Distributed coordination of emergency medical service for angioplasty patients

ARTICLE *in* ANNALS OF MATHEMATICS AND ARTIFICIAL INTELLIGENCE · NOVEMBER 2015

Impact Factor: 0.69

3 AUTHORS:

Marin Lujak
King Juan Carlos University
26 PUBLICATIONS **65** CITATIONS

Holger Billhardt
King Juan Carlos University
73 PUBLICATIONS **610** CITATIONS

S. Ossowski
King Juan Carlos University
222 PUBLICATIONS **4,875** CITATIONS

manuscript No.
(will be inserted by the editor)

Distributed coordination of emergency medical service for angioplasty patients

Marin Lujak, Holger Billhardt, Sascha Ossowski

Received: date / Accepted: date

Abstract In this paper we study the coordination of Emergency Medical Service (EMS) for patients with acute myocardial infarction with ST-segment elevation (STEMI). This is a health problem with high associated mortality. A "golden standard" treatment for STEMI is angioplasty, which requires a catheterization lab and a highly qualified cardiology team. It should be performed as soon as possible since the delay to treatment worsens the patient's prognosis. The decrease of the delay is achieved by coordination of EMS, which is especially important in the case of multiple simultaneous patients. Nowadays, this process is based on the First-Come-First-Served (FCFS) principle and it heavily depends on human control and phone communication with high proneness to human error and delays. The objective is, therefore, to automate the EMS coordination while minimizing the time from symptom onset to reperfusion and thus to lower the mortality and morbidity resulting from this disease. In this paper, we present a multi-agent decision-support system for the distributed coordination of EMS focusing on urgent out-of-hospital STEMI patients awaiting angioplasty. The system is also applicable to emergency patients of any pathology needing pre-hospital acute medical care and urgent hospital treatment. The assignment of patients to ambulances and angioplasty-enabled hospitals with cardiology teams is performed via a three-level optimization model. At each level, we find a globally efficient solution by a modification of the distributed relaxation method for the assignment problem called the auction algorithm. The efficiency of the proposed model is demonstrated by simulation experiments.

Keywords EMS Coordination · Ambulance Coordination · Angioplasty · PCI · Distributed Optimization · Auction Algorithm · Emergency Medical Assistance

1 Introduction

Based on the World Health Organization data, ischemic heart disease (IHD) is the single most frequent cause of death killing 7.4 million people in 2012, which is 13.2% of all the deaths worldwide [43]. It is a disease characterized by ischaemia (reduced blood supply) of the heart muscle, usually due to coronary artery disease. At any stage of coronary artery disease, the acute rupture of an atheromatous plaque may lead to an acute myocardial infarction (AMI), also called a heart attack [45].

University Rey Juan Carlos, Madrid, Spain, E-mail: firstname.lastname@urjc.es

Determined by electrocardiographic findings, AMI can be classified into acute myocardial infarction with ST-segment elevations (STEMI) and without ST elevation (NSTEMI). Effective and rapid coronary reperfusion is the most important goal in the treatment of patients with STEMI. One of the reperfusion methods is angioplasty or primary percutaneous coronary intervention (PCI), which is a non-surgical procedure used to treat the stenotic coronary arteries of the heart. It is the preferred treatment when feasible and when performed within 90 minutes after the first medical contact [48].

Angioplasty should be performed by a highly specialized cardiology team in a *catheterization lab*. A catheterization lab is an examination room in a hospital or clinic with diagnostic imaging equipment used to visualize the arteries and the chambers of the heart and treat any stenosis or abnormality found. Since the equipment of the lab is very costly, it is only sporadically present in hospitals. Even more, due to high costs of a cardiology team necessary for the realization of angioplasty (usually consisting of a specialized interventional cardiologist or radiologist, a cardiac physiologist, a nurse and a radiographer), angioplasty-enabled hospitals are generally sparse.

The mortality of STEMI is influenced by many factors, among them: *patient delay time*, age, treatment, history of prior myocardial infarction, and the number of diseased coronary arteries. The patient delay time, defined as the period from the onset of STEMI symptoms to the provision of reperfusion therapy, is an important determinant of the effectiveness of angioplasty and the clinical outcome of STEMI patients [11, 40]. Due to insufficient EMS coordination and organizational issues, elevated patient delay time remains a major reason why angioplasty has not become the definitive treatment in many hospitals.

In this paper, we develop a decision-support system for the coordination of EMS focusing on urgent out-of-hospital STEMI patients awaiting angioplasty (angioplasty patients) but also applicable to emergency patients of any pathology needing pre-hospital acute medical care and urgent hospital treatment. We propose a change of the centralized hierarchy-oriented organizational structure to a patient-oriented distributed organizational structure of EMS that increases the flexibility, scalability, and the responsiveness of the EMS system. The proposed decision-support system is based on the integration and coordination of all the phases EMS participants go through in the process of emergency medical assistance (EMA). The model takes into consideration ambulances', patients', hospitals', and cardiology teams' real-time positions for real-time assignment of patients to the EMS resources. The objective of the proposed system is the reduction of patient delay times by distributed real-time optimization of decision-making processes.

In more detail, we mathematically model patient delay time and present a three-level problem decomposition for the minimization of combined arrival times of multiple EMS actors necessary for angioplasty. For the three decomposition levels, we propose a distributed EMS coordination approach and modify the auction algorithm proposed by Bertsekas in [5] for the specific case. The latter is a distributed relaxation method that finds an optimal solution to the assignment problem.

On the first level, agents representing ambulances find in a distributed way the patient assignment that minimizes arrival times of available ambulances to patients. After the treatment in situ, on the second optimization level, ambulances carrying patients are assigned to available hospitals. On the third level, arrival times of cardiology teams to hospitals are coordinated with the arrival times of patients.

The proposed approach is based on a global view, not concentrating only on minimizing single patient delay time, but obtaining the EMS system's best solution with respect to the (temporal and spatial) distribution of patients in a region of interest. Simulated emergency scenarios demonstrate the efficiency of the coordination procedure and a significant reduction in the average patient delay time.

This paper is organized as follows. Section 2 treats the State-of-the-Art in EMA coordination for emergency patients in general and angioplasty patients in particular. In Section 3, we introduce the delays related with the EMS coordination for angioplasty patients. The importance of reducing delays in angioplasty procedure is demonstrated on the case of the EMS provider of the Community of Madrid in Spain, SUMMA 112, in Section 4. In Section 5, we formulate the EMS coordination problem for simultaneous EMA of multiple angioplasty patients. Section 6 describes the proposed multi-agent architecture with the modified auction algorithm for the coordination of participants in EMA for angioplasty. Section 7 contains simulation results comparing the proposed coordination approach and the benchmark First-Come-First-Served (FCFS) principle used presently in most of the Western world countries. We draw conclusions and outline the directions for future work in Section 8.

2 State-of-the-Art in emergency medical assistance coordination

In the conventional EMA coordination approach, a medical emergency coordination center (ECC) is the only EMA coordinator. The region of interest is divided into geographically separated areas for which a team of human operators coordinates the ambulance fleet and hospitals for emergency patients' assistance. Such a fragmented EMS coordination approach often presents a suboptimal coordination solution due to the lack of the system's global view and high proneness to human error and delays. Furthermore, its centralized organizational structure is a bottleneck of the EMS, which, in case of contingencies, can cause the halt of the whole system.

Usually, the ECC has real-time information of the states of ambulances and hospitals. It dispatches the nearest available (idle) ambulance with Advanced Life Support (ALS) to the patient, thus applying the FCFS strategy. Even though this strategy is optimal in the case of a single patient awaiting EMA, in the case of multiple simultaneously appearing patients, it provides suboptimal results as it discriminates against patients appearing later [8].

When an ambulance arrives to the scene, it diagnoses AMI by an electrocardiogram, confirms the diagnosis to the ECC, provides in-situ assistance, and takes the patient to the hospital assigned by the ECC. The ECC applies the FCFS strategy for hospital assignment by assigning the patient to the nearest available hospital with a catheterization lab while the hospital alerts its closest cardiology team to the case. Furthermore, the ECC transfers patient identity and clinical data, treatment and expected arrival times to all the EMS participants involved and creates a record of the patient movement.

The right choice of the ambulance to be assigned to a pending out-of-hospital patient reduces total patient delay, which can significantly improve patients' chances and reduce morbidity and mortality caused by any emergency pathology including STEMI. Consequently, EMSs seek to minimize ambulance arrival times to patients and, if hospitalization is necessary, minimize the time of the patient's transport to an adequate nearby hospital.

Not all hospitals can treat angioplasty patients due to the lack of technical means or cardiology teams or both. Usually, each hospital with a catheterization lab has assigned to it its own cardiology team(s) located at alert outside hospital and obliged to reach the hospital in the case of emergency. The reason for their outside hospital location is the cardiology teams' cost that constitutes a large portion of the costs in surgical services [22]. Furthermore, significant cost savings can be obtained if cardiology teams work for multiple hospitals [12].

When a patient arrives to the hospital, the angioplasty procedure itself generally takes about 30 minutes but in specific cases it can last up to 2 and a half hours. The duration is dependent upon the technical difficulty of the case and the number of balloon catheters that have to be employed. The procedure's duration can be forecasted with prior angiography.

The most important factor in maximizing emergency patient's survivability is a timely and effective treatment. The EMS coordination process usually implemented in the ECCs is manual and the management is based on a case-by-case principle with high human workload necessary for telephone arrangements to find a solution. In the case of the presence of an increased number of angioplasty patients and different medical teams located in multiple sites, support for optimized EMS coordination based on information updated in real time is necessary for efficient angioplasty planning and scheduling.

There is a significant volume of research on EMS efficiency improvements related to ambulance coordination models for emergency patients e.g., [8, 20] and operating room planning and scheduling, e.g., [12]. The basis of the proposed methods is mostly queuing theory, simulations and mathematical programming, e.g., [12, 23, 39].

Domnori et al. in [18] discuss the suitability of agent-based applications to managing healthcare emergencies and large scale disasters and their application to problems where the main challenge is coordination and collaboration between participants. Furthermore, López et al. in [32] propose a multiagent system using an auction mechanism based on trust to coordinate ambulances for emergency medical services and in [31] present a multi-agent system with auctions, MASICTUS, with the aim of supporting the diagnosis of acute stroke diseases while coordinating ambulance services and expert neurologists for patient attendance. The auction mechanism here is based on three patient priority cases where the winning ambulance has the best expected arrival time and a good trust degree.

In [1], Bandara et al. study optimal dispatch of paramedic units to emergency calls to maximise patients' survivability. They show that the FCFS policy is not always optimal and that dispatching ambulances considering the priority of the call leads to an increase in the average survival probability of patients. However, they do not research possible efficiency improvements within the patients of the same priority.

In [8], we propose mechanisms that dynamically reduce patient delay through efficient assignment of ambulances to patients as well as the redeployment of available ambulances in the region of interest. We test these mechanisms in different experiments using historical data from SUMMA 112. The results empirically confirm that our proposal reduces significantly the EMA average response times. Moreover, in [33] we propose an organization-based multi-agent application for EMA based on the auction algorithm [5] where EMS participants are considered with different trust levels. The simulation results confirm further improvements in shortening EMA delays. Additionally, in [6], we present an abstract event-based architecture for fleet management systems that supports tailoring dynamic control regimes for coordinating fleet vehicles and illustrate it for the case of EMS management.

Furthermore, in [7], we present an approach that addresses the problem of timely automatic transmission of complete, real-time information about the current state of the ambulance fleet that is usually transmitted by the ambulance crew members. Due to the often stressful work of those professionals, the information is frequently not sent in a timely manner. We use a Complex Event Processing architecture (e.g., [15]) to automatically identify and transmit incidents and changes in the operational states of ambulances. As a result, the availability of information in the ECC and, thus, the effectiveness of the service is improved.

Availability of real-time information about EMS participants is the base for their efficient assignment to patients. The efficiency of the assignment can be achieved through various primal, dual, and primal-dual methods that have been proposed for obtaining an optimal solution of the assignment problem. The Hungarian method was the first such approach that solves assignment problem within time bounded by a polynomial expression of the number of agents [27]. Since it is centralized, it cannot be implemented in the systems that have intrinsically distributed information and computation. Therefore, a distributed version of the Hungarian method was proposed in [21]. Here, agents autonomously perform different substeps of the Hungarian algorithm based on their own information and the one received from other agents in the system. The algorithm computes a globally optimal solution in O(n^3) cumulative time (O(n^2) for each agent), with O(n^3) messages exchanged among n agents.

The auction algorithm proposed in [5] is another distributed relaxation method for the assignment problem that, based on an appropriate choice of bids, determines prices for objects and renders them more or less attractive for the agents to bid for. It is an iterative procedure that operates like an auction in which unassigned agents bid simultaneously for objects by raising their prices. The bidding is typically executed concurrently, where each agent calculates its bids simultaneously and independently of other agents considering only its own local information. Once all bids are collected, multiple bids are iteratively compared to determine the best offer for the system and the objects are awarded to the highest bidder. Even though it implicitly tries to solve a dual problem, the algorithm actually attains a dual approximate optimal solution complying with ϵ-complementary slackness [4]. Furthermore, it can be interpreted as a Jacobi-like iterative relaxation method for solving a dual problem. The Jacobi method is an algorithm for determining the solutions of a diagonally dominant system of linear equations. The worst case complexity of the auction algorithm with ϵ scaling is $O(NAlog(NC))$, where N is the number of agents, A, the number of mutually assignable pairs of agents and objects, and C is the maximum absolute object auction value. The algorithm is competitive with existing methods, and when executed in a distributed system or on a parallel machine, the algorithm exhibits substantial speedup [5]. For all these reasons and due to its economical market-related flavor, we use a modification of the auction algorithm for the assignment of EMS resources to angioplasty patients.

In spite of an exhaustive quantity of work on the optimization of EMA, to the best of our knowledge, there is little work on optimization models for the coordination of EMS for STEMI patients. This case is specific since it includes the coordination of the assignment of idle ambulances to patients, assignment of catheterization laboratories in available hospitals to diagnosed STEMI patients, and the assignment of available cardiology teams to hospitals for the angioplasty procedure. All of the three assignments need to be combined to guarantee good arrival times for all the simultaneous patients awaiting the treatment. This EMS coordination problem is somewhat similar to the case described in [34] where we shortly describe a distributed multi-agent coordination model for after-hours urgent surgery patients that cannot be covered by the in-house surgery teams. Therefore, in this paper, we extend the model presented in [34] and adapt it to the coordination of EMS for angioplasty. Furthermore, we propose a three-level distributed optimization approach for this case.

3 Angioplasty and patient delay

STEMI is diagnosed based on the patient's history of severe chest pain lasting for 20 minutes or more, not responding to nitroglycerine and having an evident ST-segment elevation in the electrocardiogram [48]. It is treated as an emergency where an immediate goal is to open blocked arteries and reperfuse the heart muscles as soon as possible either through urgent coronary angiography and primary angioplasty or with thrombolysis, whichever is available. When the formerly mentioned therapies are unsuccessful, it is treated with bypass surgery.

Delays in treating STEMI increase the likelihood and amount of cardiac muscle damage due to localised hypoxia, which is why the recommended delay between a patient's arrival at the hospital and the time he/she receives angioplasty (also called *door-to-balloon time*) is no more than 90 minutes [44]. Also, the delay between the first medical contact and reperfusion therapy is referred to as *system delay*. It is a good service quality indicator and patient outcome predictor [47]. Regarding the connection between the system delay and mortality in the short, medium and long run, Terkelsen et al. in [47] performed a historical follow-up study of Danish medical registries of STEMI patients, who were transported by the EMS and treated with angioplasty from January 1, 2002 to December 31, 2008, at 3 high-volume angioplasty centers in Western Denmark. A total of 6209 patients underwent primary angioplasty within 12 hours of system delay. The mortality rates were as follows: a system delay of 0 through 60 minutes ($n = 347$) corresponded to a long-term mortality rate of $15.4\%(n = 43)$; a delay of 61 through 120 minutes (n=2643) to a rate of 23.3% ($n = 380$); a delay of 121 through 180 minutes (n=2092) to a rate of 28.1% ($n = 378$); and a delay of 181 through 360 minutes (n=1127) to a rate of 30.8% ($n = 275$) (with the probability of type II error $P_\beta.001$). Moreover, *patient delay time* should be reduced as much as possible by active hospital-EMS coordination since each 30 minutes of the delay increases the relative risk of 1-year mortality by 7.5% [16]. By reducing the patient delay time, we minimize the extent of heart muscle damage and preserve the pumping function of the heart. The result is lower mortality and the invalidity resulting from STEMI [45].

Different countries have legally established upper bounds on the maximal allowed patient and/or system delay time. To take a concrete example, the American College of Cardiology (ACC) and American Heart Association (AHA) recommend that the system delay for patients with STEMI should be less than 90 minutes [25, 46]. The European Society of Cardiology in [45] for angioplasty indicates a goal patient delay of 90 min while in high-risk cases with large anterior infarcts it is 60 min. Anyhow, the prognosis of STEMI patients varies greatly depending on a person's health, the extent of the heart damage influenced by the patient delay time, the treatment given, but also based on the meteorological factors and the time of appearance. It was noticed that the incidence of non-fatal STEMI is approximately 40% higher from 06:00 – 12:00 AM as compared to other times of the day [29]. Season of the year, daily atmospheric temperature, pressure, and relative humidity are meteorological factors that also affect the number of STEMI deaths per month. Many geographically dispersed studies confirm this claim, e.g., Greece [17], Belgium [36], US and Canada [10], Italy [35], Hungary [26], and Korea [28]. Temperature influences sympathetic tone, blood pressure, and blood platelet functioning [10, 24, 29, 30, 36].

Morbidity and mortality from STEMI have decreased over the years due to better treatment including greater use of angioplasty. However, more than 25 % of patients with STEMI die before receiving medical care, most often from ventricular fibrillation. To limit infarct size and to prevent infarct extension and expansion in patients with STEMI, the main issue is to coordinate efficiently EMS and initiate reperfusion therapy as soon as possible.

4 Case-study: SUMMA 112, EMS provider in the Community of Madrid, Spain

We demonstrate the criticality of the problem in the case of SUMMA 112, the emergency medical service provider in the Community of Madrid in Spain covering a population of more than 6.4 million.

The strategy that SUMMA 112 uses for the coordination of emergency AMI patient assistance is the following. When SUMMA makes the first contact with a patient, in the case of fibrinolisis and/or shock, the patient is given the highest priority and is directed towards primary angioplasty coordinating in the least time possible the EMS participants necessary for EMA. Otherwise, i.e., if there was no contraindication to fibrinolisis and/ or shock, then

- if the time from symptom onset was from 12 to 24 hours and if the patient still shows symptoms or instability, or if the time from symptom onset is from two to four hours, he/she is given the highest priority (priority zero) and is directed to the primary angioplasty;
- if the time from symptom onset is less than two hours, or it is from two to four hours, but the expected patient delay time is more than two hours, then the patient is directed towards fibrinolysis. After that, he/she is assigned priority 1 and is transferred to immediate rescue angioplasty. If reperfusion injury has occurred, then the patient is transferred to immediate angioplasty after thrombolysis.

The average time from symptom onset to contacting the emergency services in SUMMA 112 with procedures and strategies described previously on average is 70 min; SUMMA 112 attends on average in 10 min and stabilizes the patient, diagnoses ST-Elevation Acute Coronary Syndrome, and administers fibrinolysis on average in 25 min; following fibrinolysis, it takes on average 38.5 min to reach the hospital [2].

We analyzed SUMMA 112's patient data for the year 2009. Out of 33 hospitals belonging to the public health service of Madrid, 9 of them had catheterization laboratories and human cardiology teams capable of performing the angioplasty procedure. SUMMA 112 disposes of 36 ambulances with Advanced Life Support (ALS) whose (fixed) base stations are located at the 33 hospitals of the Community. Angioplasty patients share EMS resources with the rest of emergency patients. Moreover, cardiology teams are present in the hospitals during work hours and available from home after-hours; when at home, they assist patients only if called by the hospital [41].

The day with the highest number of emergency patients in Madrid in 2009 was January 21. There were 221 most sever emergency patients, 40 of whom were simultaneously waiting for ambulance assignment in the system with at least one more emergency patient [8]. Even though we do not dispose of the data on the pathology of these patients, in the Community of Madrid in 2009, cardiovascular diseases caused 11,453 deaths, representing 27.7% of all deaths (41,268) of the Community in that year [38]. Furthermore, 2146 people died of AMI, to which must be added a majority of the 1,415 deaths assigned to the diagnosis of cardiac arrest, death without assistance and unknown cause deaths [38].

Moreover, in [37] the clinical and angiography data in the database of the Hospital Puerta de Hierro in Madrid from January 2005 to October 2007 was analyzed. The hospital covers a local population of 635,495 inhabitants generally concentrated in the urban area nearby and well connected by road. The goal door-to-balloon time in this study performed over 389 patients was 60 min or less. 84.7% of STEMI patients were treated with angioplasty with a median door-to-balloon time of 79 (53-104) min, which was 30 (60-90) min lower when the EMS gave prior warning to the hospital ($p < 0.01$). Reducing this time was especially relevant in the cases seen outside working hours. Although the first-medical-contact-to-balloon time is not mentioned, presumably it was rather long for the patients that were transferred to hospital since the beginning-of-symptoms-to-balloon median time was 235 (percentiles 25-75, 170-335) min. However, patients who came to the emergency department by their own means had the longest door-to-balloon (100 min, $p < 0.01$).

In addition, 32% of STEMI patients in Madrid do not receive reperfusion therapy and for patients who do, the delay times are greater than those recommended in clinical practice guidelines. Therefore, the high frequency of this pathology, together with a greater probability of occurrence during morning hours and on days with bad climate conditions, leads to increased patient delay times. This calls for an improved approach to the coordination of simultaneous EMA of multiple patients assigned for angioplasty.

5 Problem formulation

In this paper, we consider the problem of the dynamic real-time assignment of multiple simultaneously appearing urgent out-of-hospital angioplasty patients to ambulances and consequently to angioplasty-enabled hospitals with out-of-hospital cardiology teams. The objective is to minimize overall patient delay to angioplasty while respecting as much as possible the limits on allowed maximal individual patient delay.

From the system's efficiency and fairness point of view, EMS coordination should result in as high utilitarian social welfare as possible while respecting the requirements on maximal individual patient delay. A high utilitarian social welfare means a significant reduction of patient delay for most of the patients, with (usually a few) worst off patients (see, e.g., [42]). By introducing the constraint on maximum patient delay, the worst off patients are assigned acceptable delay times. If, on the other hand, we decided to optimize the EMA worst-off behavior or egalitarian welfare, we would deteriorate the system efficiency and thus the utilitarian welfare, which in the EMS coordination case would mean higher patient delay for most of the simultaneous patients (see, e.g., [13]).

Hesitation of patients to search for medical help sometimes might average several hours and thus can prevent the early application of life-saving procedures and contribute substantially to a diminished effectiveness of treatment. Since we cannot influence this hesitation time, in the development of the coordination model, we concentrate on the minimization of the expected patient delay intended as the time elapsed from the moment a patient contacts the ECC to the moment (s)he is treated with angioplasty in the hospital.

The expected patient delay defined in this way is made of the following parts, Figure 1:

T1 Emergency call response and decision-making for the assignment of EMS resources;
T2 Mobilization of an idle ambulance and its transit from its momentary position to the patient's out-of-hospital position;
T3 Patient's treatment in-situ by ambulance staff;
T4 Patient's transport in the ambulance to an assigned hospital;
T5 Cardiology team's transport from their momentary out-of-hospital location to the hospital;
T6 Expected waiting time due to previous patients in the catheterization lab (if any).

Considering the aforementioned EMS tasks' delays, four distinct agent sets have been identified. Let P be a pending patient set. Let C be a set of cardiology teams and $C_{av} \subset C$ the subset of available cardiology teams. Furthermore, let A be the set of identical, capacitated ambulances to be scheduled to assist patients based on one-to-one assignment and $A_{av} \subset A$ the subset of available such ambulances. Moreover, let H be a set of hospitals with catheterization lab and $H_{av} \subset H$ the subset of available such hospitals.

Availability of every catheterization lab in each hospital $h \in H$ depends on previous patients (if any) booked for that catheterization lab with higher or the same urgency as the patient in question. Therefore, let $\rho_{h,p}$ represent expected delay times of free time windows of catheterization lab(s) of hospital h for patient p. Moreover, all agent sets are represented

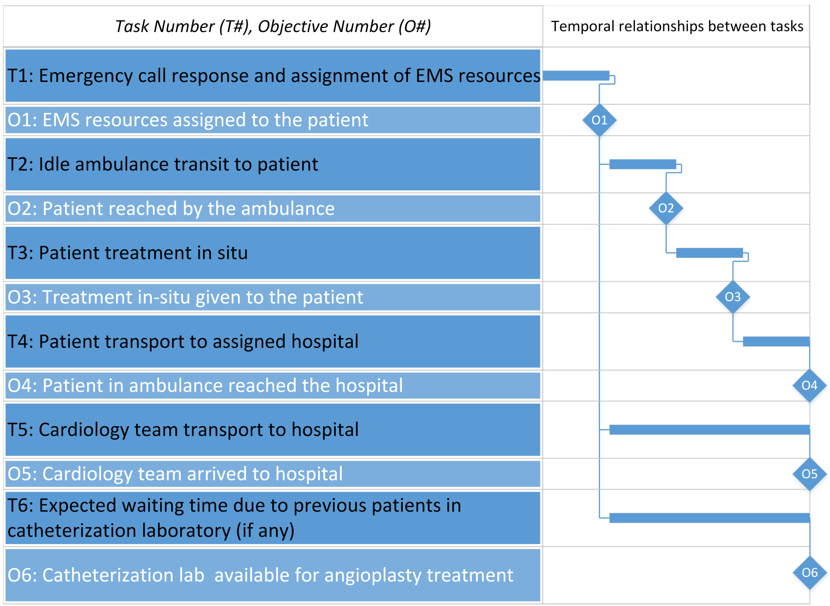

Fig. 1 Diagram showing temporal sequence and relation of six medical emergency tasks (T) and objectives (O) necessary in coordinating EMS for a patient in the need of angioplasty treatment

by points in the plane. The abbreviation pos_a is used for the position of any kind of agent a while $t(x, y)$ is the expected arrival time from position x to position y.

In the case there is only one pending patient in the system, the problem is to find ambulance $a \in A_{av}$, cardiology team $c \in C_{av}$ and hospital $h \in H_{av}$ that in combination minimize patient's delay time, $\varDelta t_p$:

$$\min \varDelta t_p = \min_{a \in A_{av}} t(a,p) + t(p) + \min_{h \in H_{av}} \left(\max \left(t(p,h), \min \rho_{h,p}, \min_{c \in C_{av}} t(c,h) \right) \right) \tag{1}$$

subject to:

$$\varDelta t_p \leq t_p^{max} \ , \forall p \in P \ , \tag{2}$$

where $t(p)$ is the expected in-situ patient assistance duration, assumed to depend on the patient pathology but not on the assigned ambulance. It can be estimated based on the initial ECC communication with the patient with reasonable precision. Furthermore, t_p^{max} is the maximal allowed patient delay time for patient $p \in P$ depending on the patient's pathology, while $\min \rho_{h,p}$ represents the expected shortest delay time until hospital h will be free for patient p. Additionally, the expected patient's $t(p, h)$ and cardiology team's arrival times to hospital $t(c, h)$ are considered as the delay times until the arrival to the catheterization lab in the hospital. Then, the objective for each patient $p \in P$ is to choose a triple $\langle a \in A_{av}, h \in H_{av}, c \in C_{av} \rangle$ optimizing (1) subject to (2). Hospital h_p chosen for patient $p \in P$ is thus

$$h_p = \arg \min_{h \in H_{av}} \left(\max \left(t(p,h), \min \rho_{h,p}, \min_{c \in C_{av}} t(c,h) \right) \right) \ . \tag{3}$$

Therefore, the optimal patient delay time for a single patient is the lowest among the highest values of the following three times for all available ambulances and angioplasty-enabled hospitals, Figure 1:

- *the expected patient delay time to hospital* (the sum of times T2, T3, and T4),
- *the expected minimal arrival time* among cardiology teams to the same hospital (T5), and
- *the expected shortest waiting time* until hospital h will be free for patient p, $\min \rho_{h,p}$, (T6).

For simplicity, we let $t_{php} = \max_{h \in H_{av}}(t(p,h), \min \rho_{h,p})$ for all patients $p \in P$. Then, from the global point of view, considering all pending out-of-hospital patients, the problem transforms into:

$$\min \Delta t_P = \sum_{p \in P} \Delta t_p = \sum_{p \in P} t(a,p) + \sum_{p \in P} t(p) + \sum_{p \in P} \left(\max_{h \in H_{av}} \left(t_{php}, \min_{c \in C_{av}} t(c,h) \right) \right) \quad (4)$$

subject to

$$\Delta t_p \leq t_p^{max}, \forall p \in P. \quad (5)$$

The overall patient delay time Δt_P in (4) is an additive function. Since the minimum arrival times cannot be always guaranteed for all patients due to the limited number of EMS resources, a sum of the EMA tasks' durations should be minimized for each patient individually and for the system globally considering individual constraints. This gives an underlying linear programming structure to the EMS coordination problem. Therefore, it is possible to guarantee optimal outcomes even when the optimization is performed separately on individual sum components, i.e., when ambulance assignments are negotiated separately from the hospital and cardiology team assignment, e.g., [9, 14, 19]. This fact significantly facilitates the multi-agent system's distribution and enables a multi-level optimization.

Hence, we decompose (4) as follows. On the first level, we assign ambulances to patients such that the expected arrival time of ambulances to patients $t(a,p)$ is minimized. Note that since $t(p)$ in (4) is a constant for every patient p depending only on the patient's pathology and not on the assigned ambulance, we can exclude it from the optimization. We have:

$$\min \sum_{p \in P} \sum_{a \in A_{av}} t(a,p) x_{ap} \quad (6)$$

subject to:

$$\sum_{a \in A_{av}} x_{ap} = 1, \forall p \in P \quad (7)$$

$$\sum_{p \in P} x_{ap} \leq 1, \forall a \in A_{av} \quad (8)$$

$$\sum_{a \in A_{av}} t(a,p) x_{ap} \leq t_{ap}^{max}, \forall p \in P \quad (9)$$

$$x_{ap} \in \{0,1\}, \forall p \in P, a \in A_{av}. \quad (10)$$

where (7) and (8) are constraints on the one-on-one assignment of patients to ambulances assuming that the number of available ambulances is larger than or equal to the number of patients, i.e., $|P| \leq |A_{av}|$. Moreover, (10) defines a value of 0 or 1 for binary decision variable x_{ap}. All of the mentioned constraints are hard constraints, while (9) is a soft constraint on the allowed maximal patient waiting time for ambulance t_{ap}^{max} recommended by legal requirements.

Then, on the second optimization level, we approach the second part of (4):

$$\min \sum_{p \in P} \left(\max_{h \in H_{av}} \left(t_{php}, \min_{c \in C_{av}} t(c, h) \right) \right) \tag{11}$$

which is an NP-hard combinatorial problem. However, by approximating (11) with a sequence of problems where we first decide on the assignment of hospitals to pending patients and then assign cardiologists to patients already assigned to hospitals, we obtain two linear programs to which we can apply tractable optimal solution approaches as is the auction algorithm [5]. Moreover, if we introduce constraints on the expected arrival times of individual EMS actors with respect to maximal allowed patient delay time in this approximation process and gradual relaxation of the restrictions in case of a nonexistent feasible solution, we can achieve efficient and computationally fast solutions. While we explain the relaxation approach in more detail in Section 6.2, in continuation we present the approximation of (11) by the following two sequential problems. On the second level, we first optimize the arrival of patients to hospitals, i.e.,

$$\min \sum_{p \in P} \sum_{h \in H_{av}} t_{php} x_{ph} \tag{12}$$

subject to:

$$\sum_{h \in H_{av}} x_{ph} = 1, \forall p \in P \tag{13}$$

$$\sum_{p \in P} x_{ph} \leq 1, \forall h \in H_{av} \tag{14}$$

$$\sum_{h \in H_{av}} t(php) x_{ph} \leq t_p^{max}, \forall p \in P \tag{15}$$

$$x_{ph} \in \{0, 1\}, \forall p \in P, h \in H_{av}. \tag{16}$$

where (13) and (14) are hard constraints on the one-on-one assignment of patients to hospitals assuming that the number of available hospitals is larger than or equal to the number of patients, i.e., $|P| \leq |H_{av}|$. (15) is a soft constraint for overall patient delay time to hospital constrained by the maximal allowed patient delay time t_p^{max}. Moreover, (16) is a hard constraint on binary decision variable x_{ph}.

Then, the optimization problem on the third level is:

$$\min \sum_{h \in H_{as}} \sum_{c \in C_{av}} \left(t(c, h) - t_{php} \right) x_{ch} \tag{17}$$

subject to:

$$\sum_{h \in H_{as}} x_{ch} \leq 1, \forall c \in C_{av} \tag{18}$$

$$\sum_{c \in C_{av}} x_{ch} = 1, \forall h \in H_{as} \tag{19}$$

$$x_{ch} \geq 0, \tag{20}$$

where (18) ensures that no cardiology team is allocated more than once and (19) assigns one cardiology team to each hospital in H_{as}. H_{as} is a solution set of hospitals assigned in (12)–(16) while (20) is a constraint on nonnegativity of decision variable x_{ch}.

6 Proposed EMS distributed multi-agent coordination model

The EMS system can be seen as an interconnected geographically distributed system where EMS participants are capable of processing local real-time information independently while interacting in a cooperative way. Moreover, the EMS participants' interactions are mostly local in nature, thus leading to the setting where most of the participants usually have intense interaction within some range based on the geographical position, but almost no interaction outside this range. This is true both for ambulances as for the hospitals with out-of-hospital cardiology teams. In this context, distributing the computation among ambulances and hospitals while balancing their communication load can increase the overall throughput, robustness, and flexibility of the system. Additionally, in such a distributed setting, the ECC is unnecessary for coordination.

For the aforementioned reasons, we propose a dynamic and distributed EMS resource assignment model for angioplasty patients applicable to all emergency out-of-hospital patients whose time to hospital treatment is of the utmost priority. The key of our proposed model lies in the distribution of the EMS related decisions to allow for as high autonomy as possible regarding local decisions.

The proposed solution is founded on the collaborative multi-agent system organization with four classes of agents described in Section 6.1. In the proposed model, patient assignment decisions are performed on three levels via the proposed modification of the auction algorithm presented in Section 6.2. On the first optimization level, arrival times of ambulances to out-of-hospital patients are optimized for all emergency patients who arrive at the hospital by ambulance. On the second level, the optimization of arrival times of out-of-hospital patients to hospitals considering the hospital availability is performed for all patients assigned for angioplasty. Moreover, on the third level, out-of-hospital cardiology teams are assigned to the hospitals that have been assigned to patients on the second level.

6.1 EMS multi-agent system

All agents in the proposed multi-agent system are autonomous and independent decision makers that feature a determined sequence of steps and message exchanges in order to resolve each emergency case as described in the following.

- **Medical emergency-coordination center (ECC):** ECC receives emergency calls from patients and informs the rest of the EMS system of the new pending patients. In the case of centralized EMS organization, it assigns an ambulance and an angioplasty-enabled hospital for each case, thus performing the high-level management of the STEMI EMA procedure. In the distributed case, it is not necessary for the coordination since ambulances and hospitals assign patients on their own in a distributed way and mutually communicate and coordinate their actions if necessary.

- **Patient:** Each pending patient agent $p \in P$ represents a real out-of-hospital (angioplasty) emergency patient requiring ambulance EMA and angioplasty. After calling ECC from his/her out-of-hospital location, (s)he gets assisted in-situ by an ambulance crew and, if necessary, gets transferred to a hospital where (s)he receives angioplasty. Moreover, new patient requests continuously unfold through time and must be assigned in real time to ambulances such that the patients get assisted with as low delay as possible while respecting the maximal individual patient delay t_p^{max}. Each patient $p \in P$ is described as follows:
$$p = \{pos_p, t_p^{in}, t_p^{max}\}, \tag{21}$$
where pos_p is the patient's position and t_p^{in} is the patient's detection time defined as the time when the ECC is informed about the incident.
- **Ambulance:** Ambulance agent $a \in A$ represents an ambulance with ALS, together with the human crew assigned to it. The task of each ambulance is, once it has arrived to a patient, to assist him/her in-situ and, if necessary, to transfer him/her to a hospital in the shortest time possible. Furthermore, ambulances communicate to the ECC to obtain information on patients awaiting ambulances, to hospitals for patient transfer, and to an assigned patient for his/her first medical treatment in-situ and transport to hospital. Each ambulance $a \in A$ is described as follows:
$$a = \{pos_a, S_a\}, \tag{22}$$
where pos_a indicates position and S_a the ambulance's status that defines subsets of ambulances A and can be: idle A_i, moving to incident position A_{mip}, in-situ assistance A_{isa}, and ambulances moving to a hospital A_{mth}. Moreover, the set of available ambulances is denoted by $A_{av} = A_i \cup A_{mip}$. Available ambulances $a \in A_{av}$ are considered for the assignment to pending patients. Furthermore, each ambulance route originates from its momentary location to the patient's location and terminates at the assigned hospital if necessary. We assume that an ambulance is committed to a given patient only when it gets to the location of the patient. After arriving at the patient's location, the ambulance cannot be redirected elsewhere until transferring the patient to the hospital if necessary. When the patient is transferred to the hospital, the ambulance returns to its base hospital where it waits for a next patient assignment.
- **Angioplasty-enabled hospital:** Hospital agents $h \in H$ represent angioplasty-enabled hospitals within the region of interest. They are responsible of managing their catheterization labs while they coordinate with cardiology team(s) and ambulances' agents to assist patients. Each hospital $h \in H$ is described as:
$$h = \{pos_h,\ \rho_{h,p},\ S_{h,p}\}, \tag{23}$$
where pos_h is the position and $S_{h,p}$ is the temporal availability of a hospital $h \in H$ for a patient $p \in P$ after assisting patients who are booked before and whose severity of the pathology is equal to or worse than the one of the patient in question. The availability states can be either available av or unavailable un. Each hospital has at its disposal also updated information on cardiology team availability $\rho_{c,h}$ and incoming patients.
- **Cardiology team:** $c \in C$ is responsible of the performance of angioplasty. Each cardiology team $c \in C$ is described as:
$$c = \{pos_c,\ \rho_{c,h},\ S_{c,h}\}, \tag{24}$$
where pos_c is the cardiology team's position and $\rho_{c,h}$ is its availability for hospital $h \in H$ that can be either available av or unavailable un. Moreover, the team's status

S_c defines its state that can be: idle C_i, moving to an assigned hospital C_{mh}, or in the assigned hospital C_{ih}. Then, the set of available cardiology teams C_{av} comprises idle C_i and the cardiology teams moving to an assigned hospital C_{mh}, i.e., $C_{av} = C_i \cup C_{mh}$. Additionally, it is assumed that the cardiology team's members are positioned at different locations out of hospital and move to hospital when needed. Their combined expected arrival time to hospital is the maximum of the individual arrival times.

Besides the aforementioned agents, we assume the existence of additional services accessible on the Internet that may provide patients' medical information (e.g., health records). Such services may be provided and consulted by different medical professionals involved in the process in order to obtain medical data of an angioplasty patient.

6.2 Proposed EMS coordination algorithm

In this Section, we present the EMS distributed coordination approach that assigns ambulances and hospitals with cardiology teams to angioplasty patients and, therefore, solves problem (4)-(5). The solution approach is presented in Algorithms 1–3.

In Algorithm 1, the overall decision for the assignment of each patient to the triple made of ambulance, hospital, and cardiology team is divided into the assignment of patients to ambulances on the first level, where $Resource\ Assignment(P, A_{av})$ resolves (6)–(10), the assignment of patients to hospitals on the second level, where $Resource\ Assignment(P, H_{av})$ computes (12)–(16), and the assignment of hospitals with allocated patients from the second level to cardiology teams on the third level at which $Resource\ Assignment(H_p^{as}, C_{av})$ finds the solution to (17)–(20), (see Figure 2).

Algorithm 1: Coordination of EMS resources for urgent angioplasty assistance

```
Input: $P; A_{av}; H_{av}; C_{av}$
Output: $\{p, a_p : a_p \in A_{av}, h_p : h_p \in H_{av}, c_p : c_p \in C_{av}\}, \forall p \in P$
initialization: $AS \leftarrow \emptyset$
$\{p, a_p\} \leftarrow ResourceAssignment(P, A_{av}), \forall p \in P$
$AS \leftarrow AS \bigcup \{p, a_p\}, \forall p \in P$
$\{p, h_p^{as}\} \leftarrow ResourceAssignment(P, H_{av}), \forall p \in P$
$AS \leftarrow AS \bigcup \{p, h_p^{as}\}, \forall p \in P$
$\{h_p^{as}, c_p\} \leftarrow ResourceAssignment(H_p^{as}, C_{av}), \forall h_p^{as} \in H_p^{as}$
$AS \leftarrow AS \bigcup \{h_p^{as}, c_p\}, \forall h_p^{as} \in H_p^{as}$
return $AS$
```

For the distributed assignment of EMS resources to patients, at every decision-making level, each agent included in the assignment of that level independently decides its assignment based on its local information and the communication only with other relevant agents on the same level, as can be seen in Algorithms 2–4.

Algorithm 2 is a modified version of Bertsekas' auction algorithm [5]. The reason for the application of the auction algorithm is its intuitive economical flavor, low computational complexity, and the distribution of the decision-making procedure which is performed between a set of bidder agents $b \in B$, and a set of bid object agents $o \in O$. In the algorithm,

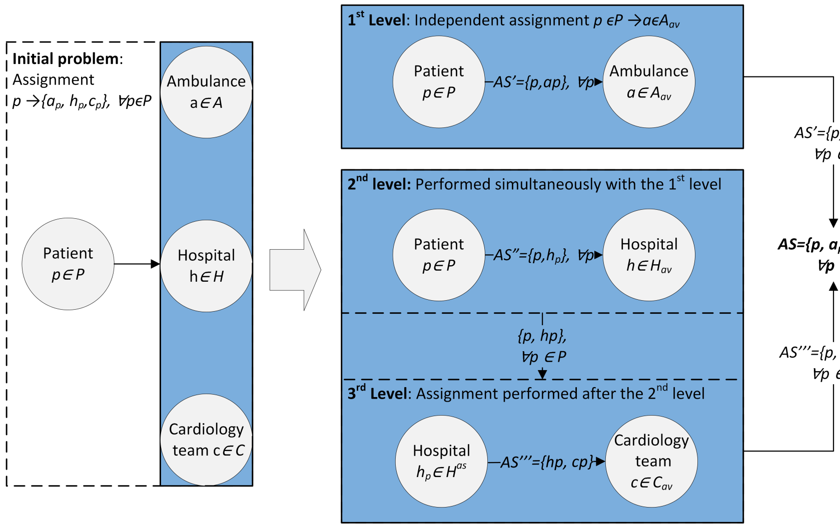

Fig. 2 Proposed three-level decomposition of the problem of EMS coordination for urgent out-of-hospital STEMI patients awaiting angioplasty

it is important that the number of bidders $|B|$ is equal to or less than the number of bid objects $|O|$, i.e., $|B| \leq |O|$. Moreover, the assignment of bidders to bid objects is performed in iterations that are composed of a bidding (Algorithm 3) and an allocation phase (Algorithm 4). In Figure 3, we present the distributed functioning of the modified auction algorithm.

In more detail, each bidder agent $b \in B$ keeps in its memory the value $price_o^b(k)$, i.e., its most recent knowledge about the actual price of each object $o \in O$, and the set AS'_b of its most recent knowledge about all the bidders' assignments at the respective optimization level (Algorithm 3). On the other hand, each bid object keeps in its memory the bids received from bidders in the last two iterations $Bids_o(k-1)$ and $Bids_o(k)$ and the set AS'_o of its most recent knowledge about all the bidders' assignments at the respective optimization level. None of the aforementioned local copies of variables have to coincide with the actual values; they may also differ from one agent to another due to the dynamics of their previous communication and local interaction with the environment.

At each time period, all three levels of Algorithm 1 are performed by executing algorithm 2. The first and the second level can be performed in parallel, while the third level is performed after the second since the results of the second are the input to the third. Therefore, when pending patients get assigned to hospitals, they are assigned to cardiology teams.

Similarly, in each iteration of Algorithm 2, the allocation phase is performed sequentially after the bidding phase.

Each agent keeps a local copy of Algorithms 3 and 4, and, depending on its role, performs the bidding or the allocation phase of the auction algorithm. In every phase, all participating agents perform their individual computations in parallel (computation of a bid by all unassigned bidders on the bidding phase and the computation of the object allocation by bid objects at the allocation phase). After updating the local information, at the bidding phase, each unassigned bidder independently finds the best bid object (the object with the highest value) given the actual auction prices, and issues its bid to that object. This stage is over when all unassigned bidders issue their respective bids. Furthermore, at the allocation phase

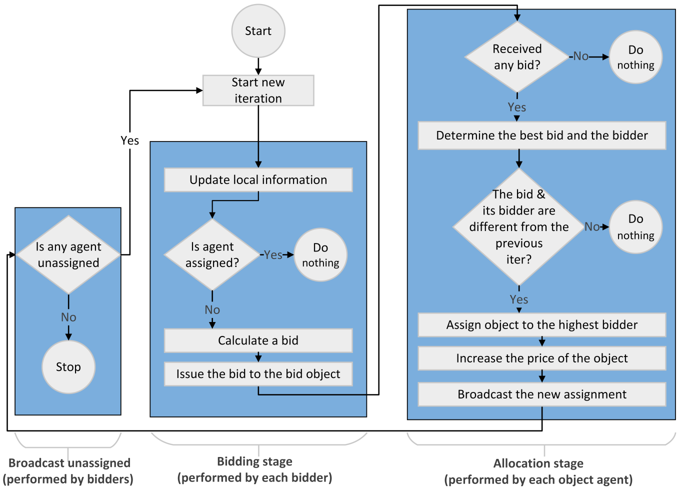

Fig. 3 Proposed modification of the auction algorithm for distributed EMS coordination

(Algorithm 4), each bid object with at least one received bid determines if the highest received bid is higher than the one in the previous iteration and if the present bidder is different from the winning one in the previous iteration (if any). If this is the case, the object cancels the previous assignment communicating it to the previous assigned bidder (if any), assigns itself to the present bidder with the highest bid, increments its price, and broadcasts its new assignment. Otherwise, it does not update its assignment and remains with the previously assigned price and bidder (if any).

If at the end of the allocation phase, a bid object received more than one bid and/or there was a different bidder previously assigned to the object, then there is still at least one unassigned bidder present. The bid object broadcasts the message that unassigned bidding agents are still present, and a new iteration begins. The unassigned bidding agents update the objects' prices and the algorithm continues in iterations until all bidders are assigned and all conflicting bids are resolved.

Algorithm 2: Resource Assignment(b,o) based on the auction algorithm

```
Input: $B, O$
Output: $\{b, o_b : o_b \in O\}, \forall b \in B$
initialization: $k \leftarrow 1$
repeat
    $k \leftarrow k + 1$
    for each $b \in B$ do
        $BiddingPhase(b, o)$
    for each $o \in O$ do
        $AllocationPhase(b, o)$
until $\{b, o_b\} \in AS', \forall b \in B$;
return $AS$'
```

Algorithm 3: BiddingPhase(b,o) executed by each bidder agent $b \in B$

```
Input: $b, O, k$
Output: $\{b, bid_{o,b}(k)\}$
Initialization
if $k = 1$ then
    $AS'^{b} \leftarrow \emptyset, bid_{o,b}(k) \leftarrow \emptyset$
    for each $o \in O$ do
        $price_o^b \leftarrow \emptyset$
if $b \notin AS'^{b}$ then
    Local information update
    Broadcast $AS'^{b}$ and $price_o^b(k), \forall o \in O$
    $Received \leftarrow Receive\{AS'^{b'}, price_o^{b'}(k) | b' \neq b \wedge \forall o \in O\}$
    $price_o^b(k) \leftarrow \max_{b \in \{Received \bigcup local\}} \{price_o^b, \forall o \in O\}$
    $AS'^{b} \leftarrow \arg\max_{b \in \{Received \bigcup local\}} \{price_o^b(k), \forall o \in O\}$
    Find objects that offer the least $c_o$ and the second least cost $c_{o'}$
    $o = \arg\min_{o \in O} \{t(o,b) + price_o^b(k)\}$
    $c_o = \min_{o \in O} \{t(o,b) + price_o^b(k)\}$
    $c_{o'} = \min_{o' \in O \wedge o' \neq o} \{t(o',b) + price_{o'}^b(k)\}$
    Calculate bid for the object $o$ with the least cost
    $bid_{o,b} = c_o - c_{o'} + \epsilon$
    Issue bid $bid_{o,b}$ to object $o$
    return $\{b, bid_{o,b}(k)\}$
```

If we assume that the cardinality of the patient set $|P|$ is smaller than the one of ambulances, i.e., $|P| \leq |A_{av}|$ and hospitals, $|P| \leq |H_{av}|$, then pending patients bid for ambulances at the first level, and for hospitals at the second level. Furthermore, if the cardinality of the set of the available cardiology teams is smaller than the set of hospitals assigned to patients at the second level, i.e., $|C_{av}| \leq |H_{as}|$, then on the third level, cardiology teams bid for the hospitals $h_p^{as} \in H^{as}$. The particularity of the third optimization level, which resolves problem (17)-(20) is the individual cardiology team's bidding cost that, to minimize the team's tardiness with respect to the patient's arrival to hospital, does not include the expected arrival time from the bidder to the object $t(o,b)$, as is the case for the first two levels, but the difference between the arrival times of cardiology teams and patients to hospitals assigned on the second level, i.e., for each hospital $h_p^{as} \in H_{as}$ assigned to patient $p \in P$, the bid in the bidding phase of Algorithm 2 is calculated as:

$$c_h = \min_{h \in H_{as}} \Big(t(c,h) - t_{php} \Big) + price_h(k), \forall c \in C_{av}. \tag{25}$$

Here, if a cardiology team is assignable to more hospitals, preferably all, then the optimization over multiple arrival times gives a globally optimal solution. Oppositely, if each cardiology team can work only at one or a subset of hospitals, the expected arrival time might be significantly longer, which may jeopardize patients' outcomes.

A global termination condition of Algorithm 2 is that all patients are assigned to EMS resources based on a one-to-one assignment. It is decomposed into a collection of local termination conditions, one for each bidder. Each bidder monitors its own computations and

Algorithm 4: AllocationPhase(b,o) executed by each object agent $o \in O$

```
Input: $B, o, k$
Output: $\{AS'^{o}, price_{o'}(k)\}$
initialization
if $k = 1$ then
    $AS'^{o} \leftarrow \emptyset, bid_o^{max} \leftarrow \emptyset, b_o^{max} \leftarrow \emptyset$
$Bids_o(k) \leftarrow \emptyset$
Receive bids from the bidders and save them
$Bids_o(k) \leftarrow Bids_o(k) \cup \{bid_{o,b}\},\ \forall b \in B_o$
if $Bids_o(k) \neq \emptyset$ then
    Determine the highest bid
    $bid_o^{max}(k) \leftarrow \max\{bid \in Bids_o(k)\}$
    Determine the corresponding highest bidder
    $b_o^{max}(k) \leftarrow \arg\max\{bid \in Bids_o(k)\}$
    if $(bid_o^{max}(k) \geq bid_o^{max}(k-1)) \wedge (b_o^{max}(k) \neq b_o^{max}(k-1))$ then
        Assign object $o$ to the highest bidder $b_o^{max}(k)$
        $AS'^{o} \leftarrow AS'^{o} - \{\langle b_o, o\rangle \in AS'^{o}\}$
        $AS'^{o} \leftarrow AS'^{o} \bigcup \{b_o^{max}(k), o\}$
        Increment the price for object $o$
        $price_o(k) \leftarrow price_o(k-1) + bid_o^{max}(k)$
    else
        $AS'^{o}$ and $price_o$ remain unchanged
Broadcast local assignment $AS'^{o}$ and price $price_{o'}(k)$
return $AS'^{o}, price_{o'}(k)$
```

checks whether a local termination condition on the local assignment holds. Termination occurs at some iteration k if k is the shortest time for which the local assignment holds for all bidders and no message is in transit between any two agents [3].

For the overall system's solution to comply with soft constraints, the matrix representing the arrival times from bidders (e.g., patients) to bid objects (e.g., ambulances or hospitals) reachable within the maximal allowed patient delay time t_p^{max} for all $p \in P$ should be at least triangular with at least $|P|(|P|+1)/2$ actual arrival times. Otherwise, an optimal solution to the problem may include times higher than t_p^{max} . In other words, if patients decide sequentially on the EMS resource assignment, there should be at least one decision-making ordering of patients that will give an acceptable EMS resource assignment to every patient. Otherwise, a solution respecting the maximal delays does not exist and the EMS resources with the arrival times higher than t_p^{max} should be introduced for a minimal number of patients. This is done in a non-decreasing order of the arrival times until at least a triangular matrix shape is achieved for actual patient-EMS resource arrival times. The EMS resources with combined arrival times higher than t_p^{max} for every patient $p \in P$ are penalized with weights varying depending on the distance from t_p^{max} such that higher deviations get avoided in the optimal solution as much as possible. In this way, we relax the constraints on the maximal allowed patient delay for a minimal number of patients. Furthermore, the delay time deviations of those patients are minimal at the system's level.

Lastly, to assure a correct functioning of the proposed approach, there should be a coordinator agent that will coordinate the sequencing of the three EMS resource assignment levels in Algorithm 1. This algorithm is initially performed in the first task, i.e., when the emergency call is received. It is relaunched with every new relevant event - changes of availability of the EMS resources, delay and arrival times, or of the patients' states. By allowing for reassignment of resources based on the adaptation to contingencies in real time, we obtain a flexible EMS coordination solution.

7 Simulation experiments

In this Section, we describe the simulation settings, experiments, and results. We test the proposed approach for the coordination of EMS resources in angioplasty patients' assistance focusing on the average patient delay time in the case of multiple pending patients. We compare the performance of our approach with the FCFS method since it is applied by most of the medical emergency-coordination centers in Western world countries (e.g., SUMMA 112 in Madrid, Spain).

Ambulances share the transport infrastructure with other vehicles. In the experiments, we use a so-called mesoscopic traffic simulation model where traffic flow is described with high level of detail, but at the same time flow behaviour is presented at a low level of description, i.e., we concentrate on the statistical description of the same. We introduce a further simplification by substituting the function of travel time $t(\mathbf{x}, \mathbf{y})$ between two neighboring nodes of a transport network $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ in Euclidean 2-space with Euclidean distance $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{2} (x_i - y_i)^2}$, and thus we convert the travel time minimization problem to the Euclidean distance minimization problem.

7.1 Simulation settings

To demonstrate the scalability of our solution and its potential application to small, medium and large cities and regions, in the experiments we vary the number of EMA ambulances from 5 to 100 with increment 5 and the number of angioplasty-capable hospitals from 2 to 50 with increment 2. The number of cardiology teams $|C|$ in each experiment equals the number of hospitals $|H|$. Thus, the number of setup configurations used, combining different numbers of ambulances and hospitals with cardiology teams, sums up to 500.

For each configuration, we perform the simulation on 3 different instances of random EMS participants' positions since we want to simulate a sufficiently general setting applicable to any urban area that does not represent any region in particular. The EMS participants are distributed across the environment whose dimensions are 50×50 kms. In each instance, we model hospital positions and the initial positions of ambulances, out-of-hospital cardiology teams, and patients based on a continuous uniform distribution. Therefore, each configuration can be considered as a unique virtual city with its EMS system. Assuming that the EMS system is placed in a highly dense urban area, this kind of modelling of the positions of EMS participants represents a general enough real case since the election of the hospital positions in urban areas is usually the result of a series of decisions developing over time with certain stochasticity, influenced by multiple political and demographical factors.

In the simulations, ambulances are initially assigned to the base stations in the hospitals of the region of interest. Additionally, we assume that after transferring a patient to the hospital, an ambulance is redirected to the base station where it waits for the next patient assignment. Furthermore, we assume that the hospitals have at the disposal a sufficient number of catheterization laboratories so that the only optimization factor from the hospital point of view is the number of available cardiology teams. If there are more patients with the same urgency already assigned waiting for treatment in the same hospital, they are put in a queue.

The simulation of each instance is run over a temporal horizon in which new patients are generated based on a certain appearance frequency. The EMS resources are dynamically coordinated from the appearance of a patient until the time he/she is assisted in hospital by a cardiology team. Each instance simulation is run over the total of 300 patients whose appearance is distributed equally along the overall time horizon based on the following two predetermined frequency scenarios: low (1 new patient every 10 time periods) and high (1 new patient every 2 time periods). In more detail, in the first scenario, we test our solution over minimally 3000 consecutive periods in each instance of each configuration. In the second scenario, the experiment is performed over at least 600 consecutive periods for each instance of each configuration.

The period between two consecutive executions of the EMS coordination algorithm is considered here as a minimum time interval in which the assignment decisions are made; usually it ranges from 1 to 15 minutes. In each period, the actual state of EMS resources and pending patients is detected and the EMS coordination is performed such that the EMS resources are (re)assigned for all patients. To achieve an efficient dynamic reassignment of ambulances, the execution of the EMS coordination algorithm is furthermore performed with every new significant event, i.e., any time there is a significant change in the system due to new patients, or the significant change of travel time or state of any of the EMS participants.

Let us assume that the update of information in such an EMS system is performed every 5 minutes. Then at least 3000 consecutive periods of simulation in one configuration respond to at least 15000 minutes of simulation, which is more than 10 days of simulation per instance. Three such instances result in more than a month of simulation for each such virtual urban area defined by every setup configuration.

7.2 Simulation results

In the experiments, we test the performance of the proposed EMS coordination method with respect to the FCFS benchmark approach. The comparison is based on the relative performance function P defined as:

$$P = \frac{\bar{t}_{FCFS} - \bar{t}_{OR}}{\bar{t}_{FCFS}} \cdot 100 \ , \ [\%], \tag{26}$$

where $\bar{t}_{FCFS}$ and $\bar{t}_{OR}$ are average patient delay times of the benchmark FCFS approach and the proposed model, respectively.

The simulation results of the performance function P for the two simulated cases of frequency of new patient appearance of 1 and 5 new patients every 10 time periods are presented in Figures 4, 5, 6, 7, 8, and Table 1. Figures 4 – 7 show the average patient delay improvement of proposed EMS coordination solution versus the FCFS strategy considering varying number of ambulances and hospitals in the EMS system. The performance of the proposed approach increases as the number of angioplasty enabled hospitals increases from almost identical average patient delay in the configuration setup with 2 hospitals up to 87,14 $\%$ with 50 hospitals, as can be seen in Figure 7 and Table 1.

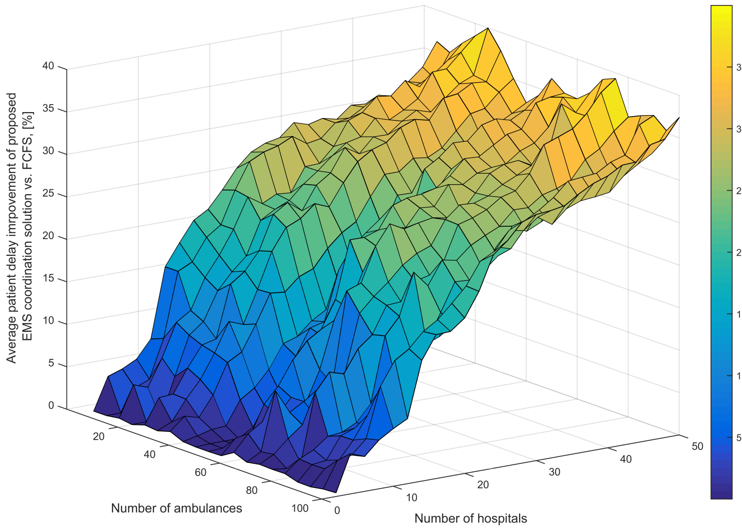

Fig. 4 Average patient delay time performance of the proposed EMS coordination approach vs. the FCFS strategy [%] for the frequency of appearance of 1 new patient every 10 time periods

Moreover, Figure 8 shows mean patient delay improvement of proposed EMS coordination solution in respect to the FCFS strategy as the number of hospitals increases. This improvement is a mean value for each experimented number of hospitals over all the experimented scenarios with the number of ambulances ranging from 5 to 100.

Observing the performance dynamics with respect to the varying number of hospitals, it is evident from Figures 5, 7, and 8 that the performance of the proposed EMS coordination method increases on average proportionally to the increase of the number of hospitals. With a relatively low number of angioplasty-enabled hospitals (less than 15), our proposed EMS coordination approach performs on average better than FCFS up to 15%. As the number of hospitals increases, the performance improves on average up to the maximum of 39,98% for the first case, Figure 5, and up to 87,14 % , for the second case, Figure 7. However, mean patient delay improvement for the two cases is 35% and 45.5% respectively, Figure 8.

Looking at the EMS coordination algorithm's performance dynamics with respect to the varying number of ambulances, in Figure 6, two regions are evident: the first one appears with a very low number of ambulances where the performance of the EMS coordination algorithm with respect to the FCFS method grows significantly faster with the increase of the number of the hospitals than the second region with a higher number of ambulances where the performance' improvement growth tendency is steadier. The performance values of the first region steeply decrease to the steady values of the valley region as the number of ambulances changes from 15 to 20. Moreover, the comparison of the overall values of Figure 4 and 6 implies that the proposed EMS coordination method's performance improves on average with respect to the FCFS method as the frequency of patient appearance increases, which is also evident from Figure 8.

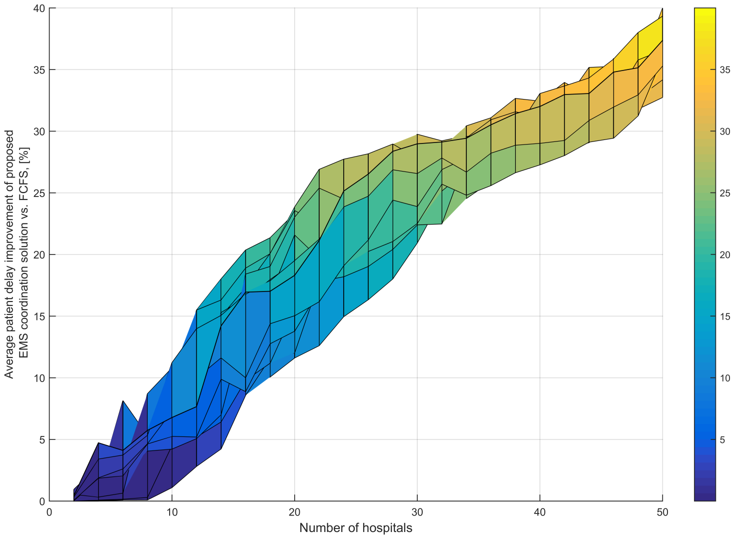

Fig. 5 Average patient delay time performance of the proposed EMS coordination approach vs. the FCFS strategy [%] for the frequency of appearance of 1 new patient every 10 time periods, horizontal view with respect to the number of hospitals

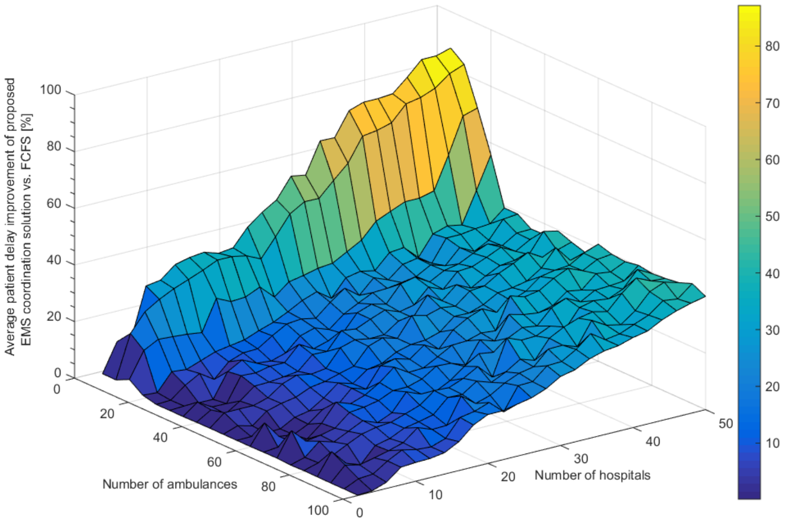

Fig. 6 Average patient delay time performance of the proposed EMS coordination approach vs. the FCFS strategy [%] for the frequency of appearance of 1 new patient every 2 time periods

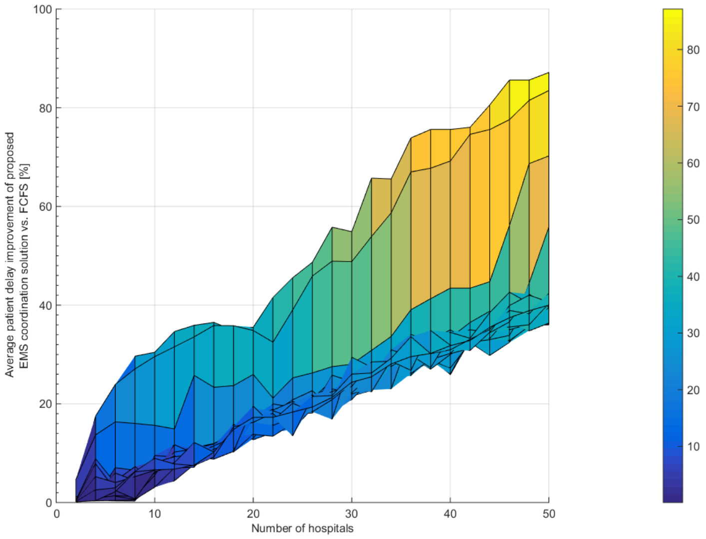

Fig. 7 Average patient delay time performance of the proposed EMS coordination approach vs. the FCFS strategy [%] for the frequency of appearance of 1 new patient every 2 time periods, horizontal view with respect to the number of hospitals

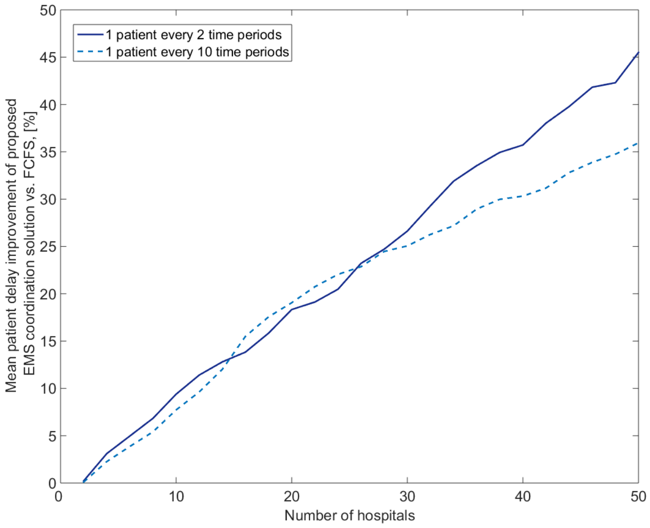

Fig. 8 Mean patient delay time improvement of the proposed EMS coordination approach vs. the FCFS strategy [%] for the frequencies of appearance of 1 new patient every 2 and 10 time periods

Table 1 Minimum and maximum values of performance function P in the simulation

Frequency of patient appearance	1/10	5/10
P min.value, [%]	0,001	0,002
P max. value, [%]	39,98	87,14

The static assignment of the FCFS principle discriminates against patients appearing later. Since ambulances are not equally distributed in the area, the proposed EMS coordination method compensates for the lack of EMS resources and their unequal distribution by reassigning them dynamically to pending patients. Dynamically optimized reassignment of EMS resources in real time is the main key to the improvement of the system's performance. Thus, proportional to the increase of the number of hospitals, there is a constant improvement of performance up to the maximum values of the simulated experiments as seen in Table 1.

Even though the velocity of the EMS actors is not a relevant factor in the comparison of the performance of our proposed EMS coordination solution and the FCFS method, looking individually at the performance of each one of these methods, it is evident that the assignment cost accumulated through the time will be lower when the velocity of the EMS actors is higher. The efficiency of the solution is limited from above by the time period duration, i.e., the proposed EMS algorithm should be executed with every important event that might change the system's assignment solution.

8 Conclusions

In this paper, we proposed a distributed and optimized EMS coordination model that facilitates a seamless coordination among the emergency medical assistance participants for the minimization of delay times of angioplasty patients. The proposed model implies the change of the current functioning based on a manual coordination through communications via phone calls, towards an automated coordination process where the basic decisions are taken (or proposed) by software agents.

In order to reduce patient delay times, we proposed a distributed coordination tool based on a three-level optimization in which the assignment of EMS resources to patients is performed via iterative auctions for the minimization of ambulance arrival times to patients, and the patients' and cardiology teams' arrival times to hospitals. The proposed EMS coordination approach enables globally optimized EMS resource assignment even in cases when multiple patients have to be assisted at the same time and provides an increased flexibility and responsiveness of the emergency system. Additionally, the ambulance assignment can be optimized both in combination or independently of the hospital and cardiology team assignment, depending on the patients' needs and the means of transport to the hospital.

Our simulation results show the efficiency of the proposed solution approach, resulting in significantly lower delay times for angioplasty on average. Of course, the effectiveness of the proposed model depends on the initial classification of patients, and the related determination of the urgency of their cases, as well as on the timely availability of cardiology teams and hospitals. Still, as the current experience in the region of Madrid shows, good quality patient assessments and the EMS resource availabilities can be assured in practice.

To implement our approach in practice, a patient's location needs to be known to the system. Ideally, patients should contact the ECC through a mobile phone with GPS for easier location. In addition, ambulances should have a GPS and a navigator for localizing the patient and navigating the way to him/her, as well as a means of communication with the rest of the EMS participants, and a digitalized map showing ambulances, patients and hospitals. Moreover, hospitals should have a digitalized receptionist service to receive and process relevant data of a patient before his/her arrival. None of these requirements go significantly beyond the current state of affairs in major cities (such as Madrid). Moreover, there are intrinsic uncertainties present in the EMS coordination. In our experiments, we assume that travel times can be accurately forecasted, which, of course, is an important factor for the performance of the proposed system. In reality, this may not always be the case, as real-world traffic conditions are notoriously hard to predict. However, there is abundant literature on traffic-aware vehicle route guidance systems tackling this problem, and we believe that such systems can be easily integrated into our approach. Still, an effective proof of this conjecture is left to future work.

There are some additional prerequisites for the practical application of the proposed approach. The EMS provider, especially if it currently implements hierarchically structured procedures, should be willing to move towards a patient-oriented organizational structure. In addition, the parameters of the proposed EMS coordination approach should be chosen with care for each particular application scenario and area, since overly fast reassignments of EMS resources could cause instability in the EMS system. Additionally, as a future work, we plan to develop an EMS coordination model that considers both efficiency and equity issues among EMS actors, while forecasting future patients across a receding horizon.

Acknowledgements This work was supported in part by the Spanish Ministry of Economy and Competitiveness through the projects iHAS (grant TIN2012-36586-C03-02) and SURF (grant TIN2015-65515-C4-4-R), and by the Community of Madrid through the programme MOSI-AGIL-CM (grant P2013/ICE-3019, co-funded by EU Structural Funds FSE and FEDER).

References

1. D. Bandara, M. E. Mayorga, and L. A. McLay. Optimal dispatching strategies for emergency vehicles to increase patient survivability. *International Journal of Operational Research*, 15(2):195–214, 2012.
2. N. Behzadia, M. A. Salinero-Fortb, A. de Blasc, M. Taboadaa, L. Pérez de Islad, and J. L. López-Sendóne. Prehospital thrombolysis: Two years experience of the community of madrid emergency services (summa 112). *Rev Esp Cardiol (Engl Ed).*, 65(10):960–961, 2012.
3. Dimitri P Bertsekas. Auction algorithms for network flow problems: A tutorial introduction. *Computational Optimization and Applications*, 1(1):7–66, 1992.
4. Dimitri P. Bertsekas. Auction algorithms. In *Encyclopedia of Optimization*, pages 128–132. 2009.
5. D.P. Bertsekas. The auction algorithm: A distributed relaxation method for the assignment problem. *Annals of Operations Research*, 14(1):105–123, 1988.
6. Holger Billhardt, Alicia Fernandez, Lissette Lemus, Marin Lujak, Nodar Osman, Sascha Ossowski, and Carles Sierra. Dynamic coordination in fleet management systems: Toward smart cyber fleets. *Intelligent Systems, IEEE*, 29(3):70–76, 2014.
7. Holger Billhardt, Marin Lujak, Sascha Ossowski, Ralf Bruns, and Jürgen Dunkel. Intelligent event processing for emergency medical assistance. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 200–206. ACM, 2014.
8. Holger Billhardt, Marin Lujak, Vicente Sánchez-Brunete, Alberto Fernández, and Sascha Ossowski. Dynamic coordination of ambulances for emergency medical assistance services. *Knowledge-Based Systems*, 70:268–280, 2014.
9. Stephen Bradley, Arnoldo Hax, and Thomas Magnanti. Applied mathematical programming. 1977.

10. Steven C Brooks, Robert H Schmicker, Thomas D Rea, et al. Out-of-hospital cardiac arrest frequency and survival: evidence for temporal variability. *Resuscitation*, 81(2):175–181, 2010.
11. C.P. Cannon, C.M. Gibson, C.T. Lambrew, D.A. Shoultz, D. Levy, W.J. French, J.M. Gore, W.D. Weaver, W.J. Rogers, and A.J. Tiefenbrunn. Relationship of symptom-onset-to-balloon time and door-to-balloon time with mortality in patients undergoing angioplasty for acute myocardial infarction. *JAMA: the journal of the American Medical Association*, 283(22):2941–2947, 2000.
12. Brecht Cardoen, Erik Demeulemeester, and Jeroen Beliën. Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, 201(3):921–932, 2010.
13. Yann Chevaleyre, Paul E Dunne, Ulle Endriss, et al. Issues in multiagent resource allocation. *Informatica*, 30(1), 2006.
14. Yann Chevaleyre, Ulle Endriss, Sylvia Estivie, and Nicolas Maudet. Multiagent resource allocation with k-additive utility functions. In *In Proc. DIMACS-LAMSADE Workshop on Computer Science and Decision Theory, Annales du LAMSADE*, pages 83–100, 2004.
15. Gianpaolo Cugola and Alessandro Margara. Processing flows of information: From data stream to complex event processing. *ACM Computing Surveys (CSUR)*, 44(3):15, 2012.
16. Giuseppe De Luca, Harry Suryapranata, Jan Paul Ottervanger, and Elliott M Antman. Time delay to treatment and mortality in primary angioplasty for acute myocardial infarction every minute of delay counts. *Circulation*, 109(10):1223–1225, 2004.
17. P Dilaveris, Andreas Synetos, Georgios Giannopoulos, et al. Climate impacts on myocardial infarction deaths in the athens territory: the climate study. *Heart*, 92(12):1747–1751, 2006.
18. Elton Domnori, Giacomo Cabri, and Letizia Leonardi. Coordination issues in an agent-based approach for territorial emergence management. In *Advanced Information Networking and Applications (WAINA), 2011 IEEE Workshops of International Conference on*, pages 184–189. IEEE, 2011.
19. Ulrich Endriss, Nicolas Maudet, Fariba Sadri, and Francesca Toni. On optimal outcomes of negotiations over resources. In *AAMAS*, volume 3, pages 177–184, 2003.
20. Michel Gendreau, Gilbert Laporte, and Frédéric Semet. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel computing*, 27(12):1641–1653, 2001.
21. S. Giordani, M. Lujak, and F. Martinelli. A distributed algorithm for the multi-robot task allocation problem. In *Trends in Applied Intelligent Systems: Proc. of the 23rd internat. conf. on Ind. eng. and other applications of applied int. systems*, volume 6096 of *Lecture Notes in Artificial Intelligence*, pages 721–730. Springer-Verlag, 2010.
22. Francesca Guerriero and Rosita Guido. Operational research in the management of the operating theatre: a survey. *Health care management science*, 14(1):89–114, 2011.
23. Shane G Henderson. Operations research tools for addressing current challenges in emergency medical services. *Wiley Encyclopedia of Operations Research and Management Science*, 2011.
24. Johan Herlitz, Mikael Eek, Mikael Holmberg, and Stig Holmberg. Diurnal, weekly and seasonal rhythm of out of hospital cardiac arrest in sweden. *Resuscitation*, 54(2):133–138, 2002.
25. James G Jollis, Christopher B Granger, Timothy D Henry, et al. Systems of care for st-segment–elevation myocardial infarction: A report from the american heart associations mission: Lifeline. *Circulation: Cardiovascular Quality and Outcomes*, 5(4):423–428, 2012.
26. Ildikó Kriszbacher, Imre Boncz, Miklós Koppán, and József Bódis. Seasonal variations in the occurrence of acute myocardial infarction in hungary between 2000 and 2004. *International Journal of cardiology*, 129(2):251–254, 2008.
27. Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.
28. Jang Hoon Lee, Shung Chull Chae, Dong Heon Yang, et al. Influence of weather on daily hospital admissions for acute myocardial infarction (from the korea acute myocardial infarction registry). *International journal of cardiology*, 144(1):16–21, 2010.
29. Jose Ramon Garmendia Leiza, Jesus Maria Andres de Llano, Juan Bautista Lopez Messa, et al. New insights into the circadian rhythm of acute myocardial infarction in subgroups. *Chronobiology international*, 24(1):129–141, 2007.
30. Donald Lloyd-Jones, Robert J Adams, Todd M Brown, Mercedes Carnethon, Shifan Dai, Giovanni De Simone, T Bruce Ferguson, Earl Ford, Karen Furie, Cathleen Gillespie, et al. Heart disease and stroke statistics2010 update a report from the american heart association. *Circulation*, 121(7):e46–e215, 2010.
31. Beatriz López, Silvana Aciar, Bianca Innocenti, and Isabel Cuevas. How multi-agent systems support acute stroke emergency treatment. In *IJCAI Workshop A2HC*, pages 51–59, 2005.
32. Beatriz López, Bianca Innocenti, and Dídac Busquets. A multiagent system for coordinating ambulances for emergency medical services. *Intelligent Systems, IEEE*, 23(5):50–57, 2008.
33. Marin Lujak and Holger Billhardt. Coordinating emergency medical assistance. In *Agreement Technologies*, pages 597–609. Springer, 2013.

34. Marin Lujak, Holger Billhardt, and Sascha Ossowski. Optimizing emergency medical assistance coordination in after-hours urgent surgery patients. In Nils Bulling, editor, *Multi-Agent Systems*, volume 8953 of *LNCS*, pages 316–331. Springer, 2015.
35. Roberto Manfredini, Fabio Manfredini, Benedetta Boari, et al. Seasonal and weekly patterns of hospital admissions for nonfatal and fatal myocardial infarction. *The American journal of emergency medicine*, 27(9):1097–1103, 2009.
36. PR Martens, Paul Calle, B Van den Poel, P Lewi, Belgian Cardiopulmonary Cerebral Resuscitation Study Group, et al. Further prospective evidence of a circadian variation in the frequency of call for sudden cardiac death. *Intensive care medicine*, 21(1):45–49, 1995.
37. Susana Mingo, Javier Goicolea, and Luis et al. Nombela. Angioplastia primaria en nuestro medio. análisis de los retrasos hasta la reperfusión, sus condicionantes y su implicación pronóstica. *Revista española de cardiología*, 62(1):15–22, 2009.
38. Instituto nacional de estadistica. Defuncines según la causa de muerte, año 2009. `http://www.ine.es/prensa/np664.pdf`, February 2011.
39. Abdur Rais and Ana Viana. Operations research in healthcare: a survey. *International Transactions in Operational Research*, 18(1):1–31, 2011.
40. S.S. Rathore, J.P. Curtis, J. Chen, Y. Wang, B.K. Nallamothu, A.J. Epstein, H.M. Krumholz, et al. Association of door-to-balloon time and mortality in patients admitted to hospital with st elevation myocardial infarction: national cohort study. *BMJ: British Medical Journal*, 338, 2009.
41. Joaquín Álvarez Rodríguez, Isabel Casado Flores, Javier Botas Rodríguez, et al. Reperfusión del infarto agudo de miocardio con elevación del segmento st en la comunidad de madrid.
42. Amartya Kumar Sen. *Collective choice and social welfare*, volume 11. Elsevier, 2014.
43. WHO Fact sheet No. 310. `http://www.who.int/mediacentre/factsheets/fs310/en/`, May 2014.
44. CY Soon, WX Chan, and HC Tan. The impact of time-to-balloon on outcomes in patients undergoing modern primary angioplasty for acute myocardial infarction. *Singapore medical journal*, 48(2):131, 2007.
45. P.G. Steg, S.K. James, D. Atar, et al. Esc guidelines for the management of acute myocardial infarction in patients presenting with st-segment elevation the task force on the management of st-segment elevation acute myocardial infarction of the european society of cardiology (esc). *European Heart Journal*, 2012.
46. Masashi Takahashi, Shun Kohsaka, Hiroaki Miyata, et al. Association between prehospital time interval and short-term outcome in acute heart failure patients. *Journal of Cardiac Failure*, 17(9):742–747, 2011.
47. Christian Juhl Terkelsen, Jacob Thorsted Sørensen, Michael Maeng, et al. System delay and mortality among patients with stemi treated with primary percutaneous coronary intervention. *JAMA: the journal of the American Medical Association*, 304(7):763–771, 2010.
48. Frans Van de Werf, Jeroen Bax, Amadeo Betriu, et al. Management of acute myocardial infarction in patients presenting with persistent st-segment elevation. *European heart journal*, 29(23):2909–2945, 2008.