Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.

SIMULATING NEW YORK CITY HOSPITAL LOAD BALANCING DURING COVID-19

Enrique Lelo de Larrea
Henry Lam
Elioth Sanabria
Jay Sethuraman

Department of IE&OR
Columbia University
500 West 120th Street
New York, NY 10027, USA

Sevin Mohammadi
Audrey Olivier
Andrew W. Smyth

Department of Civil Eng. & Eng. Mechanics
Columbia University
500 West 120th Street
New York, NY 10027, USA

Edward M. Dolan
Nicholas E. Johnson
Timothy R. Kepler
Afsan Quayyum
Kathleen S. Thomson

Bureau of Management Analysis and Planning
Fire Department, City of New York
9 MetroTech Center
Brooklyn, NY 11201, USA

ABSTRACT

In most emergency medical services (EMS) systems, patients are transported by ambulance to the closest most appropriate hospital. However, in extreme cases, such as the COVID-19 pandemic, this policy may lead to hospital overloading, which can have detrimental effects on patients. To address this concern, we propose an optimization-based, data-driven hospital load balancing approach. The approach finds a trade-off between short transport times for patients that are not high acuity while avoiding hospital overloading. In order to test the new rule, we build a simulation model, tailored for New York City's EMS system. We use historical EMS incident data from the worst weeks of the pandemic as a model input. Our simulation indicates that 911 patient load balancing is beneficial to hospital occupancy rates and is a reasonable rule for non-critical 911 patient transports. The load balancing rule has been recently implemented in New York City's EMS system.

1 INTRODUCTION

In New York City (NYC), the emergency medical services (EMS) system is operated by the city's Fire Department (FDNY). Each year, the system receives approximately 1.5 million medical incident calls, 1.1 million of which result in a patient transport to a hospital. Given the scale of this operation, it is beneficial for the FDNY to rely on quantitative tools, especially when considering the implementation of policy changes. These tools currently include the tracking of several performance metrics via computational dashboards, which get the necessary inputs from the ambulance EMS (computer-aided) dispatch system.

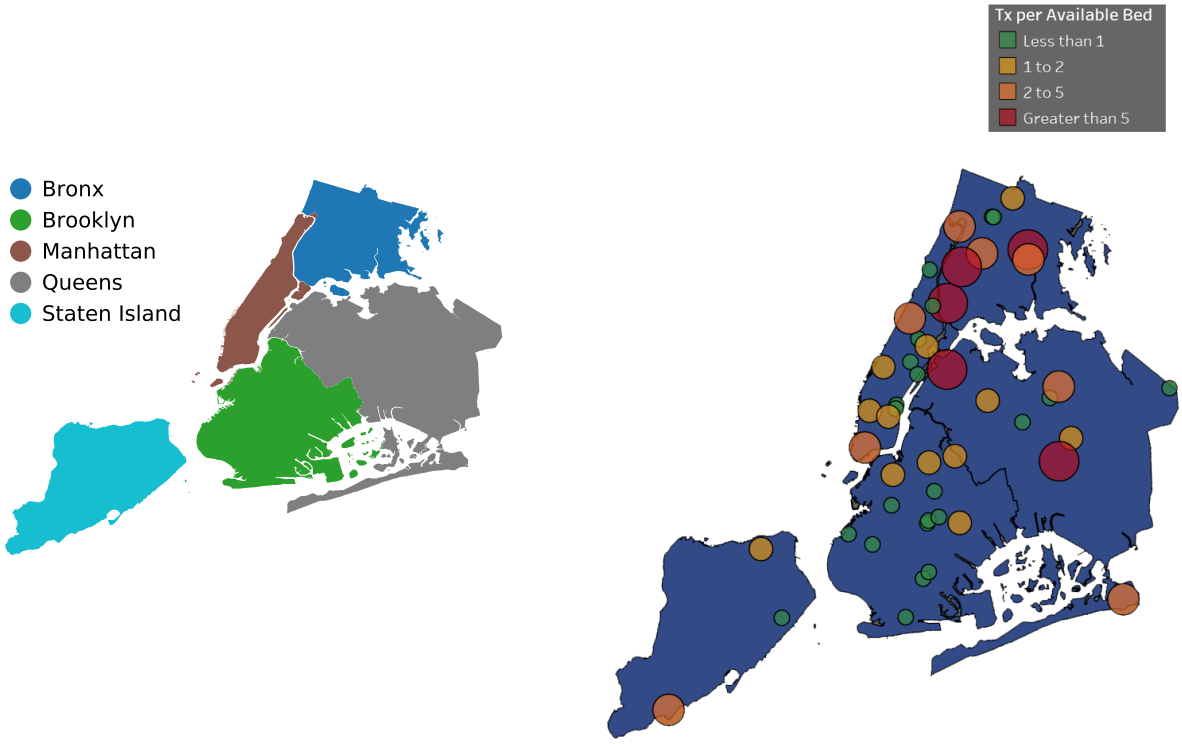

Figure 1: (Left) The 5 boroughs of NYC. (Right) The ratio of number of patient transports to number of available beds for NYC hospitals during the first wave of the COVID-19 pandemic in spring 2020. Larger circles indicate more transports per bed.

This system logs historical details of EMS dispatch operations, and its current configuration is not integrated with real-time data analytics or optimization. Therefore, for testing new policies, simulation and statistical methods need to be developed. In this paper, we describe an in-house EMS simulation model to help the FDNY assess a change in the hospital assignment rule.

The desire for a change in the hospital assignment rule was driven by the COVID-19 pandemic. In normal times, the system recommends to transport patients to the closest most appropriate hospital (closest in terms of time, not distance). In this paper, we refer to this rule as the *closest hospital* rule. However, during the first wave of the pandemic, in March and April 2020, the EMS system experienced a spike in incident calls, and certain hospitals suffered a considerable overload due to both EMS transports and walk-in patients. Two dynamics during this time contributed to an inefficiency for hospital capacity: (1) The downtown core of Manhattan and Brooklyn was vacated of its normal working population and therefore medical incident density reduced. (2) Outer-borough hospitals in residential areas became COVID-19 hotspots and quickly overwhelmed. To illustrate this overload, we plot in Figure 1 the number of EMS patient transports per available hospital bed during one day in April 2020. Larger values are indicators of possible hospital overloading. This is the case for some hospitals in Queens and the Bronx. To address this concern, the FDNY partnered with Columbia University in the late spring to explore an alternative hospital assignment rule which takes into account both the closeness of the hospital and its capacity level in preparation for a second pandemic surge. We refer to this approach as the *load balancing* rule and we assess its impact using the simulation model. Similar results to the ones presented here assisted in the FDNY's decision to actually implement the load balancing rule in response to lessons learned during the spring 2020 pandemic surge.

The use of simulation in the EMS context has a long and rich history. See, for instance, the survey article by Aboueljinane et al. (2013) for a detailed comparison of EMS simulation models in terms of input data, assumptions, and performance metrics. Here, we mention only a few relevant examples. In NYC, for example, Savas (1969) was one of the first to use simulation and a cost analysis to assess the benefits of creating a satellite ambulance station in Brooklyn and dispersing ambulance standby locations across the city (a practice that is still in use today). Ingolfsson et al. (2003) analyze, via simulation, the impact of consolidating all of Edmonton's ambulance stations into a "single start" station, especially in regards to ambulance utilization and ambulance coverage (defined as the proportion of incidents to which an ambulance can respond in less than a specific target time). Henderson and Mason (2005) build a simulation model for the Auckland area (which was later expanded for Melbourne). This model features a complex network-based travel time model, a visualization interface, and the use of trace simulation (historical data) for the incident call process. More recently, for the French Val-de-Marne department, Aboueljinane et al. (2014) evaluate several strategies to improve ambulance coverage. These strategies include adding ambulances in certain shifts, relocating ambulances to new stations, and reducing dispatch times. They also consider a simulation optimization technique to find better ambulance standby locations according to the time of day. Olave-Rojas and Nickel (2021) implement a hybrid simulation model for the EMS system in an area of northern Germany. They use machine learning techniques to predict ambulance average travel speeds.

The work previously mentioned focuses on improving the dispatch side of the EMS operations. This relates to setting ambulance standby locations, selecting the staffing levels, and determining dispatch (ambulance selection) policies. Less attention has been given to hospital selection rules. As mentioned by Aboueljinane et al. (2013), many authors assume the closest hospital rule in their models. That being said, some simulation studies have been done on hospital selection. Wears and Winton (1993) study, in the northeast Florida and southeast Georgia area, the effect of modifying (a) the necessary severity level for a trauma patient to be transported to a specialized hospital (maybe bypassing closer hospitals), and (b) the helicopter dispatch policy. Unlike us, their approach does not take into account the capacity of the hospitals. Wang et al. (2012) analyze patient mortality by comparing twelve hospital selection rules in the aftermath of a single mass casualty incident in Pittsburgh. Some of these rules take into account the capacity of hospitals and the length of waiting queues. This analysis, however, focuses on a single event, not on normal day-to-day operations. Finally, Aringhieri et al. (2018) test several dispatching, routing, and redeployment policies in a simulation scheme. Some of these modify the closest hospital rule by considering waiting times at the hospital. These strategies resemble the current practice of hospital redirection and hospital diversion policies present in NYC's EMS system. Our load balancing rule can work together with both of these existing practices; while redirection and diversion are reactive measures to stabilize hospital capacities, load balancing can be seen as proactive, since it anticipates the behavior during the next day, thereby creating an opportunity to mitigate real-time overload.

The rest of the paper is organized as follows. In Section 2, we describe the main elements of the EMS simulation model. Section 3 presents the simulation-based comparison between the closest hospital and load balancing rules. We conclude in Section 4. For more technical details, Appendix A includes the procedure to calibrate the hospital discharge process and Appendix B has the optimization formulation of the load balancing problem.

2 THE SIMULATION MODEL

The simulation model is a computational program which represents the main objects, agents, and dynamics of an EMS system. Although this model is tailored to the structure, operations, and needs of NYC's EMS system, it can be modified accordingly to other cities or geographies. The code is written in Python, which allows for user-friendly visualizations.

2.1 New York City Geography

For EMS operations, the FDNY divides NYC into nearly 2,400 geographical areas (polygons), also known as atoms. Certain decisions, such as ambulance dispatch to an incident or hospital selection for a transport, are made at the atom level. For instance, two incidents in the same atom, that require a hospital transport with the same (critical) care category, will usually be assigned to the same hospital. Given the relevance of atoms in the EMS system, the simulation model reads the atom geographical information (boundaries, and a user-defined centroid) from an external file. All the virtual objects of the model will be located within the boundaries established by this file (with the exception of some EMS hospitals which are located outside NYC proper).

2.2 EMS Objects

The simulation model is built using the discrete event simulation methodology and has an object-oriented structure. We define several classes representing real-life EMS objects. The objects' attributes can be read by the model from external files. The following is a brief description of the main objects of the model.

Ambulances: also called units (we use either term indistinguishably from now on). There are two types of ambulances: basic life support (BLS) and advanced life support (ALS). BLS units treat low acuity incidents, while ALS units are designated to high acuity ones. For the most critical incidents, two ambulances (one BLS and one ALS) will be dispatched to the scene. Ambulances work in schedules (also known as tours) which are either 8 or 12 hours long. Each unit has an assigned standby position, known as the cross-street location (CSL). Between incident responses, the ambulance remains at its CSL, waiting for the next incident. Once the tour ends, the ambulance goes back to its assigned station, where a crew shift occurs and a new tour begins. Finally, the EMS system includes both ambulances managed either directly by the FDNY (municipal units) or by the hospitals (voluntary units). Both municipal and voluntary units are coordinated for dispatch and hospital transport through NYC's 911 system via the EMS dispatch system. The model contains the information and CSL of around 500 units distributed across the city.

Hospitals: In general, hospitals receive patients via walk-in, inter-facility transport, and 911 transport. In this model however, we only keep track of 911 transports, which are the ones directly related to EMS operations. Each hospital has a different capacity level which evolves through time. On one hand, new patient transports arrive to the hospital, a percentage of which (currently set to 40%) will occupy a bed, thus reducing the hospital capacity. On the other hand, we assume a random hospital discharge process which simulates patient discharges, increasing the hospital capacity. See Section 2.4 for more details on the discharge process. Even when the hospital is at zero capacity, the model allows for new patients to be transported to it. In that case, the hospital becomes overloaded and the hospital's bed occupancy surpasses 100%. This behavior is reflective of real-life situations (see Figure 1). Additionally, not all hospitals have the resources to accept all patient types. For instance, not all hospitals are equipped to treat severe burns or stroke patients. All 911 receiving hospitals in NYC do accept General Emergency Department patients. The model contains the information of more than 60 hospitals that are part of the EMS system (not all of them in NYC proper).

Ambulance Stations: These buildings house FDNY ambulances when they are out of service or between tours. Municipal units are assigned to an ambulance station, whereas voluntary units use their hospital as base. The model contains the information of 40 ambulance stations.

Incidents: After a medical incident is reported to the 911 system, unit(s) will be dispatched based on the location and severity of the incident. Upon arrival, the crew spends time on-scene tending to and evaluating the patient. If the patient needs to go to a hospital, the unit will proceed with the transport. The incident's acuity or severity is assessed in two stages. An initial severity level (coded as 1–8), which in practice is selected by a medically trained call taker utilizing computerized triage, determines the resource type (BLS or ALS) and number of units that are dispatched. Once at the scene, the EMS crew will reassess the situation and, if a hospital transort is necessary, assign a care category to the patient. The atom location

and care category enable the EMS dispatch system to provide a list of recommended hospitals in order of closest to furthest from the incident location. The model allows for the use of either historical or random incidents. Feeding historical incidents into the model is particularly useful when the user wishes to "replay" certain periods of time, while changing certain policies or model inputs. See Section 2.4 for more details on how to simulate random incidents.

Dispatcher: This object represents a fictional "main" agent who acts as the system decision maker. The dispatcher decides which unit(s) will be dispatched to an incoming incident and, if a transport is required, to which hospital the patient will be sent. In real life, the role of the dispatcher is a joint effort between human dispatchers, the EMS dispatch system, and a set of policies.

EMS System: a parent object that contains all of the other EMS objects. It defines a simulation window, sets up an initial configuration for the objects, and starts the simulation clock. During a simulation run, it also gathers several metrics and statistics for posterior analysis.

2.3 Model Dynamics

The main dynamics of the model are standard and similar across different EMS systems; see Figure 2. We summarize them as follows:

1. Incident arrival: A new incident is generated and its location and severity level are informed to the dispatcher.
2. Unit assignment (dispatch): According to the incident characteristics, the dispatcher assigns one or more ambulances to it. As a general rule, low acuity incidents (severity levels 4–8) get assigned to one BLS unit, higher acuity incidents (severity levels 2–3) get assigned to one ALS unit, and the most severe incidents (severity level 1) get assigned to two units (one of each). The dispatcher selects the necessary units according to shortest estimated time of arrival. If no units are available, the incident is pushed into a waiting queue.
3. On-scene treatment: Once the unit arrives to the incident location, the care category of the patient is determined. If the patient does not require to be transported to a hospital, then the unit finishes the job and returns to its CSL.
4. Hospital assignment: If the patient needs to go to a hospital, the dispatcher will proceed to assign an appropriate hospital. The default behavior is to follow the closest hospital rule. This is the hospital with the necessary equipment to treat the incident's care category and with shortest estimated time of arrival from the atom location. For an alternative to the closest hospital rule, we define in Section 3 a load balancing rule that takes into account the citywide hospital capacities.
5. At-hospital procedure: Once the ambulance arrives to the hospital, it delivers the patient and finishes the current job. It travels back to its CSL and becomes available for a new incident assignment (if required, it can be assigned to a new job before reaching the CSL). The patient (with some probability) gets admitted and occupies a bed (effectively lowering the hospital bed capacity by one). Some (random) time later, the patient is discharged and the hospital bed capacity increases by one.

2.4 Sub-models

Besides the main dynamics discussed above, the simulation system requires additional sub-models that will inject the "randomness" into the model. These models address dynamics that are, in their own right, interesting and complex. In this paper, we often opt to use simple models, but the framework is flexible enough so that some or all of these sub-models can be expanded upon by the user to make them more realistic.

Incident generation: While we can feed the simulation model with historical incidents (also known as trace simulation), an incident generation model can also be used to sample random incidents. The current

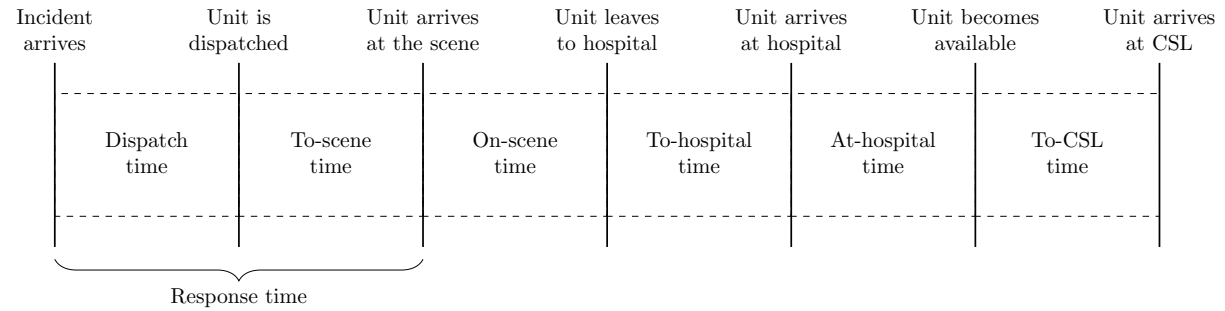

Figure 2: A simplified scheme of the EMS main dynamics.

version of the model uses a simple spatio-temporal model which is uniform in space (in the NYC geography) and follows a non-homogeneous Poisson process for the time component. After the time and location have been simulated, the care category is simulated independently according to user-defined probabilities.

Dispatch time: In practice, emergency medical incidents are not dispatched immediately. Callers to the 911 system are briefly interviewed by a police call-taker and, if the emergency is of medical nature, are then transferred to an EMS call taker. To model the dispatch time, defined as the length of time between when the incident is first received by the EMS call taker to the moment when a unit is assigned to the incident, we assume a multi-server queue with 23 identical call takers and an exponentially-distributed call duration time with mean equal to 3 minutes. These numerical values are close to the ones observed by the FDNY in practice. A more complex model would be a multi-server, priority-based queue. See, for instance, the EMS call center model of van Buuren et al. (2015).

Travel time: The model constantly requires travel times between two locations; for instance, the time for a unit to get to an incident scene, the transport time to the hospital, or the time to go back to a station or CSL. By default, the model generates travel times using the geodesic distance in miles between two locations and by assuming a constant ambulance speed. This simple approach has been used before (Silva and Pinto 2010). We assume that the speed is higher when the unit is on a job (i.e. going to a scene or transporting patients to a hospital). In addition, for travel times between incident locations and hospitals, the model can use an external travel time matrix. If this matrix is properly calibrated with ambulance traffic data, it can improve the fidelity of the simulation model. An important difference between these two approaches is the granularity: geodesic distances can be computed between any two pairs of latitude-longitude coordinates, whereas the travel time matrix uses the incident and hospital atoms as input.

On-scene time: The time that an EMS crew takes to treat the patient can depend on the severity level and care category, among other factors. We currently assume a uniform distribution.

At-hospital time: Once the ambulance arrives to the hospital, it has to wait some time while the hospital personnel admits the patient into the emergency department. This time most likely depends on the saturation level of the hospital. We currently assume a uniform distribution.

Hospital patient discharge: Modeling the hospital-side of an EMS system is particularly challenging, especially from the standpoint of the EMS system manager. Whereas EMS managers have access to data from their dispatchers and ambulances, having precise and frequent feedback from the hospitals is not always a reality. In our setting, we are interested in modeling the rate at which hospitals discharge patients, effectively increasing their capacity. We do so by considering an initial estimated capacity level and a *pure-death* process for the discharge times. We calibrate this model to achieve a long-term "equilibrium" of hospital occupancy during periods where the number of incidents and hospital transports are relatively stable. See Appendix A for more details on the calibration of this model.

Out of service: Ambulances can go out of service for a variety of reasons (such as unit maintenance, staffing issues or mechanical failures). This will partially or totally interrupt a scheduled tour, reducing

ambulance availability. We assume that units can randomly go out of service, with a certain probability, only after finishing a hospital transport.

3 HOSPITAL LOAD BALANCING APPLICATION

We test our simulation model in the context of a new hospital load balancing rule. The standard rule for hospital assignment sends the patient to the closest appropriate hospital. Intuitively, this rule makes sense because the patient is provided with fast medical care and the ambulance is able to finish the current job and report back for duty as soon as possible. This approach, however, does not take into account the capacity level of hospitals.

In the presence of critical external events, such as natural disasters or pandemics, the number of patient transports to a hospital might increase drastically. This could lead to the hospital being overloaded which in turn could compromise patient care. To avoid this scenario, we propose a new hospital assignment rule which attempts to achieve load balancing across the hospital system.

The standard closest hospital rule can be summarized by a function that maps each atom to its closest hospital. The output of the load balancing rule is another function, which still assigns hospitals to atoms, but does so by considering both (a) the distance between atom and hospital and (b) the hospital capacity. The procedure to determine the load balancing function is to solve an integer optimization problem which minimizes the system-wide hospital transport time subject to hospital capacity constraints. See Appendix B for the mathematical formulation of this problem and the details of its inputs.

Solving the load balancing problem requires as an input, among others, an estimate of the daily bed capacity levels of each hospital in the system. In real life, the FDNY has access to daily updates on these values, so we can run the load balancing optimization and update the hospital assignment function each day (the availability and frequency of this data might be different for other cities).

3.1 Simulation Setup

We now conduct a simulation study to assess the impact of implementing the load balancing rule on the EMS system. The simulation setup is as follows. We select a simulation window of 14 days. Since we have access to real-life incident data, we feed into the model all the incidents received from March 25th, 2020 and April 7th, 2020. This period corresponds to some of the weeks with highest incident counts during the COVID-19 epidemic in NYC. For reference, during this period, the system received on average more than 5,500 incidents per day. In contrast, during regular times, this quantity hovers around 4,000. For simplicity, we assume that all incidents requiring a transport had a General Emergency Department care category. In reality, this care category amounts to around 80% of total transports.

When computing travel times, we differentiate between hospital transport times and other travel times. To increase accuracy, for hospital transport times we use a travel time matrix calibrated with historical ambulance data and network analysis. For the rest of the travel times, we use the geodesic model with constant speed of 12 miles per hour for units on a job and 9 miles per hour for off-duty units.

To simulate the evolution of the hospitals' capacity levels, we first assume a city-wide hospital occupancy level of 80%. The total number of beds per hospital is estimated using publicly available data from the New York State Department of Health. The patient discharge process is then calibrated using the procedure described in Appendix A. We assume that in a normal day, there are approximately 2,900 hospital transports (this is assuming an estimated 1.5 million yearly incidents, 70% of which result in a transport).

3.2 Hospital Assignment Rule Comparison

We run two simulations using the same setup described in the previous section. The difference between them is the hospital assignment rule that is assumed. The first simulation uses the closest-hospital rule. The second one uses the load balancing rule, where the hospital assignment per atom is updated at the beginning of each simulated day.

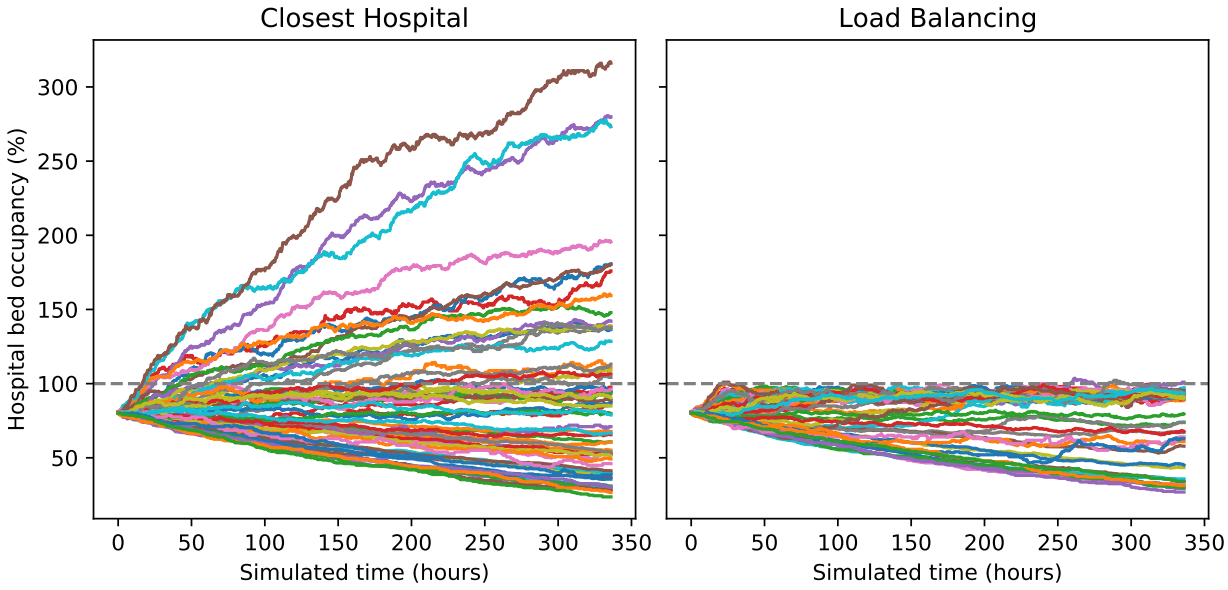

Figure 3: Comparison of the simulated hospital capacities using both hospital assignment rules: closest hospital (left) and load balancing (right). Each line corresponds to the proportion of occupied beds of a NYC hospital. The dashed gray line corresponds to a 100% occupancy level. The data is illustrative only.

Several metrics are of interest when evaluating these rules. The first one is the hospital occupancy level as a function of time. The load balancing rule should keep these levels in check, since it is designed to deviate transports to hospitals which are not saturated. Figure 3 presents a side-by-side comparison of both rules, for one simulation run. When the closest hospital rule is followed (left panel of Figure 3), several hospitals reach their maximum capacity level and become overloaded. This occurs to hospitals that are close to a high number of incidents and do not have the capacity to meet the demand. In contrast, when the load balancing rule is in place, the transports to hospitals are better distributed (right panel of Figure 3). In this scenario there are still hospitals that might reach their maximum capacity levels and might even briefly surpass them. This is due to the randomness of the hospital discharge model and the fact that our incident prediction for the next day is not exact (see Appendix B). Even with these factors into consideration, the load balancing rule is able to correctly keep hospital capacity overloads in check.

The simulated values in Figure 3 are for illustration purposes only and should not be taken as reflective of a real-life situation. In particular, we note that such high occupancy levels would be prevented in real-time via hospital redirection, diversion, and internal load balancing policies. A redirection occurs when the EMS dispatch system detects that a certain hospital shows signs of overloading. For instance, a redirection might be triggered when a certain number of units spend too long stationed in the hospital's emergency department. Similarly, a hospital staff might request a hospital diversion directly to the EMS system, if they consider they are temporarily unable to receive more patients. Finally, a hospital is capable of load balancing itself via inter-facility transports. In this simulation scenario, we do not consider such real-time policies. That being said, in practice, load balancing would work jointly with the other policies to better load balance the hospital system.

Recall that, with the load balancing rule, there is a trade-off between hospital transport time and hospital capacities. Indeed, if ambulances are dispatched to hospitals with capacity, that may not be the closest ones, the overall transport times will inevitably increase. In addition, sending units to farther away hospitals may have an indirect impact on the average incident response time (defined as the time between the incident arrival and the unit arrival at the scene; see Figure 2). If the ambulances are transporting patients away

Table 1: Summary statistics for the hospital transport and unit response times. All the times are in minutes.

	Hospital Transport Time		Response Time	
	Closest Hospital	Load Balancing	Closest Hospital	Load Balancing
mean	9.2	10.8	6.6	6.7
std	3.0	4.4	4.6	4.7
median	8.9	9.9	5.5	5.6
95-percentile	14.3	19.1	15.0	15.4
99-percentile	17.8	23.4	21.9	22.2

from their usual area, they might not be able to respond as fast to new incidents. We explore the impact of imposing the load balancing rule on both metrics (hospital transport and response times) in Table 1. The statistics presented are based on one simulation run for each rule.

In this simulation, the average transport time increases by 1.6 minutes, when we use the load balancing rule. This increase is by no means negligible, but for low acuity patients, it should not be life-threatening. We do see a deterioration in the tail of the distribution; the 99th percentile increases considerably from 17.8 to 23.4 minutes. This reinforces the belief that load balancing should only be used for non-critical transports. If an upper bound on the admissible transport times is required (say, for instance 30 or 40 minutes), it can be imposed by modifying the load balancing optimization problem. On the other hand, ambulance response times were not so adversely affected by the load balancing rule. The mean response time increased by less than half a minute and, although the response times do tend to be larger, the 99th percentile only increased in 0.3 minutes. It should be noted that the underlying travel time model used to compute response times is too simplistic (it is based on the geodesic distance and a constant unit speed), and therefore, the numerical results might not reflect the reality. In this study, however, we are more interested in the difference in the shapes of the distributions and not so much in particular numerical values.

3.3 Main Limitations

The main limitation of the simulation model comes from the simplifying assumptions made on the hospital patient arrival and discharge processes. While they are only part of our entire model, these processes are incredibly complex and deserve their own individual studies. Another limitation comes from the simple geodesic travel time model. We corrected the travel times between locations and hospitals using an external matrix that was calibrated with ambulance traffic data and network analysis, but we did not make the extension to general travel times between two arbitrary locations. Finally, we opted for simple sub-models for dispatch, on-scene, and at-hospital times.

The primary objective of using the simulation model to assess the new load balancing rule was to check the high-level dynamics and detect any potential issues. Due to the urgency of real-life implementation, we have not presented a formal validation of the model. To do the latter task, we first would need to refine all the sub-models mentioned above and calibrate them using real data. All these improvements are important ongoing work.

4 CONCLUSION

The FDNY requires analytical tools to better understand the impact of potential policy changes to their EMS system. To this goal, we implemented an EMS simulation model for NYC. The model captures the main objects and dynamics of the system. Like similar models in the literature, our model can use trace simulation (i.e. using historical data) for the incident arrival process. This allows the user to test several scenarios and changes in policy, subject to incidents observed in the past.

We used the simulation model to evaluate a new hospital load balancing approach, which differs from the standard and widely-used rule of sending patients to the closest hospital. This approach minimizes the

overall patient transport time, but also takes into account the capacity levels of the hospitals. Assuming a particular patient discharge model, we showed in our simulation that the load balancing rule effectively keeps the individual hospital occupancy rates below the at-capacity level. In contrast, using load balancing would result in longer transport times (unit response times were not so clearly affected). This suggest that load balancing might be beneficial for the EMS system as a whole, but it should only be used for patients that are not high acuity.

Simulation is an essential tool when considering the impact of a new policy, especially in critical areas such as EMS. It helps decision makers to quantify and visualize several scenarios, before making changes in real life. Our analysis assisted in the FDNY's decision to implement the load balancing rule in response to the COVID-19 pandemic. The new rule works in addition to preexisting hospital redirection and diversion policies and will hopefully serve as an extra safeguard for balancing system-wide hospital capacity levels.

ACKNOWLEDGMENTS

We are grateful to the members of the Bureau of Management Analysis and Planning at the FDNY for the many enlightening conversations and for providing their expert knowledge and data on the EMS system. We appreciate the guidance of Commissioner Elizabeth Cascio, Chief Lilian Bonsignore, Chief Jonathan Pistilli, Dr. David Prezant, James Saunders, and Matt Talty. We thank Jason Qian for implementing the first version of the simulation model and Yuanlu Bai and Haoxian Chen for providing help on the load balancing optimization. We gratefully acknowledge support from Google and the Tides Foundation under the grant "EMS Resource Deployment Modeling" and the Columbia University Urban Technology Pilot Award.

A HOSPITAL DISCHARGE PROCESS CALIBRATION

Consider a hospital in the EMS system. We model the discharge processes one day at a time (we measure the time in hours). Let $X(t)$, $0 \leq t \leq 24$, denote the number of patients occupying a bed during a given time of the day. For a moment, we assume that there are no incoming patients that day. Then, we can model $X(\cdot)$ as a (decreasing) pure-death process with rate μ. That is, given a patient occupancy of $X(t)$ at time t, the time of the next discharge is distributed as an exponential random variable with rate $X(t)\mu$. In other words, the discharge rate is proportional to the hospital occupancy level. It is a well-known property of this process that, at the end of the day, $\mathbb{E}[X(24) \mid X(0)] = \exp(-24\mu)X(0)$.

We now propose a simple procedure to calibrate μ. Let c be the total capacity of the hospital and let a be the number of patients that are expected to arrive to the hospital during an average (baseline) day. Normally, patients arrive throughout the day, but for this exercise we assume that they arrive at the end of the day. The calibration is based on a bed occupancy equilibrium condition. Simply put, on a regular day, the number of patient discharges roughly matches the number of patient arrivals, and the hospital occupancy level remains constant. Assuming a target occupancy rate of $\rho \in (0,1)$, we have, at the beginning of the day $X(0) = \rho c$, and at the end of the day $X(24) = \exp(-24\mu)\rho c + a$. In equilibrium, we then have $X(24) = X(0)$. Solving for the rate μ, we get $\mu = -(1/24)\log(1 - a/(\rho c))$, as long as $a < \rho c$. With this calibration, we would expect the bed occupancy to remain more or less constant throughout the days, as long as the number of daily arrivals remains close to a. During stress periods when the number of arrivals is much larger than a, we would expect the bed occupancy to increase.

B LOAD BALANCING OPTIMIZATION PROBLEM

Given an EMS system with m atoms and n hospitals, our objective is to assign one hospital to each atom by minimizing the total (estimated) travel time subject to the hospital capacity constraints. The decision variables can be seen as indicators $x_{ij} \in \{0,1\}$, where $x_{ij} = 1$ if and only if hospital j is assigned to atom i.

At the beginning of each simulated day, we gather the following inputs:

- The predicted daily hospital transports originating from atom i for the next day, denoted as $f_i \in \mathbb{Z}_+$. In this application, we naively estimate f_i as the running average of transports of the last $d = 3$ days.
- The current bed availability at hospital $j \in \{1,\ldots,n\}$, denoted as $c_j \in \mathbb{Z}$. A negative c_j indicates that the hospital is overwhelmed and that it currently has $-c_j > 0$ patients above its total capacity.
- The estimated time of arrival from atom i to hospital j, denoted as $T_{ij} \in \mathbb{R}_+$.
- The expected proportion of transported patients that are admitted to the hospital, denoted as $\delta \in (0,1)$. We currently use $\delta = 0.4$.

With these inputs, we solve the following integer optimization problem to update the hospital assignment rule:

$$\begin{aligned}
\underset{x}{\text{minimize}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} f_i T_{ij} x_{ij} \\
\text{subject to} \quad & \sum_{i=1}^{m} \delta f_i x_{ij} \leq \max(c_j, 0), \text{ for all } j, \\
& \sum_{j=1}^{n} x_{ij} = 1, \text{ for all } i, \\
& x_{ij} \in \{0,1\}, \text{ for all } i, j.
\end{aligned}$$

REFERENCES

Abouelji nane, L., E. Sahin, and Z. Jemai. 2013. "A review on simulation models applied to emergency medical service operations". *Computers & Industrial Engineering* 66(4):734–750.

Abouelji nane, L., E. Sahin, Z. Jemai, and J. Marty. 2014. "A simulation study to improve the performance of an emergency medical service: Application to the French Val-de-Marne department". *Simulation Modelling Practice and Theory* 47:46–59.

Aringhieri, R., S. Bocca, L. Casciaro, and D. Duma. 2018. "A Simulation and Online Optimization Approach for the Real-Time Management of Ambulances". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 2554–2565. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Henderson, S. G., and A. J. Mason. 2005. "Ambulance Service Planning: Simulation and Data Visualisation". In *Operations Research and Health Care: A Handbook of Methods and Applications*, edited by M. L. Brandeau, F. Sainfort, and W. P. Pierskalla, 77–102. Boston, MA: Springer US.

Ingolfsson, A., E. Erkut, and S. Budge. 2003. "Simulation of single start station for Edmonton EMS". *Journal of the Operational Research Society* 54(7):736–746.

Olave-Rojas, D., and S. Nickel. 2021. "Modeling a pre-hospital emergency medical service using hybrid simulation and a machine learning approach". *Simulation Modelling Practice and Theory* 109:102302.

Savas, E. S. 1969. "Simulation and Cost-Effectiveness Analysis of New York's Emergency Ambulance Service". *Management Science* 15(12):B608–B627.

Silva, P. M. S., and L. R. Pinto. 2010. "Emergency medical systems analysis by simulation and optimization". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 2422–2432. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

van Buuren, M., G. J. Kommer, R. van der Mei, and S. Bhulai. 2015. "A Simulation Model for Emergency Medical Services Call Centers". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 844–855. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Wang, Y., K. L. Luangkesorn, and L. Shuman. 2012. "Modeling emergency medical response to a mass casualty incident using agent based simulation". *Socio-Economic Planning Sciences* 46(4):281–290.

Wears, R. L., and C. N. Winton. 1993. "Simulation Modeling of Prehospital Trauma Care". In *Proceedings of the 1993 Winter Simulation Conference*, edited by G. W. Evans, M. Mollaghasemi, E. Russel, and W. Biles, 1216–1224. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

ENRIQUE LELO DE LARREA is a Ph.D. student in Operations Research at Columbia University. Before joining Columbia, he worked as a credit risk analyst at BBVA Mexico. He holds a double Bachelor's degree in Applied Mathematics and Actuarial Science from ITAM and a Master's degree in Operations Research from Columbia. His research interests include stochastic simulation, applied probability, and financial engineering. His email address is enrique.lelodelarrea@columbia.edu.

EDWARD DOLAN is the Deputy Commissioner for Strategic Initiatives and Policy at the Fire Department of the City of New York (FDNY). His email address is edward.dolan@fdny.nyc.gov.

NICHOLAS JOHNSON is the Director of Operations Research for the FDNY. Prior to joining the FDNY, he was a Postdoctoral Associate at NYU's Marron Institute of Urban Management where his research focused on modeling real-time urban populations using mobility data. He earned a Ph.D. in Urban Science from the University of Warwick in the United Kingdom and holds a Master's degree from NYU's Interactive Telecommunications Program where he used physical computing and interaction design to explore the impact and pervasiveness of waste streams in urban environments. His email address is nicholas.johnson@fdny.nyc.gov.

TIMOTHY KEPLER is the Director of Data Quality for the FDNY. He has a Master's degree in Public Administration from the City University of New York at Baruch College. For the past five years he has worked as an analyst for the FDNY, working intensively with data from computer aided dispatch systems. His email address is timothy.kepler@fdny.nyc.gov.

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. His research focuses on simulation and optimization under uncertainty. His email address is khl2114@columbia.edu.

SEVIN MOHAMMADI is currently pursuing a Ph.D. degree in Civil Engineering and Engineering Mechanics at Columbia University. Prior to her current position, she obtained a Master's degree in Civil Engineering, majoring in Transportation, from the University of Tennessee Knoxville. Her research interests include data-driven and machine learning approaches in Transportation Engineering. Her email address is sm4894@columbia.edu.

AUDREY OLIVIER is currently an Associate Research Scientist in Civil Engineering at Columbia University, and will be joining the University of Southern California as an Assistant Professor in fall 2021. She holds a Ph.D. in Civil Engineering and Engineering Mechanics from Columbia University and a Diplôme d'Ingénieur from Ecole Centrale de Nantes, France. Her research interests revolve around probabilistic data analytics and physics-based modeling for civil engineering applications. Her email address is audreyol@usc.edu.

AFSAN QUAYYUM is a Data Scientist at the Bureau of Management Analysis and Planning at the FDNY. He has an M.S. in Mathematics and a B.S. in Mathematics with a minor concentration in Applied Physics from New York University. His work focuses on implementing statistical learning techniques for inference and prediction to help EMS and Fire Operations. His email address is afsan.quayyum@fdny.nyc.gov.

ELIOTH SANABRIA is a Ph.D. student in Operations Research at Columbia University. He is primarily interested in the interplay between simulation, machine learning and optimization from a probabilistic point of view, as well as their broad applications in healthcare to improve patients outcomes. His email address is m.elioth@columbia.edu.

JAY SETHURAMAN is a Professor of Industrial Engineering and Operations Research at Columbia University. Currently, he serves as the chair of the IEOR department at Columbia. His research interests are in discrete optimization and applications, game theory, mechanism design, and applied probability. His email address is jay@ieor.columbia.edu.

ANDREW SMYTH is the Carleton Professor of Civil Engineering & Engineering Mechanics and also serves as the Co-Chair of the Smart Cities Center of the Data Science Institute at Columbia University. His research focuses on infrastructure monitoring, dynamic system identification and modeling. He received his Ph.D. in Civil Engineering from the University of Southern California as well as an M.S. in Electrical Engineering, an M.S. from Rice University, and a Sc.B. and A.B. from Brown University. His email address is smyth@civil.columbia.edu.

KATHLEEN THOMSON is the Assistant Commissioner for the Bureau of Management Analysis and Planning at the FDNY. Her email address is kat.thomson@fdny.nyc.gov.