Hindawi
Advances in Operations Research
Volume 2021, Article ID 7120291, 11 pages
https://doi.org/10.1155/2021/7120291

Research Article

An Optimization Model to Address Overcrowding in Emergency Departments Using Patient Transfer

Zeynab Oveysi,[1] **Ronald G. McGarvey**,[1,2] **and Kangwon Seo**[1,3]

[1]*Department of Industrial and Manufacturing Systems Engineering, E3437 Lafferre Hall, University of Missouri, Columbia 65211, Missouri, USA*
[2]*Truman School of Public Affairs, E3437 Lafferre Hall, University of Missouri, Columbia 65211, Missouri, USA*
[3]*Department of Statistics, E3437 Lafferre Hall, University of Missouri, Columbia 65211, Missouri, USA*

Correspondence should be addressed to Ronald G. McGarvey; mcgarveyr@missouri.edu

Received 22 April 2021; Revised 7 July 2021; Accepted 29 July 2021; Published 23 August 2021

Academic Editor: Panagiotis P. Repoussis

Overcrowding of emergency departments (EDs) is a problem that affected many hospitals especially during the response to emergency situations such as pandemics or disasters. Transferring nonemergency patients is one approach that can be utilized to address ED overcrowding. We propose a novel mixed-integer nonlinear programming (MINLP) model that explicitly considers queueing effects to address overcrowding in a network of EDs, via a combination of two decisions: modifying service capacity to EDs and transferring patients between EDs. Computational testing is performed using a Design of Experiments to determine the sensitivity of the MINLP solutions to changes in the various input parameters. Additional computational testing examines the effect of ED size on the number of transfers occurring in the system, identifying an efficient frontier for the tradeoff between system cost (measured as a function of the service capacity and the number of patient transfers) and the systemwide average expected waiting time. Taken together, these results suggest that our optimization model can identify a range of efficient alternatives for healthcare systems designing a network of EDs across multiple hospitals.

1. Introduction

Overcrowding in hospital emergency departments (EDs) is recognized as a serious problem in many countries around the world [1] and affected many hospitals especially during the response to emergency situations such as pandemics or disasters [2]. Overcrowding occurs when the arrival rate of patients exceeds the ED's available capacity [3]. One contributing factor to overcrowding in the USA is a reduced supply of EDs; from 1995 to 2016, although the number of ED visits increased by 51 percent, the number of EDs decreased by 12 percent [4]. Other key factors which cause overcrowding in the USA are the aging population, limited access to medical care from other providers, the safety net, seasonal illness, surgical scheduling, and high utilization of ED for nonemergency care [5,6]. Overcrowding has some negative outcomes for both patients and service providers [7]. Patients may face prolonged pain and long waiting times which leave them unsatisfied [1, 8, 9]. According to the National Hospital Ambulatory Medical Care Survey in 2017 in the United States, the average waiting time at emergency departments for a patient to visit a physician, physician assistant, or nurse is about 40 minutes and around 17 percent of patients waited more than an hour [10]. In fact, up to 10% of patients can become frustrated from long waiting times and may leave the ED without treatment [11], which increases such patients' risk of death or hospital readmission within the next seven days [12]. ED staff frustration is recognized as one of the negative effects of overcrowding on healthcare providers [13].

One potential solution to overcome the overcrowding problem and have quick and high-quality services in EDs is adding more resources to EDs [14]. However, such an approach is limited not only by operating budget constraints but also by limitations on available personnel and the size of the ED facility [14].

Introducing an incentive policy to accident and emergency departments by the UK government in 2000 was another attempt to reduce patient waiting time in such departments [15]. This policy requires 98% of patients to be discharged, transferred, or admitted to inpatient care within 4 hours of their arrival [15]. Large penalties were imposed for failing to meet the 98% target and some hospital managers even lost their jobs for this reason [16]. Gruber et al. showed that this policy reduced patient waiting time by 19 minutes and it also decreased mortality by 14% [17].

Transferring patients has also been discussed in some studies as an option to help address overcrowding [3, 18, 19] and was utilized in some areas facing large numbers of patients due to emergency situations, such as New York City [2]. Nezamoddini and Khasawneh [3] proposed a mathematical model to quantify the effect of transferring patients between hospitals on patients' waiting time in a multihospital system.

In this study, we propose a novel mathematical model to capture the effects of transferring patients between hospitals on patients' waiting time. Similar to Nezamoddini and Khasawneh [3], the objective of our model is to determine the number of servers in each ED and the rates of patient transfer between EDs, in such a way that the cost of the system is minimized. However, unlike [3], who did not explicitly account for queueing effects, our model includes concepts from queueing theory (QT) to account for delays in patients' receiving service due to overcrowding.

The remainder of the paper is organized as follows. Section 2 presents a literature review on research examining ED overcrowding. Section 3 provides our mathematical modeling approach. Section 4 presents the results of numerical testing and sensitivity analysis. Section 5 provides a conclusion and suggestions for future work.

2. Literature Review

Overcrowding in EDs has been recognized as a problem for many years. Various solutions and methods have been applied to improve patient flow in EDs. In many operations research studies examining the overcrowding problem in EDs, the main question was how many resources should be allocated to each queue in an ED or to each hospital in a multihospital system, to reduce the patients' waiting times. Some researchers have found that an optimized manpower allocation can reduce the patients' waiting time in ED by up to 20% [3, 20]. Daldoul et al. [5] proposed a stochastic mixed-integer linear programming (MILP) model to optimize the number of staff and beds in each queue (six queues for six main activities) in an ED to minimize patients' waiting time. El-Rifai et al. [21] also proposed a stochastic MIP model to find the optimal number of personnel for each shift to minimize patients' waiting time. Izady and Worthington [15] proposed a heuristic algorithm that combined queueing and simulation models to determine the required number of each type of medical staff during each "staffing interval" to meet a 4-hour sojourn time target (98% of patients must be discharged, transferred, or admitted to inpatient care within 4 hours of arrival), where a "staffing interval" is the time interval utilized for analysis, during which the number of staff is constant (a representative time might be one or two hours in an ED) [15]. Izady and Worthington [15] applied their method in a generic ED and showed that significant improvement with respect to this target can be made even without an increase in total staff hours. Sinreich et al. [22] introduced two iterative heuristic algorithms, which combined simulation and optimization models for scheduling the work shifts of the ED medical staff. These authors' algorithm shifted the resource capacity from low-demand hours to peak demand hours, and as a result, there was a significant reduction in patient waiting time as well as the peak utilization values of the ED medical staff.

Some studies examined the effects of transferring patients between hospitals in a multihospital setting [3, 18]. Nezamoddini and Khasawneh [3] found that transferring patients between hospitals can be an effective way to reduce patients' waiting time. They used the concept of a capacitated network to model a multihospital system and allowed the nonemergency patients to be transferred between hospitals subject to capacity constraints on the maximum number of transfers allowed per unit time. Soni [18] developed rule-based patient transfer protocols and tested the protocols in a multihospital patient flow simulation model and found that effective patient transfer protocols can optimize the patient flow in a hospital system.

Regarding the solution techniques utilized, some researchers have used simulation models to capture the complexity and dynamic nature of processes in EDs. Cabrera et al. [14] used an agent-based simulation to model EDs. They concluded that although their simulation experiments helped to generate a better understanding of the problem, they were time-consuming even for a small problem. Hung and Kissoon [19] used discrete-event simulation (DES) to evaluate the effect of using an Observation Unit (OU) and patient transfer to other inpatient units on overcrowding in a pediatric emergency department (PED). They considered four scenarios representing combinations of regular PED operations with and without a five-bed OU and transfer mandate. They concluded that a combination of an OU and patient transfer mandate improved the waiting time compared to PED with neither an OU nor a transfer mandate. Moreover, their results showed that the simulated OU without transfer mandate had an occupancy rate of 73.1%; this rate dropped to 48.1% by applying the transfer policy, indicating a significant improvement in the occupancy rate of OU. Gul et al. [23] also used DES to analyze the effect of the patient surge in EDs after an earthquake. They first used Artificial Neural Networks (ANNs) to estimate earthquake causalities and generate inputs for the DES model. Then, the DES model used the ANN outputs to simulate a network of EDs and generate performance outputs for the corresponding EDs. After constructing the simulation model, a Design of Experiments (DOE) was conducted to assess the effects of different factors on the LoS in the ED and the utilization of ED resources. To show their framework, Gul et al. [23] used a network with five EDs located in one of the regions with the highest estimated injury rate after an

earthquake in Istanbul, Turkey. The results from their study can be helpful for planning for the expected earthquake in Istanbul.

Some researchers have combined QT concepts and simulation to analyze patient flow [24, 25]. Alavi-Moghaddam et al. [26] showed that by using QT analysis (with discrete-event simulation to model and validate patient flow metrics), one can identify solutions that improve patients' flow and reduce waiting times in EDs. Hu et al. [7] compared the use of QT with discrete-event simulation in modeling EDs. They reported that QT models had lesser data requirements and computational cost, due to QT models' tendency to simplify the problems, while simulation models captured more details in systems but were more sensitive to changes of parameters. Thus, they suggested that a combination of both was the ideal approach to model such problems.

3. Modeling

Our model attempts to reduce the negative impacts of ED overcrowding in a multihospital system by making optimal allocation decisions in two areas: (1) the number of servers at each hospital's ED and (2) the rate of nonemergency patient transfers between hospitals. To capture the nonlinear queueing effects, we utilize an approach based on that of [27], which allows for an MILP model to represent each ED as an M/M/C queueing system. Our research extends the model of [27] in that it allows for each queueing system (ED) to potentially transfer some patients to other EDs, which requires that we utilize a mixed-integer nonlinear programming (MINLP) model. Figure 1 presents such a notional three-ED system, showing the arrivals of patients into the system and transfers of patients between EDs.

The sets and indices, data parameters, and decision variables used in the MINLP model are as follows:

Sets and indices

(i) I: set of EDs, indexed by i
(ii) M: set of values considered for a number of servers, indexed by m
(iii) N: set of values considered for server utilization, indexed by n
(iv) K: set of patient types, indexed by k, where $k = 1$ denotes emergency, $k = 2$ denotes nonemergency
(v) T: set of time periods, indexed by t

Data parameters

(i) ζ_m: number of servers associated with the set element m
(ii) κ_n: server utilization associated with the set element n
(iii) α_i: cost per unit service capacity at ED i
(iv) $\gamma_{i\bar{i}}$: cost per patient transferred ED i to ED $\bar{i}$
(v) β_i: waiting penalty cost, per unit time spent waiting (time in queue plus time in transfer), for patients treated at ED i
(vi) δ_{ik}: queueing penalty cost, per patient type k in a queue, at ED i

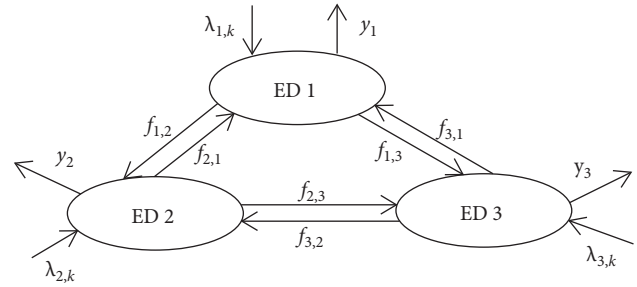

Figure 1: Three-ED network.

(v) μ_i: service rate at ED i
(vi) λ_{ik}: arrival rate, from outside of the system, of patient type k at ED i
(vii) $\theta_{i\bar{i}}$: travel time required to transfer a patient from ED i to ED $\bar{i}$
(viii) π_i: maximum number of patients allowed to be transferred from ED i
(ix) η_i: total budget available for servers at ED i
(x) ϕ_{mn}: expected waiting time in queue for an ED having ζ_m servers operating at a utilization rate κ_n

Decision variables

(i) v_i: expected number of patients in a queue awaiting service at ED i
(ii) r_{ik}: expected number of patient type k in a queue awaiting service at ED i
(iii) w_i: expected time in queue per patient treated at ED i
(iv) z_i: expected waiting time in system (time in queue plus time in transfer) per patient treated at ED i
(v) p_i: maximum utilization allowed at ED i
(vi) s_i: number of servers at ED i
(vii) y_i: effective arrival rate of patients into ED i
(viii) $f_{i\bar{i}}$: rate at which nonemergency patients are transferred from EDi to ED$\bar{i}$; note $f_{ii} = 0$ by assumption

(ix) $x_{mni} = \begin{cases} 1 & \text{if ED } i \text{ operates with } \zeta_m \text{ servers at utilization rate } \kappa_n \\ 0 & \text{otherwise} \end{cases}$

Note that this model makes the following assumptions:

(i) The patient interarrival times follow an exponential distribution
(ii) All patients who enter the system from the outside must be treated at some ED
(iii) Due to the potential risks of transferring emergency patients such as heart rate changes, increased intracranial pressure, and respiratory rate changes [28], it is assumed that they are admitted immediately after arrival to the ED. However, nonemergency patients can be transferred between EDs
(iv) The service time per patient follows an exponential distribution, and to simplify the model, it is not differentiated by patient type. However, it can be differentiated by patient type for future research

(v) Each patient departs the system following service.

Objective function

$$\text{Min} \sum_i \alpha_i s_i + \sum_i \sum_{\bar{i}} \gamma_{i\bar{i}} f_{i\bar{i}} + \sum_i \beta_i z_i + \sum_i \sum_k \delta_{ik} r_{ik}. \quad (1)$$

Objective function (1) minimizes the total system cost, defined as the sum of each ED's service capacity cost, the cost of transferring patients between EDs, along with penalty costs associated with the average waiting time per patient at each ED, and the average number of patients waiting in a queue at each ED.

Constraints

$$\sum_m \sum_n x_{mni} = 1, \quad \forall i, \quad (2)$$

$$\sum_m \sum_n \zeta_m x_{mni} = s_i, \quad \forall i, \quad (3)$$

$$\sum_m \sum_n \kappa_n x_{mni} = p_i, \quad \forall i. \quad (4)$$

Constraints (2)–(4) assign a unique number of servers and utilization levels to each ED.

$$w_i = \sum_m \sum_n \phi_{mn} x_{mni}, \quad \forall i. \quad (5)$$

Constraint (5) calculates the expected waiting time in queue for patients at ED i, based on the M/M/C queueing system with ζ_m servers and utilization level of κ_n. Note that this can be computed *a priori* for all pairs (ζ_m, κ_n) utilizing the standard M/M/C formulae.

$$\left[\sum_m \sum_n \zeta_m \kappa_n x_{mni}\right] \mu_i \geq y_i, \quad \forall i. \quad (6)$$

Constraint (6) ensures that the utilization level at ED i does not exceed p_i.

$$y_i = \sum_k \lambda_{ik} + \sum_{\bar{i}} f_{\bar{i}i} - \sum_{\bar{i}} f_{i\bar{i}}, \quad \forall i. \quad (7)$$

Constraint (7) computes the effective arrival rate into EDi, comprised of both patients arriving into ED i from outside of the system, and the net patients transferred into ED i.

$$v_i = y_i w_i, \quad \forall i, \quad (8)$$

$$r_{i1} = \frac{\lambda_{i1}}{y_i} * v_i, \quad \forall i, \quad (9)$$

$$r_{i2} = \frac{y_i - \lambda_{i1}}{y_i} * v_i, \quad \forall i. \quad (10)$$

Constraints (8)–(10) compute the total number of patients in the queue and then disaggregate this into the number of emergency patients and nonemergency patients in the queue, respectively.

$$\sum_{\bar{i}} f_{i\bar{i}} \leq \lambda_{i2}, \quad \forall i. \quad (11)$$

Constraint (11) allows only nonemergency patients to be transferred between EDs.

$$f_{i\bar{i}} \leq \pi_i, \quad \forall i, \forall \bar{i}. \quad (12)$$

Constraint (12) permits at most π_i patients to be transferred from ED i to ED i'.

$$\alpha_i s_i \leq \eta_i, \quad \forall i. \quad (13)$$

Constraint (13) limits the total cost for servers at ED i to not exceed η_i.

$$z_i = \frac{\left(\left[\sum_{\bar{i}} f_{\bar{i}i}\left(\theta_{\bar{i}i} + w_i\right)\right] + \left[\left(\sum_k \lambda_{ik} - \sum_{\bar{i}} f_{i\bar{i}}\right) w_i\right]\right)}{y_i}, \quad \forall i. \quad (14)$$

Constraint (14) computes the expected waiting time in the system (time in queue plus time in transfer) per patient treated at ED i.

$$v_i, r_{ik}, w_i, z_i, p_i, s_i, y_i, f_{i\bar{i}} \geq 0, \quad \forall i, \forall \bar{i}, \forall k,$$
$$x_{mni} \in \{0, 1\}, \quad \forall m, \forall n, \forall i. \quad (15)$$

4. Experimental Results

4.1. Example Problem. Consider the following example problem, similar in many respects to that presented in [3]. It is assumed that there are three emergency departments in the system, each having identical arrival rates of 5 and 5.5 emergency and nonemergency patients per hour, respectively. Each unit of service capacity costs \$30 per hour. Transferring one patient between any pair of EDs costs \$10 and takes 0.25 hours. The service rate at each ED is 0.5 patients per hour. A penalty cost of \$2 per hour is assumed for patient waiting time in the system (time in queue plus time in transfer). Penalty costs are also incurred based on the average number of patients waiting in a queue at each ED, at a cost of \$5 and \$2 per emergency and nonemergency patient, respectively. Table 1 presents the sets of utilization values and the number of servers considered for each ED. The available budget for servers at each ED is assumed to be \$1700, which is greater than the expense incurred if the maximum number of servers (56, from Table 1) was selected.

4.2. Computational Results. The mathematical model presented in Section 3 was coded in the GAMS 27.2.0 modeling environment and solved using the MINLP solver SCIP 27.2.0. The optimal solution has an objective function value of \$2,195. No patients are transferred between EDs in this optimal solution. Table 2 presents the optimal values for the decision variables p_i, s_i, w_i, z_i, and r_{ik} at each ED; note that these values are identical at each ED in this solution.

Table 1: Sets of utilization values and number of servers considered for each ED.

ζ_m	κ_n
20	0.60
24	0.64
28	0.68
32	0.72
36	0.76
40	0.80
44	0.84
48	0.88
52	0.92
56	0.96

Table 2: Optimal variable values for each ED.

ED i	p_i	s_i	w_i	z_i	r_{i1}	r_{i2}
1	0.88	24	0.307	0.307	1.535	1.689
2	0.88	24	0.307	0.307	1.535	1.689
3	0.88	24	0.307	0.307	1.535	1.689

4.3. Sensitivity Analyses. To determine the effects of the various input parameters on the optimal solutions obtained by our MINLP, a Design of Experiments (DOE) was conducted; all statistical analyses were performed utilizing Minitab 17. In this DOE, input parameters were varied at only one emergency department (denoted ED1), and all parameters at the other two EDs remained unchanged from their previously tested baseline values, with one exception: values η_2 and η_3 were set equal to \$1400, such that at the assumed value of $\alpha_2 = \alpha_3 = \$30/$ server hour, up to 44 servers would be feasible at each of ED2 and ED3. In total, ten input parameters were examined in this DOE, with a resolution V fractional factorial design (2_V^{10-3}) utilized for screening, using a single replicate for each point and zero center points. Table 3 presents the high and low levels tested for each input parameter in this DOE for ED1 (the values for the other two EDs correspond to the center point of the values in Table 3). For each of these 128 experiments, the MINLP model was solved using GAMS/SCIP to obtain the optimal values for all decision variables. Appendices A and B present the designs and responses, respectively, for these 128 experiments.

The following responses were tracked with respect to ED1: $s_1, r_{11}, r_{12}, w_1, z_1, p_1$, and the number of patients transferred from and to ED1 ($f_{12} + f_{13}$ and $f_{21} + f_{31}$, respectively). The regression model specification considered all potential first and two factor interaction terms. The regression model selection was performed using a stepwise procedure, with the p value threshold to enter and depart the model set equal to 0.05, with the necessary first-order terms retained to produce a hierarchical model. Appendices A, B, and C (see Supplementary Materials) present the fractional factorial designs, table of coded coefficients, and significant main effects and interaction terms for all responses. Appendix D presents a table of significant factors for each response containing the level of significance and the direction of effects, along with plots of significant two-way interactions (see Supplementary Materials, Figures D.1–D.8).

Table 3: DOE design factors and their levels.

Factors	Units	Levels	
		−1	+1
α_1	\$/server hour	0	60
$\gamma_{1\bar{i}}$	\$/patient	0	20
β_1	\$/hour	0	4
δ_{11}	\$/patient	0	10
δ_{12}	\$/patient	0	4
μ_1	Patient(s)/hour	0.36	0.65
λ_{11}	Patient(s)/hour	1	9
λ_{12}	Patient(s)/hour	1	10
$\theta_{1\bar{i}}$	Hour	0	0.5
η_1	\$/hour	1920	3360

The remainder of this section presents a detailed examination of two responses of particular importance to ED overcrowding: z_1 (expected waiting time in queue plus time in a transfer per patient treated at ED1) and $f_{21} + f_{31}$ (number of patients transferred to ED1).

4.3.1. Sensitivity Analysis for z_1. Consider the response z_1, the expected waiting time in queue plus time in a transfer per patient treated at ED1. The stepwise regression procedure described above returned the regression model (in uncoded units) presented in equation (17); this regression model had an adjusted R-squared value of 71%. Table 4 presents statistics on this (coded) regression model's coefficients. According to this analysis, there are seven main effects and nine interaction terms significant at the $p = 0.05$ level (factors $\gamma_{1\bar{i}}$ and λ_{12}, while not significant individually, are included to retain a hierarchical model, since they appear in statistically significant interaction terms). Three of these main effects, α_1, $\theta_{1\bar{i}}$, and μ_1, are significant at the $p = 0.001$ level, indicating that the expected waiting time plus time in the transfer is impacted considerably by changes to the cost per unit service capacity and the travel time between EDs (with time in system increasing as each of these parameters increases) and to the service rate (with time in system decreasing as this parameter increases). Figure D.5 in the Supplementary Materials presents interaction plots for the nine significant interaction terms. Observe that three interaction terms are significant at the $p = 0.001$ level, namely, $\alpha_1 * \mu_1$, $\theta_{1\bar{i}} * \lambda_{11}$, and $\theta_{1\bar{i}} * \lambda_{12}$. The latter two of these interaction terms somewhat mediate the effects of the travel time on the expected time in the system; on average, the reduced level of the arrival rate of patients from outside of the system accelerates the increase of the expected time in the system when the travel time increases. This would only be reasonable if this increased arrival rate of patients from outside of the system is impacting the likelihood of patient transfers between EDs, which will be examined next.

Table 4: Coded regression model coefficients.

Term	Effect	Coef.	SE coef.	T-value	p value	VIF
Constant	—	0.18861	0.00766	24.61	≤0.001	—
$\gamma_{1\bar{i}}$	0.02225	0.01113	0.00766	1.45	0.149	1.00
α_1	0.14602	0.07301	0.00766	9.53	≤0.001	1.00
β_1	−0.03066	−0.01533	0.00766	−2.00	0.048	1.00
δ_{11}	−0.04873	−0.02437	0.00766	−3.18	0.002	1.00
δ_{12}	−0.04056	−0.02028	0.00766	−2.65	0.009	1.00
$\theta_{1\bar{i}}$	0.12642	0.06321	0.00766	8.25	≤0.001	1.00
μ_1	−0.09382	−0.04691	0.00766	−6.12	≤0.001	1.00
λ_{11}	−0.04142	−0.02071	0.00766	−2.70	0.008	1.00
λ_{12}	0.00740	0.00370	0.00766	0.48	0.630	1.00
$\gamma_{1\bar{i}} * \delta_{12}$	0.03588	0.01794	0.00766	2.34	0.021	1.00
$\alpha_1 * \theta_{1\bar{i}}$	−0.04605	−0.02303	0.00766	−3.01	0.003	1.00
$\alpha_1 * \mu_1$	−0.05341	−0.02670	0.00766	−3.48	0.001	1.00
$\alpha_1 * \lambda_{11}$	−0.03084	−0.01542	0.00766	−2.01	0.047	1.00
$\alpha_1 * \lambda_{12}$	0.03736	0.01868	0.00766	2.44	0.016	1.00
$\beta_1 * \delta_{11}$	0.03866	0.01933	0.00766	2.52	0.013	1.00
$\theta_{1\bar{i}} * \lambda_{11}$	−0.06884	−0.03442	0.00766	−4.49	≤0.001	1.00
$\theta_{1\bar{i}} * \lambda_{12}$	−0.09034	−0.04517	0.00766	−5.89	≤0.001	1.00
$\lambda_{11} * \lambda_{12}$	0.03320	0.01660	0.00766	2.17	0.032	1.00

$$\begin{aligned} z_1 = {} & 0.1379 - 0.00068\gamma_{1\bar{i}} + 0.00618\alpha_1 - 0.01733\beta_1 - 0.00874\delta_{11} - 0.01911\delta_{12} + 0.7379\theta_{1\bar{i}} - 0.1394\mu_1 \\ & + 0.00221\lambda_{11} + 0.00210\lambda_{12} + 0.000897\gamma_{1\bar{i}} * \delta_{12} - 0.00307\alpha_1 * \theta_{1\bar{i}} - 0.00614\alpha_1 * \mu_1 - 0.000128\alpha_1 * \lambda_{11} \\ & + 0.000138\alpha_1 * \lambda_{12} + 0.001933\beta_1 * \delta_{11} - 0.03442\theta_{1\bar{i}} * \lambda_{11} - 0.04015\theta_{1\bar{i}} * \lambda_{12} \\ & + 0.000922\lambda_{11} * \lambda_{12}. \end{aligned} \tag{16}$$

4.3.2. Sensitivity Analysis for $f_{21} + f_{31}$. Consider the response $f_{21} + f_{31}$, the number of patients transferred into ED1. The stepwise regression procedure described above returned the regression model presented in equation (17); this regression model had an adjusted R-squared value of 78%. Table 5 presents statistics on this (coded) regression model's coefficients. According to this analysis, there are four main effects and six interaction terms significant at the $p = 0.05$ level. Each main effect is significant at the $p = 0.001$ level, indicating that the number of patients transferred into ED1 is impacted considerably by changes to the cost per unit service capacity and the arrival rate of both emergency and nonemergency patients from outside of the system (with the number of transferred patients decreasing as each of these parameters increases) and to the service rate (with the number of transferred patients increasing as this parameter increases). Figure D.8 in the Supplementary Materials presents interaction plots for the six significant interaction terms. Observe that the interaction terms $\alpha_1 * \lambda_{11}$ and $\alpha_1 * \lambda_{12}$ all magnify the main effects of these individual terms, with even greater decreases in the number of patients transferred into ED1 when either pair of these parameters are jointly increased. In aggregate, an increase in the arrival rate of emergency or nonemergency patients into ED1 from outside the system is associated with a decreased number of patients transferred into ED1, which partly explains the interaction effect discussed in the previous section, in which the reduced level of the arrival rate of patients from outside of the system accelerates the increase of the expected time in the system when the travel time increases. Recall that z_1, the expected waiting time in queue plus time in a transfer per patient treated at ED1, does not account for the time in the system spent by patients transferred from ED1 to other EDs; the only transfer time that it accounts for is that of patients transferred into ED1.

$$\begin{aligned} f_{21} + f_{31} = {} & 5.367 - 0.0036\alpha_1 - 0.22\mu_1 - 0.4009\lambda_{11} - 0.3273\lambda_{12} - 0.0427\alpha_1 * \mu_1 \\ & - 0.001712\alpha_1 * \lambda_{11} - 0.001811\alpha_1 * \lambda_{12} + 0.422\mu_1 * \lambda_{11} + 0.315\mu_1 * \lambda_{12} + 0.01032\lambda_{11} * \lambda_{12}. \end{aligned} \tag{17}$$

4.4. Sensitivity Analyses on ED Size. To assess the effect of ED size on the number of transfers occurring in the system, a sensitivity analysis was performed examining three EDs of different sizes. The large ED has arrival rates of 10 and 11 emergency and nonemergency patients per time unit, respectively. The medium ED has arrival rates of 5 and 5.5

Table 5: Coded regression model coefficients.

Term	Effect	Coef.	SE coef.	T-value	p value	VIF
Constant	—	2.3642	0.0849	27.85	≤0.001	
α_1	−2.6216	−1.3108	0.0849	−15.44	≤0.001	1.00
μ_1	0.6784	0.3392	0.0849	4.00	≤0.001	1.00
λ_{11}	−1.4609	−0.7305	0.0849	−8.60	≤0.001	1.00
λ_{12}	−1.5391	−0.7695	0.0849	−9.06	≤0.001	1.00
$\alpha_1 * \mu_1$	−0.3716	−0.1858	0.0849	−2.19	0.031	1.00
$\alpha_1 * \lambda_{11}$	−0.4109	−0.2055	0.0849	−2.42	0.017	1.00
$\alpha_1 * \lambda_{12}$	−0.4891	−0.2445	0.0849	−2.88	0.005	1.00
$\mu_1 * \lambda_{11}$	0.4891	0.2445	0.0849	2.88	0.005	1.00
$\mu_1 * \lambda_{12}$	0.4109	0.2055	0.0849	2.42	0.017	1.00
$\lambda_{11} * \lambda_{12}$	0.3716	0.1858	0.0849	2.19	0.031	1.00

emergency and nonemergency patients per time unit, respectively. The small ED has arrival rates of 3.75 and 4.125 emergency and nonemergency patients per time unit, respectively. The optimization model is modified slightly here, to include only constraints (2)–(7), (11), and (14). The objective function is modified as represented in equation (18), deleting the final two summation penalty terms from objective (1). Rather than associating a financial penalty with delay times, we introduce a new constraint (19) which imposes an upper bound, denoted by σ, on the systemwide average expected waiting time in queue plus time in transfer, which can be computed as $\sum_i y_i z_i / \sum_i \sum_k \lambda_{ik}$. We varied this upper bound σ across a range of values, from a minimum value of 0.0284 to a maximum value of 1.3913 (the systemwide average for the minimum cost solution if constraint (19) is not considered). In total, 26 different solutions were identified, constituting an efficient frontier for the tradeoff between objective function (18) and the left-hand side of constraint (19). All parameters were assumed to take the baseline values from Section 4.1 with two exceptions: we assume that the cost per unit service capacity at each ED is equal to 10 times the cost per patient transferred between EDs, say, \$10 and \$1, respectively. The potential numbers of servers considered at each ED were also modified from the values presented in Table 1; for this sensitivity analysis, ζ_m was varied to include all integer values between 2 and 60. Table 6 presents the sensitivity analysis' objective values. As it can be seen, the objective value decreases as the upper bound value increases. In fact, it implies that as the average expected waiting time in queue plus time in transfer in the system becomes more flexible, a fewer number of servers and fewer patient transfers are required in EDs. Therefore, the associated costs (equation (18)) decrease. The following figures present the sensitivity analysis' expected waiting time in queue plus time in a transfer per patient treated (Figure 2), number of servers (Figure 3), ED utilization (Figure 4), and percent of nonemergency patients transferred (Figure 5).

$$\text{Min} \sum_i \alpha_i s_i + \sum_{\bar{i}} \sum_i \gamma_{\bar{i}i} f_{\bar{i}i}, \tag{18}$$

$$\frac{\sum_i y_i z_i}{\sum_i \sum_k \lambda_{ik}} = \frac{\sum_{\bar{i}} \left(\left[\sum_{\bar{i}} f_{\bar{i}i} \left(\theta_{\bar{i}i} + w_i \right) \right] + \left[\left(\sum_k \lambda_{ik} - \sum_{\bar{i}} f_{\bar{i}i} \right) w_i \right] \right)}{\sum_i \sum_k \lambda_{ik}} \leq \sigma. \tag{19}$$

Table 6: Objective values.

Solution #	Objective value
1	1020.24
2	970.10
3	960.00
4	940.00
5	921.87
6	920.00
7	910.00
8	900.00
9	890.68
10	881.53
11	880.00
12	873.87
13	871.15
14	870.00
15	863.91
16	860.05
17	854.91
18	852.97
19	851.16
20	845.39
21	842.60
22	840.05
23	837.79
24	833.03
25	830.05
26	830.00

These results demonstrate how the optimization model utilizes a variety of strategies to achieve a constrained systemwide average expected waiting time at minimum cost. Consider, for example, solutions 22 and 23. They achieve relatively similar performance, with respective objective function values of 840.06 and 837.79 and respective systemwide average expected waiting times of 0.7988 and 0.8733. The utilization at each ED is essentially unchanged across solutions 22 and 23, with 96%, 92%, and 92% utilization, respectively, at the large, medium, and small ED in each solution. However, the underlying structure has changed significantly, with solution 22 utilizing 44, 23, and 17 servers, respectively, at the large, medium, and small EDs, and very little patient transfer (1% of the nonemergency patients transferred from the small ED to each of the

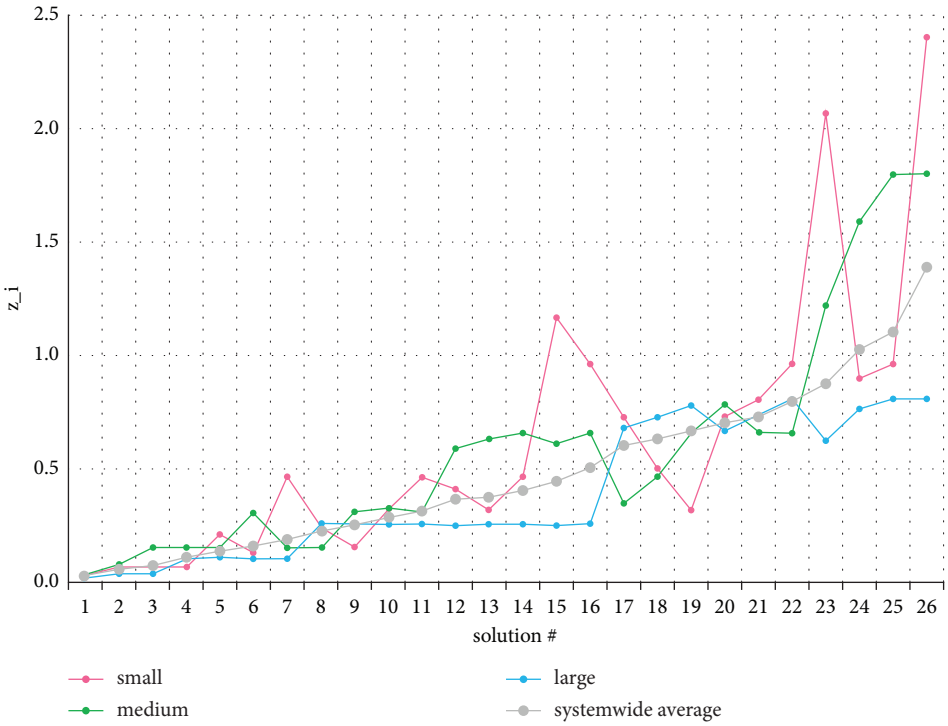

Figure 2: Expected waiting time in queue plus time in a transfer per patient treated.

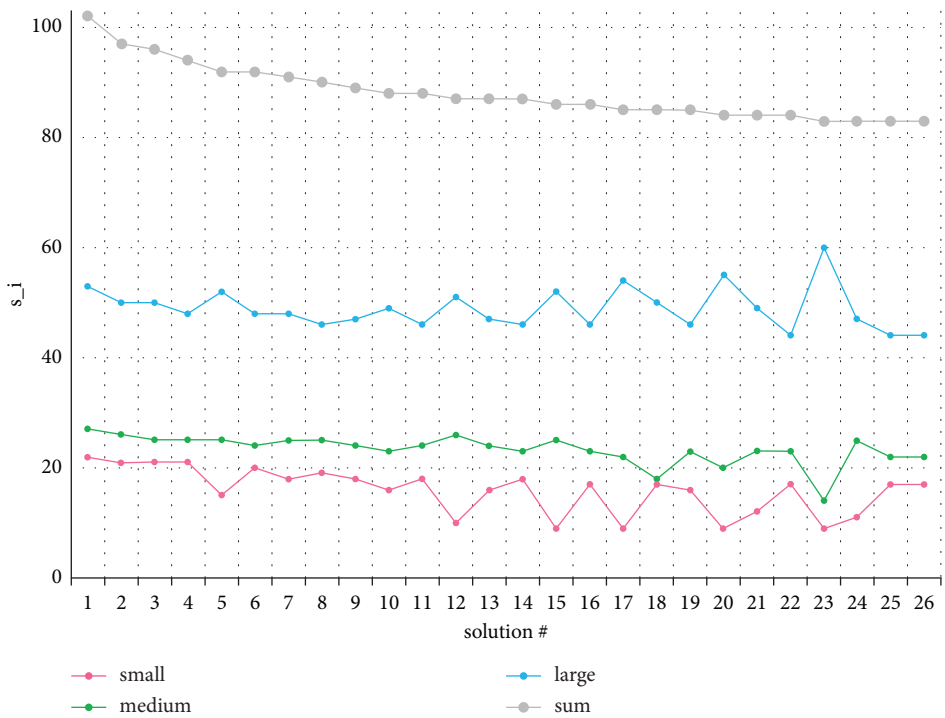

Figure 3: Number of servers.

medium and large EDs). By contrast, solution 23 utilizes 60, 14, and 9 servers, respectively, at the large, medium, and small EDs (one fewer server, in total, than does solution 22), but extensive patient transfer (91% and 74% of the nonemergency patients from the small and medium EDs, respectively, are transferred to the large ED).

Across all 26 solutions identified, the optimization model utilized patient transfer extensively for nonemergency patients arriving at the small ED; on average, 25.3% and 4.4% of such patients were transferred to the large and medium EDs, respectively. The patient transfer was utilized less frequently for nonemergency patients arriving at the medium ED; on average,

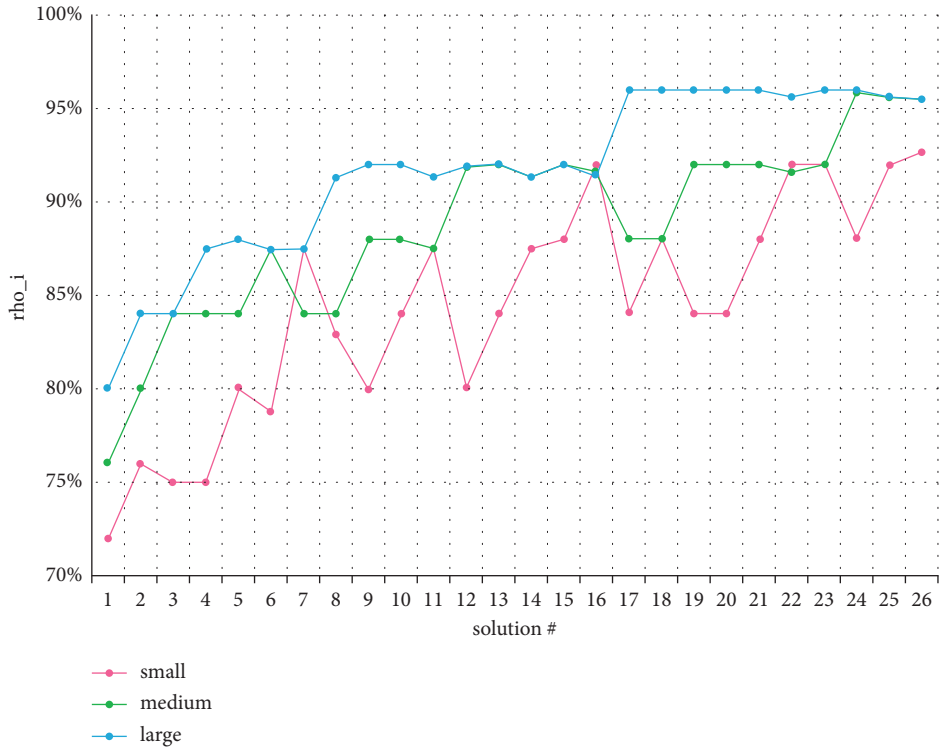

Figure 4: ED utilization.

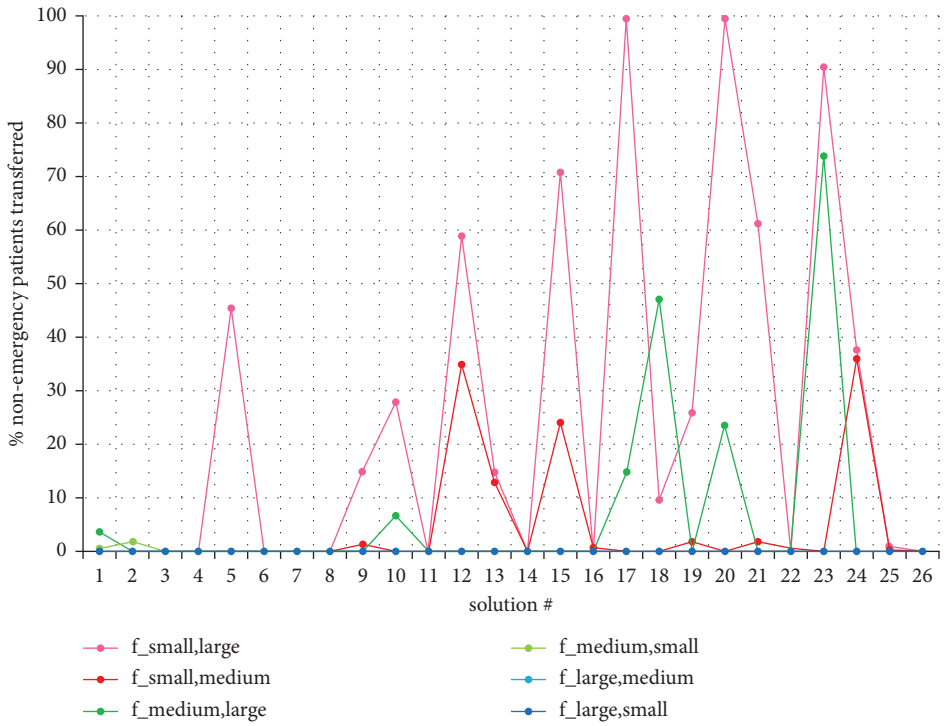

Figure 5: Percent of nonemergency patients transferred.

6.5% and 0.1% of such patients were transferred to the large and small EDs, respectively. There were no instances across all 26 solutions in which nonemergency patients were transferred from the large ED to another ED.

5. Conclusions and Future Work

Overcrowding in hospital emergency departments (EDs) is a problem that affected many hospitals especially during the response to emergency situations such as pandemics or disasters. In this study, we propose a novel optimization model to address overcrowding in a network of EDs via a combination of two decisions: modifying service capacity to EDs and transferring patients between EDs. This model is similar to that presented in [3]; however, whereas the authors in [3] did not account for queueing effects, our model includes queueing considerations in a MINLP, capitalizing on the closed-nature form of M/M/C queueing effects, similar to the approach utilized in [27].

Computational testing was performed, using a Design of Experiments to determine the effects of changes to the various input parameters for a single ED (denoted ED1) on the optimal solutions obtained by our MINLP. Regarding the *expected waiting time in queue plus time in a transfer per patient treated*, the most significant main effects indicated that this response is impacted considerably by changes to the cost per unit service capacity and the travel time between EDs (with time in the system increasing as each of these parameters increases) and to the service rate (with time in system decreasing as this parameter increases), with interaction terms somewhat mediating the effects of the travel time on the expected time in system; on average, the reduced level of the arrival rate of patients from outside of the system accelerates the increase of the expected time in the system when the travel time increases. This would only be reasonable if this increased arrival rate of patients from outside of the system is impacting the likelihood of patient transfers between EDs. Examining this further, we find that for the *number of patients transferred into ED1*, the most significant main effects indicated that this response is affected significantly by changes to the cost per unit service capacity and the arrival rate of both emergency and nonemergency patients from outside of the system (with the number of transferred patients decreasing as each of these parameters increases) and to the service rate (with the number of transferred patients increasing as this parameter increases), with interaction terms between the cost and each arrival rate magnifying the main effects of each these individual terms. In aggregate, an increase in the arrival rate of emergency or nonemergency patients into ED1 from outside the system is associated with a decreased number of patients transferred into ED1, which partly explains the aforementioned interaction effect, in which expected time in the system is found to increase with increases in the travel time between EDs only when the arrival rate of patients from outside of the system into ED1 is at its reduced level.

Additional computational testing examined the effect of ED size on the number of transfers occurring in the system, considering three EDs of different sizes (denoted large, medium, and small). The MINLP was modified slightly here; rather than including a financial penalty for delay times in the objective, we introduce a new constraint imposing an upper bound on the systemwide average expected waiting time in queue plus time in the transfer. Computational testing varied this upper bound across a range of values, identifying an efficient frontier for the tradeoff between the modified objective function and the systemwide average expected waiting time. This optimization model utilizes a variety of strategies to achieve a constrained systemwide average expected waiting time at minimum cost, balancing changes to the numbers of servers at each ED with patient transfers across EDs. Across all points identified on the efficient frontier, the MINLP utilizes patient transfer extensively for nonemergency patients arriving at the small ED, somewhat infrequently for arrivals to the medium ED, and in no instances for arrivals to the large ED. Taken together, these results suggest that our optimization model can identify a range of efficient alternatives for healthcare systems designing a network of EDs across multiple hospitals. Moreover, the model can be helpful to have more balanced EDs with respect to the number of patients and patient waiting time in a network of EDs in case of emergency situations such as natural disasters.

Future work could extend this analysis by considering queueing systems other than M/M/C to represent the stochastic nature of patient arrivals and service times at EDs. Further, while this analysis models steady-state performance, which is useful for network design, an extension to transient system performance in nonsteady-state would allow for similar models to be used in a real-time dispatching environment. Finally, a more nuanced differentiation between patient types, which are modeled as being either emergency or nonemergency patients in this research, could allow for such an MINLP approach to be used to allocate special types of ED service (e.g., pandemic virus testing).

Data Availability

Supplementary materials, including data, will be posted at the University of Missouri's data repository https://hdl.handle.net/10355/86722.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Supplementary Materials

Appendix A: a table of fractional factorial designs. Appendix B: a table of responses for fractional factorial designs. Appendix C: detailed statistical model outputs for responses. Appendix D: detailed statistical model outputs for responses. (*Supplementary Materials*)

References

[1] R. W. Derlet and J. R. Richards, "Overcrowding in the nation's emergency departments: complex causes and disturbing effects," *Annals of Emergency Medicine*, vol. 35, no. 1, pp. 63–68, 2000.

[2] M. Rothfeld, "13 deaths in a day: an 'apocalyptic'coronavirus surge at an NYC hospital," *The New York Times*, vol. 24, 2020.
[3] N. Nezamoddini and M. T. Khasawneh, "Modeling and optimization of resources in multi-emergency department settings with patient transfer," *Operations Research for Health Care*, vol. 10, pp. 23–34, 2016.
[4] American Hospital Association (AHA), *Trendwatch Chartbook 2018: Trends Affecting Hospitals and Health Systems*, American Hospital Association (AHA), Chicago,IL, USA, 2018.
[5] D. Daldoul, I. Nouaouri, H. Bouchriha, and H. Allaoui, "A stochastic model to minimize patient waiting time in an emergency department," *Operations Research for Health Care*, vol. 18, pp. 16–25, 2018.
[6] R. Salway, R. Valenzuela, J. Shoenberger, W. Mallon, and A. Viccellio, "Emergency department (ED) overcrowding: evidence-based answers to frequently asked questions," *Revista Médica Clínica Las Condes*, vol. 28, no. 2, pp. 213–219, 2017.
[7] X. Hu, S. Barnes, and B. Golden, "Applying queueing theory to the study of emergency department operations: a survey and a discussion of comparable simulation studies," *International Transactions in Operational Research*, vol. 25, no. 1, pp. 7–49, 2018.
[8] D. W. Spaite, F. Bartholomeaux, J. Guisto et al., "Rapid process redesign in a university-based emergency department: decreasing waiting time intervals and improving patient satisfaction," *Annals of Emergency Medicine*, vol. 39, no. 2, pp. 168–177, 2002.
[9] S. W. Rodi, M. V. Grau, and C. M. Orsini, "Evaluation of a fast track unit," *Quality Management in Health Care*, vol. 15, no. 3, pp. 163–170, 2006.
[10] Centers for Disease Control and Prevention, *National Hospital Ambulatory Medical Survey 2017 Emergency Department Summary*, Centers for Disease Control and Prevention, Atlanta, GI, USA, 2017.
[11] A. Guttmann, M. J. Schull, M. J. Vermeulen, and T. A. Stukel, "Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada," *British Medical Journal*, vol. 342, no. 1, Article ID d2983, 2011.
[12] R. Konrad, K. DeSotto, A. Grocela et al., "Modeling the impact of changing patient flow processes in an emergency department: insights from a computer simulation study," *Operations Research for Health Care*, vol. 2, no. 4, pp. 66–74, 2013.
[13] J. S. Olshaker and N. K. Rathlev, "Emergency department overcrowding and ambulance diversion: the impact and potential solutions of extended boarding of admitted patients in the emergency department," *The Journal of Emergency Medicine*, vol. 30, no. 3, pp. 351–356, 2006.
[14] E. Cabrera, M. Taboada, M. L. Iglesias, F. Epelde, and E. Luque, "Optimization of healthcare emergency departments by agent-based simulation," *Procedia Computer Science*, vol. 4, pp. 1880–1889, 2011.
[15] N. Izady and D. Worthington, "Setting staffing requirements for time dependent queueing networks: the case of accident and emergency departments," *European Journal of Operational Research*, vol. 219, no. 3, pp. 531–540, 2012.
[16] A. Frakt, "Improve emergency care? pandemic helps point the way," *The New York Times*, vol. 43, 2020.
[17] J. Gruber, T. P. Hoe, and G. Stoye, *Saving Lives by Tying Hands: The Unexpected Effects of Constraining Health Care Providers*, National Bureau of Economic Research, Cambridge, MA, USA, 2018.
[18] P. Soni, *Evaluation of Rule-Based Patient Transfer Protocols in a Multi-Hospital Setting Using Discrete-Event Simulation*, State University of New York at Binghamton, Binghamton, NY, USA, 2014.
[19] G. R. Hung and N. Kissoon, "Impact of an observation unit and an emergency department-admitted patient transfer mandate in decreasing overcrowding in a pediatric emergency department," *Pediatric Emergency Care*, vol. 25, no. 3, pp. 160–163, 2009.
[20] A. Komashie and A. Mousavi, "Modeling emergency departments using discrete event simulation techniques," 2005.
[21] E.-R. Omar, "A stochastic optimization model for shift scheduling in emergency departments," *Health Care Management Science*, vol. 18, no. 3, pp. 289–302, 2015.
[22] D. Sinreich, O. Jabali, and N. P. Dellaert, "Reducing emergency department waiting times by adjusting work shifts considering patient visits to multiple care providers," *Iie Transactions*, vol. 44, no. 3, pp. 163–180, 2012.
[23] M. Gul, A. Fuat Guneri, and M. M. Gunal, "Emergency department network under disaster conditions: the case of possible major Istanbul earthquake," *Journal of the Operational Research Society*, vol. 71, no. 5, pp. 733–747, 2020.
[24] H. Zhu, J. Gong, and J. Tang, "A queuing network analysis model in emergency departments," in *Proceedings of the 2013 25th Chinese Control and Decision Conference (CCDC)*, IEEE, Guiyang, China, May 2013.
[25] S. Au-Yeung, P. Harrison, and W. Knottenbelt, "A queueing network model of patient flow in an accident and emergency," in *Proceedings of 2006 European Simulation and Modelling Conference*, Bonn, Germany, September 2006.
[26] M. Alavi-Moghaddam, R. Forouzanfar, S. Alamdari et al., "Application of queuing analytic theory to decrease waiting times in emergency department: does it make sense?" *Archives of Trauma Research*, vol. 1, no. 3, pp. 101–7, 2012.
[27] R. G. McGarvey, *Supporting Air and Space Expeditionary Forces: Analysis of CONUS Centralized Intermediate Repair Facilities*, Rand Corporation, Santa Monica, CL, USA, 2008.
[28] C. Waydhas, "Intrahospital transport of critically ill patients," *Critical Care*, vol. 3, no. 5, pp. R83–R89, 1999.