n_0^s = \bar{n}_s \qquad \forall s \in S \tag{10}$$

$$n_t^s = \sum_{\substack{r \in R: \\ t_r = t}} x_{rs} \qquad \forall s \in S,\, \forall t \in T \tag{11}$$

$$z_t^s \geq a_h^s \left(\alpha_t n_{t-1}^s + n_t^s\right) + b_h^s \qquad \forall h \in \{1, \ldots, k_s\},\, \forall s \in S,\, \forall t \in T \tag{12}$$

$$z_t^s \leq \bar{z}_t + M y_t^s \qquad \forall s \in S,\, \forall t \in T \tag{13}$$

$$x_{rs} \in \mathcal{A} \cap \{0,1\}^{|R| \times |S|} \qquad \forall r \in R,\, \forall s \in S \tag{14}$$

$$n_t^s \in \mathbb{Z} \qquad \forall s \in S,\, \forall t \in T \cup \{0\} \tag{15}$$

$$z_t^s \in \mathbb{R} \qquad \forall s \in S,\, \forall t \in T \tag{16}$$

$$y_t^s \in \{0,1\} \qquad \forall s \in S,\, \forall t \in T \tag{17}$$

where M is a suitably-large real-valued constant and $\mathcal{A} \subset \mathbb{R}^{|R| \times |S|}$ is the polytope given by assignment constraints (1)-(3).

Let us observe that any constraint (12) forces the corresponding z_t^s variable to assume the appropriate value of function w by exploiting property (8). In particular, z_t^s estimates the total waiting time of s during t by considering all requests assigned at time slot t and the partial number of requests assigned at time slot $t-1$, obtained from ratio α_t.

Moreover, it is easy to check that any constraint (13) forces the associated y_t^s to 1 if the total waiting time z_t^s exceeds $\bar{z}_t$. Let us observe that y_t^s can get value 1 also when the previous condition is not satisfied: in that case, the corresponding solution could be feasible but not optimal because the objective function (9) is in minimization form.

3.2 Integrality Property of the Assignment Variables

In the following we show how to simplify model (9)–(17) by exploiting some integrality property of the assignment part of the model. First of all, we need a preliminary result that characterizes the polytope associated with assignment variables and constraints.

Proposition 1. *Given $|S| \cdot |T|$ integers ν_t^s, let polytope $\mathcal{P} \subset \mathbb{R}^{|R|\cdot|S|}$ be the intersection between $\mathcal{A}$ and the polyhedron in $\mathbb{R}^{|R|\cdot|S|}$ defined by the inequalities*

$$\sum_{\substack{r\in R:\\ t_r=t,\ f_{p_r}\in F_s}} x_{rs} = \nu_t^s \quad \forall s \in S,\ \forall t \in T. \tag{18}$$

Then, $\mathcal{P}$ is integral and the problem of optimizing a linear function over $\mathcal{P}$ is strongly polynomial.

Proof. Let $H(N, A)$ be a digraph with node set $N := R \cup (S \times T)$ and arc set A such that: (i) each node is in bijection with either a request $r \in R$ or a pair $(s, t) \in S \times T$; (ii) each arc is in bijection with an ordered pair $(r, (s, t))$ that satisfies both conditions $d(u_r, v_s) \leq t^{\max}(p_r)$ and $f_{p_r} \in F_s$. Moreover, consider the following formulation of $\mathcal{P}$:

$$\sum_{\substack{s\in S:\\ f_{p_r}\in F_s,\\ d(u_r,v_s)\leq t^{\max}(p_r)}} x_{rs} = 1 \qquad \forall r \in R \tag{19}$$

$$\sum_{\substack{r\in R:\\ t_r=t,\ f_{p_r}\in F_s,\\ d(u_r,v_s)\leq t^{\max}(p_r)}} x_{rs} = \nu_t^s \qquad \forall s \in S,\ \forall t \in T \tag{20}$$

$$0 \leq x_{rs} \leq 1 \qquad \forall t \in R,\ s \in S \tag{21}$$

where (19) and (20) are obtained by combining (2) respectively with (1) and (18). Now, it is easy to check that the constraints matrix associated with (19)-(21), called B, corresponds to the incidence matrix of H, thus B is totally unimodular, so it follows that $\mathcal{P}$ is integral. In particular, $\mathcal{P}$ corresponds to the feasible region of a flow problem associated with digraph H with demands $d_r = -1$ for each $r \in R$, $d_{(s,t)} = \nu_t^s$ for each $(s, t) \in S \times T$. Then, by [31], we can conclude that optimizing a linear function over $\mathcal{P}$ is strongly polynomial. □

Now, we are able to define an improved formulation in which the number of integer variables is reduced from $O(|R| \cdot |S| + |S| \cdot |T|)$ to $O(|S| \cdot |T|)$.

Theorem 1. *Let* **MIP′** *be the mixed integer program obtained from* **MIP** *by relaxing the integrality of variables x_{rs}. Then,* **MIP** *and* **MIP′** *have the same optimum value and an optimal solution to* **MIP** *can be obtained from an optimal solution to* **MIP′** *in strongly polynomial time.*

Proof. Let ω' and ω be the optimal solution values of **MIP′** and **MIP**, respectively. In general, $\omega' \leq \omega$ holds because **MIP′** is a relaxation of **MIP**. Let $\chi' := (x', n', z', y')$ be an optimal solution of **MIP′** and let consider polytope $\mathcal{P}$ with $\nu_t^s = n'^s_t$ for all $s \in S$, $t \in T$. Then, let $x^\star$ be an optimal solution obtained by maximizing function $\sum_{r\in R} \sum_{s\in S} d(u_r, v_s) x_{rs} - \gamma \sum_{s\in S} \sum_{p\in P} q_{sp} \sum_{\substack{r\in R:\\ p_r=p}} x_{rs}$ over $\mathcal{P}$. Due to Proposition 1, $x^\star$ is integral and it can be computed in strongly polynomial time. Because $\chi^\star := (x^\star, n', z', y')$ is feasible for **MIP′** and its corresponding objective function value is less or equal to ω', we have that $\chi^\star$ is an optimal solution of **MIP′**. Moreover, because $\chi^\star$ is feasible for **MIP**, we can conclude that $\chi^\star$ is optimal also of **MIP**. □

3.3 Relaxing Workload Balance

Let $\mathbf{MIP_0}$ be the mixed-integer programming problem obtained from $\mathbf{MIP'}$ by relaxing constraints (13) (and assuming $\lambda = 0$). In particular, $\mathbf{MIP_0}$ models the relaxation of MIP (9)–(17) in which emergency departments workloads are not required to be balanced. In the following, we present a reformulation of $\mathbf{MIP_0}$ as a generalized min-cost flow problem on a suitable network.

Let $D(N, A)$ be a digraph with node set N and arc set A, let $b : N \to \mathbb{R}$ be a demand function associated with nodes, let $l, \mu : A \to \mathbb{R}_+$ and $a : A \to \mathbb{R}$ be capacity, gain and cost functions associated with arcs, respectively. A pseudoflow is a function $\varphi : A \to \mathbb{R}$ such that $0 \leq \varphi(i,j) \leq l(i,j)$ holds for all arcs $(i,j) \in A$. The generalized min-cost flow problem consists of finding a pseudoflow that minimizes the overall cost $\sum_{(i,j)\in A} a(i,j)\varphi(i,j)$ subject to the generalized flow-conservation constraints

$$\sum_{(i,j)\in A} \varphi(i,j) - \sum_{(j,i)\in A} \mu(j,i)\varphi(j,i) = b(i) \qquad \forall i \in N.$$

For each $e = (i,j) \in A$, let $\bar{e} := (j,i)$ be the reverse arc corresponding to e and let $\bar{A}$ denote the set of reverse arcs associated with A. For reverse arcs, gain and cost functions satisfy $\gamma(\bar{e}) = 1/\gamma(e)$ and $a(\bar{e}) = -a(e)/\gamma(e)$, respectively. Moreover, given a pseudoflow φ, the residual capacity function $l_\varphi : A \cup \bar{A} \to \mathbb{R}$, is defined as $l_\varphi(e) = l(e) - \varphi(e)$ for each $e \in A$ and $l_\varphi(\bar{e}) = \gamma(e)\varphi(e)$. Then, let $D_\varphi(N, \bar{A}, b, l_\varphi, \gamma, a)$ be the residual network associated with φ. The gain of a cycle belonging to D_φ is the product of the gains of arcs that compose the cycle. A cycle of D_φ whose gain is strictly greater (resp. less) than one unit is called flow-generating (resp. flow-absorbing). A bicycle is composed by a flow-absorbing cycle and a flow generating cycle that are arc-disjoint and connected by a path containing at least one node. We recall that a feasible pseudoflow φ is optimal if and only if D_φ does not contain any unit-gain cycle or bicycle. For further details, the reader is referred to [16, 1].

The generalized min-cost flow is a well-known optimization problem that has a wide range of applications in many scientific area, as discussed in [1]. It belongs to the field of generalized flow, so it reduces to min-cost flow by assuming $\gamma(e) = 1$ for all $e \in A$. Since generalized min-cost flow is a special case of linear programming, it can be solved in polynomial time by the ellipsoid method [20]. In the literature, many other polynomial algorithms have been addressed, which are based on linear programming as reported in [19, 32], or exploit combinatorial approaches, like in [15, 34]. While min-cost flow can be solved in strongly polynomial time [31], it is unknown whether the generalized min-cost flow problem admits strongly polynomial algorithms. However, in [10] it is shown that the problem is strongly polynomial if there is a fixed number of arcs whose gain is either than one unit.

In the following we characterize an instance of generalized min-cost flow, denoted by $D(N, A, b, l, \gamma, a)$, which gives a combinatorial description of $\mathbf{MIP_0}$.

Let $K_s := \{1, \ldots, k_s\} \times T$ for each $s \in S$, $R_t := \{r \in R : t_r = t\}$, $R'_t := R_t$, $R' := R$, $S' := S$ and $T' := T$. Then, let $D(N, A)$ be a digraph with node set

$$N = R \cup S \cup (S \times T) \cup K_1 \cup \cdots \cup K_{|S|} \cup (S' \times T') \cup R' \cup S',$$

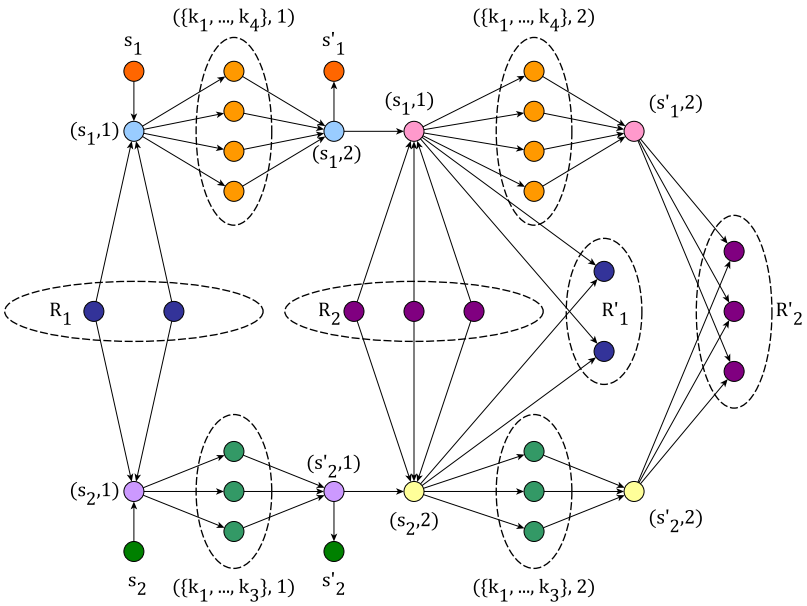

Figure 3: Example of network $D(N, A)$.

and arc set $A = \bigcup_{j=1}^{8} A_j$ such that

$$A_1 := \{(r, (s, t)) : r \in R_t,\ s \in S,\ t \in T \text{ with } d(u_r, v_s) \leq t^{\max}(p_r),\ f_{p_r} \in F_s\}$$
$$A_2 := \{(s, (s, t)) : s \in S,\ t = 1\}$$
$$A_3 := \{((s, t), (k, t)) : s \in S,\ t \in T,\ (k, t) \in K_s\}$$
$$A_4 := \{((k, t), (s', t')) : s' \in S',\ t' \in T',\ (k, t) \in K_s \text{ with } s = s',\ t = t'\}$$
$$A_5 := \{((s', t'), (s, t + 1)) : s' \in S',\ t' \in T' \setminus \{\tau\} \text{ with } s = s',\ t = t'\}$$
$$A_6 := \{((s', t' + 1), r') : s' \in S',\ t' \in T' \setminus \{\tau\},\ r' \in R'_{t'}\}$$
$$A_7 := \{((s', t'), r') : s' \in S',\ t' = \tau,\ r' \in R'_{\tau}\}$$
$$A_8 := \{((s', t'), s') : s' \in S',\ t' = 1\}.$$

Figure 3.3 shows an example of $D(N, A)$ for $T = \{1, 2\}$, $S = \{s_1, s_2\}$, $R = R_1 \cup R_2$, $k_{s_1} = 4$, $k_{s_2} = 3$. Furthermore, let us consider the following functions:

$$b(i) = \begin{cases} -1 & i = r \in R \\ +1 & i = r' \in R' \\ -\alpha_1 \bar{n}^s & i = s \in S \\ +\alpha_1 \bar{n}^{s'} & i = s' \in S' \\ 0 & i = (k,t) \in K_s \quad \forall s \in S \\ 0 & i = (s,t) \in (S \times T),\ (s',t') \in (S' \times T') \end{cases}$$

$$l(e) = \begin{cases} 1 & e \in A_1 \cup A_7 \\ \alpha_1 \bar{n}^s & e \in A_2 \cup A_8 \\ n_k - n_{k-1} & e \in A_3 \cup A_4, \forall k \in \{1, \ldots, k_s - 1\} \text{ with } n_0 = 0 \\ |R_t| - n_k & e \in A_3 \cup A_4 \text{ for } k = k_s \\ |R_t| & e \in A_5 \\ \alpha_t & e \in A_6 \end{cases}$$

$$\mu(e) = \begin{cases} \alpha_t & e \in A_5 \\ 1/\alpha_t & e \in A_6 \\ 1 & e \in A \setminus (A_5 \cup A_6) \end{cases}$$

$$a(e) = \begin{cases} d(u_r, v_s) - \gamma q_{sp} & e \in A_1 \\ a_k^s & e \in A_3 \\ 0 & e \in A \setminus (A_1 \cup A_3) \end{cases}$$

The following result states that the generalized min-cost problem associated with $D(N, A, b, l, \mu, a)$ is a relaxation of $\mathbf{MIP_0}$.

Lemma 1. *For each feasible solution χ to $\mathbf{MIP_0}$ there exists a feasible pseudoflow φ_χ associated with $D(N, A, b, l, \mu, a)$ such that χ and φ_χ have the same cost.*

Proof. Let $\chi := (x, n, z)$ be a feasible solution of $\mathbf{MIP_0}$. For each $s \in S$, let h_s be the largest integer in $\{1, \ldots, k_s\}$ such that $h_s \leq n_t^s + \alpha_t n_{t-1}^s$. Then, let φ_χ be a pseudoflow associated with $D(N, A, b, l, \mu, a)$ such that, for each $r' = r \in R$, $s' = s \in S$, $t' = t \in T$

$$\varphi_\chi(r, (s,t)) = x_{rs} \tag{22}$$

$$\varphi_\chi(s, (s,1)) = \alpha_1 \bar{n}^s \tag{23}$$

$$\varphi_\chi((s,t),(k,t)) = \begin{cases} n_k - n_{k-1} & k \in \{1, \ldots, h_s\} \\ n_t^s + \alpha_t n_{t-1}^s - n_{h_s} & k = h_s + 1 \\ 0 & k \in \{h_s + 2, \ldots, k_s\} \end{cases} \tag{24}$$

$$\varphi_\chi((k,t),(s',t')) = \varphi_\chi((s,t),(k,t)) \tag{25}$$

$$\varphi_\chi((s',t'),(s,t+1)) = n_t^s \quad \text{with} \quad t \in T \setminus \{\tau\} \tag{26}$$

$$\varphi_\chi((s',t'+1),r') = \alpha_t x_{rs} \quad \text{with} \quad t \in T \setminus \{\tau\} \tag{27}$$

$$\varphi_\chi((s',\tau),r') = x_{rs} \quad \text{with} \quad r \in R_\tau \tag{28}$$

$$\varphi_\chi((s',1),s') = \alpha_1 \bar{n}^s \tag{29}$$

$$\tag{30}$$

It is easy to check that φ_χ is feasible: capacity and pseudoflow conservation constraints are satisfied. Moreover, observe that (22) implies

$$\sum_{r\in R}\sum_{s\in S}\left(d(u_r,v_s)-\gamma\sum_{\substack{p\in P:\\ p_r=p}}q_{sp}\right)x_{rs}=\sum_{r\in R}\sum_{s\in S}a(r,(s,t))\varphi_\chi(r,(s,t)). \tag{31}$$

By relation (8), the following condition holds:

$$z_t^s=a_{h_s+1}^s\left(\alpha_t n_{t-1}^s+n_t^s\right)+b_{h_s+1}^s \quad \forall s\in S,\,\forall t\in T. \tag{32}$$

Then, by substituting (5)-(7) in (32), it follows that

$$\begin{aligned} z_t^s &= a_{h_s+1}^s(n_t^s+\alpha_t n_{t-1}^s)+\sum_{h=1}^{h_s}(a_h^s-a_{h+1}^s)n_h= \\ a_{h_s+1}^s(n_t^s+\alpha_{t-1}n_{t-1}^s-n_{h_s})&+\sum_{h=1}^{h_s}(n_h-n_{h-1})a_h^s= \\ a((s,t),(h_s+1,t))&\varphi_\chi((s,t),(h_s+1,t)) \\ &+\sum_{h=1}^{h_s}a((s,t),(h,t))\varphi_\chi((s,t),(h,t)). \end{aligned} \tag{33}$$

Thus, relations (31) and (33) imply that χ and φ_χ have the same cost. □

In general, the reverse is not true, i.e., there exist feasible solutions of generalized min-cost flow over $D(N,A,b,l,\mu,a)$ that cannot be mapped into feasible solutions of $\mathbf{MIP_0}$. However, latter problems are equivalent under certain conditions, e.g., assuming $\alpha_t=1$ for each $t\in T$. In this case, the generalized min-cost flow over $D(N,A,b,l,\mu,a)$ reduces to the min-cost flow problem over $D(N,A,b,l,a)$. Since demands and capacities are integer, there exist integral optimal flows corresponding to feasible solutions of $\mathbf{MIP_0}$ that are optimal by Lemma 1. Furthermore, we can show the following result.

Theorem 2. *Let us assume*

$$\min_{\substack{r\in R\\ s,\tilde{s}\in S,\,s\neq\tilde{s}}}\left\{d(u_r,v_s)-d(u_r,v_{\tilde{s}})\right\}\geq\max_{\substack{s,\tilde{s}\in S\\ s\neq\tilde{s}}}\left\{a_{k_s}^s-a_1^{\tilde{s}}\right\}. \tag{34}$$

Then, optimal solution to $\mathbf{MIP_0}$ *can be computed in strongly polynomial time.*

Proof. Let $\varphi^\star$ be an optimal pseudoflow to generalized min-cost flow problem associated with $D(N,A,b,l,\mu,a)$. In general, $\varphi^\star$ is not integral. Since the residual network $D_{\varphi^\star}$ corresponding to $\varphi^\star$ does not contain negative cycles, it is easy to check that vertices (s,t), (k,t), (s',t') form strictly positive cost cycles for each $s\in S$, $t\in T$ with $s'=s$ and $t'=t$. Thus, it follows that relation (33) is satisfied. Moreover, (34) ensures that for each $(r,(s,t))\in A$ such that $0<\varphi^\star(r,(s,t))<1$, there exists at least a null cost cycle in $D_{\varphi^\star}$ that contains arc $(r,(s,t))$ and $((s,t),(k,t))$ with residual capacity greater than or equal to $1-\varphi^\star(r,(s,t))$. Thus, an optimal integer pseudoflow φ' can be obtained from $\varphi^\star$ by saturating $O(|R|)$ null-cost cycles. Then, we have that φ' corresponds to a feasible solution χ' of $\mathbf{MIP_0}$, so we conclude that χ' is optimal by Lemma 1. Finally, by [10], it follows that χ' can be computed in strongly polynomial time. □

4 Computational Results

The computational experience focuses on a wide set of instances that are based on real-world data from the Lazio emergency department system during the entire year 2012.
First-aid requests characteristics have been retrieved from the Hospitals Information System (HIS)[1] and the Emergency Room Information System (ERIS)[2] of regional healthcare services authority. In particular, HIS manages the Hospital Discharge Register (HDR)[3] database, which maintains information of all hospital admissions and discharges, by integrating patients personal details, healthcare services supplied and medical treatment results. Lazio's HDR provides additional medical treatments information for STEMI, AMI and FF, which are the pathologies of interest associated with our analysis. The ERIS integrates HIS database by supplying specific and detailed information only for emergency departments. The description of regional emergency departments, with associated quality of service information, has been retrieved from statistical studies carried out by DEP-Lazio, which are based on regional and national evaluation programs for medical operations results. For more details, we refer the reader to [14, 26].
The time horizon T naturally refers to a day of the year 2012 and it has been discretized into the following three time slots: *daytime* (8:00 am - 15:59 pm), *evening* (16:00 pm - 23:59 pm), *night* (0:00 am - 7:59 am). Such decomposition has suitable operational relevance in terms of service level and expected number of requests. Moreover, the effects of emergency department workloads balancing have been evaluated by considering two different classes of values for λ_t. A first class referring to $\mathbf{MIP_0}$ allows to consider solutions with unbalanced workloads ($\lambda_t = 0$ for all $t \in T$), whereas a second class referring to $\mathbf{MIP'}$ guarantees workload-balanced solutions. In latter case, we fix $\lambda_t = \bar{z}_t$ for all $t \in T$, in order to avoid requests allocations which cause undesirable overloads of emergency departments with small capacity.

The computational experience has been carried out on a x86-64 GNU/Linux machine (CentOS 6.3) with 8 cores @2GHz and 16GB of RAM. We have generated instances of $\mathbf{MIP'}$ and $\mathbf{MIP_0}$ for each day of year 2012 by considering all 50 operating emergency departments of Lazio. Then, we have achieved optimal solutions for all instances by using IBM ILOG Cplex 12.5.1.
Table 1 summarizes computational results for $\mathbf{MIP'}$ instances by reporting average values for each month: in particular, *i*) the second column reports the average number of emergency requests occurred in each day; *ii*) the third column reports the number of infeasible instances, i.e., the number of days of the month in which at least one request could not be correctly assigned according to the constraints of our model; *iii*) the forth column shows the average optimal solution value of each day, while columns fifth and sixth indicate cost contributions of waiting time functions (the sum of the z_t^s variables in (9)) and overall penalty time value (the sum of $\lambda_t y_t^s$ terms), respectively; *iv*) the last two columns report the average Cplex performance (in terms of elapsed real computational time and total branch&bound nodes) that has been observed for solving each (feasible) instance of the month.

[1] Corresponding Italian acronym is SIO: "Sistema Informativo Ospedaliero".
[2] Corresponding Italian acronym is SIES: "Sistema Informativo per l'Emergenza Sanitaria".
[3] Corresponding Italian acronym is SDO: "Scheda di dimissione ospedaliera".

Month (2012)	Mean |R| (day)	Infeas. MIPs	Optimum (min)	Waiting time (min)	Penalty time (min)	Cplex time (sec)	Cplex nodes (average)
1	4,300.4	8	60,297.63	25,360.77	2,899.85	111.316	3,209.1
2	3,980.9	6	55,943.78	23,570.66	2,838.61	109.878	3,387.1
3	4,323.1	5	61,418.85	25,859.72	2,888.61	122.203	3,231.3
4	4,287.2	5	60,861.64	25,599.72	2,894.28	131.929	3,384.2
5	4,493.5	7	63,015.01	26,568.41	2,924.67	137.239	3,409.3
6	4,590.0	7	64,929.66	27,427.35	2,990.86	148.866	3,371.1
7	4,409.1	8	62,458.13	26,341.63	2,977.74	139.629	3,311.0
8	4,229.0	4	59,870.97	25,180.69	2,932.98	132.760	3,257.4
9	3,996.3	0	56,082.89	23,604.62	2,859.93	113.067	3,305.0
10	4,154.2	10	58,111.27	24,425.29	2,874.46	112.856	3,395.1
11	4,112.7	11	57,930.31	24,405.94	2,902.58	114.509	3,382.1
12	4,025.8	8	56,399.67	23,862.90	2,822.65	119.214	3,295.8

Table 1: Computational experience.

The results in Table 1 show that $\mathbf{MIP'}$ can be solved relatively easily by a sophisticated MIP solver like Cplex 12.5.1. The number of instances that turn out to be infeasible is relatively small, namely around 20%. The influence of the penalty term associated with workload unbalance amounts at 10% of the term associated with the waiting time. In order to evaluate how important is the penalization of such an unbalance we also solved $\mathbf{MIP_0}$ and the results are rather easy to interpret. Because $\mathbf{MIP_0}$ is a relaxation of $\mathbf{MIP'}$, as discussed in Section 3.1, optimal values to $\mathbf{MIP_0}$ are on average better of 5.36% than those of $\mathbf{MIP'}$, but at the price of an increased waiting time cost contribution, on average of 0.74%, due to the absence of workload balance. We omit detailed results on $\mathbf{MIP_0}$ instances but it is worth mentioning that they are very easy to solve both by using a combinatorial algorithm for generalized min-cost flow or by solving $\mathbf{MIP_0}$ with a general-purpose MIP solver like Cplex. In the latter case, no branching is ever necessary.

Finally, concerning the impact of the property studied in Section 3.2 that led to a simplification of formulation **MIP** into $\mathbf{MIP'}$, we have a rather interesting situation. The simplified formulation $\mathbf{MIP'}$ is faster than **MIP**, namely 6.01% in geometric mean over the 270 solved instances. However, the number of explored nodes is much higher, namely 24.29%. In other words, in the case of $\mathbf{MIP'}$ Cplex is much faster in exploring nodes, which is quite strange because the only difference between the two formulations is that some of the variables are declared continuous instead of binary, thus the LP relaxations should be identical. However, as an example, a generic instance (January 17, 2012) goes from 186,006 binary variables to only 150, which might explain the issue. Indeed, the presence of 186,006 binary variables potentially lead to many time-consuming computational steps, for example, in the preprocessing, node preprocessing, probing, cutting plane generation, branching selection, etc. A "cleaner" formulation is overall preferable and, in practice, leads to speed up the computation. Note, however, that it is a matter of tradeoff because the node increase of $\mathbf{MIP'}$ with respect to **MIP** shows that the above time-consuming steps are in fact effective.

As pointed out in the Introduction, the aim of the present study is to compute the optimal solution of a (unrealistic) fully centralized allocation of first-aid requests to EDs, so as to be able to in the evaluation of both the

state of the art and future reorganization ideas, some sort of Price of Anarchy evaluation. To achieve this we compare in Table 2 real (observed) first-aid request assignments during year 2012 with the optimal solutions of model **MIP′**. Specifically, Table 2 is organized as follows: *i*) second and third columns indicate the number of infeasible assignments with respect to the violation of constraints (2) (each patient has to reach an emergency department within a suitable time according to kind of health emergency) and (1) (each request has to be assigned to an emergency department with a suitable specialization that allows to supply appropriate medical treatments), respectively; *ii*) the forth column exhibits the average objective function value for each day, while columns fifth and sixth specify cost contributions of waiting time functions and overall penalty time value, respectively (analogously to Table 1); *iii*) the last three columns report the average relative gaps between values of observed assignments and optimal solution for the overall value and cost contributions of waiting and penalty times, respectively.

Month (2012)	**Violated constr.** (2)	**Violated constr.** (1)	**Overall value (min)**	**Waiting time (min)**	**Penalty time (min)**	**Overall value (gap%)**	**Waiting time (gap%)**	**Penalty time (gap%)**
1	2.0	360.4	101,437.07	47,964.39	3,851.36	40.18	46.69	24.58
2	2.0	299.2	91,175.49	42,848.48	3,818.35	38.86	45.01	25.79
3	1.6	362.6	102,101.16	48,415.55	3,870.28	40.05	46.73	25.38
4	2.0	361.1	101,898.65	48,231.42	3,862.21	40.57	47.24	25.13
5	1.5	370.4	105,827.41	50,405.02	3,886.48	40.37	47.19	24.80
6	2.0	383.1	110,125.15	51,671.08	3,923.45	41.27	47.19	23.76
7	1.8	343.5	108,814.64	48,679.28	3,944.40	42.90	46.07	24.47
8	2.5	361.5	109,621.28	46,288.93	3,949.02	45.15	45.39	25.72
9	1.9	301.8	93,799.83	43,309.37	3,874.65	40.19	45.43	26.19
10	2.0	320.5	96,755.77	45,338.37	3,852.34	39.32	45.33	25.32
11	1.7	335.2	95,445.33	45,136.85	3,838.66	40.06	46.79	24.52
12	1.8	357.5	94,010.21	44,578.50	3,818.23	40.00	46.22	26.16

Table 2: Comparing observed request allocations with optimized solutions.

The numbers in Table 2 immediately show that the solutions naturally obtained without a centralized allocation strategy (for example a remote triage) violate many of the constraints of **MIP′**, especially constraints (1) associated with suitable specialization. This information is especially interesting from a strategic (and practical) standpoint: such a remote triage conducted in an effective way could have a remarkable impact to significantly reduce these violations that correspond to dangerous inefficiency of the system. The rest of the numbers of Table 2 are instead interesting but not easy to interpret. In a sense the objective function (9) is completely disregarded by the observed request allocation system but maybe the part of it associated with the minimization of the travel time. Thus, the absolute and relative difference of the components of the objective function are less meaningful at this point in time, while they will be more and more so when different reorganization settings will be evaluated.

5 Optimized real-time first-aid requests assignment

The offline assignment of all first-aid requests at once is very interesting from a strategic viewpoint but does not immediately provide a tool that can support health managers during the day-by-day operations. In order to implement a regional remote triage approach it is then fundamental to define a real-time approach that can manage the assignment of incoming requests (almost) online.

Section 3.3 has introduced a reformulation of the first-aid assignment problem with unbalanced emergency department workloads $\mathbf{MIP_0}$ as a generalized min-cost flow problem on a suitable instance $D(N, A, b, l, \gamma, a)$. For this purpose, the computational experience has shown that optimal solutions to one-day instances can be efficiently computed requiring computational time less than one second. As a consequence, the high performance of flow-based method give rise naturally to the investigation of *real-time* optimization approaches for the first-aid requests assignment problem. In this section, we discuss a basic real-time paradigm that exploits the addressed flow-based reformulation.

The main idea consists in considering three sets of first-aid requests, namely $R_{\vartheta-1}$, R_{ϑ} and $R_{\vartheta+1}$, where $\vartheta \in T$ represents the current time slot. In particular, $R_{\vartheta-1}$ is the set of requests that have been assigned during previous time slot, R_{ϑ} contains all requests which have been assigned during the current time slot, whereas $R_{\vartheta+1}$ is the set of forecast requests that are expected to occur during the next time slot. We motivate this design feature by remarking that closest past and future assignments have a notable impact on current decision because of emergency department workload functions. Moreover, observe that forecasted requests can be suitably computed by analyzing the associated time series and considering stochastic perturbation effects due to periodical and exceptional events.

A basic real-time scheme can be initialized by a generalized min-cost flow instance $D(N, A, b, l, \gamma, a)$ with $T := \{\vartheta - 1, \vartheta, \vartheta + 1\}$, $R_{\vartheta} := \emptyset$ and $R_{\vartheta-1}$, $R_{\vartheta+1}$ containing requests assigned during $\vartheta - 1$ and forecast requests expected during $\vartheta+1$, respectively. The main loop of the real-time scheme is activated for each incoming request ρ, so instance $D(N, A, b, l, \gamma, a)$ is consistently updated by considering both ρ and workload balance information. Then, the instance is solved in order to compute and fix the assignment of ρ. The workload balance penalty λ can be considered by computing the set P of emergency departments that are close to saturation condition, expressed by time threshold $\bar{z}$. In particular, given a current requests assignment, characterized by variables n^s_{ϑ} for each $s \in S$, we say that $s \in P$, i.e., s is close to saturation, if n^s_{ϑ} belongs to interval $[n^s_{\bar{z}} - 1, n^s_{\bar{z}}]$ with $n^s_{\bar{z}} := \min_{h=1,\dots,k_s}\{\frac{\bar{z}-b^s_h}{a^s_h}\}$ for each $s \in S$ (in other words, $n^s_{\bar{z}}$ is the threshold for requests number of s such that $w_s(n^s_{\bar{z}}) = \bar{z}$). Thus, at each main loop iteration, the workload unbalance penalty can be treated by temporary adding the fixed penalty cost λ to the cost of arc $(\rho, (s, \vartheta)) \in A(D)$ for each $s \in P$. Algorithm 1 reports such basic real-time approach.

Algorithm 1: Real-time first-aid requests assignment algorithm

Input: $R_{\vartheta-1}$, $R_{\vartheta+1}$
Result: Optimal real-time first-aid assignments
$R_\vartheta \leftarrow \emptyset$;
$N \leftarrow R_{\vartheta-1} \cup R_{\vartheta+1}$;
Generate $D(N, A, b, l, \gamma, a)$;
$P \leftarrow \emptyset$;
foreach incoming request ρ **do**
  $R_\vartheta \leftarrow R_\vartheta \cup \{\rho\}$;
  $N \leftarrow R_{\vartheta-1} \cup R_\vartheta \cup R_{\vartheta+1}$;
  Update $D(N, A, b, l, \gamma, a)$;
  $a' \leftarrow a$;
  while $P \neq \emptyset$ **do**
    Select s from P;
    $a(\rho, (s, \vartheta)) \leftarrow a(\rho, (s, \vartheta)) + \lambda$;
    $P \leftarrow P \setminus \{s\}$;
  end
  Solve generalized min cost flow over $D(N, A, b, l, \gamma, a)$;
  Assign request ρ accordingly;
  $a \leftarrow a'$
end

Algorithm 1 exploits several ingredients proposed in the previous section, namely, the original MIP model, its relaxation $\mathbf{MIP_0}$, and the availability of a quick combinatorial solution of it, for proposing the first online first-aid requests allocation based on a centralized triage. We have not yet investigated the computational performance of this approach because remote triage policies are still under development in Lazio but this basic algorithm constitutes the starting point of the follow-up of our project.

6 Conclusions and Future Work

The assignment of service requests to emergency departments is of paramount importance both from a life-threatening and an economical viewpoints. Within a more general project that aims at defining optimal allocation policies of patients to regional hospital network facilities, the Department of Epidemiology of the Regional Health Service of Lazio (Italy) was interested in obtaining a completely offline picture of the effect of an optimal assignment of requests to emergency departments. This is in the spirit of evaluating the so-called Price of Anarchy, where the fully centralized (admittedly unrealistic) allocation is used as a reference for both the state of the art completely decentralized approach and future reorganization ideas.

We have implemented and tested with real-world data of all service requests of 2012 a Mixed Integer Linear Programming model that computes such an optimal request allocation by minimizing travel and waiting times and penalize workload unbalance among emergency departments in the region. Within the development process we have studied special cases and relaxations of the complete model showing interesting mathematical properties that are, in turn, useful from a practical viewpoint. Finally, one of those special cases allowed us to devise a real-time version of the first-aid requests allocation approach,

which can be used as a Decision Support System for the Triage Center daily operations.

The present study is an important, quantitative step in the evaluation of centralized allocation strategies like remote triage that could have a remarkable impact in making the allocation process much more efficient and effective. More precisely, the developed methodology as well as the software tools are currently used by the DEP-Lazio for the reorganization of the regional networks of emergency healthcare. Our findings will be shared with the Regional Directorate for health and social-health integration and the Regional Healthcare Emergency Unit, which operatively manages the first aid requests in Lazio. The joint analysis of the results by those who plan emergency healthcare programs and by those who operationally run them in the territory are expected to be helpful to develop and quantitatively evaluate strategies to: (a) improve health assistance for the population living in disadvantaged areas, (b) reduce waiting times in emergency departments and (c) balance workload among EDs of the Lazio region. A future step in the direction of improving the above goals is to include scheduling rules and patient priorities according to suitable functions measuring single patient waiting times. A possibility under investigation is to adjust the mathematical model by taking into account the workload of each first-aid request based on its priority code, thus obtaining a weighted version of constraints (11).

More generally, considering that the technical equipment is known for each hospital, this type of optimization (possibly coupled with simulation) techniques can be effectively used to reorganize the emergency networks in accordance with the hierarchical levels of the hospitals equipment complexity. In addition, the fleet of emergency vehicles currently in use, namely 3 helicopters to support first-aid activities (located in Viterbo, Rome and Latina) and 219 between ambulances and medical cars, must be taken into account more accurately. The emergency vehicles are located in 149 stations throughout the region, grouped into 5 Operative Centers. The fleet management is the topic of the follow up project "Optimization of the cardiac network in the Lazio region: appropriateness, timeliness and equity in access to emergency care".
The presented project and the follow up one are likely to result in optimization of the current "Hub and Spoke" model, based on the distinction of the emergency departments in basic EDs, first level EDs and second level EDs, depending on the provided intensity of care.

Acknowledgments

We are indebted with Danilo Fusco and all the members of DEP-Lazio for the overall work developed together. In addition, we would like to thank Sven Wiese and Andrea Tramontani for useful discussions on the MIP properties. Finally, the work has been supported by MIUR, Italy, under the PRIN 2009 grant. We are indebted to two anonymous referees for raising helpful questions and remarks that helped in improving the paper.

References

[1] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin. *Network Flows: Theory, Algorithms, and Applications.* Prentice Hall, Englewood Cliffs, NJ, 1993.

[2] E Almehdawe, B Jewkes, and Q. He. A markovian queueing model for ambulance offload delays. *European Journal of Operational Research*, 226(3):602 – 614, 2013.

[3] I.K. Altınel and E. Ulaş. Simulation modeling for emergency bed requirement planning. *Annals of Operations Research*, 67(1):183–210, 1996.

[4] A. Bagust, M. Place, and J.W. Posnett. Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. *BMJ*, 319(7203):155–158, 7 1999.

[5] C.C. Branas, E.J. MacKenzie, and C.S. ReVelle. A trauma resource allocation model for ambulances and hospitals. *Health Serv Res*, 35(2):489–507, Jun 2000.

[6] C.W. Burt, L.F. McCaig, and R.H. Valverde. Analysis of ambulance transports and diversions among {US} emergency departments. *Annals of Emergency Medicine*, 47(4):317 – 326, 2006.

[7] A. Chockalingam, K. Jayakumar, and M.A. Lawley. A stochastic control approach to avoiding emergency department overcrowding. In *Proceedings of the Winter Simulation Conference*, WSC '10, pages 2399–2411. Winter Simulation Conference, 2010.

[8] S.C.K. Chu and L. Chu. A modeling framework for hospital location and service allocation. *International Transactions in Operational Research*, 7(6):539–568, 2000.

[9] J.K. Cochran and K.T. Roche. A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Computers & Operations Research*, 36(5):1497 – 1512, 2009.

[10] E. Cohen and N. Megiddo. Algorithms and complexity analysis for some flow problems. *Algorithmica*, 11(3):320–340, 1994.

[11] L.G. Connelly and A.E. Bair. Discrete event simulation of emergency department activity: a platform for system-level operations research. *Acad Emerg Med*, 11(11):1177–1185, Nov 2004.

[12] S. Deo and I. Gurvich. Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science*, 57(7):1300–1319, 2011.

[13] P. Enders. *Applications of stochastic and queueing models to operational decision making*. PhD thesis, Tepper school of Business, Carnegie Mellon University, 2010.

[14] D. Fusco, A.P. Barone, C. Sorge, M. D'Ovidio, M. Stafoggia, A. Lallo, M. Davoli, and C.A. Perucci. P.re.val.e.: outcome research program for the evaluation of health care quality in lazio, italy. *BMC Health Serv Res*, 12:25, 2012.

[15] A.V. Goldberg, S.A. Plotkin, and É. Tardos. Combinatorial algorithms for the generalized circulation problem. *Mathematics of Operations Research*, 16:351–379, 1989.

[16] A.V. Goldberg, É. Tardos, Stanford University. Computer Science Department, and R.E. Tarjan. *Network Flow Algorithms*. Number No. 1252 in Computer Science Department: Report STAN-CS. Department of Computer Science, Stanford University, 1989.

[17] M.M. Gunal and M. Pidd. Understanding accident and emergency department performance using simulation. In *Simulation Conference, 2006. WSC 06. Proceedings of the Winter*, pages 446–452, Dec 2006.

[18] P.R. Harper, A.K. Shahani, J.E. Gallagher, and C. Bowie. Planning health services with explicit geographical considerations: a stochastic location–allocation approach. *Omega*, 33(2):141 – 152, 2005.

[19] S. Kapoor and P.M. Vaidya. Fast algorithms for convex quadratic programming and multicommodity flows. In *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*, STOC '86, pages 147–159, New York, NY, USA, 1986. ACM.

[20] N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4):373–395, 1984.

[21] A. Komashie and A. Mousavi. Modeling emergency departments using discrete event simulation techniques. In *Proceedings of the 37th Conference on Winter Simulation*, WSC '05, pages 2681–2685. Winter Simulation Conference, 2005.

[22] R. Konrad, K. DeSotto, A. Grocela, P. McAuley, J. Wang, J. Lyons, and M. Bruin. Modeling the impact of changing patient flow processes in an emergency department: Insights from a computer simulation study. *Operations Research for Health Care*, 2(4):66 – 74, 2013.

[23] E. Koutsoupias and C.H. Papadimitriou. Worst-case equilibria. In *Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science*, pages 404–413, 1999.

[24] J. N. Lavis and G. M. Anderson. Appropriateness in health care delivery: definitions, measurement and policy implications. *CMAJ: Canadian Medical Association Journal*, 154(3):321–328, 02 1996.

[25] S. Lee. The role of hospital selection in ambulance logistics. *IIE Transactions on Healthcare Systems Engineering*, 4(2):105–117, 2014.

[26] C. Renzi, C. Sorge, D. Fusco, N. Agabiti, M. Davoli, and C.A. Perucci. Reporting of quality indicators and improvement in hospital performance: the p.re.val.e. regional outcome evaluation program. *Health Serv Res*, 47(5):1880–1901, Oct 2012.

[27] V. Schmid. Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research*, 219(3):611 – 621, 2012.

[28] D. Sinreich and O. Jabali. Staggered work shifts: a way to downsize and restructure an emergency department workforce yet maintain current operational performance. *Health Care Management Science*, 10(3):293–308, 2007.

[29] D. Sinreich and Y. Marmor. Ways to reduce patient turnaround time and improve service quality in emergency departments. *Journal of Health Organization and Management*, 19(2):88–105, 2005.

[30] D. Sinreich and Y.N. Marmor. A simple and intuitive simulation tool for analyzing emergency department operations. In *Proceedings of the 36th Conference on Winter Simulation*, WSC '04, pages 1994–2002. Winter Simulation Conference, 2004.

[31] É. Tardos. A strongly polynomial minimum cost circulation algorithm. *Combinatorica*, 5(3):247–255, July 1985.

[32] P.M. Vaidya. Speeding up linear programming using fast matrix multiplication. In *Proceedings of 30th IEEE Annual Symposium on Foundations of Computer Science*, pages 332–337, 1989.

[33] J. Wang, J. Li, and P.K. Howard. A system model of work flow in the patient room of hospital emergency department. *Health Care Management Science*, 16(4):341–351, 2013.

[34] K.D. Wayne. A polynomial combinatorial algorithm for generalized minimum cost flow. *Mathematics of Operations Research*, 27(3):445–459, 2002.

[35] J.-Y. Yeh and W.-S. Lin. Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department. *Expert Syst. Appl.*, 32(4):1073–1083, May 2007.