Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2018) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Young Scientist Conference on Computational Science

Coupling Game Theory and Discrete-Event Simulation for Model-Based Ambulance Dispatching

Xinyu Fu[1,*], Alva Presbitero[1,2] , Sergey V. Kovalchuk[1] ,Valeria V. Krzhizhanovskaya[1,2]

[1]*ITMO University, 197101, 49 Kronverksky pr., St Petersburg, Russia*
[2]*University of Amsterdam, 1012 WX Amsterdam, The Netherlands*

Abstract

A critical part of the ambulance dispatching system for patients with acute diseases is analyzing the interactions between different stakeholders such as hospitals, patients, and healthcare authorities. Allocating medical resources becomes crucial for two main reasons typical to acute cases: 1) patients need professional facilities and 2) acute states occur unexpectedly. In this research, firstly we employed a queuing model as an analytical model to see the essential interactions between hospitals and patients and then extended this model into a hybrid model using game theory and discrete event simulation in the case of multi-hospitals scenario. The experimental settings and corresponding data were developed for Acute Coronary Syndrome (ACS) which is one of the leading causes of death worldwide. The primary goal of our work is to develop a city-scale model for analyzing and optimizing the dynamics of ambulance dispatching in Saint Petersburg having ACS as working example of critical disease affecting modern society.

© 2018 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/)
Peer-review under responsibility of the scientific committee of the 7th International Young Scientist Conference on Computational Science.

Keywords: Queuing Theory; Game Theory; Acute Coronary Syndrome; Ambulance Dispatching; Discrete Event Simulation

[*] Corresponding author. Tel.: +7-981-846-5026.
E-mail address: hsinyu.fu@niuitmo.ru

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/)
Peer-review under responsibility of the scientific committee of the 7th International Young Scientist Conference on Computational Science.

1. Introduction

Medical resource allocation is fundamental to the concept of “smart cities”. Hospitals, especially emergency departments (ED), give a huge contribution to public service. However, there remains an imbalance in terms of services allotted to patients. For instance, some hospitals are overcrowded, while others are free. According to the data of the medical resource allocations in Saint Petersburg, the northern region of the city has more hospitals but lesser inhabitants. The case is, however, opposite in the southern part of the city, where there are less hospitals but more inhabitants. Hence, a problem of loading arises when more people visit an ED having limited number of facilities, like for instance, in the southern part of Saint Petersburg, in spite of other EDs being relatively “free” that. This may lead to economic waste, health risks, and the eventual decline of public trust to medical systems.

In this paper, we introduce a solution to the load problem for hospitals by efficiently pooling the current medical resources through the effective dispatch of patients by ambulances. For instance, overcrowded hospitals can redirect patients to other hospitals that are free so that all the hospitals in the system can share the load effectively. In fact, in some regions in the United States of America, EDs can declare *diversion* status to indicate whether they can accommodate incoming patients’ requests or not. In this way, the local emergency medical system (EMS) may reduce ambulances going to hospitals in the *diversion* status [1]. However, R.M. N Mihal [2] claimed in their work that it is not advisable to redirect patients due to 1) extra travelling time 2) poor health outcomes 3) little benefits in the reduction of waiting time. More so, EMS prompted that the ambulances’ resources may be insufficient due to too many redirections. It even comes to a point where ambulances are unable to accept new emergency calls [3]. Eckstein and Chan et al. [4] pointed out 21,240 incidents in Los Angeles, where ambulances waited in front of a hospital for even 1 hour. This strengthened the general opposition towards ambulance diversions. Massachusetts banned this norm in 2009[4]. However, our work aims to demonstrate the contrary. That is, we offer a possible explanation why ambulance diversion could actually help balance the system’s load in some cases.

Many studies suggested solutions for overcrowding. The obvious and simplest solution here is to increase hospital capacity. It seems straightforward and useful. However, R.A. McCain et al. [5] proposed a verified Nash equilibrium hypothesis for overcrowding and they suggested that increasing ED capacity has little or fewer impact to this situation.

The treatment process to Acute Coronary Syndrome (ACS) patients has similar properties with emergency patients because ACS patients need professionals specializing on this domain, whose services are expensive and sometimes also limited to a specific hospital. Occurrences of ACS are usually unexpected and ACS patients are badly in need of specialized and professional treatment as soon as possible. A delay may be too risky, which could result in critical health output [6].

Our work is an extension of the work in [7]. To concretely analyse and optimize ambulance dispatching, our work begins by formulating an "*M/M/c*" queuing-network model for the scenario with two hospitals. The model is a deviation from the analytical model used in [7]. In our case, a hospital can choose an “Accepting (A)” strategy with the "*M/M/c/∞*" model or a “Redirecting(R)” strategy with the "*M/M/c/N*" model. We assume that the hospitals are playing a non-cooperative game where each hospital switches its strategy based on its benefits (denoted by the score function, which is the ratio between incoming requests and time spent by each patient). In previous studies such as in K.Y. Lin et al. [8], the model "*M/M/1*" was used. We then extend the queuing-network model to a multi-hospitals scenario via discrete event simulation on a two-dimensional map. Finally, we added a game theory-based model to facilitate the interplay happening between hospitals.

2. Methodology for coupled queuing and game theory modelling

2.1. One-dimensional analytical model for 2 hospitals

We formulated a queuing-network model for analysing the interactions between two hospitals and patients in a one-dimensional map as shown in Figure 1. The derivation of the analytical model deviates from [7]. We have similar assumptions but the methods for defining the parameters are different. We also derived new set of equations.

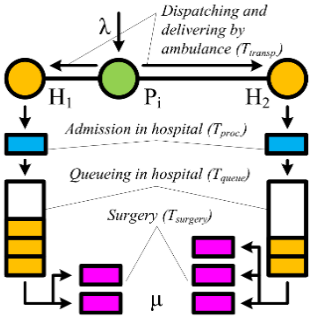
λ
Dispatching and delivering by ambulance ($T_{transp.}$)
H_1
P_i
H_2
Admission in hospital (T_{proc})
Queueing in hospital (T_{queue})
Surgery ($T_{surgery}$)
μ

Figure 1 Ambulance dispatching and queuing in one-dimensional map. Source: [4]

We firstly assigned two hospitals (H1 and H2) to each side of the one-dimensional map. Each hospital can have a strategy of “Accepting” (A, accepting all the patients’ requests) or “Redirecting” (R, redirecting patients to opposite hospital if the current number of patients in the hospital exceeds the predefined threshold). Patients are initialized uniformly between H1 and H2. Patients are delivered to a hospital according to the principle of proximity. That is, whenever patients arrived at any hospital, they will be served as soon as possible given that there are free facilities (servers or medical practitioners), otherwise they will have to wait. The serving process follows the principle of “First in First Out” (FIFO).

The patient flow in hospitals is modelled by the multi-servers queuing model "*M/M/c*" [9], where the first two Ms denote the Markov exponential distributions with inter-arrival time and service time respectively, and “c” is the number of parallel servers in hospitals (stands for the number of professional facilities for ACS patients). The "accepting" strategy of a hospital is modelled by the "*M/M/c/∞*" model with an infinite system capacity. The "redirecting" strategy is described by the "*M/M/c/N*" model with a limited system capacity N. The arrivals of patients and hospital services are governed by the Poisson process and Exponential process respectively. We define λ as request rate which is the mean value of arriving process (Poisson process). μ is the serving rate ($\mu = 1/T_{surgery}$) which is the mean time for the serving process. t_c is defined as the time distance between two hospitals. The total time is given by equation (1):

$$T = T_{trans} + T_{que} + T_{surgery} \tag{1}$$

In the analytic model, each hospital has a strategy “A” or “R”. Therefore, there are 4 combinations of strategies namely: AA (H1 is taking Accepting strategy and H2 is taking Accepting strategy, the same norm for RR/RA/AR situation), RR and RA/ARin system. The load parameter is $a = \lambda/\mu$ and load parameter with multiple servers is denoted by $\rho = \lambda/c\mu$

For the hospital taking the redirecting strategy, we first introduce the probability where there are no patients in the system [10]:

$$P_0 = \left[1 + \sum_{n=1}^{c} \frac{a^n}{n!} + \frac{a^c}{c!\sum_{n=c+1}^{N} \rho^n - c}\right] \tag{2}$$

We then expand it to the probability of n and *N* (system capacity) patients in the system. This also corresponds to the probability of redirecting the patients in the system:

$$P_n = \rho^n P_0$$

$$P_{rej} = P_N = P_{max} = \frac{a^N}{c!\,c^{N-c}} P_0 \tag{3}$$

In the RA case, the effective λ in H1 is summarized by equation (4):

$$\lambda_{e1} = \lambda_1 \left(1 - P_{rej}(\lambda_1)\right) \tag{4}$$

The effective λ in H2 is

$$\lambda_{e2} = \lambda_2 + \lambda_1 P_{rej}(\lambda_1) \tag{5}$$

In RR case, it is possible that both hospitals reject the request. Then the nearest hospital has to accept the patient without passing admission control. Consequently, the request rate is updated as λ_e :

$$\lambda_e = \lambda^* = \sum_{j=1}^{3} \lambda_{i,j} \tag{6}$$

$$\lambda_{i,1} = \lambda_i \left(1 - P_{rej}(\lambda_i)\right) = \lambda_i - \lambda_i^{R1} \tag{7}$$

$$\lambda_{i,2} = \lambda_{i-1}^{R1} \left(1 - p_{rej}(\lambda_i + \lambda_{1-i}^{R1})\right) = \lambda_{i-1}^{R1} - \lambda_i^{R2} \tag{8}$$

$$\lambda_{i,3} = 0.5(\lambda_i^{R2} + \lambda_{i-1}^{R2}) \tag{9}$$

Because of rerouting, travel times for different strategies vary [7] for different strategies applied in system. They are summarized below:

$$T_{transp}^{AA} = T_{transp}^{RA} = 0.25t_c \tag{10}$$

$$T_{transp}^{AR} = \frac{0.25 + 0.75p_{rej}(\lambda_R)}{1 + P_{rej}(\lambda_R)} \tag{11}$$

$$T_{transp}^{RR} = \frac{0.25\lambda_{i,1} + 0.75\lambda_{i,2} + 0.5\lambda_{i,3}}{\lambda_i^*} t_c \tag{12}$$

Using the equations above, we could find some new definitions summarized below:

Definition 1. Average number [10] of patients in queue and average time of each patient spent in queue are :

$$L_Q = \frac{P_0 a^c \rho}{c!\,(1-\rho)^2)} [1 - \rho^{N-c} - (N-c)\rho^{N-c}(1-\rho)] \tag{13}$$

$$W_Q = \frac{L_Q}{\lambda_e} \tag{14}$$

Definition 2. Average number of patients in server is:

$$L_s = \sum_{n=1}^{c} P_n n + \sum_{n=c+1}^{N} P_n c \tag{15}$$

Definition 3. Global time [7] is:

$$T_{global} = \frac{T_1 * \lambda_1 + T_2 * \lambda_2}{T_1 + T_2} \tag{16}$$

The equations presented are not only used in the 1-dimensional analytical model, but also in the 2-dimensional discrete event simulation. We will now focus more on 2-D model, because it is more representative of the real-world scenario.

2.2. Two-dimensional multi-hospital coupled model

As shown in Figure 2, in the multi-hospitals scenario, we placed K (3 in this case) hospitals in a two-dimensional map with known fixed locations in a circular area. Each hospital has its own service coverage. Patients "appearing" in a certain area will be delivered to the hospital responsible for that area. Patients' locations are generated uniformly in the entire 2-dimensional map. Similar to the two-hospital scenario, each hospital can choose between two strategies: 1) to accept patients without restriction or 2) to redirect patients when the hospital is overcrowded.

We summarized the general methodology below:

(a) Once patients are fetched by the ambulance, the nearest hospital (target hospital) is chosen as top priority. Patients will then be sent to the target hospital and accepted if the target hospital has an accepting strategy or redirecting strategy with "non-overcrowding" status. However, if the target hospital's strategy is redirecting and hospital status is "overcrowding", the ambulance then brings patients to the second nearest hospital. We repeat this process with all the hospitals. If all hospitals are busy and reject patients' requests, then the patients are sent to the best-suitable hospital (the minimum expected time spent in traveling and serving). Here we present a discrete event simulation, which is different from a deterministic model.

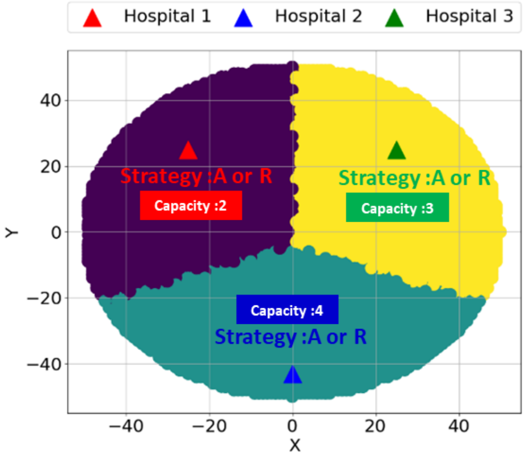

Figure 2 An example of simulated two-dimensional area with 3 hospitals. Shaded sectors are the "service coverage" zones. Patients from each zone are normally delivered to the nearest hospital if it accepts the request. Hospital capacity is predefined according to the number of professionals in the hospital.

(b) The sequence of interval times of patients' appearance, is generated by an exponential distribution with expectation $\lambda_{appearing}$. The locations of patients being fetched have uniform distribution inside the map. The service area of each hospital is determined by the shortest distance from the location where the patients are fetched and the hospital. The crowding level of a hospital is positively correlated to the current queuing length of the hospital. In addition, if the queue length surpasses a predefined threshold N (the capacity of buffer/boarding in hospital, which is positively correlated to the number of serving facilities in that hospital) then the hospital taking the Redirecting/*Rejecting* strategy will decline the request. If not, the hospital will accept the patient and put him or her in the queue. The serving time is also modelled as an exponential distribution via parameter $\lambda_{serving}$. When all hospitals reject the patient's request, patients will be sent to the hospital with less expected time (sum of travelling and queuing time).

(c) L_{served} is the number of patients being served in each hospital over a period (2000 timesteps in our simulation). We assigned it as payoff for the non-cooperative game because it explained two key benefits for hospitals and patients: 1) The number of patients being served in hospital and 2) average total time spent in the system. Consequently, the payoff matrix can be introduced in Table 1 (L_{1_A} denoted as the short name of L_{served} in Hospital 1 with Accepting strategy, same norm to L_{2_A}, L_{3_R} etc.):

Table 1 Payoff matrix of three hospitals

		Hospital 3			
		A		R	
		Hospital 2		Hospital 2	
		A	R	A	R
Hospital 1	A	$L_{1\_A}, L_{2\_A}, L_{3\_A}$	$L_{1\_A}, L_{2\_R}, L_{3\_A}$	$L_{1\_A}, L_{2\_A}, L_{3\_R}$	$L_{1\_A}, L_{2\_R}, L_{3\_R}$
	R	$L_{1\_R}, L_{1\_A}, L_{1\_A}$	$L_{1\_R}, L_{2\_R}, L_{3\_A}$	$L_{1\_R}, L_{2\_A}, L_{3\_R}$	$L_{1\_R}, L_{2\_R}, L_{3\_R}$

3. Results

3.1. 2-hospitals scenario

Figure 3 shows an example of simulation results with different combinations of strategies by 1-D model. The global time spent in both hospitals reflects the overall system performance. A larger number of medical professionals and facilities (such as catheterization labs for angiography) corresponds to a shorter time a patient spends in the system. The drawback, however, is that the idle time grows. Therefore, the system becomes less cost-effective. If both hospitals follow a "rejecting" strategy then the global time is obviously the highest. It is interesting that the hospital with fewer facilities (N= [2,3] on the right of Figure 3) can save the system time by selecting the "accepting" strategy.

In the scenarios with equal capacity of facilities in both hospitals (N = [2,2]), different strategies (accepting strategy in Hospital 1 and rejecting strategy in Hospital 2) provide the same global time.

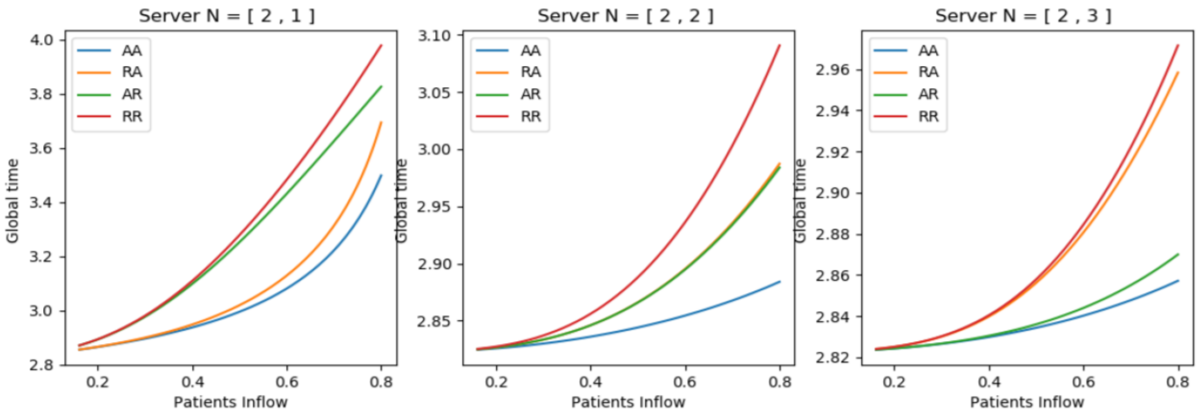

Figure 3 Simulation results for a two-hospital system: Global time spent by all patients in both hospitals (queue, transportation, and service). A pair [X, Y] denotes the number of medical facilities in each hospital. Hospital strategies are indicated by "A" for "accepting and "R" for "rejecting/redirecting". The legend describes the combinations of strategies, e.g. "RA" means Hospital 1 takes rejecting strategy, and Hospital 2 takes accepting strategy.

3.1. Multi-hospitals scenario

First, we experimented with 3 hospitals placed in a two-dimensional map with a radius of 50 km. We set (0,0) as the origin as shown in Figure 2. Each hospital has a similar area of service coverage. Therefore, we assigned the locations of hospitals at coordinates [-25, 25], [0, -43], [25, 25] in terms of the Cartesian coordinate system for Hospitals 1, 2 and 3 respectively. The travelling velocity is pre-defined as 2 km per time step. We run 2000 time steps as 1 iteration (for each of the patient incoming flow), gradually adding more patient requests by reducing the mean of time interval of the appearances of patients (e.g. decreasing average time interval from 24 to 6 in Figure 4), and monitored the number of patients served by all the hospitals.

As it is shown in Figure 4, when the incoming flow is low (e.g. only 100 patients appeared in total), the number of patients served is approximately the same for all strategies. This is because only a few patients are redirected to other hospitals when the system is free (no queues in the hospitals). It also explains that the serving ratio (the ratio between patients served and patients appeared in the system) reaches to 93%. The remaining 7% of patients are still being served or travelling. With the number of patient requests growing to around 150, the RRA and RRR strategies become most beneficial. When there is a heavy incoming flow of patients (e.g. 300 patients appeared in total), the system is saturated, because all the hospitals are overcrowded. Therefore, when the system is busy, redirecting to other hospitals is not necessary since the requests will be rejected mostly.

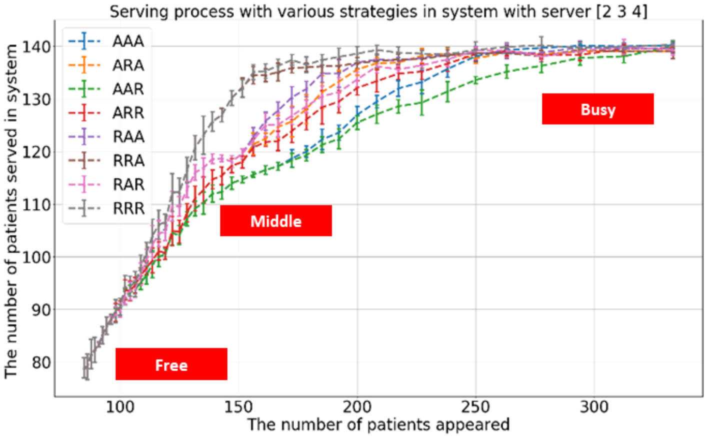

Figure 4. The number of patients served by 3 hospitals taking different strategies. "A" in the legend stands for accepting strategy, "R" denotes redirecting strategy. The numbers of servers/facilities in hospitals are [2,3,4] and queuing lengths are [2,3,4] (This configuration is applied throughout other results in the paper). The average serving rate is 120 timesteps. For each value of the patient inflow, 10 simulations were run with random initialization of the patient locations and serving times. The results are presented by the mean and error bars indicating 1 standard deviation.

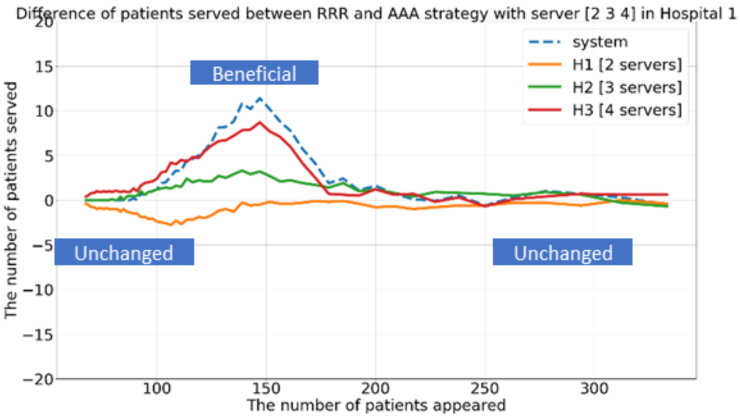

Figure 5 The difference between the number of total patients served (all values are mean value of 10 times' simulations) by RRR and AAA strategy. The number of facilities/server is [2,3,4]. When the number of patients in the system (see blue dashed line) reaches the transition value (around 230), the difference stays unchanged or as we put it, the system is in the de-pooling/unchanged state. Redirecting strategy does not help balance the loads of system.

In Figure 5, at around 230 patient requests, the number of patients being served stays unchanged or *de-pooling* happens, which states that the system is at its maximum capacity and cannot serve more patients even though the number of incoming patients keeps increasing. In addition, when incoming flow level is intermediate (labelled *Beneficial*), the redirecting strategy greatly helps in decreasing system jam (e.g. when the number of patients is approximately 150, the difference between RRR and AAA is around 11 patients). Additionally, the hospital with more servers (red line) served much more additional patients than the hospitals with less servers. In other words, the most "benefits" of system come from the hospital with more servers when RRR strategy is taken by the system. Hence, we can deduce from here that it will be beneficial for system if the hospital with more servers can accept more patients

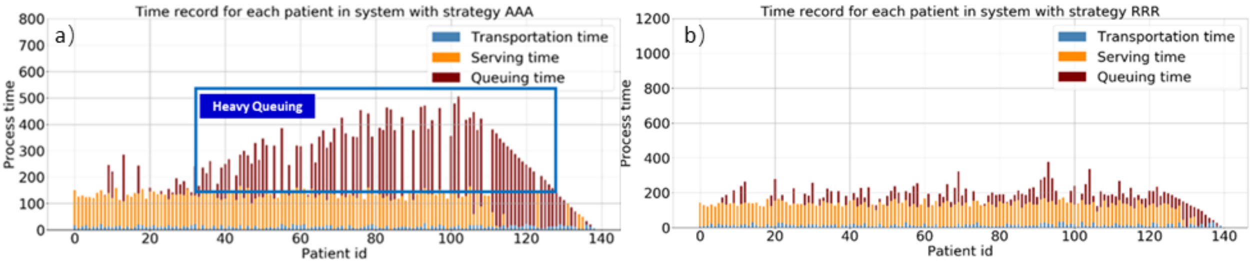

Figure 6 Time consumption of a single simulation with 140 patients in the system (the numbers of servers for Hospital 1, 2 and 3 are 2, 3 and 4 respectively]). The ID of patient is also the order of patients being fetched. a) AAA Strategy (b) RRR strategy.

Next, we analysed the processing time for each patient (see Figure 4). In Figure 6, queuing becomes heavy in AAA as emphasized by the blue rectangle in (a). The system's load is better balanced in RRR since we see here that each patient has similar time consumption.

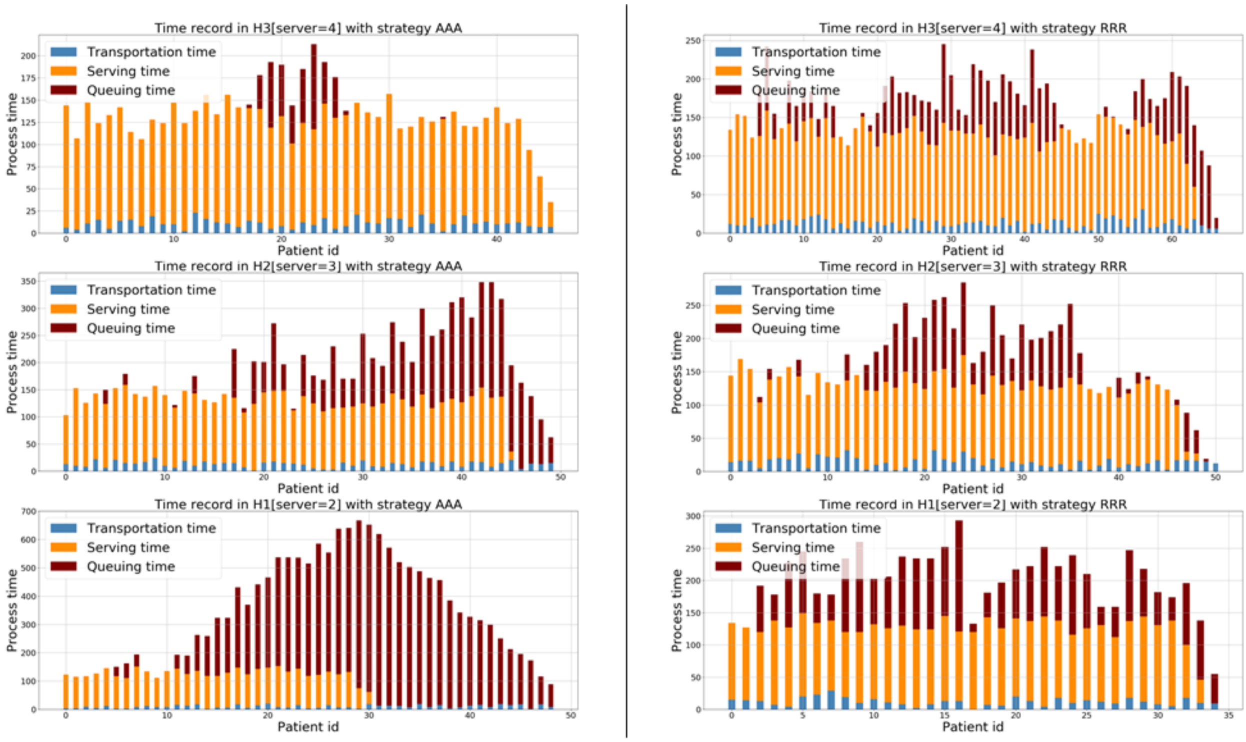

Figure 7 Time consumption of each patient ID served in AAA (left with 3 figures) and RRR situation (right with 3 figures)

As expected, the hospital with more facilities (or servers) becomes more efficient (generally with less idle time) when the system is in RRR. In Figure 7, in the AAA situation (left), many patients are queuing in Hospital 1 with 2 servers, but at the same time, Hospital 3 with 4 servers is relatively free. On the contrary, in RRR, the medical system is able to balance the load of patients equally among hospitals. The main reason behind the inefficiency of the system is that the hospital with less servers becomes overloaded, while hospital with more servers are free of patients.

In Figure 8, the left figure is the result of Nash equilibriums with strategy initialized at AAA. Here we show that AAA strategy almost dominated the entire map. One can say that it reaches *defensive equilibrium.* However, if the hospitals start with the RRR strategy (as shown in the right figure), with increasing patient's incoming rate, the strategies shift from RRR to ARR, AAR and AAA with increasing patients' incoming rate.

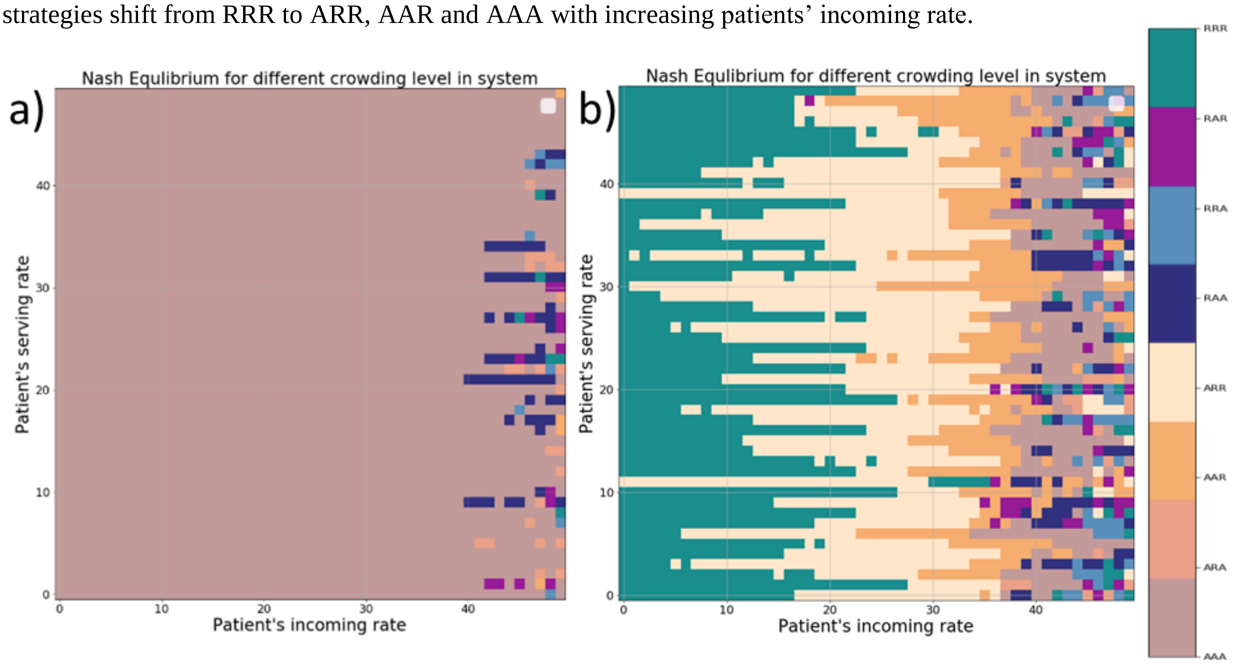

Figure 8 Nash equilibriums of different patient's incoming rate and serving rate for AAA (a) to the left) or RRR (b) to the right) as initialized situation. Each equilibrium (also each pixel) is the one with highest frequencies with 10 times simulations.

4. Conclusions and future work

The results of our two-dimensional discrete event model coupled with game theory show that redirecting strategy, in fact, helps in balancing the system's load by distributing the patients equally to all hospitals. Our results suggest that hospitals with fewer servers should redirect patients to other hospitals, while hospitals with more servers are recommended to accept more patients - a result that is in fact intuitive. We also looked at the Nash equilibrium of the system by initializing strategies from all accepting (AAA) and all redirecting strategies (RRR) and observed how the system evolves to an eventual strategy. We showed that if the initial strategy of the hospitals is AAA, it is likely that AAA becomes a dominant strategy regardless of patients' incoming flow. However, if RRR is the initial strategy, then the hospitals are inclined to switch to an accepting strategy as patients coming into the system increases. Our model serves as a stepping stone and may be used for the development of sophisticated models that are applicable in the real-world scenario. Concretely, we aim to create a city-scale simulation of 13 hospitals with known number of facilities and population densities in Saint Petersburg, Russia.

Acknowledgements. This research is financially supported by The Russian Science Foundation, Agreement #14-11-00823.

References

[1] Deo S, Gurvich I. Centralized vs. Decentralized Ambulance Diversion: A Network Perspective. Manage Sci 2011;57:1300–19. doi:10.1287/mnsc.1110.1342.

[2] N Mihal RM. WHEN EMERGENCY ROOMS CLOSE : Ambulance Diversion in the West San Fernando Valley Natasha Mihal Renee Moilanen The Ralph and Goldy Lewis Center for Regional Policy Studies 2005.

[3] McCarter J. Office of the Auditor General of Ontario, 2010 Annual Report 2010:448.

[4] Liz Kowalczyk. State’s ER policy passes checkup - The Boston Globe n.d.

[5] McCain RA, Hamilton R, Linnehan F. Emergency Department Overcrowding as a Nash Equilibrium: Hypothesis and Test by Survey Methodology 2014:1–12.

[6] WHO | Health in 2015: from MDGs to SDGs. WHO 2015.

[7] Kovalchuk S V, Moskalenko MA, Yakovlev AN. Towards Model-based Policy Elaboration on City Scale using Game Theory : Application to Ambulance Dispatching n.d.:1–14.

[8] Lin KY. Decentralized admission control of a queueing system: A game-theoretic model. Nav Res Logist 2003;50:702–18. doi:10.1002/nav.10085.

[9] Kendall DG. Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. Ann Math Stat 1953;24:338–54. doi:10.1214/aoms/1177728975.

[10] Banks J, S. Carson J, L. Nelson B. Discrete event system simulation / Jerry Banks, John S. Carson II, Barry L. Nelson. SERBIULA (Sistema Libr 20) 2018.