Treatment Planning of Victims with Heterogeneous Time-sensitivities in Mass Casualty Incidents

Yunting Shi
Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200030, China, sherryshi@sjtu.edu.cn
Nan Liu
Carroll School of Management, Boston College, Chestnut Hill, MA 02467, nan.liu@bc.edu
Guohua Wan
Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200030, China, ghwan@sjtu.edu.cn

The current emergency response guidelines suggest giving priority of treatment to those victims whose initial health conditions are more critical. While this makes intuitive sense, it does not consider potential deterioration of less critical victims. Deterioration may lead to longer treatment time and irrecoverable health damages, but could be avoided if these victims were to receive care in time. Informed by a unique timestamps dataset of surgeries operated in a field hospital set up in response to a large-scale earthquake, we develop scheduling models to aid treatment planning for mass casualty incidents (MCIs). A distinguishing feature of our modeling framework is to simultaneously consider victim health deterioration and wait-dependent service times in making decisions. We identify conditions under which victims with a less critical initial condition have higher or lower priority than their counterparts in an optimal schedule—the priority order depends on victim deterioration trajectories and the resource (i.e., treatment time) availability. Leveraging these structural insights, we develop efficient solution algorithms. A counterfactual analysis based on our data shows that adopting our model would significantly reduce both the total number of deteriorated victims (by 32%) and the surgical makespan (by 8%) compared to using the then-implemented treatment plan; care coordination among surgical teams could further reduce the number of deteriorated. By demonstrating the value of adopting data-driven approaches in MCI response, our research holds strong potentials to improve emergency response and to inform its policy making.

Key words: Mass Casualty Incident, Treatment Planning, Patient Deterioration, Data-driven Modeling, Scheduling, Optimization Algorithms
History:

1. Introduction

A mass casualty incident (MCI) is defined as "an event which generates more patients at one time than locally available resources can manage using routine procedures" (WHO 2007). MCIs can result from a variety of events: disasters (both natural and man-made), terrorist attacks, and traffic accidents etc. The characterizing feature of an MCI is that emergency response resources are overwhelmed by a sudden jump in demand, making the rationing of resources inevitable. After an MCI occurs, victims are evacuated and triaged to determine their priority levels (we use "victims" and "patients" interchangeably in this article). Then, victims are transported to nearby medical facilities that participate in the response effort. However, some MCIs caused by large-scale natural disasters such as earthquake, tsunami, and tropical storm can significantly damage local infrastructure, including roads, hospitals, and communication and utility networks.

Hospitals that remain functional are far away and difficult to reach, rendering it extremely difficult, if not impossible, to quickly transport victims there. The 2004 Indian Ocean Earthquake and Tsunami, 2010 Haiti Earthquake, and 2014 Typhoon Yolanda are examples of such catastrophic incidents.

When these large-scale disasters strike, field hospitals play a critical role in disaster response. A field hospital is a temporary medical facility that takes care of casualties on-site before they can be safely transported to more permanent facilities (Driscoll 2004). Take earthquake as an example. After an earthquake, victims are rescued under the debris of buildings or from underground. Field hospitals are set up within a very short amount of time to provide sophisticated care to the extent possible, including imaging, surgeries, orthopedics, intensive care, pediatrics, and OBGYN etc. They can be launched by nearby hospitals that remain operational, military medical units, and domestic or international humanitarian missions. After the 2008 Sichuan Earthquake (Richter scale 8.0), field hospitals set up by the West China Hospital became operational the next day after the earthquake (Chen et al. 2010). After the 2010 Haiti Earthquake (Richter scale 7.0), the Israel Defense Forces Medical Corps Field Hospital was launched from a distance of 6000 miles and fully operational on site in 89 hours (Kreiss et al. 2010). Most recently, after the 2020 Aegean Sea Earthquake (Richter scale 7.0), Turkey's Disaster and Emergency Management Presidency set up medical tents near areas with the highest damage very shortly after the incident (CNBC 2020).

Given limited resources, the emphasis of field hospitals is to "provide life-saving medical care to as many people as possible" (Merin et al. 2010). Indeed, a commonly adopted principle in MCI response efforts is to *do the greatest good for the greatest number* (Argon et al. 2010). Thus, field hospitals face two critical questions: (1) who to admit (and who to deny) for treatment; and (2) for those admitted, in what sequence and at what time they should be treated. Our research is motivated by these operational challenges in field hospitals and develops a decision model to directly answer the second question. Our model prescribes a treatment plan that aims for the greatest good for the greatest number of victims admitted for care. Moreover, using our model as an evaluation tool to compare the outcomes from different combinations of victims admitted can inform an answer to the first question.

The operations research (OR) literature on the planning and optimization of response efforts to MCIs has been growing quickly during the last decade (Argon et al. 2010). A large volume of this literature develops stylized models to identify structural insights and provide high-level guidance to the emergency response community; see, e.g., Jacobson et al. (2012), Mills et al. (2013, 2018a,b), Sun et al. (2018). While stylized models inform helpful rules of thumb for practice, there remains a critical need for "decision-oriented, operations research models to improve preparations for and response to major emergencies" (Larson et al. 2006). Leveraging the insights gleaned from a unique timestamps dataset recorded during the 2008 Sichuan earthquake, we develop a data-driven model to aid decision making for treatment planning in field hospitals.

Our dataset contains surgical data of 101 victims operated by 13 surgical teams in a field hospital. A key phenomenon captured by the previous OR literature on MCI response is that victim health conditions

(e.g., survival probabilities) deteriorate over time. Analyzing our data reveals *new* empirical evidences on emergency response operations. Specifically, during the patient deterioration process, patient surgical times may increase while they wait for treatment; furthermore, a patient's surgical time may increase at a faster rate after he deteriorates. It is evident that such dependence of surgical procedure time on victim wait time influences how field hospitals should plan their surgical operations.

If victims do not deteriorate or their surgical times do not depend on wait times, then the previous models developed in the classic literature on job scheduling and surgical scheduling may apply. If victims deteriorate but their surgical times do not depend on wait times, then insights generated by the previous literature on MCI response are applicable. One important motivation for the MCI response literature is to solve the patient-distribution problem (i.e., how to distribute patients from the incident scene to nearby medical facilities), where it appears reasonable to assume (transport) service time does not change over time. In field hospitals, however, patients deteriorate and their surgical times may increase while they wait. To the best of our knowledge, no OR models exist to inform treatment planning in field hospitals where these complicating factors jointly present. We fill the gap in the literature and develop one such model, demonstrating potentials to save more lives and do more good in the aftermath of MCIs.

We adopt a modeling framework similar to that of a traditional job scheduling problem, but with some important differences. Upon arrival at a field hospital, victims have been triaged based on their health conditions. Simple triage and rapid treatment (START) is a commonly used triage system in emergency medicine in the US. To facilitate discussion, we follow the terminologies from START to classify victims. START categorizes victims into four types based on health conditions (Lerner et al. 2008).

- Expectant: victims unlikely to survive given their conditions or level of available care;
- Immediate: victims who require medical attention as soon as possible;
- Delayed: victims who have serious injuries but their status are not expected to deteriorate significantly over several hours; and
- Minor (or "walking wounded"): victims with relatively minor injuries.

Following the widely accepted and practiced emergency response principle of doing the greatest good for the greatest number, only *immediate* and *delayed* victims will be admitted to field hospitals and our model concerns how to schedule surgical care of these two types of victims.

In our base model, we consider a single surgical team, with a certain number of immediate victims and delayed victims to operate. Both types of victims have their respective due times, which can be viewed as the "feasible" time window to treat these victims in the field hospital due to clinical and operational concerns. The procedure time of an immediate victim increases linearly in his wait time. The health condition of a delayed victim deteriorates after he waits for S units of time. A deteriorated delayed victim may suffer from significant organ damage or amputation which could be possibly avoided if he receives treatment before S; in addition, a delayed victim's procedure time increases at a faster rate after deterioration. We use an

increasing piecewise linear function to model how a delayed victim's procedure time changes with his wait time. After deterioration, a delayed victim's procedure time increases at the same rate as an immediate one in wait time. The goal of the provider is to minimize the number of deteriorated delayed victims subject to the due time requirements of all victims.

The objective of minimizing the number of deteriorated patients drives to serve delayed patients earlier, but the due time requirements which entail controlling the makespan may promote treating immediate patients first, partly because their procedure times increase faster in wait time than their delayed counterparts. Analyzing this tradeoff allows us to fully characterize the structure of the optimal surgical schedule and develop an efficient solution algorithm for the base model. In a sharp contrast to what has been suggested by START and commonly followed in practice—to always treat immediate patients first, we find that sometimes it is better off to give priority to delayed patients. The optimal priority order depends on how fast victims deteriorate and the resource (i.e., treatment time) availability.

In addition to this base model, we consider a setting with victims whose procedure times may–or–may not depend on wait times. Therefore, based on health conditions (delayed vs. immediate) and the dependence of procedure times on wait times, we have four different types of victims. To incorporate more practical considerations, we also investigate the implementation of our base model if (1) the provider needs to take a mandated rest after working for a certain amount of time, (2) a second batch of victims arrive, and (3) the field hospital has multiple surgical teams to coordinate.

A key assumption in our model is that surgical time, given a victim's wait time, is deterministic. This assumption allows us to derive algorithms that can quickly generate the optimal surgical schedule. Timely action is essential in MCI response. Our numerical study based on real data demonstrates that the surgical schedule derived from our model has robust performance even when surgical times have variability: our schedule is mostly on time and if not has very limited tardiness. In a counterfactual analysis performed on the 13 surgical teams in our data, we find that instead of using the actually adopted surgical schedules, following our model would save 8 more delayed victims from deterioration, marking a 32% reduction in the total number of deteriorated cases; at the same time, the makespan of 13 surgical teams could be reduced by 8% on average, creating opportunities to save more victims. These improvements are made possible by refining the operations of each team assuming the same set of victims are served by each team. If victims, upon their arrival, could be allocated in a better way among teams, additional victims could be saved from deterioration. A complete care coordination could save 6 more and even a partial one could save 3 more. As a demonstration for practical applications, we have developed a prototype of web-based tool accessible at www.tinyurl.com/mci-rescue to implement our base model for emergency response.

Besides providing operational decision support, our work has important policy implications for MCI response. Our modeling framework is informed by analyzing data recorded during an MCI and our work demonstrates significant value of data-driven modeling. As in other health sectors which have used data

extensively to drive decision making, policies for MCI response need to be made based on data and scientific modeling approaches rather than intuition and simple heuristics. At a high level, our research generates two particular useful insights to inform policy making for MCI response. First, an MCI response policy to do the greatest good for the greatest number needs *not* always prioritize treatment solely based on victim initial health conditions, like what the current policies suggest. Victim deterioration is another critical factor to look into. Second, care coordination can significantly improve health outcomes from MCI response efforts and it should be considered in emergency response processes.

The remainder of this article is organized as follows. Section 2 reviews the relevant literature. Section 3 describes our data and empirical findings. Sections 4 and 5 present our model, its extensions, and their analysis. Section 6 discusses the numerical study and results. Finally, Section 7 draws concluding remarks. All proofs of the technical results are shown in Appendix A.

2. Literature Review

Our research draws upon several streams of literature. From the perspective of functional domains, our work is related to the healthcare operations management literature that studies how to allocate limited medical resources among patients with heterogeneous health conditions that may deteriorate over time; see, e.g., Deo et al. (2013). We do not intend to do a comprehensive review of this literature and we shall focus on those studies closely related to our work in the context of MCI response. The literature on MCI response has been growing recently (Argon et al. 2010). We draw attention to three recent articles that are most related to our work, namely Jacobson et al. (2012), Mills et al. (2013), and Hu et al. (2021). Jacobson et al. (2012) study how to determine patient priorities in a resource-constrained environment. They focus on analyzing a clearing system with two types of jobs, which have type-specific lifetimes, service times, and reward distributions. Based on the remaining jobs in the system, the decision maker dynamically determines which type of job to serve next in order to maximize the expected total reward, e.g., the expected number of survivors. Mills et al. (2013) construct a fluid model of patient triage in a setting where all patients are present at time zero. In their main analysis, patients are categorized into two classes and each class has an associated non-negative reward function, which is monotone non-increasing in time. The objective is to maximize the total rewards collected by specifying the fraction of service capacity allocated to each patient class at any time. Hu et al. (2021) study the use of proactive service for inpatients who may deteriorate while staying in hospital. They develop a multi-server queueing model with two customer classes: moderate and urgent. Customers have class-dependent arrival rates, abandonment rates, and service rates. Customers may transition classes while waiting. Hu et al. (2021) analyze a deterministic fluid approximation to derive the optimal control policy which minimizes the total cost of waiting, abandonment, and class transitioning.

Our work significantly departs from these three articles in the following ways. First, the modeling frameworks and application contexts are different. Our model seeks to solve the surgical scheduling problem in

field hospitals. Jacobson et al. (2012) and Mills et al. (2013) are motivated by the transportation problem of moving victims to medical facilities that participate in the response effort. The model of Hu et al. (2021) naturally finds applications in managing hospital inpatient beds, such as intensive care units (ICUs). We adopt the framework of a job scheduling model with pre-specified due times, while Jacobson et al. (2012) and Mills et al. (2013) formulate their models as clearing systems without explicit end times and Hu et al. (2021) develop a multi-server queueing model. Second, Mills et al. (2013) assume constant service time and Jacobson et al. (2012) and Hu et al. (2021) assume service times are patient class-dependent. However, in our setting patient surgical times increase in their wait times. To capture this, our model allows service time to be both class- and time-dependent. Third, patient deterioration is modeled differently. These three papers consider decreasing patient rewards, patient class transition costs, abandonment, or changes in service rate. We explicitly model both the clinical and operational impact of deterioration: patient health condition degrades to a more critical level and patient service times increase in a faster rate post deterioration. Instead of maximizing rewards or minimizing costs, we have a perhaps more tangible objective, which is to minimize the number of deteriorated patients.

From the modeling perspective, our work is related to the appointment and surgical scheduling literature that investigates how to schedule patients over time in a clinic session. A commonly-used objective in this literature is to optimize the tradeoff among patient waiting, provider idling, and overtime. Extensive work has been devoted to developing mathematical programming models to minimize a weighted sum of costs due to these components, by scheduling the treatment start time of each patient. The previous literature has considered a wide range of factors in constructing a schedule, such as patient heterogeneity, random service times, walk-ins, and no-shows. A recent sample of works include, but are not limited to, Hassin and Mendel (2008), Robinson and Chen (2010), Jung et al. (2019), Zacharias and Yunes (2020), and Wang et al. (2020). See, e.g., Gupta (2007) and Ahmadi-Javid et al. (2017) for in-depth reviews of this literature. Though we also study surgical scheduling, several key features distinguish our model from those considered previously. In our model, patient service time increases over time. In addition, patients deteriorate and their service times increase at a faster rate once they wait beyond a certain threshold. Our objective is not cost-based, but to minimize the number of deteriorated cases subject to requirements on due times. These lead to different system dynamics and tradeoffs to be considered, rendering the previous appointment and surgical scheduling models largely inapplicable in our study context.

From the methodological point of view, our research draws upon the literature on traditional job scheduling in two directions, namely, multi-agent scheduling (see, e.g., Baker and Smith 2003, Agnetis et al. 2004, and Leung et al. 2010) and scheduling with deteriorating jobs (for comprehensive surveys on this topic, see, e.g., Alidaee and Womer 1999 and Cheng et al. 2004). Multi-agent scheduling is concerned with scheduling multiple classes of jobs and each class has its own objective, possibly different from each other. (This is different from scheduling with multi-objectives where there is only one class of jobs that have the same

set of multiple objectives.) In our study, we employ the framework of multi-agent scheduling, where we regard the two classes of patients (i.e., immediate and delayed) as the jobs of two agents, each carrying a different objective (i.e., the number of deteriorated delayed patients and the lateness of treatment). Due to the time sensitivity of victims in MCIs, we model surgeries to be scheduled as deteriorating jobs whose procedure durations increase with time. Our models integrate both the features of multi-agent scheduling and scheduling with deteriorating jobs. We study the structural properties of these new scheduling models and develop efficient algorithms to solve them.

3. Exploratory Study: 2008 Sichuan Earthquake

In this section, we use a dataset recorded during the early-stage rescue in the 2008 Sichuan earthquake to conduct an exploratory study on the relationship between victim wait times and procedure times. The impact of patient wait time on service time has been studied in other healthcare settings. For instance, it has been shown that delay to ICU admission leads to longer stay in hospital (Chan et al. 2017, Renaud et al. 2009). Our data present a unique setting of MCI. In contrast to other settings, our data are relatively limited due to the challenges in data recording as discussed below. However, such limited availability of data represents reality faced by decision makers in MCIs. Our purpose is to exploit the data available to us and to inform optimization models that can aid decision making in such data-limited environments.

The 2008 Sichuan earthquake in China, also known as Wenchuan earthquake, struck around 2:28pm local time on May 12, 2008. (Wenchuan is a county in the Sichuan Province where the epicenter of the earthquake located.) Measuring at a magnitude of 8.0 on the Richter scale, this earthquake is one of the deadliest earthquakes to hit China and is the 18th deadliest earthquake of all time in the world. Approximately 15 million people lived in the affected area. Over 69,000 people lost their lives in the quake, 374,176 were reported injured, and 18,222 were listed as missing as of July 2008. The economic loss amounted to over Chinese ¥ 8.45 billion (approximately US $ 1.2 billion).

In response to the event, many domestic and international organizations launched humanitarian missions. Our data were recorded in a field hospital dispatched by West China Hospital located in Chengdu, Sichuan Province which is about 80km (approximately 50 miles) southwest of Wenchuan. The field hospital became operational and started to admit surgical patients in the following day of the event. Our data consist of 101 victims, who received care from 13 surgical teams. All these surgeries were conducted in the following day of the earthquake. Our dataset includes the surgical sequence, surgical time, and clinical type of surgery of each victim. Table 1 below shows the summary statistics of our data.

It was very challenging to collect these timestamps data in such an austere and chaotic environment and they were all recorded manually. Due to challenges in data collection, we have limited patient-level data. But the data available to us were extremely valuable because they represent what decision makers

Table 1 Summary Statistics of All Data

	Victims ($n = 101$)		Surgical Teams ($n = 13$)	
	Procedure time (min)	Wait time (min)	# of cases	Total shift length (min)
Mean	90.297	309.208	7.769	732.692
Median	85	305	7	725
Stdev	41.128	221.589	1.527	27.076
Maximum	245	710	11	790
Minimum	25	0	6	680

actually face in early-stage rescue during an MCI. How to extract useful information in such a data-limited environment to inform decision making is our focus in this section.

We learn from the rescue teams that victims transported to the field hospital were quickly triaged following guidelines similar to START; see also Zhang et al. (2012). Expectant victims were provided palliative care and not sent to surgical teams; minor victims were waiting to be transported to base hospitals far away from the incident scene; immediate and delayed victims received surgeries in field hospitals and our data contain these two types of patients.

Our main goal in this section is to investigate the impact of victim wait times on their procedure times. The wait time is measured by the difference between a victim's arrival at the surgical team and his surgery start time. We hypothesize that

H1: If victims receive surgeries without any delay, immediate ones have a longer procedure time than delayed ones;

H2: The procedure times for both types of victims increase in their wait times for surgeries;

H3: After waiting for some time, a delayed patient's procedure time increases at a faster rate in his wait time; and

H4: The rate at which an immediate patient's procedure time increases in his wait time is higher than that of a deteriorated delayed patient.

To test these hypotheses, we consider the following linear regression model (1). Let p_i be the procedure time of patient i, w_i be the wait time, and x_i be an indicator variable which takes value 1 if patient i belongs to the immediate type and 0 otherwise. We let $S \geq 0$ be the wait time after which a delayed victim's procedure time increases faster in his wait time. Note that we do not know whether such an S exists *a priori*; we intend to identify the value of S, if it exists, following the maximum likelihood principle described below.

$$p_i = \alpha_0 + \theta \times x_i + \alpha \times w_i + \gamma \times (1 - x_i) \times (w_i - S)^+ + (\gamma + \eta) \times x_i \times w_i + \epsilon_i, \tag{1}$$

where α_0, θ, α, γ, and η are model parameters to be estimated for a given S and ϵ_i is the random noise term assumed to be normally distributed with mean 0. Depending on the victim type, a more explicit form of our regression model (1) can be written as follows.

$$p_i = \begin{cases} \alpha_0 + \alpha \times w_i + \gamma \times (w_i - S)^+ + \epsilon_i, & \text{if the victim is a delayed type,} \\ (\alpha_0 + \theta) + (\alpha + \gamma + \eta) \times w_i + \epsilon_i, & \text{otherwise.} \end{cases} \tag{2}$$

If $\theta > 0$, $\alpha > 0$, $\gamma > 0$ and $\eta > 0$ with statistical significance, then we have evidences to support hypotheses H1 through H4, respectively. Naturally we would expect the intercept $\alpha_0 > 0$ as a delayed patient should have a positive procedure time even if his wait time is zero. This can serve as a simple sanity check on our model specification.

Before fitting the regression model (1) to our data, we conduct some preliminary analysis on the impact of victim wait times on their procedure times. Among the 101 surgeries, there are 7 surgical types. The ranges of victim wait times and procedure times are similar across all 7 types. For each type, we run a simple linear regression to check if there appears to be some dependence between wait time and procedure time. Table 2 shows the sample size and the p-value of simple linear regression per surgical type. We observe a strong linear relationship between wait time and procedure time for types 1 and 4 patients; for other types of patients, wait time does not seem to have a significant impact on procedure time. Next, we focus our analysis on types 1 and 4 patients.

Table 2 Surgical Types and Preliminary Analysis Results

	Type	# of observations	p-value
1	Limb amputation and trimming	9	0.046
2	Soft tissue injury treatment	5	0.303
3	Craniocerebral injury surgery	15	0.781
4	Severe open fracture surgery	32	<0.0001
5	Spine and pelvic injury surgery	8	0.893
6	Large blood vessel repair, anastomosis, ligation	5	0.272
7	Internal surgery	27	0.098

In our data, we do not know which victim is immediate or delayed. However, based on our communication with rescue teams, the practice in the field hospital was to always prioritize immediate victims over delayed ones. Therefore, we deduce that the first few victims operated by each surgical team were immediate, though we do not know the exact number. Neither do we know the switching time S for delayed patients.

To deal with these issues, we adopt the widely-accepted maximum likelihood principle and use the following sequential testing procedure. We let t be a cutoff time. If a victim's wait time is no larger than t minutes in our data, we consider him as an immediate one; otherwise, we consider him as a delayed one. Thus for a fixed t, we can determine the value of the indicator variable x_i for each victim i. Now, for each $t \in [10, 400]$ and $S \in [t, 650]$ with a step-size 10 minutes, we fit the regression model (2) to our data; in our iterations, we require that $S \geq t$ because all victims served before t are immediate ones and have no switching time. The model with the largest likelihood is considered the "true" model which best represents our data and hence reality. It turns out that the best model has $t = 140$ minutes and $S = 470$ minutes. Table 3 presents the regression results of the full model which contains all regression coefficients and the reduced model with only statistically significant coefficients.

Table 3 Regression Results of the Best Model with $t = 140$ minutes and $S = 470$ minutes

	Full model			Reduced model		
Variable	Coefficient	Std. error	p-value	Coefficient	Std. error	p-value
α_0	51.840	13.120	0.000	57.619	5.230	0.000
α	0.118	0.047	0.016	0.109	0.019	0.000
θ	14.674	14.828	0.329			
γ	0.355	0.185	0.062	0.244	0.075	0.003
η	-0.266	0.183	0.154			
Number of observations	41			41		
Log-likelihood	-178.21			-180.30		
R-squared	0.620			0.579		
Adjusted R-squared	0.578			0.557		

We observe that α and γ are statistically significant, in support of hypotheses H2 and H3 above. That is, victim procedure times increase in wait times, and for delayed victims, their procedure times increase at a faster rate once their wait times bypass a certain threshold. However, given that θ and η are not significant, we do not have strong support for H1 and H4 above. In other words, in our dataset delayed and immediate victims are likely to have the same procedure time if they do not encounter any delay in treatment; the rate at which a deteriorated delayed victim's procedure time increases in his wait time is likely the same as that of an immediate patient.

In sum, our exploratory study reveals several important findings. First, procedure times of victims may be independent of or increase in their wait times. Second, for those wait-sensitive delayed victims, procedure time appears to be an increasing piecewise linear function of wait time. Third, the rate at which a delayed victim's procedure time increases may catch up with that of an immediate patient if the wait is sufficiently long. We should note that our exploratory study, which is based on limited data from a single MCI, is by no means comprehensive and does not reveal the exact mechanism via which wait times may impact procedure times; but it does generate sensible empirical observations to inform the development of our scheduling models that follow.

4. Model

In this section, we focus on the operation of a single surgical team in the field hospital and we will consider settings with multiple surgical teams in Section 5. As discussed above, admitted victims to a field hospital belong to two types, immediate and delayed, based on their health conditions. Our model concerns prioritizing these victims. As demonstrated by our exploratory study in Section 3, victim procedure times may or may not increase over time. Therefore, based on health condition and the dependence of procedure times on wait times, we have four different types of victims, illustrated in Table 4. For convenience, we call victims whose procedure times depend on wait times as *unstable* victims, and others as *stable* ones. We use N^i, N^d, N^{si}, and N^{sd} to represent the number of unstable immediate victims, unstable delayed victims, stable

immediate victims, and stable delayed victims at time 0, respectively. Time 0 is the time when the surgical team is ready to operate. For convenience, we denote the set of victim types to be $\mathcal{T} := \{i,\ d,\ si,\ sd\}$ and let $\mathcal{J}$ be the set of all victims so that $|\mathcal{J}| = \sum_{k\in\mathcal{T}} N^k$. Define a function $\phi : \mathcal{J} \to \mathcal{T}$ such that $\phi(j)$ returns the type of victim j.

Table 4 Types and Numbers of Victims

Number of victims		**Health condition** Immediate	Delayed
Procedure time	Unstable (wait-dependent)	N^i	N^d
	Stable (wait-independent)	N^{si}	N^{sd}

These victims need to receive surgeries from one surgical team. (We assume that no victims would leave without receiving treatment—this is also supported by our data discussed in Section 3 where all 101 victims got surgeries and none of them left or died in the rescue process.) Clearly, not all victims will be able to get care immediately and some need to wait. Our goal is to determine the schedule of serving these victims with the limited medical resources available. We first analyze the model with unstable victims only and then we investigate the model with both unstable and stable victims.

4.1. Model with Unstable Victims Only

In this section, we assume that we only have unstable victims at time 0, i.e., $N^i, N^d > 0$ while $N^{si} = N^{sd} = 0$. Motivated by empirical observations in Section 3, we make the following assumptions on the functional relationship between procedure time and wait time for unstable patients. If an unstable immediate patient waits w units of time before his surgery, his procedure time, denoted by $p^i(w)$, is

$$p^i(w) = \beta_0 + \beta w, \tag{3}$$

where $\beta_0 \geq 0$ is the setup time for the procedure and $\beta \geq 0$ is the "wait sensitivity" of his procedure time, i.e., for each additional unit of time an unstable immediate patient waits, his procedure time increases by β units of time.

For unstable delayed patients, their procedure times increase in their waiting as well. But there is a *switch time* S, after which delayed patients deteriorate to a more critical level and their procedure times increase at a faster rate (supported by H3 above). For simplicity, we call deteriorated delayed patients as deteriorated patients. Specifically, we assume that if an unstable delayed patient waits w units of time before his surgery, his procedure time, denoted by $p^d(w)$, is

$$p^d(w) = \begin{cases} \alpha_0 + \alpha w, & \text{if } w \leq S, \\ \alpha_0 + \alpha S + \beta(w - S), & \text{if } w > S, \end{cases} \tag{4}$$

where $\beta_0 \geq \alpha_0$ and $\beta > \alpha$. It is easy to see that (4) can be equivalently written in a more concise form below.

$$p^d(w) = \alpha_0 + \alpha w + (\beta - \alpha)(w - S)^+, \tag{5}$$

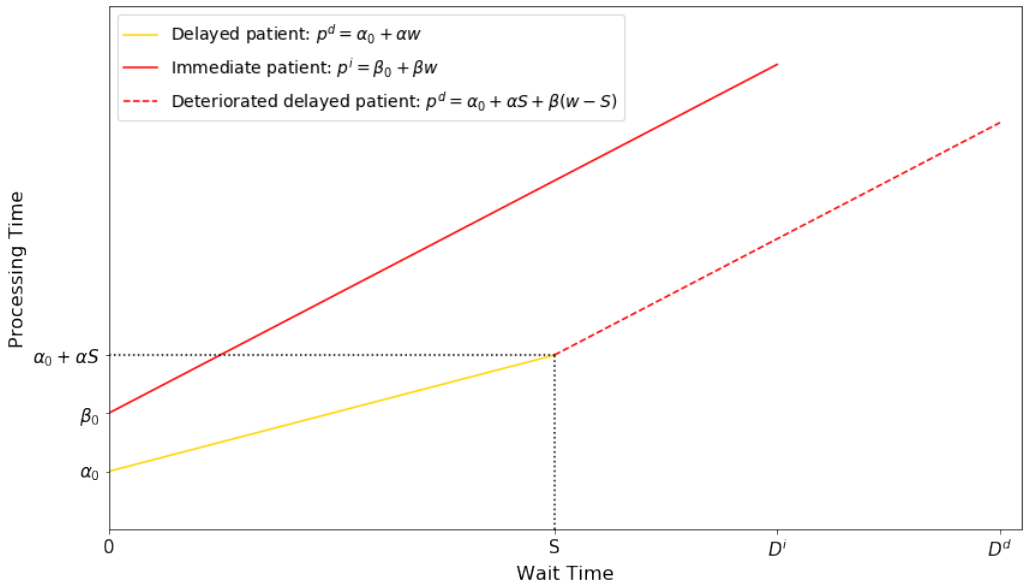
where $(\cdot)^+ = \max(\cdot, 0)$. Note that after S, deteriorated patients have the same wait sensitivity of procedure time as immediate patients. In addition to having a higher level of wait sensitivity, deteriorated patients are exposed to a higher medical risk and they may suffer from significant organ damage or amputation which could be possibly avoided if they were to receive treatment before S. (As discussed later, an important objective of the decision maker is to minimize the number of deteriorated patients.) Since in this subsection we are only concerned with unstable patients, we will refer to unstable immediate (resp. unstable delayed) victims as immediate (resp. delayed) victims, when the context is clear.

For both types of unstable patients, we assume that they have their respective *due times*. The time interval between time 0 and the due time can be viewed as the "feasible" time window to treat patients in the field hospital. We let D^d and D^i denote the due times for unstable delayed and immediate patients, respectively. Patients of the same type have a common due time. We assume that $D^i \leq D^d$, i.e., immediate patients have a shorter feasible time window to receive treatment than delayed patients. Figure 1 shows how the processing time changes with wait time for both types of patients as well as their respective due times. Due times exist for both clinical and operational reasons. Immediate patients who were to receive treatment after their due times are likely to develop complications that present a severe medical risk. The due time of delayed patients can be viewed as the maximum amount of time that a surgical team can work for in a clinic session. Hence another goal of the decision maker, to be elaborated below, is to ensure all victims will be treated before their due times to the extent possible.

In our model, we assume that given patient wait time, his procedure time is deterministic. We also assume that patient procedure time increases linearly or piecewise-linearly in wait time. While in general patient procedure time may contain random noise and may also change non-linearly with patient wait time, our assumptions capture the first-order impact of patient procedure time as well as the key trade-offs that affect the optimal schedule. These assumptions also make the model tractable with optimal schedules that can be easily and quickly derived—in MCIs speedy response is essential. Our numerical experiments in Section 6 show that, even when some of these assumptions are relaxed, the optimal schedules obtained from our model still perform well.

Next, we introduce the objectives of our model in detail and discuss how to derive the optimal schedule. Given $N^i > 0$ unstable immediate victims and $N^d > 0$ unstable delayed victims at time 0, the decision maker needs to (quickly) specify the surgical start times for each victim and she is concerned with two objectives simultaneously. The first objective is to minimize the number of delayed patients who deteriorate,

Figure 1 Processing Times and Due Times of Unstable Victims



as such deterioration may lead to irrecoverable damages to patient health as discussed above. For a patient j, we let w_j represent his wait time and recall that $\phi(j)$ returns his type. Define

$$U_j^d = \begin{cases} 1, & \text{if } w_j > S \text{ and } \phi(j) = d, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

So the total number of deteriorated cases is $\sum_{j\in\mathcal{J}} U_j^d$.

The second objective of the decision maker is to control the lateness of treating patients. Let C_j be the treatment completion time of patient j and L_j be his lateness, i.e.,

$$L_j = \max\{0,\ C_j - D^{\phi(j)}\}. \tag{7}$$

We use L_{max}^k to represent the maximum lateness of type k victims. Specifically, we have

$$L_{max}^k = \max\{L_j :\ \phi(j) = k,\ j \in \mathcal{J}\},\ k \in \{d,\ i\}.$$

So, L_{max}^i (L_{max}^d resp.) is the difference between the treatment completion time of the last immediate (delayed resp.) patient and his respective due time D^i (D^d resp.). Note that if the last patient finishes treatment before his due time, then the lateness is zero. To capture the second objective, the decision maker would like to ensure that both $L_{max}^i \le Q$ and $L_{max}^d \le Q$, where $Q \ge 0$ is a tuning parameter under her control. If $Q = 0$, the decision maker would like all patients to receive surgeries before their due times, but she could also be more flexible by choosing a positive Q.

The classic notation used in the machine scheduling literature is based on a triplet $\alpha|\beta|\gamma$, where α, β, and γ represent the number of machines/servers, the characteristics of jobs, and the objective, respectively.

Recently, this notation has been extended to settings where there are two types of jobs and two objectives, one for each type of jobs (Agnetis et al. 2007, Leung et al. 2010). To facilitate the presentation, we follow this notation to concisely represent our problem as follows.

$$1\big|p_j^d,\ S,\ d_j^d = D^d:\ p_j^i,\ d_j^i = D^i\big|(\sum U_j^d, L_{max}^d \le Q):\ L_{max}^i \le Q, \tag{8}$$

where p_j^k and d_j^k respectively represent the processing time and due time of a patient j who belongs to type $k \in \{i, d\}$ and S is the switching time of delayed patients. Our problem is fundamentally different from those considered in the previous scheduling literature. One unique challenge of our problem comes from the deterioration of delayed patients over time. Patient procedure times increase in their wait times, but when delayed patients deteriorate, how their procedure times change with wait times becomes different. Consequently, the final set of deteriorated patients and the resulting L_{max}^i and L_{max}^d depend on the schedule through a rather complex relationship.

Prior to deriving the optimal schedule for (8), we present two ancillary results below which facilitate our analysis: Proposition 1 and Lemma 1.

Proposition 1 *(Browne and Yechiali 1990) Consider an scheduling problem $1|p_j = a_j + b_j w_j|C_{max}$, where p_j is the processing time of job j, $a_j, b_j > 0$ are job-specific constants, and w_j is the wait time of job j before its service. The objective is to minimize C_{max}, i.e., the makespan. It is optimal to process jobs in an increasing order of the ratio a_j/b_j, i.e., a job with a smaller ratio of a_j/b_j goes first.*

This result is intuitive: the optimal scheduling policy should first process jobs with shorter setup times (i.e., smaller a_j's) and those of which service times are more sensitive to wait times (i.e., with larger b_j's). A particularly noteworthy point here is that the ratio a_j/b_j is the only statistic one needs to track in order to determine the optimal sequence of job processing.

To facilitate understanding, we can view a_j as a measure of patient j's initial condition and b_j as his "deterioration" speed. The larger the value of a_j, the worse the patient j's initial condition and the longer the setup time. The larger the value of b_j, the quicker the patient j's procedure time increases as he waits. For ease of discussion, we call a_j/b_j as the *initial condition-to-deterioration speed ratio*, or I2D ratio for short. Proposition 1 says that to minimize the makespan, one should prioritize a patient with a smaller I2D ratio. Note that the key premise for Proposition 1 is that I2D ratios are constant over time, which, however, is not true in our model due to the deterioration of delayed victims.

Lemma 1 *Deteriorated delayed victims always have a smaller I2D ratio than immediate victims.*

To see this, note that at time S the intercepts of the procedure time function for delayed and immediate patients are $\alpha_0 + \alpha S$ and $\beta_0 + \beta S$, respectively, but their slopes are the same β after S. Then it is evident that

$$\frac{\alpha_0 + \alpha S}{\beta} < \frac{\beta_0 + \beta S}{\beta}$$

since $\alpha_0 \leq \beta_0$ and $\alpha < \beta$. This fact, together with Proposition 1, suggests that to minimize makespan, after time S deteriorated victims should have priority than immediate victims.

Next we present the key result of this section. We first describe the structure of the optimal schedule to problem (8) and then present an algorithm to derive such a schedule. The latest time to start treating a delayed patient before he deteriorates is S, the switch time. For a delayed patient who starts service at S, we denote the finish time of his service by $C = S + \alpha_0 + \alpha S$. Let $\bar{D}^i = D^i + Q$, before which all immediate patients should be served for a schedule to be feasible. Similarly, let $\bar{D}^d = D^d + Q$, before which all delayed patients should be served for a schedule to be feasible. We say "one type of patients have higher (lower resp.) priority than the other type" implying that no patients from the latter (former resp.) type will be served unless all patients from the former (latter resp.) type have received treatment, provided that both types of patients are present in the context of discussion.

Theorem 1 *If there exists an optimal schedule for problem (8), then it is non-idling and has the following properties.*

(a) If $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$, then all immediate patients should be served consecutively before $\bar{D}^i$ and as many delayed patients as possible should be served before immediate patients.

(b) If $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$ and $\bar{D}^i \leq C + \alpha_0\beta - \beta_0\alpha$, then all immediate patients should be served consecutively from time 0 and they have higher priority than delayed patients.

(c) If $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$ and $\bar{D}^i > C + \alpha_0\beta - \beta_0\alpha$, then immediate patients, if any, completing treatment before $C + \alpha_0\beta - \beta_0\alpha$ are served consecutively from time 0 and no delayed patients are served before them. Immediate patients who complete treatment after $C + \alpha_0\beta - \beta_0\alpha$, if any, are served consecutively and as many delayed patients as possible should be served before these immediate patients.

Figure 2 illustrates the structural properties of the optimal schedule discussed in Theorem 1. We use red to represent immediate patients, yellow for delayed patients, and orange for deteriorated delayed patients, respectively. The corresponding parameter setting is shown under each sub-figure. One interesting, and perhaps striking, observation is that sometimes it can be optimal to give delayed patients higher priority than immediate patients! This is contrary to what common wisdom would suggest and practice would do—to always prioritize immediate patients over delayed ones. In fact, giving strict priority to immediate patients is only optimal in one of the three sub-figures in Figure 2.

A natural question arises: what are the key driving forces to schedule delayed patients ahead of immediate ones? Theorem 1 and Figure 2 shed some light on this question. The first critical factor is the I2D ratio. Prioritizing patients with a smaller I2D ratio reduces the makespan and thus helps the provider keep the lateness of patients to be within the tolerable range. When $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$, i.e., delayed patients have a smaller I2D ratio, they are likely to be scheduled before immediate patients; see Figure 2(a). However, when $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$, i.e., delayed patients have a larger I2D ratio, it *does* not mean that they would always have lower priority

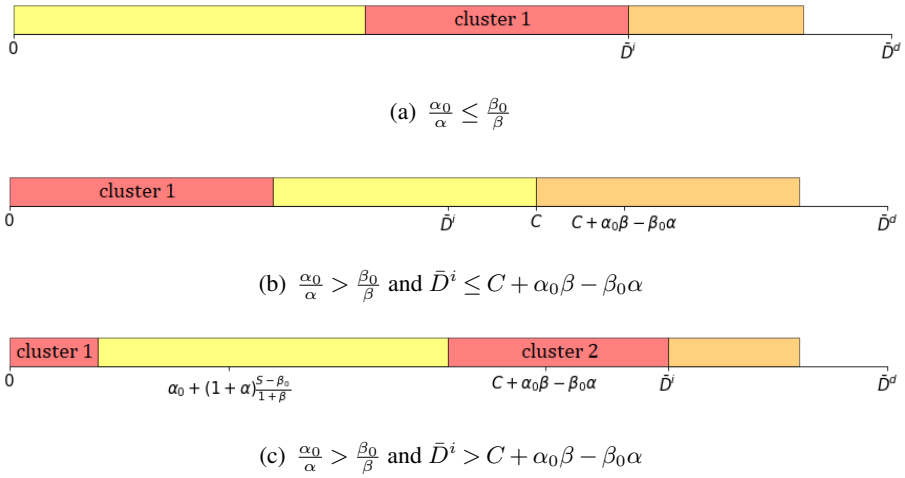


Figure 2 The structure of optimal policy in the unstable victims-only model

than immediate patients. Recall that the provider's objective is to minimize the number of deteriorated delayed patients. This drives her to "squeeze" in as many delayed patients as possible so that they will be served before deterioration. If the due time of immediate patients D^i is relatively long, it is possible for the provider to serve some, if not all, delayed patients before immediate ones, without exceeding the due time requirement for immediate patients; see Figure 2(c). In sum, the optimal schedule depends on the tradeoff of the two possibly competing objectives of the provider: minimizing the number of deteriorated cases calls for serving delayed patients earlier, but the due time requirements entail controlling the makespan which may drive the provider to serve delayed patients later when delayed patients have a larger I2D ratio.

Next, we discuss in detail the optimal schedule for each case shown in Figure 2. Figures 2(a) shows how an optimal schedule looks like when $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$. Specifically, all immediate patients are served within one cluster and as late as possible before $\bar{D}^i$. (We call a time period in which a same type of patients are served as a "cluster".) This not only minimizes the service completion time of all patients, but also helps minimize the number of deteriorated delayed patients. After scheduling all immediate patients, if there is any time left before $\bar{D}^i$, one should serve as many delayed patients as possible; if not all delayed patients could be served before $\bar{D}^i$, the rest of them will be served after $\bar{D}^i$ but before $\bar{D}^d$, shall a feasible schedule exist. To see why this structure is optimal, consider otherwise. Suppose in an optimal schedule, a delayed patient is served within the cluster of immediate patients. Then switching this delayed patient with the immediate one right before him, while maintaining the sequence of other patients, clearly does not increase the number of deteriorated patients. At the same time, this swap may reduce both of the makespans for serving delayed and immediate patients, resulting in a feasible schedule. Iterating such swaps will result in an optimal schedule which has the structure shown in Figures 2(a). To see why such a swap reduces the makespans, consider two cases. If the delayed patient does not change his deterioration status after the swap, then Proposition 1 and Lemma 1 directly explain it. Otherwise, this delayed patient must be deteriorated before the swap but

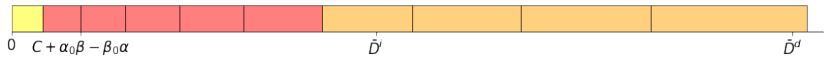
not deteriorated after the swap. While Proposition 1 and Lemma 1 cannot be applied directly, we are able to show that the makespan is indeed shortened after the swap—this case also highlights a key difference between our model and those considered in the previous scheduling literature, where the I2D ratio remains the same for all jobs throughout.

When $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$, immediate patients have a smaller I2D ratio than that of delayed patients served before S. Setting due times aside, compared with delayed patients who do not deteriorate, immediate patients have a higher priority to be served in order to minimize the makespan (Proposition 1); equivalently, this priority order maximizes the number of patients served within a given time window. The optimal schedule, however, becomes more complex as it depends on whether $\bar{D}^i$ is small or large. If $\bar{D}^i \leq C + \alpha_0\beta - \beta_0\alpha$ as shown in Figure 2(b), then $\bar{D}^i$ is sufficiently small such that if a delayed patient were served after all immediate patients had been served, this delayed patient would have not deteriorated. In other words, it would be better off serving this delayed patient after all immediate patients have been served, because this shortens the makespan and may even help reduce the number of deteriorated delayed patients.

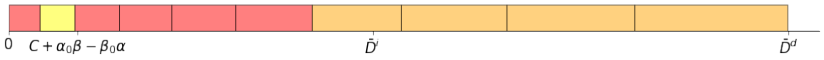
If $\bar{D}^i > C + \alpha_0\beta - \beta_0\alpha$, this implies that some immediate patients could have service completed after $C + \alpha_0\beta - \beta_0\alpha$. From the discussion above, we know that immediate patients whose service completion time $< C + \alpha_0\beta - \beta_0\alpha$, if any, have priority than delayed patients. Now, for immediate patients completing service after $C + \alpha_0\beta - \beta_0\alpha$, we show that delayed patients have priority over these immediate patients if any. To understand this, let us tag an immediate patient who completes service after $C + \alpha_0\beta - \beta_0\alpha$. Suppose that a delayed patient starts service right after this tagged immediate patient. It turns out that swapping this delayed patient with the tagged immediate patient may or may not save this delayed patient from deterioration; but regardless, this swap shortens the overall makespan. Therefore, giving delayed patients higher priority than immediate patients whose service completion time $> C + \alpha_0\beta - \beta_0\alpha$ does not lead to more deteriorated cases but actually makes the solution more likely to be feasible. As a result, the optimal schedule before D^i in this case has a simple sandwich-like structure: immediate–delayed–immediate, as shown in Figure 2 (c).

To minimize the number of deteriorated patients as shown in the case of Figure 2 (c), we should reserve as much time as possible to serve delayed patients before $C + \alpha_0\beta - \beta_0\alpha$. One intuitive way is to schedule as many immediate patients as possible from D^i backward, so that the first immediate patient's service completion time $> C + \alpha_0\beta - \beta_0\alpha$, i.e., to maximize cluster 2 in Figure 2 (c). Then, schedule the rest of immediate patients in cluster 1 and insert as many delayed patients as possible between cluster 1 and cluster 2. But, as demonstrated in the following example, this intuitive schedule may not even be feasible. Consider a setting with 5 immediate and 5 delayed patients at time 0, i.e., $N^i = N^d = 5$. Suppose that $p^i(w) = 30 + 0.2w$, $p^d(w) = 30 + 0.1w$, $S = 30$, $\bar{D}^i = 350$, and $\bar{D}^d = 750$. It follows that $C = 66$ and $C + \alpha_0\beta - \beta_0\alpha = 69$. Figure 4 (a) shows the intuitive schedule, where all 5 immediate victims can be scheduled from D^i backwards such that the first immediate victim's finish time does not pass 69, but this intuitive

schedule is not even feasible. The optimal schedule is shown in Figure 4 (b) where only 4 immediate victims are scheduled in cluster 2 and 1 immediate victim is scheduled in cluster 1, leaving space for 1 delayed victim served before deterioration.



(a) Infeasible schedule with 5 immediate patients scheduled from time $\bar{D}^i$ backwards



(b) Optimal schedule

Figure 3 Comparison of different schedules in a numerical example with $N^i = N^d = 5$ (each block represents one patient)

As the above example shows, the optimal schedule when $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$ and $\bar{D}^i > C+\alpha_0\beta-\beta_0\alpha$ is not as simple as other cases. Theorem 1(c) provides perhaps the most structural results one could obtain. To derive the optimal schedule, one needs to do an exhaustive search by varying the number of immediate patients served in cluster 1 (or cluster 2) in Figure 2 (c). We use ST(cluster) and FT(cluster) to denote the start time and finish time of a cluster, respectively. Algorithm 1 details the steps to derive the optimal schedule.

Algorithm 1: Optimal schedule in the unstable victims-only model when $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$ and $\bar{D}^i > C+\alpha_0\beta-\beta_0\alpha$ (see Figure 2(c) for an illustration)

Schedule N^i immediate patients consecutively from time 0 forward. Let $\bar{m} \le N^i$ be the largest number of immediate patients so that the last one's service completes no later than $C+\alpha_0\beta-\beta_0\alpha$.

Initialization: $m \leftarrow 0$;

while $m \le \bar{m}$ **do**

1. Schedule m immediate patients forwards from time 0—denote this cluster as cluster 1;
2. Schedule $N^i - m$ immediate patients backwards from $\bar{D}^i$—denote this cluster as cluster 2;
3. Insert delayed patients forwards from FT(cluster 1). If necessary, move the next cluster backwards to ensure no idle time in the schedule;

if *The resulting schedule is feasible* **then** 4. Record this schedule;

5. $m \leftarrow m+1$;

end

6. Compare all feasible schedules and find the optimal one.

4.2. Model with Both Unstable and Stable Victims

In this section, we consider the model with both unstable and stable victims, i.e., $N^i,\ N^d,\ N^{si},\ N^{sd} > 0$. For a stable immediate patient, his due time is the same as that of an unstable immediate patient, i.e., D^i, but his procedure time is a constant p^{si}. To be consistent, we follow our earlier notation for procedure times. Thus, for an immediate patient who waits w units of time before service, his procedure time, denoted by $p^i(w)$, is

$$p^i(w) = \begin{cases} \beta_0 + \beta w, & \text{if the immediate patient is unstable,} \\ p^{si}, & \text{otherwise.} \end{cases} \tag{9}$$

For a stable delayed patient, he has the same switch time S and due time D^d as unstable delayed patients. While a stable delayed patient deteriorates after S, his procedure time remains a constant p^{sd}. Thus, if a delayed patient waits w units of time before his surgery, his procedure time, denoted as $p^d(w)$, is

$$p^d(w) = \begin{cases} \alpha_0 + \alpha w + (\beta - \alpha)(w - S)^+, & \text{if the delayed patient is unstable,} \\ p^{sd}, & \text{otherwise.} \end{cases} \tag{10}$$

The objective of the provider remains the same. She wants to minimize the number of delayed patients (both stable and unstable) who receive treatment after their switch time S, while ensuring that the lateness of serving patients (both stable and unstable) is no longer than $Q \geq 0$. Following the earlier notation, we can concisely represent this problem as follows.

$$1\big|p_j^d,\ S,\ d_j^d = D^d :\ p_j^i,\ d_j^i = D^i\big|(\sum U_j^d, L_{max}^d \leq Q) :\ L_{max}^i \leq Q, \tag{11}$$

where p_j^d is defined in (10) and p_j^i is defined in (9). Note that the objective $\sum U_j^d$ now includes both stable and unstable delayed patients who receive treatment after S; L_{max}^d (L_{max}^i resp.) is the maximum lateness of all delayed (immediate resp.) patients including both stable and unstable ones.

With stable patients included in the model, the optimal schedule becomes more complicated than before. If the sole objective is to minimize the makespan, then all unstable patients should be served before all stable patients. However, the provider's objective is about maximizing the number of delayed patients who start service before their switch time S, while controlling the makespan. This drives the provider to serve some stable delayed patients early in the schedule when possible, especially if they have a short procedure time. Recall that in the base model, we differentiate three cases depending on the parameter setting (see Figure 2). When stable patients enter into the equation, we need to consider one more dimension. Specifically, we differentiate whether p^{sd}, the procedure time of a stable delayed patient, is longer than $\alpha_0 + \alpha S$ or not, for reasons to be discussed next.

Next, we focus on the structure of the optimal schedule to shed light on the key insights in this model. Theorem 2 summarizes perhaps all provable structural results for the optimal schedule. We use ST and FT to represent the start time and finish time of patient service, respectively. To facilitate understanding, Figure

4 shows some representative structures of the optimal schedule for each of the eight parameter settings. We use red to denote unstable immediate patients, pink for stable immediate patients, yellow for unstable delayed patients, and beige for stable delayed patients. For simplicity, we do not differentiate deteriorated delayed patients from those who have not deteriorated. (Algorithms to derive the exact optimal schedule depend on the parameter setting and are tedious; we defer them to Appendix B. For ease of understanding, these algorithms make references to different patient clusters marked in Figure 4.)

Theorem 2 *If there exists an optimal schedule for problem* (11)*, then it is non-idling and has the following properties.*

(a) Unstable immediate patients have higher priority than stable immediate patients.

(b) Unstable immediate patients with FT $\leq S + (1+\beta)p^{sd}$ *have higher priority than stable delayed patients.*

(c) Unstable delayed patients with FT $\leq S + p^{sd}$ *have higher priority than stable delayed patients.*

(d) Unstable delayed patients have higher priority than stable delayed patients if $p^{sd} \geq \alpha_0 + \alpha S$.

(e) Among patients with ST $> S$*, unstable ones (both immediate and delayed) have higher priority than stable delayed ones.*

(f) Consider the situation where $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$ *and focus on patients with FT* $\leq \bar{D}^i$*, then*

- *if* $p^{sd} \geq \alpha_0 + \alpha S$*, unstable delayed patients have higher priority than unstable immediate patients and stable patients (both immediate and delayed patients).*
- *if* $p^{sd} < \alpha_0 + \alpha S$*, then among patients with FT* $\leq S + p^{sd}$*, unstable delayed patients have higher priority than unstable immediate patients; among patients with ST* $> S$*, unstable delayed ones have deteriorated and have priority over unstable immediate ones.*

(g) Consider the situation where $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$*, then*

- *if* $\bar{D}^i \leq \min\{C + \alpha_0\beta - \alpha\beta_0,\ S + (1+\beta)p^{sd}\}$*, all unstable immediate patients should be served consecutively from time 0 and thus have higher priority than unstable delayed patients.*
- *if* $\bar{D}^i > \min\{C + \alpha_0\beta - \alpha\beta_0,\ S + (1+\beta)p^{sd}\}$*, then among patients with FT* $\leq \bar{D}^i$*, unstable immediate ones with FT* $< \min\{C + \alpha_0\beta - \alpha\beta_0,\ S + (1+\beta)p^{sd}\}$ *have priority over unstable delayed ones, while unstable immediate ones with FT* $\geq C + \alpha_0\beta - \alpha\beta_0$ *are served consecutively and have lower priority than unstable delayed ones if any.*

In general, Theorem 2(a)-(e) characterize the priority order between unstable and stable patients while (f) and (g) concern that among unstable patients. We start by explaining Theorem 2(a). Immediate patients need to be served before $\bar{D}^i$. If in a feasible schedule a stable immediate patient is served before an unstable immediate one and no delayed patients are served between them, then flipping this pair of patients while keeping the service order of other patients unchanged can reduce the makespan and still leads to a feasible schedule; the swap may even lower the number of deteriorated delayed patients. If there are delayed patients

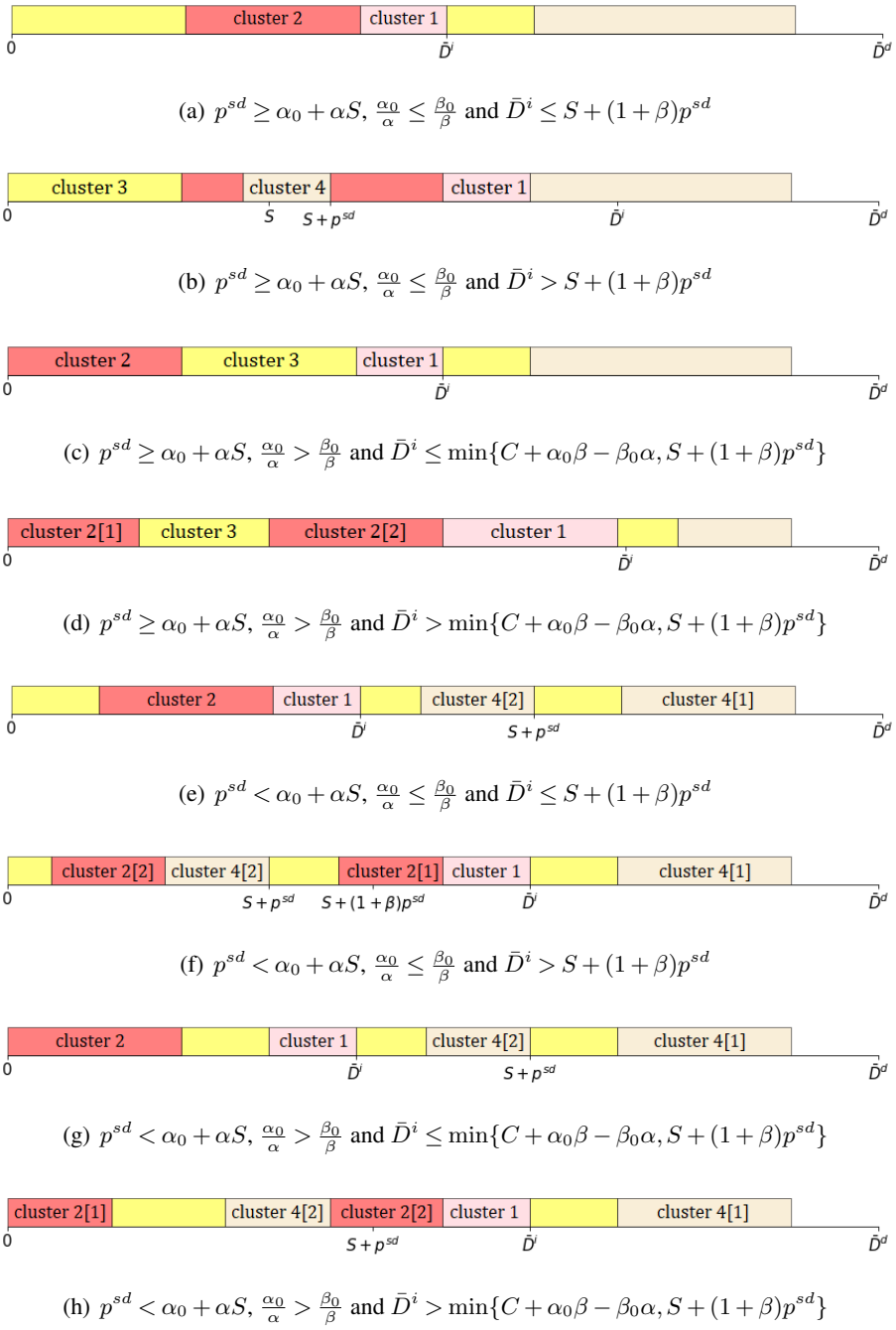


Figure 4 Representative Structures of the Optimal Schedule in the Model with Both Stable and Unstable Victims

served between them, swapping the positions of the stable immediate patient and delayed patients allows the delayed patients to receive earlier treatment, without lengthening the overall completion time of these patients. Now, no delayed patients are served between the stable immediate patient and the unstable immediate one. Following the earlier discussion to do another swap results in the desired structure. Thus, there

exists an optimal schedule where unstable immediate patients have higher priority than stable immediate patients.

Theorem 2(b) says that unstable immediate patients, if placed sufficiently early in the schedule, should be served before any stable delayed patient. If in a feasible solution, a stable delayed patient is served before one of such early-placed unstable immediately patients, then according to Theorem 2(a), there are no stable immediate patients served between them. If there is no unstable delayed patients between the stable delayed patients and this early-placed unstable immediate patient, then flipping this pair of patients would clearly reduce the makespan, but would not increase the number of deteriorated delayed patients, because this stable delayed patient would be served sufficiently early and not deteriorate in the newly flipped schedule. If there are unstable delayed patients served between them, swapping the positions of these unstable delayed patients and the stable delayed patient decreases the overall completion time. We can further show that after the swap, both the stable and unstable delayed patients would not deteriorate because they are served sufficiently early. After this swap, no unstable delayed patients exist between the stable delayed and the early-placed unstable immediate one; following the earlier discussion to do another swap leads to the priority order described in Theorem 2(b).

Next, let us look at the priority order of unstable delayed patients versus others. Theorem 2(c) says that unstable delayed patients, if placed sufficiently early in the schedule, should be served before any stable delayed patients. The rationale behind this result is similar to that of Theorem 2(b).

If $p^{sd} \geq \alpha_0 + \alpha S$, then for delayed patients whose services start before S, the processing time of stable delayed patients is always longer than unstable delayed patients. Therefore, to maximize the number of delayed patients served before S, unstable delayed patients should have higher priority than stable delayed patients before S. After S, however, all delayed patients would deteriorate; serving stable delayed patients after unstable delayed ones shortens the makespan and helps the provider serve all patients before their respective due times. This explains Theorem 2(d).

Theorem 2(e) is intuitive. Changing the service order of patients in a schedule after the switch time S does not affect the number of deteriorated delayed patients. Therefore, after S, unstable patients (both immediate and delayed) should be served before stable delayed patients, because this reduces the makespan and helps to finish service on time.

Theorem 2(f) discusses the structure of the schedule when $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$. Without stable patients, we know that unstable delayed patients have higher priority than unstable immediate patients before $\bar{D}^i$; see Theorem 1(a). With stable patients, we want to know if the same priority order remains. Consider $p^{sd} \geq \alpha_0 + \alpha S$ and suppose that in an optimal schedule an unstable immediate patient is served before an unstable delayed patient, then we know that no stable delayed patients are served between them (Theorem 2(d)). To minimize the makespan, stable immediate patients should be served as late as possible, so before $\bar{D}^i$ no stable immediate patients are served before unstable immediate patients and delayed ones (both stable and unstable).

However, if we keep the positions of other patients unchanged but only swap the positions of the aforementioned unstable immediate patient and unstable delayed patient so that the unstable delayed one is served first, we shorten the makespan and may also reduce the number of deteriorated (Theorem 1(a)), leading to an improvement of the schedule. This reasoning implies that giving unstable delayed patients the highest priority before $\bar{D}^i$ is the optimal choice here.

If $p^{sd} < \alpha_0 + \alpha S$, then it is more complicated because stable delayed patients may have shorter processing times than unstable delayed ones. Thus it might be better off serving some stable delayed patients before serving all unstable delayed patients. While this may increase the overall makespan, it can reduce the number of deteriorated delayed patients. This is not the only complicating issue brought by stable delayed patients. Without stable delayed patients, unstable delayed patients always have higher priority than unstable immediate patients before $\bar{D}^i$ if $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$ (Theorem 1(a)). However, with stable delayed patients, the strict priority order between unstable delayed and unstable immediate patients before $\bar{D}^i$ as pressed for by the condition $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$ does not hold. What we can show is that in the early portion and the late portion of the schedule before $\bar{D}^i$, unstable delayed patients have higher priority than unstable immediate ones, but this priority order may not prevail throughout; see Figure 4(f) where this priority only holds before and after the stable delayed patient cluster 1[2]. The key reason is that, with this cluster of stable delayed patients in the middle, only serving unstable immediate patients after this cluster and giving service priority to unstable delayed patients before it may actually increase the completion time of all patients. Appendix C provides a detailed explanation on this.

Theorem 2(g) presents results similar to those in Theorem 1 (b) and (c). When $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$ and $\bar{D}^i$ is sufficiently small, then unstable immediate patients have the highest priority. If $\bar{D}^i$ becomes large, then some delayed patients can be inserted among immediately patients to reduce the number of deteriorated delayed patients; see Figures 4(d) and (h).

In sum, stable patients usually should have lower priority than their unstable counterparts due to their time-invariant service times; see Theorem 2(a), (c), and (d). However, as discussed above, when stable delayed patients have sufficiently short service times, i.e., $p^{sd} < \alpha_0 + \alpha S$, this priority does not hold in the optimal schedule and the provider may want to schedule some stable delayed patients early (and perhaps earlier than some immediate ones) to reduce deteriorated cases; see Figure 4(e)-(h). Among unstable patients, delayed ones may have higher priority than (some of the) immediate ones, contrary to the common wisdom. The main driving forces are similar to those discussed early in the model with unstable patients only. Delayed patients tend to gain higher priority than immediate ones in an optimal schedule when 1) they have a smaller I2D ratio (see Figure 4(a), (b), (e), and (f) and Theorem 2(f)); or 2) the due time of immediate patients $\bar{D}^i$ is relatively long compared to the switch time of delayed ones S, so the provider can squeeze in some delayed ones before finishing the service of all immediate ones in time (see Theorem 2(g) and Figure 4(d) and (h)).

5. Model Extensions

In this section, we discuss three extensions to our model. For ease of presentation, we shall assume that all victims are unstable, i.e., their procedure times increase in their wait times. Model extensions with both stable and unstable victims included can be analyzed in a similar way. We first investigate the situation where providers need to take a mandated break during service to reduce fatigue, then study how to handle a potential second wave of patient arrivals, and finally consider a setting with multiple surgical teams.

5.1. Mandated Rest for Providers

In our base model, we assume that providers work continuously from time 0 until the service completion of the last patient. However, long work hours lead to provider fatigue, which may result in worse clinical outcomes (Gaba and Howard 2002). Therefore, providers are often advised, and sometimes mandated, to take breaks for mental and physical rest and regeneration after working for long hours (Janhofer et al. 2019).

Next, we consider how to incorporate mandated rest for providers in our model. Rescue surgical teams usually take no more than one single break during their whole shift. Accordingly, we suppose that the providers take a rest of r units of time immediately after working for τ or more units of time (surgeries are non-preemptive). This rest period divides the schedule into two parts: pre-rest schedule and post-rest schedule. Interestingly, the structural properties of the optimal schedule established in Theorem 1 still hold if we only look at the pre-rest schedule—or—the post-rest one alone. Proposition 2 formalizes this result.

Proposition 2 *If there exists an optimal schedule for the model with a mandated rest period, then the following holds.*

(a) When $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$, among patients with wait times $< \tau$, immediate patients should be served consecutively before $\bar{D}^i$ and as many delayed patients as possible should be served before immediate patients. The same priority result also holds for patients with wait times $\geq \tau$.

(b) When $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$, among patients with wait times $< \tau$, then immediate patients, if any, completing treatment before $C + \alpha_0\beta - \beta_0\alpha$ are served consecutively from time 0 and no delayed patients are served before them. Immediate patients who complete treatment after $C + \alpha_0\beta - \beta_0\alpha$, if any, are served consecutively and as many delayed patients as possible should be served before immediate patients. The same priority results also hold for patients with wait times $\geq \tau$.

One may wonder why the same structure for the optimal schedule described in Theorem 1 continues to hold for the pre-rest and post-rest schedules, respectively. Here is an intuitive explanation. The structural results established in Theorem 1 hold for patients with any wait time throughout the schedule, as long as the corresponding conditions are met. Therefore, if we only focus on the set of patients treated in the pre-rest schedule, i.e., those with wait times $< \tau$, the same structural results should still apply for this set of patients.

And the same argument works for patients treated in the post-rest schedule. While Proposition 2 now seems intuitive, it is the stepping stone for deriving a simple solution algorithm in this setting.

Figure 5 shows a general stricture for the optimal schedule with a mandated rest period when $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$. We use n_1 to denote the number of delayed patients in cluster 1, i.e., those with wait times $< \tau$ and service finish times $\leq \bar{D}^i$. We let m_1 be the number of immediate patients in cluster 2, i.e., those with wait times $< \tau$, so the number of immediate patients in cluster 3 is $N^i - m_1$. The green block "r" represents the rest period. One just needs to compare all feasible schedules with this structure to obtain the optimal schedule.

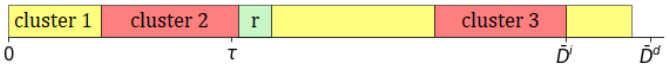


Figure 5 The structure of the optimal schedule with a mandated rest period when $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$

Next, we move to discuss the situation when $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$. Depending on whether $\tau \leq C + \alpha_0\beta - \beta_0\alpha$ or not, the structure of the optimal schedule is different; see Figure 6. We use m_1 to represent the number of immediate patients with wait times $< \tau$, including m_1^s immediate patients with service finish times $\leq C + \alpha_0\beta - \beta_0\alpha$ and $m_1 - m_1^s$ immediate patients with service finish times $> C + \alpha_0\beta - \beta_0\alpha$. Given m_1, the number of immediate patients with wait times $\geq \tau$ is $N^i - m_1$. We use m_2^s to denote number of immediate patients with wait times $\geq \tau$ and service finish times $\leq C + \alpha_0\beta - \beta_0\alpha$, so $m_2^s \leq N^i - m_1$. Note that either $m_1 - m_1^s = 0$ or $m_2^s = 0$ by definition. Finally, we use n_1 to denote the number of delayed patients with wait times $< \tau$ and service finish times $\leq \bar{D}^i$. A combination of $(m_1,\ m_1^s,\ m_2^s,\ n_1)$ uniquely defines a schedule. Then, to search for the optimal schedule, one just needs to compare schedules with structures shown in Figure 6 by properly varying m_1, m_1^s, m_2^s, and n_1. The detailed algorithms can be found in Appendix D.

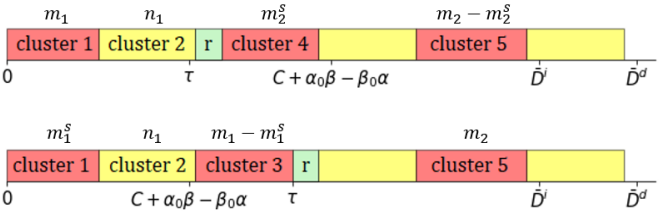


Figure 6 The structure of the optimal schedule with a mandated rest period when $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$

5.2. Arrival of a Second Batch

In MCIs, especially in earthquakes, rescued victims to be treated by surgical teams on site usually arrive in a single batch before their surgical operations. While it is rare, a surgical team may receive a second batch of victims during its operations. We consider how to handle a potential second batch of arrivals of victims.

Since the chance of having a second batch is very low, we do not plan the original schedule in anticipation of such a batch. Instead, we will re-optimize the schedule shall a second batch arrive.

Suppose that the second batch is ready for surgeries at the same time as the first batch, but arrives at time $t_w > 0$ due to delay in transportation. Similar to the first batch, the second batch has two types of victims at time 0: immediate victims and delayed victims, who have the same characteristics as their counterparts in the first batch. Specifically, in the second batch, immediate and delayed victims have procedure times described by (3) and (4), respectively; their due times are D^i and D^d, respectively; delayed patients have a switch time of S. The decision maker has the same objective as before; she seeks to minimize the number of deteriorated delayed patients in total while ensuring the lateness of all victims not to exceed Q.

Before our analysis, we first note that both types of victims in the second batch have already waited t_w units of time upon their arrival at the surgical team. We assume no preemption in surgical operations. We let $t_0 \geq t_w$ be the earliest time the surgical team can take on a new patient. Therefore, t_0 is the service completion time of the patient who is being served at time t_w if there is such a patient, or, $t_0 = t_w$ if the surgical team has already served the whole first batch by time t_w. Depending on t_0, we consider the following three cases. We let $W^i \geq 0$ and $W^d \geq 0$ to represent the number of immediate and delayed victims in the second batch at t_0, respectively.

First, if $t_0 \leq S$, then delayed victims in the second batch, if any, have not deteriorated yet. At t_0, we just need to update the number of immediate and delayed victims in the system by adding the second batch to the remainders in the first batch, and then re-calculate the optimal schedule based on Theorem 1 and the algorithms discussed in Section 4.1.

Second, if $S < t_0 < \bar{D}^i$, then delayed victims in the second batch, if any, have already deteriorated. In this case, the number of deteriorated delayed patients have already been determined, and the provider should only be concerned with meeting the requirements on due times. All immediate patients need to be served before $\bar{D}^i$. To save more time for other patients, the provider should try to minimize the makespan of all patients served before $\bar{D}^i$. In order to do that, deteriorated delayed patients should have priority than immediate ones before $\bar{D}^i$ by Lemma 1. Following this logic, Algorithm 2 below re-optimizes the schedule in this case.

Algorithm 2: Re-optimize the model with a second batch of arrivals when $S < t_0 < \bar{D}^i$

Initialization: add W^d to the set of delayed patients and W^i to the set of immediate victims;

1. Schedule all remaining immediate patients backwards from $\bar{D}^i$ to time t_0;
2. Insert all delayed patients from t_0 to $\bar{D}^d$. Remove idling time by moving patients backwards if necessary.

Lastly, if $\bar{D}^i \leq t_0 < \bar{D}^d$, then all immediate patients in the first batch should have already been served in a feasible schedule. We shall assume that $W^d \geq 0$ and $W^i = 0$ (for otherwise we would not have a feasible

schedule). In this case, only deteriorated delayed patients are left to be served and they should be served one by one.

5.3. Multiple Surgical Teams

In this section, we consider a setting with multiple surgical teams to coordinate. We assume that each team is identical. To differentiate from the single-team setting, we assume that M^d delayed victims and M^i immediate victims arrive at time 0. The question is how to allocate these victims upon their arrival to different teams so that the total number of deteriorated delayed patients is minimized, while for each team the lateness requirements are also satisfied. Leveraging our earlier analysis of single-team operations, we develop the following dynamic program to solve the case with multiple teams.

Suppose that there are $K \geq 2$ provider teams. Let x_j^i (resp. x_j^d) be the number of immediate (delayed) patients allocated to team j, $j = 1, 2, \ldots, K$. Define

$$M_j^i = M^i - \sum_{k=1}^{j-1} x_k^i \quad \text{and} \quad M_j^d = M^d - \sum_{k=1}^{j-1} x_k^d.$$

That is, M_j^i (resp. M_j^d) represents the number of immediate (resp. delayed) patients left to be allocated among teams j through K, $j = 1, 2, \ldots, K$. Let $\sigma(x^i, x^d)$ be the minimum number of deteriorated delayed patients in a provider team if x^i immediate patients and x^d delayed patients are allocated to this team. Specifically, we define

$$\sigma(x^i, x^d) = \begin{cases} \text{the objective of Problem (8) with } (N^i, N^d) = (x^i, x^d), & \text{if feasible,} \\ \infty, & \text{otherwise.} \end{cases} \tag{12}$$

Finally, we let $f_j(M_j^i, M_j^d)$ be the minimal number of deteriorated delayed patients if M_j^i immediate patients and M_j^d delayed patients are allocated among teams j through K, $j = 1, 2, \ldots, K$. Then, we can evaluate $f_j(M_j^i, M_j^d)$ recursively for $j = K-1, K-2, \ldots, 1$ as follows.

$$f_j(M_j^i, M_j^d) = \min_{0 \leq x_j^i \leq M_j^i,\ 0 \leq x_j^d \leq M_j^d} \left[f_{j+1}(M_j^i - x_j^i, M_j^d - x_j^d) + \sigma(x_j^i, x_j^d) \right], \quad 0 \leq M_j^i \leq M^i,\ 0 \leq M_j^d \leq M^d, \tag{13}$$

with the boundary condition

$$f_K(M_K^i, M_K^d) = \sigma(M_K^i, M_K^d), \quad 0 \leq M_K^i \leq M^i,\ 0 \leq M_K^d \leq M^d.$$

Solving for $f_1(M^i, M^d)$ provides the optimal objective for the model with multiple surgical teams and also the optimal schedule for each team.

6. Robustness Test and Case Study

Our model assumes that surgical procedure times are deterministic given the wait time of victims. In this section, we first conduct a robustness test of our model in settings where procedure times have variability. The test results show that our model can provide quality solutions in environments with uncertainties. Then, using our data recorded in the field, we carry out a counterfactual analysis to compare the performance of our model with that under the then-implemented policies. This case study demonstrates the value of adopting our model in practice.

6.1. Robustness Test

To populate model parameters for our robustness tests, we use regression results in Section 3. Specifically, we use the reduced model, where the standard deviation of the noise term is estimated to be 19.66 minutes. Therefore, the procedure times of victims in our tests are specified as follows.

$$p_i = \begin{cases} 58 + 0.11 \times w_i + 0.24 \times (w_i - 470)^+ + \epsilon_i, & \text{if the victim is a delayed type,} \\ 58 + 0.35 \times w_i + \epsilon_i, & \text{otherwise,} \end{cases} \tag{14}$$

where w_i is the wait time of victim i and $\{\epsilon_i,\ i = 1, 2, 3 \dots\}$ is assumed to be a sequence of i.i.d. normal random variables with mean 0 and standard deviation 19.66.

In the tests, we generate 1000 samples, each of which contains 7 delayed victims and 3 immediate victims. Their procedure times are randomly sampled based on (14). We assume that delayed victims have no due times, i.e., $D^d = \infty$. To set a sensible due time for immediate victims, we use 60% of the expected makespan of serving these 10 victims under the START policy, which treats 3 immediate victims first followed by the delayed ones. The due time for immediate victims D^i is thus set as 795 minutes. We vary Q, the maximum tardiness, from 0 to 60 minutes with a step-size of 10 minutes to observe the impact of Q. For each of these 1000 samples, we compare various performance metrics under the START policy and the schedule informed by our model (8).

Under the START policy, immediate victims are prioritized and it turns out that immediate victims are always treated before $D^i + Q$ across all samples in our experiments. To reduce the number of deteriorated delayed victims, our model (8) suggests different from START and calls for serving 2 immediate victims, followed by 4 delayed victims, 1 immediate victim, and finally 3 delayed victims. Our schedule turns out to be the same for different values of Q we tested.

Since the value of Q does not affect the schedule in our tests, the random components in procedure times realized in each sample are a key driving force for the number of deteriorated victims in that sample. It turns out that for the same sample and schedule, the number of deteriorated victims is also the same for different values of Q. This is likely resulting from the fact that the variability in procedure times is relatively small compared with the magnitude of procedure times. Across 1000 samples, the average number of deteriorated

delayed victims under START is 3.97 out of 7 delayed ones, while that under our schedule is 3.06, marking a 23% reduction. This result is consistent for all values of Q. A paired t-test shows that the reduction made by our schedule compared with START is statistically significant with p-value < 0.01.

Due to randomness in procedure times, immediate patients may be treated after $D^i + Q$ under our schedule—if this happens we call the schedule *tardy*. We next evaluate how likely our schedule is tardy and the extent of the tardiness if the schedule is indeed tardy. Row 2 in Table 5 shows the percentage of our schedules being tardy in 1000 samples for different Q's. If our schedule is tardy for one sample, we further evaluate, among those immediate victims who receive late treatment after $D^i + Q$, the percentages of those who get served within 20, 40, and 60 minutes after $D^i + Q$. Rows 3-5 in Table 5 show the average of these percentages from samples where our schedule is tardy for each Q.

Table 5 Analysis of Schedule Tardiness due to Variability in Procedure Times

Q (min)	0	10	20	30	40	50	60
% of non-tardy schedule under model (8)	81.80%	84.10%	86.70%	89.50%	91.90%	93.10%	94.50%
% of victims with tardiness $\leq$ 60 minutes	69.78%	73.58%	74.44%	75.24%	75.31%	72.46%	72.73%
% of victims with tardiness $\leq$ 40 minutes	55.49%	56.60%	58.65%	60.00%	58.02%	62.32%	63.64%
% of victims with tardiness $\leq$ 20 minutes	26.92%	33.96%	39.10%	34.29%	32.10%	39.13%	38.18%

As expected, when Q increases, our schedule is more likely to be on time. Even when $Q = 0$, our schedule is on time for more than 80% of the times despite the variability in procedure times; if one can tolerate 1 hour lateness after D^i to serve immediate victims, then our schedule can meet the target with 95% probability. What is more assuring is that, even if immediate victims receive treatment later than the target time $D^i + Q$, the tardiness is quite limited under our schedule. We find that more than a quarter of these immediate victims get treated within 20 minutes, more than half treated within 40 minutes, and about three quarters treated within 1 hour, after the target time.

These experiments based on real data provide strong evidences that our model can generate schedules that have robust performances in settings beyond modeling assumptions. While procedure times have variability, our schedule is almost always on time and if not has fairly limited tardiness. This lends support to our modeling choice that assumes procedure times are deterministic (given the wait time of victims) in our study context. It also suggests that our model can be a potentially useful tool in MCI response. The next section evaluates the effectiveness of this tool.

6.2. Case Study

In this section, we carry out a counterfactual analysis to evaluate the potential impact of adopting our model in the 2008 Sichuan earthquake. Recall that in Section 3 we analyze the data recorded from the rescue scene. There are 13 surgical teams providing care to 101 victims. Our analysis has categorized these victims into

four types: stable delayed (SD), unstable delayed (UD), stable immediate (SI), and unstable immediate (UI). Our data also document the actual sequence of surgeries performed by each team. Next we will evaluate, for each surgical team, what would happen if we were to alter the surgical sequence based on our model.

To be fair, we use the following surgical procedure times when evaluating both the schedule actually used and the schedule suggested by our model. When the surgical sequence is altered, surgical procedure times remain the same for stable victims. For these victims, we estimate that the average procedure times for delayed and immediate victims are 93 minutes and 91 minutes, respectively. For unstable victims, their procedure times might change if the surgical sequence becomes different. We use (14) to calculate the procedure time of an unstable victim depending on his type and wait time, assuming the noise term $\epsilon_i = 0$. We do not use the observed procedure times in our evaluation of the actually used schedules, because we have no means to estimate their counterparts for unstable victims when the surgical sequence is altered without resorting to our regression analysis. So a reasonable and fair way appears to use procedure times estimated by our regression model.

For each surgical team, we first evaluate the makespan for the surgical schedule actually used. Then, similar to the robustness test in Section 6.1, we set D^i, i.e. the due time for immediate victims, as 60% of this makespan. We assume $Q = 0$ and that delayed victims have no due times, i.e., $D^d = \infty$. Based on these parameters, we derive the surgical schedule proposed by our model. Table 6 compares the originally implemented schedule and our proposed schedule for each surgical team. We observe that our schedule simultaneously reduces the number of deteriorated victims and shortens the makespan for all teams. In total, the number of deteriorated victims is reduced by 32%, from 25 to 17; the makespan for each team is reduced by 8% on average. These improvements are achieved by better utilization of existing medical resources without adding additional ones; they demonstrate significant potential values of adopting our model in response to an MCI.

6.2.1. Value of Care Coordination In the analysis above, we reorganize the surgical sequence assuming that each surgical team still treats the same set of victims. Among 13 surgical teams, we observe that the number of victims served ranges from 6 to 11. In particular, team A serves 6 victims and none is deteriorated, while team F serves 11 victims and five are deteriorated. If victims could be allocated among teams in a more balanced way, the number of deteriorated may be further reduced. Next, we consider two possibilities of care coordination: partial coordination and complete coordination. In partial coordination, we partition surgical teams into three clusters so that the average number of victims served by a team is roughly the same across clusters. In complete coordination, we pool all victims together and then allocate them to each team. In both cases, the allocation and treatment plans are derived by the dynamic program (13).

Table 7 shows the optimal allocation of victims and the number of deteriorated for each surgical team, under both partial coordination and complete coordination. Compared to no coordination (but with optimized surgical schedule for each team), the partial coordination further reduces the number of deteriorated

Table 6 Summary of Results for Counterfactual Analysis

	Victim Mix				Original Schedule		Proposed Schedule	
Surgical Team	SD	UD	SI	UI	Makespan (mins)	# of deteriorated	Makespan (mins)	# of deteriorated
A	2	2	1	1	522	0	470	0
B	5	2	2	0	810	3	769	3
C	6	2	1	1	863	4	842	3
D	4	3	1	1	768	3	735	2
E	3	2	2	1	775	3	654	1
F	6	2	0	3	987	5	977	5
G	3	3	1	0	722	1	563	0
H	4	3	1	0	689	2	656	1
I	1	4	1	1	604	1	553	0
J	2	2	1	1	550	0	470	0
K	5	0	0	2	601	1	601	1
L	2	1	1	2	578	1	485	0
M	4	0	1	2	599	1	599	1
Total	47	26	13	15	9068	25	8374	17

SD: stable delayed; UD: unstable delayed; SI: stable immediate; UI: unstable immediate.

from 17 to 14, marking another 12% reduction. (Note that we do not optimize the partition of surgical teams and this improvement only reflects the outcome from one particular partition. The outcome can be even better under the optimal partition.) Under the complete coordination, 11 victims would be deteriorated—this is actually the best outcome one could hope for given patient demand (101 victims) and resources available (13 surgical teams). These results demonstrate significant values from care coordination among surgical teams. In particular, a partial coordination makes a significant improvement in patient outcomes than no coordination and achieves most benefits that would have resulted from a complete coordination.

Table 7 Summary of Results under Care Coordination

	Partial Coordination					Complete Coordination				
Team	SD	UD	SI	UI	# of Deteriorated	SD	UD	SI	I	# of Deteriorated
A	4	4	0	0	1	4	3	0	1	1
C	4	3	0	1	1	4	3	0	1	1
F	4	2	1	1	1	4	2	1	1	1
J	3	0	1	3	1	4	2	1	1	1
L	3	0	2	3	2	4	2	1	1	1
B	4	4	0	0	1	4	2	1	1	1
D	3	4	1	0	1	4	2	1	1	1
G	3	2	2	1	1	4	2	1	1	1
I	3	2	2	1	1	4	2	1	1	1
E	7	0	0	0	1	4	2	1	1	1
H	4	3	0	1	1	4	2	1	1	1
K	3	2	2	1	1	2	2	2	1	0
M	2	0	2	3	1	1	0	2	3	0
Total #	47	26	13	15	14	47	26	13	15	11

In reality, complete coordination may be difficult but partial coordination is quite feasible. Victims usually arrive in batches. Triage nurses often present to coordinate care. Therefore, triage nurses could help allocate victims, upon their arrival, to the surgical teams they are in charge of. Our model offers a way to improve this allocation process and suggests that even some coordination can make a remarkable difference.

7. Concluding Remarks

In this paper, we develop a scheduling model to guide treatment planning for field hospitals in response to MCIs. We consider two types of patients differing in their initial health conditions: immediate and delayed. Both types of patients have their respective due times. The assumptions on patient service times are informed by our analysis of a real earthquake dataset. The service time of an immediate patient increases linearly in his wait time. A delayed patient's service time is an increasing piecewise linear function of his wait time. In particular, after waiting for a certain threshold, delayed patients deteriorate and their service times increase at a faster rate. The goal of the scheduling model is to minimize the number of deteriorated cases subject to the due time requirements. In addition to this base model, we study several extensions to incorporate practical considerations. We identify conditions under which treatment priority should (and should not) be given to delayed patients rather than immediate ones in order to do the greatest good for the greatest number. Our numerical study demonstrates the robustness of our model in settings that go beyond the modeling assumptions on service time. A counterfactual analysis based on our data of 13 surgical teams shows that adopting our model would significantly reduce the number of deteriorated cases as well as the surgical makespan and care coordination among surgical teams, even partially, can lead to significant improvement in patient outcomes.

A distinguishing feature of our modeling framework, in contrast to those considered in the previous literature, is to simultaneously consider patient deterioration and wait-dependent service times in making scheduling decisions. By capturing these essential features of surgical operations in field hospitals, our models hold strong potentials to improve emergency response and its policy making. Our work demonstrates the value of adopting data-driven approaches in MCI response and suggests that MCI response policies and guidelines should rely more on data and scientific modeling approaches rather than common wisdom and simple heuristics. The current policies suggest prioritizing treatment of victims solely based on their initial health conditions. Our research emphasizes that victim deterioration trajectory is another key factor that should be explicitly taken into account. Furthermore, we demonstrate significant improvement in health outcomes that can result from better allocation of workload among rescue teams in MCI response and therefore call for guidelines designed to fully take advantage of care coordination.

Our work points to several avenues for future research. First, it would be interesting to conduct counterfactual analyses using other MCI data to evaluate the performance of our model. Second, it would be

valuable to build a clinical database from previous MCIs and to extend efforts in further developing data-driven solutions for MCI response. Third, given that data from MCIs may be limited, it would be meaningful to study if clinical data collected in regular hospital operations can be used to inform patient characteristics in MCIs, e.g., to help predict victim type and procedure time. We leave these research topics for the future.

References

A. Agnetis, P. B. Mirchandani, D. Pacciarelli, and A. Pacifici. Scheduling problems with two competing agents. *Operations research*, 52(2):229–242, 2004.

A. Agnetis, D. Pacciarelli, and A. Pacifici. Multi-agent single machine scheduling. *Annals of Operations Research*, 150(1):3–15, 2007.

A. Ahmadi-Javid, Z. Jalali, and K. J. Klassen. Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, 258(1):3–34, 2017.

B. Alidaee and N. K. Womer. Scheduling with time dependent processing times: review and extensions. *Journal of the Operational Research Society*, 50(7):711–720, 1999.

N. T. Argon, S. Ziya, and J. E. Winslow. Triage in the aftermath of mass-casualty incidents. *Wiley Encyclopedia of Operations Research and Management Science*, 2010.

K. R. Baker and J. C. Smith. A multiple-criterion model for machine scheduling. *Journal of scheduling*, 6(1):7–16, 2003.

S. Browne and U. Yechiali. Scheduling deteriorating jobs on a single processor. *Operations Research*, 38(3):495–498, 1990.

C. W. Chan, V. F. Farias, and G. J. Escobar. The impact of delays on service times in the intensive care unit. *Management Science*, 63(7):2049–2072, 2017.

G. Chen, W. Lai, F. Liu, Q. Mao, F. Tu, J. Wen, H. Xiao, J.-c. Zhang, T. Zhu, B. Chen, et al. The dragon strikes: lessons from the wenchuan earthquake. *Anesthesia & Analgesia*, 110(3):908–915, 2010.

T. E. Cheng, Q. Ding, and B. M. Lin. A concise survey of scheduling with time-dependent processing times. *European Journal of Operational Research*, 152(1):1–13, 2004.

CNBC. Strong earthquake kills 28 people in turkey and greek islands. 2020. https://www.cnbc.com/2020/10/30/strong-earthquake-strikes-aegean-sea-shaking-turkey-greece.html, accessed Feb 1, 2021.

S. Deo, S. Iravani, T. Jiang, K. Smilowitz, and S. Samuelson. Improving health outcomes through better capacity allocation in a community-based chronic care model. *Operations Research*, 61(6):1277–1294, 2013.

R. S. Driscoll. A grateful heart: The history of a world war I field hospital. *Military Review*, 84(4):98, 2004.

D. M. Gaba and S. K. Howard. Fatigue among clinicians and the safety of patients. *New England Journal of Medicine*, 347(16): 1249–1255, 2002.

D. Gupta. Surgical suites' operations management. *Production and Operations Management*, 16(6):689–700, 2007.

R. Hassin and S. Mendel. Scheduling arrivals to queues: A single-server model with no-shows. *Management Science*, 54(3): 565–572, 2008.

Y. Hu, C. W. Chan, and J. Dong. Optimal scheduling of proactive service with customer deterioration and improvement. 2021. Working paper, Columbia Business School.

E. U. Jacobson, N. T. Argon, and S. Ziya. Priority assignment in emergency response. *Operations Research*, 60(4):813–832, 2012.

D. E. Janhofer, C. Lakhiani, and D. H. Song. Addressing surgeon fatigue: current understanding and strategies for mitigation. *Plastic and Reconstructive Surgery*, 144(4):693e–699e, 2019.

K. S. Jung, M. Pinedo, C. Sriskandarajah, and V. Tiwari. Scheduling elective surgeries with emergency patients at shared operating rooms. *Production and Operations Management*, 28(6):1407–1430, 2019.

Y. Kreiss, O. Merin, K. Peleg, G. Levy, S. Vinker, R. Sagi, A. Abargel, C. Bartal, G. Lin, A. Bar, et al. Early disaster response in haiti: The israeli field hospital experience. *Ann Intern Med*, 153:45–48, 2010.

R. C. Larson, M. D. Metzger, and M. F. Cahn. Responding to emergencies: Lessons learned and the need for analysis. *Interfaces*, 36(6):486–501, 2006.

E. B. Lerner, R. B. Schwartz, P. L. Coule, E. S. Weinstein, D. C. Cone, R. C. Hunt, S. M. Sasser, J. M. Liu, N. G. Nudell, I. S. Wedmore, et al. Mass casualty triage: an evaluation of the data and development of a proposed national guideline. *Disaster medicine and public health preparedness*, 2(S1):S25–S34, 2008.

J. Y.-T. Leung, M. Pinedo, and G. Wan. Competitive two-agent scheduling and its applications. *Operations Research*, 58(2): 458–469, 2010.

O. Merin, N. Ash, G. Levy, M. J. Schwaber, and Y. Kreiss. The israeli field hospital in haiti—ethical dilemmas in early disaster response. *New England Journal of Medicine*, 362(11):e38, 2010.

A. F. Mills, N. T. Argon, and S. Ziya. Resource-based patient prioritization in mass-casualty incidents. *Manufacturing & Service Operations Management*, 15(3):361–377, 2013.

A. F. Mills, N. T. Argon, and S. Ziya. Dynamic distribution of patients to medical facilities in the aftermath of a disaster. *Operations Research*, 66(3):716–732, 2018a.

A. F. Mills, J. E. Helm, A. F. Jola-Sanchez, M. V. Tatikonda, and B. A. Courtney. Coordination of autonomous healthcare entities: Emergency response to multiple casualty incidents. *Production and Operations Management*, 27(1):184–205, 2018b.

B. Renaud, A. Santin, E. Coma, N. Camus, D. Van Pelt, J. Hayon, M. Gurgui, E. Roupie, J. Hervé, M. J. Fine, et al. Association between timing of intensive care unit admission and outcomes for emergency department patients with community-acquired pneumonia. *Critical care medicine*, 37(11):2867–2874, 2009.

L. W. Robinson and R. R. Chen. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing & Service Operations Management*, 12(2):330–346, 2010.

Z. Sun, N. T. Argon, and S. Ziya. Patient triage and prioritization under austere conditions. *Management Science*, 64(10):4471–4489, 2018.

S. Wang, N. Liu, and G. Wan. Managing appointment-based services in the presence of walk-in customers. *Management Science*, 66(2):667–686, 2020.

WHO. Mass casualty management systems: strategies and guidelines for building health sector capacity. 2007. Geneva: World Health Organization.

C. Zacharias and T. Yunes. Multimodularity in the stochastic appointment scheduling problem with discrete arrival epochs. *Management Science*, 66(2):744–763, 2020.

L. Zhang, X. Liu, Y. Li, Y. Liu, Z. Liu, J. Lin, J. Shen, X. Tang, Y. Zhang, and W. Liang. Emergency medical rescue efforts after a major earthquake: lessons from the 2008 wenchuan earthquake. *The Lancet*, 379(9818):853–861, 2012.

Appendix A: Proofs of the Results

Proof of Theorem 1:

For a given schedule q, we let $S_i(q)$ and $C_i(q)$ represent the service start and completion times of patient i in the schedule, respectively; let $C^i_{max}(q)$ ($C^d_{max}(q)$ resp.) be the C_{max} for immediate (delayed resp.) patients.

Theorem 1(a): When $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$, all immediate patients should be served consecutively before $\bar{D}^i$.

Consider an optimal schedule q_1, where the immediate patients are not served consecutively and therefore at least one delayed patient is served among immediate patients. Without loss of generality, suppose delayed patient j gets service between immediate patients i & k and immediate patient i gets service at time t. We use q_2 to denote the schedule obtained from q_1 by interchanging patients i & j. When the types of patient j are same in schedules q_1 and q_2, according to Proposition 1, we have $C_j(q_1) \geq C_i(q_2)$. In schedule q_2, since $C^i_{max}(q_2) < C^i_{max}(q_1)$ and $C^d_{max}(q_2) < C^d_{max}(q_1)$, q_2 is feasible. Besides, the number of deteriorated delayed patients in q_2 is no larger than that in q_1. Thus, q_2 is at least as good as the optimal schedule q_1.

When the types of patient j are different in schedules q_1 & q_2, then patient j must be a deteriorated delayed patient in schedule q_1 and be a delayed patient in schedule q_2. Therefore, we have $\beta_0 + (1+\beta)t > S$ and $t \leq S$, i.e., $\frac{S-\beta_0}{1+\beta} < t \leq S$.

We have

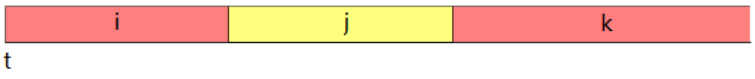
$$C_j(q_1) = [\beta_0 + (1+\beta)t](1+\beta) + \alpha_0 + \alpha S - \beta S, \tag{15}$$

$$C_i(q_2) = [\alpha_0 + (1+\alpha)t](1+\beta) + \beta_0. \tag{16}$$

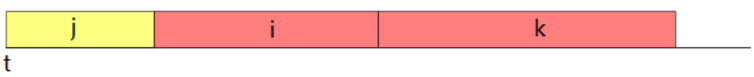
Subtracting (16) from (15) leads to

$$\begin{aligned} C_j(q_1) - C_i(q_2) &= (\beta-\alpha)(1+\beta)t + \beta_0\beta - \alpha_0\beta + \alpha S - \beta S \\ &> (\beta-\alpha)(S-\beta_0) + \beta_0\beta - \alpha_0\beta + \alpha S - \beta S \\ &> \beta_0\alpha - \alpha_0\beta \\ &> 0 \end{aligned}$$

Since $C^i_{max}(q_2) < C^i_{max}(q_1)$ and $C^d_{max}(q_2) < C^d_{max}(q_1)$, q_2 is feasible. Besides, the number of deteriorated delayed patients in q_2 is smaller than that of q_1. This contradicts the optimality of q_1 and hence no delayed patients would be served among immediate patients in an optimal schedule.



Schedule q_1



Schedule q_2

As many delayed patients as possible should be served before immediate patients.

When $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$, since immediate patients should be served consecutively, we could treat immediate patients as one "big" immediate patient f. Without of generality, suppose patient f gets service at time t. Let p_f be the total procedure time of patient f. It follows that

$$C_f = \frac{\beta_0}{\beta}[(1+\beta)^{N_i} - 1] + (1+\beta)^{N_i} t \tag{17}$$

$$p_f = \frac{\beta_0}{\beta}[(1+\beta)^{N_i} - 1] + [(1+\beta)^{N_i} - 1]t \tag{18}$$

Thus, the I2D ratio of patient f equals $\frac{\beta_0}{\beta}$.

Consider an optimal schedule q with n_1 delayed patients served before immediate patients. Suppose that there could be at most n_2 $(n_2 > n_1)$ delayed patients served before immediate patients without making $C^i_{max} > \bar{D}^i$. Without loss of generality, suppose $n_2 = n_1 + 1$. Denote the first delayed patient served after immediate patients in q as patient j. Let q^* be a scheduled obtained by swapping the positions of patients j & f. When the types of patient j remains the same in both q and q^*, according to Proposition 1, we have $C_f(q^*) \leq C_j(q)$. Since $C^i_{max}(q^*) \leq \bar{D}^i$ and $C^d_{max}(q^*) < C^d_{max}(q)$, q^* is feasible. Besides, the number of deteriorated delayed patients in q^* is no more than that in q, hence q^* is at least as good as q.

If the types of patient j are different in schedules q and q^*, then patient j is not deteriorated in q^* while deteriorated in q, i.e., $\frac{S-\beta^f_0}{1+\beta^f} < t \leq S$. Denote $\frac{\beta_0}{\beta}[(1+\beta)^{N_i} - 1]$ as β^f_0 and $(1+\beta)^{N^i} - 1$ as β^f. We have

$$C_f(q^*) = \alpha_0 + \beta^f_0 + \alpha_0\beta^f + (1+\alpha)(1+\beta^f)t, \tag{19}$$

$$C_j(q) = \alpha_0 + (\alpha-\beta)S + (1+\beta)\beta^f_0 + (1+\beta)(1+\beta^f)t. \tag{20}$$

Equation (20) - Equation (19)

$$\begin{aligned}
&= \beta\beta^f_0 - \alpha_0\beta^f + (\alpha-\beta)S + (1+\beta^f)(\beta-\alpha)t \\
&> \beta\beta^f_0 - \alpha_0\beta^f + (\alpha-\beta)S + (1+\beta^f)(\beta-\alpha)\frac{S-\beta^f_0}{1+\beta^f} \\
&> \beta^f_0\alpha - \alpha_0\beta^f \\
&> [(1+\beta)^n - 1](\frac{\beta_0}{\beta}\alpha - \alpha_0) \\
&> 0.
\end{aligned}$$

Since $C^i_{max}(q^*) \leq \bar{D}^i$ and $C^d_{max}(q^*) < C^d_{max}(q)$, q^* is feasible. Besides, the number of deteriorated delayed patients of q^* is less than that in q, contradicting the optimality of q and proving the desire results.

Proof(b): If $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$ and $\bar{D}^i \leq C + \alpha_0\beta - \beta_0\alpha$, all immediate patients should be served consecutively from time 0 and they have higher priority than delayed patients.

Consider an optimal schedule q_1 where immediate patients are not served consecutively and there exists at least one delayed patient served among immediate patients. Without loss of generality, suppose delayed

patient j is served between immediate patients i & k and delayed patient j gets service at time t. Since $\bar{D}^i \le C + \alpha_0\beta - \beta_0\alpha$, we have $C_k(q_1) \le C + \alpha_0\beta - \beta_0\alpha$. Then:

$$C_j(q_1) = S_k(q_1) = \frac{C_k(q_1) - \beta_0}{1+\beta} \le \alpha_0 + (1+\alpha)\frac{S-\beta_0}{1+\beta}.$$

It follows that

$$t = \frac{C_j(q_1) - \alpha_0}{1+\alpha} \le \frac{\alpha_0 + (1+\alpha)\frac{S-\beta_0}{1+\beta} - \alpha_0}{1+\alpha} \le \frac{S-\beta_0}{1+\beta}.$$

Let q_2 represent the schedule obtained from q_1 by interchanging patients j & k. Since $t \le \frac{S-\beta_0}{1+\beta}$, we have

$$S_j(q_2) = C_k(q_2) \le \beta_0 + (1+\beta)t \le S.$$

Thus, patient j is not deteriorated in q_2, i.e., the type of patient j in q_2 remains the same as in q_1. According to Proposition 1, $C_k(q_1) \ge C_j(q_2)$, i.e., patients after j in q_2 get service earlier than those after k in q_1. Since the number of deteriorated delayed patients in q_2 is no more than that in q_1 and the maximum completion times of immediate patients and deteriorated delayed patients in q_2 are no longer than those in q_1, q_2 is at least as good as the optimal schedule q_1, proving the desired result.

A similar argument shows that when $\bar{D}^i \le C + \alpha_0\beta - \beta_0\alpha$, immediate patients have higher priority than delayed patients.

Proof(c): If $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$ and $\bar{D}^i > C + \alpha_0\beta - \beta_0\alpha$, then immediate patients, if any, completing treatment before $C + \alpha_0\beta - \beta_0\alpha$ are served consecutively and no delayed patients are served before them.

Consider an optimal schedule q_1 where immediate patients with completion times $\le C + \alpha_0\beta - \beta_0\alpha$ are not served consecutively and there contains at least one delayed patient served among immediate patients. Without loss of generality, suppose delayed patient j is served between immediate patients i & k and delayed patient j gets service at time t. Since $C_j(q_1) = S_k(q_1) \le \alpha_0 + (1+\alpha)\frac{S-\beta_0}{1+\beta}$, we have

$$t = \frac{C_j(q_1) - \alpha_0}{1+\alpha} \le \frac{\alpha_0 + (1+\alpha)\frac{S-\beta_0}{1+\beta} - \alpha_0}{1+\alpha} \le \frac{S-\beta_0}{1+\beta}.$$

Let q_2 represent the schedule obtained from q_1 by interchanging patients j & k. Since $t \le \frac{S-\beta_0}{1+\beta}$, we have

$$S_j(q_2) = C_k(q_2) \le \beta_0 + (1+\beta)t \le S.$$

Thus patient j is not deteriorated in q_2, i.e., the type of patient j in q_2 remains the same as in q_1. Since $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$, according to Proposition 1, $C_k(q_1) \ge C_j(q_2)$. Since the maximum completion times of deteriorated delayed and immediate patients in q_2 are no more than those in q_1, and the number of deteriorated delayed patients in q_2 is no more than that in q_1, q_2 is at least as good as q_1, proving the desired result.

A similar argument shows that immediate patients, if any, completing treatment before $C + \alpha_0\beta - \beta_0\alpha$ have higher priority than delayed patients.

Immediate patients who complete treatment after $C+\alpha_0\beta-\beta_0\alpha$, if any, are served consecutively.

Consider an optimal schedule q_3 where immediate patients with completion times $> C+\alpha_0\beta-\beta_0\alpha$ are not served consecutively and there contains at least one delayed patient served between immediate patients. Without loss of generality, suppose delayed patient j is served between immediate patients i & k and immediate patient i get service at time t. Let q_4 be the schedule obtained from q_3 by interchanging patients i & j. When $t>S$, since $\frac{\beta_0+\beta S}{\beta} > \frac{\alpha_0+\alpha S}{\beta}$, according to Proposition 1, we have $C_j(q_3)\geq C_i(q_4)$. Since the maximum completion times of deteriorated delayed & immediate patients in q_4 are shorter than those in q_3, and the number of deteriorated delayed patients in q_4 is no more than that in q_3, q_4 is at least as good as q_3.

When $\alpha_0+(1+\alpha)\frac{S-\beta_0}{1+\beta}<t\leq S$, we have

$$C_j(q_3)=\alpha_0+(\alpha-\beta)S+(1+\beta)C_i(q_3)=\alpha_0+\beta_0+\beta\beta_0+(\alpha-\beta)S+(1+\beta)^2t, \tag{21}$$

$$C_i(q_4)=\beta_0+(1+\beta)C_j(q_4)=\alpha_0+\beta_0+\alpha_0\beta+(1+\alpha)(1+\beta)t. \tag{22}$$

Equation (21) − Equation (22)

$$\begin{aligned}
&=\beta(\beta_0-\alpha_0)+(\alpha-\beta)S+(1+\beta)(\beta-\alpha)t\\
&>\beta(\beta_0-\alpha_0)+(\alpha-\beta)S+(1+\beta)(\beta-\alpha)[\alpha_0+(1+\alpha)\frac{S-\beta_0}{1+\beta}]\\
&>(\beta-\alpha)\alpha S+\alpha(\beta_0-\alpha_0)+(\beta-\alpha)(\alpha_0\beta-\alpha\beta_0)\\
&>0.
\end{aligned}$$

Since the maximum completion times of deteriorated delayed and immediate patients in q_4 are no longer than those in q_3, q_4 is feasible. Besides, the number of deteriorated delayed patients in q_4 is smaller than that in q_3, contradicting the optimality of q_3. Thus, when $\frac{\alpha_0}{\alpha}>\frac{\beta_0}{\beta}$, immediate patients with FT $> C+\alpha_0\beta-\beta_0\alpha$ are served consecutively.

As many delayed patients as possible should be served before immediate patients with FT $> C+\alpha_0\beta-\beta_0\alpha$.

When $\frac{\alpha_0}{\alpha}>\frac{\beta_0}{\beta}$, since immediate patients with FT $> C+\alpha_0\beta-\beta_0\alpha$ are served consecutively, we could treat immediate patients with FT $> C+\alpha_0\beta-\beta_0\alpha$ as one "big" immediate patient f. Without loss of generality, suppose in optimal schedule q, there are m (≥ 1) immediate patients with FT $> C+\alpha_0\beta-\beta_0\alpha$ and patient f gets service at time t. We have

$$C_f=\frac{\beta_0}{\beta}[(1+\beta)^m-1]+(1+\beta)^m t \tag{23}$$

$$p_f=\frac{\beta_0}{\beta}[(1+\beta)^m-1]+[(1+\beta)^m-1]t \tag{24}$$

Thus, the I2D ratio of patient f equals $\frac{\beta_0}{\beta}$.

Consider there are n_1 delayed patients served before immediate patients with FT $> C+\alpha_0\beta-\beta_0\alpha$ in

optimal schedule q. Suppose that at most $n_2(>n_1)$ delayed patients could have priority than immediate patients with FT $> C+\alpha_0\beta-\beta_0\alpha$, without causing the schedule to be infeasible. Without loss of generality, suppose $n_2 = n_1 + 1$. Denote the first delayed patient served after immediate patients with FT $> C + \alpha_0\beta - \beta_0\alpha$ as patient j. Let q^* be the schedule obtained by swapping the positions of patient j & f. When the types of patient j remain the same in q and q^*, according to Proposition 1, we have $C_f(q^*) \le C_j(q)$. Since $C^i_{max}(q^*) \le \bar{D}^i$ and $C^d_{max}(q^*) < C^d_{max}(q)$, q^* is feasible. Besides, the number of deteriorated delayed patients in q^* is no more than that in q, and hence q^* is at least as good as q.

If the types of patient j are different between q and q^*, then patient j is not deteriorated in q^*. Denote $\frac{\beta_0}{\beta}[(1+\beta)^m-1]$ as β_0^f and $(1+\beta)^m-1$ as β^f, we have

$$C_f(q^*) = \alpha_0 + \beta_0^f + \alpha_0\beta^f + (1+\alpha)(1+\beta^f)t, \tag{25}$$

$$C_j(q) = \alpha_0 + (\alpha-\beta)S + (1+\beta)\beta_0^f + (1+\beta)(1+\beta^f)t. \tag{26}$$

Recall that immediate patients' finish times $> C + \alpha_0\beta - \beta_0\alpha$, we have $t > \alpha_0 + (1+\alpha)\frac{S-\beta_0}{1+\beta}$. Equation (26) - Equation (25)

$$\begin{aligned}
&= \beta\beta_0^f - \alpha_0\beta^f + (\alpha-\beta)S + (1+\beta^f)(\beta-\alpha)t \\
&> \beta\beta_0^f - \alpha_0\beta^f + (\alpha-\beta)S + (1+\beta^f)(\beta-\alpha)[\alpha_0 + (1+\alpha)\frac{S-\beta_0}{1+\beta}] \\
&> (\beta_0-\alpha_0)[(1+\beta)^m-1] + (\alpha-\beta)S + (1+\beta)^m(\beta-\alpha)\alpha_0 + (1+\alpha)(\beta-\alpha)(1+\beta)^{m-1}(S-\beta_0) \\
&> (\beta-\alpha)S[(1+\beta)^{m-1}(1+\alpha)-1] + (\beta_0-\alpha_0)[(1+\beta)^m-1] + (\beta-\alpha)(1+\beta)^{m-1}[\alpha_0(1+\beta)-\beta_0(1+\alpha)] \\
&> (\beta-\alpha)S[(1+\beta)^{m-1}(1+\alpha)-1] + (\beta-\alpha)(1+\beta)^{m-1}(\alpha_0\beta-\beta_0\alpha) + (\beta_0-\alpha_0)[(1+\beta)^{m-1}-1+\alpha(1+\beta)^{m-1}] \\
&> 0.
\end{aligned}$$

We see that victims following victim f in q^* get treatment earlier than those in q. Since $C^i_{max}(q^*) \le \bar{D}^i$ and $C^d_{max}(q^*) \le C^d_{max}(q)$, q^* is feasible. Besides, the number of deteriorated delayed patients in q^* is smaller than that in q, q^* is better than q, proving the desired result.

□

Proof of Theorem 2

(a) Note that the I2D ratio of stable immediate patients equals $p^{si}/0 \to +\infty$. Consider an optimal schedule q_1 where unstable immediate patients have no priority over stable immediate patients, there must be at least one stable immediate patient served before unstable immediate patients. Without loss of generality, suppose stable immediate patient i is served before unstable immediate patient j and patient i gets service at time t. Consider two cases.

Case (1): When there is no delayed patients served between patients i & j, let us obtain q_2 by swapping the positions of patients i & j. According to Proposition 1, $C_j(q_2) < C_i(q_1)$. Since $C^i_{max}(q_2)$ and $C^d_{max}(q_2)$ are no longer than those in q_1, q_2 is feasible. Besides, the number of deteriorated delayed patients in q_2 is no larger than that in q_1, q_2 is at least as good as q_1.

Case (2): When there are delayed patients served between i and j, let us denote the delayed patient served after the stable immediate patient i as patient k. Obtain q_2 by swapping the positions of patients i & k. If patient k is stable delayed patient, swapping positions of i and k has no effect on patient j and patients after j . Since $C^d_{max}(q_2) \leq C^d_{max}(q_1)$ and $C^i_{max}(q_2) = C^i_{max}(q_1)$, q_2 is feasible, Besides, since delayed patient j gets service earlier in q_2 than q_1, the number of deteriorated delayed patients in q_2 is smaller or equal to that in q_1, q_2 is at least as good as q_1. If patient k is an unstable delayed patient, we consider two sub-cases. When the types of patient k are the same in both q_1 and q_2, according to Proposition 1, $C_i(q_2) < C_k(q_1)$. Thus, patient k and those after k in q_1 get service earlier in q_2. Besides, $C_i(q_2) < C_k(q_1)$ and $C_k(q_2) < C_k(q_1)$, we have $C^i_{max}(q_2) < C^i_{max}(q_1)$ and $C^d_{max}(q_2) < C^d_{max}(q_1)$, so q_2 is feasible. Meanwhile, since the number of deteriorated delayed patients in q_2 is no more than that in q_1, q_2 is at least as good as q_1. When the types of patient k are different between q_1 and q_2, then patient k is deteriorated in q_1 while not in q_2. We have

$$\begin{aligned} C_k(q_1) &= (p^{si} + t)(1 + \beta) + \alpha_0 + \alpha S - \beta S, \\ C_i(q_2) &= \alpha_0 + (1 + \alpha)t + p^{si}. \end{aligned}$$

Recall that $S - p^{sd} < t \leq S$, we have

$$\begin{aligned} C_k(q_1) - C_i(q_2) =& (\beta - \alpha)t + \beta p^{sd} + (\alpha - \beta)S \\ >& (\beta - \alpha)(S - p^{sd}) + \beta p^{sd} + (\alpha - \beta)S \\ >& 0. \end{aligned}$$

Thus, $C^i_{max}(q_2) < C^i_{max}(q_1)$ and $C^d_{max}(q_2) < C^d_{max}(q_1)$, and q_2 is feasible. Besides, the number of deteriorated delayed patients in q_2 is smaller than that in q_1, q_2 is better than q_1. After swapping the positions of the stable immediate patient and the delayed patients after her, Case (2) is reduced to Case (1) above, showing that unstable immediate patients have higher priority than stable immediate patients.

(b) Consider an optimal schedule q_1 where unstable immediate patients with FT $\leq S + (1 + \beta)p^{sd}$ do not have priority over stable delayed patients and there exists at least one stable delayed patients i served before at least one unstable immediate patient j. Without loss of generality, suppose stable delayed patient i gets service at time t.

First, we consider when no unstable delayed patients are served between patients i & j. Since $C_j(q_1) \leq S + (1 + \beta)p^{sd}$, we have $C_i(q_1) = S_j(q_1) \leq \frac{S-\beta_0}{1+\beta} + p^{sd}$. Thus, $t = S_i(q_1) \leq \frac{S-\beta_0}{1+\beta}$. Let q_2 represent the schedule obtained from q_1 by interchanging patients i & j. Since $t \leq \frac{S-\beta_0}{1+\beta}$, we have

$$S_i(q_2) = C_j(q_2) \leq \beta_0 + (1 + \beta)t \leq S.$$

In q_2, stable delayed patient i is not deteriorated. Since $\beta_0/\beta < p^{sd}/0$, according to Proposition 1, $C_i(q_2) \leq C_j(q_1) \leq \bar{D}^i \leq \bar{D}^d$. Since $C^i_{max}(q_2) < C^i_{max}(q_1)$ and $C^d_{max}(q_2) < \bar{D}^d$, schedule q_2 is feasible. Besides, the number of deteriorated delayed patients in q_2 is no more than that in q_1, q_2 is at least as good as q_1.

Next, we consider when there are unstable delayed patients served between patients i & j, Without loss of generality, suppose unstable delayed patient k is served between patient i & j in q_1. (1) When $p^{sd} \geq \alpha_0 + \alpha S$, swapping the positions of patient i and patient k decreases the maximum completion times of immediate and delayed patients without worsening the solution, forming the same case as above. (2) when $p^{sd} < \alpha_0 + \alpha S$, (i) If $t + p^{sd} > S$, unstable delayed patient k deteriorates in q_1. Obtain q_2 by swapping the positions of i and k. we have

$$\begin{aligned} C_k(q_1) &= t + p^{sd} + \alpha_0 + \alpha S - \beta S + \beta(t + p^{sd}) \\ &= t + p^{sd} + \max\{\alpha_0 + \alpha S - \beta S + \beta(t + p^{sd}), \alpha_0 + \alpha(t + p^{sd})\} \\ C_i(q_2) &= t + \alpha_0 + \alpha t + p^{sd} < C_k(q_1). \end{aligned}$$

Swapping the positions of i and k decreases the maximum completion times of immediate and delayed patients without worsening the solution, forming the same case as above. (ii) If $t + p^{sd} \leq S$ and $\alpha_0 + (1 + \alpha)t \leq S$, swapping the positions of i and k decreases the maximum completion times of immediate and delayed patients without worsening the solution, forming the same case as above; (iii) If $t + p^{sd} \leq S$ and $\alpha_0 + (1 + \alpha)t > S$, we have $t > \frac{S - \alpha_0}{1 + \alpha}$. However, in q_1, $t = S_j(q_1) - p_k(q_1) - p^{sd} < \frac{S - \beta_0}{1 + \beta} < \frac{S - \alpha_0}{1 + \alpha}$. This case does not exist.

(c) Consider an optimal schedule q_1 where unstable delayed patients with FT $\leq S + p^{sd}$ do not have priority over stable delayed patients and there exists at least one stable delayed patient i served before at least one unstable delayed patient j. Without loss of generality, suppose stable delayed patient i gets service at time t.

We first consider the case when no immediate patients are served between patients i & j. If unstable delayed patient j is not deteriorated, then $C_j(q_1) \leq S + p^{sd} < S + (1 + \alpha)p^{sd}$, and we have $C_i(q_1) = S_j(q_1) < \frac{S - \alpha_0}{1 + \alpha} + p^{sd}$. Thus, $t = S_i(q_1) < \frac{S - \alpha_0}{1 + \alpha}$. Let q_2 represent the schedule obtained from q_1 by interchanging patient i & j. Since $t < \frac{S - \alpha_0}{1 + \alpha}$, we have

$$S_i(q_2) = C_j(q_2) = \alpha_0 + (1 + \alpha)t < S.$$

In q_2, stable delayed patient i is not deteriorated. Since $\alpha_0/\alpha < p^{sd}/0$, according to Proposition 1, $C_i(q_2) \leq C_j(q_1)$. Since the maximum completion times of delayed and immediate patients in q_2 are no more than those in q_1, q_2 is feasible. Besides, the number of deteriorated delayed patients in q_2 is no more than that in q_1, q_2 is at least as good as q_1.

If, however, unstable delayed patient j is deteriorated, i.e., $S_j(q_1) > S$. Let us obtain q_2 by swapping the positions of i and j. We have

$$\begin{aligned} C_j(q_1) &= t + p^{sd} + \alpha_0 + \alpha S - \beta S + \beta(t + p^{sd}) \\ &= t + p^{sd} + \max\{\alpha_0 + \alpha S - \beta S + \beta(t + p^{sd}), \alpha_0 + \alpha(t + p^{sd})\}, \\ C_i(q_2) &= t + \alpha_0 + \alpha t + p^{sd} < C_j(q_1). \end{aligned}$$

The maximum completion times of immediate and delayed patients in q_2 are no more than those in q_1, so q_2 is feasible. Besides, the number of deteriorated delayed patients in q_2 is no more than that in q_1. Hence, q_2 is at least as optimal as q_1.

Next, we consider the case when there are immediate patients served between patients i & j. Since $S + (1+\beta)p^{sd} > S + p^{sd}$, no unstable immediate patients are served between patients i & j. Thus, immediate patients between i & j are stable immediate patients. Without loss of generality, suppose stable immediate patient k is served between patient i and j. Consider three sub-cases below. (1) When $t \geq S$, we have $C_j(q_1) = t + p^{sd} + p^{si} + p_j^d > S + p^{sd}$. This case does not exist. (2) When $t < S$ and $p^{sd} \geq \alpha_0 + \alpha S$, let us obtain q_2 by swapping the positions of patients i and j. Since $p^{sd} \geq \alpha_0 + \alpha S > \alpha_0 + \alpha t$, we have $C_k(q_2) < C_k(q_1)$, which means $S_i(q_2) < S_j(q_1)$. If patient j gets treatment without deterioration in q_1, so does patient i in q_2. Since the maximum completion times of immediate and delayed patients in q_2 are smaller than those in q_1 and the number of deteriorated delayed patients in q_2 is no more than that in q_1, q_2 is at least as good as q_1. (3) When $t < S$ and $p^{sd} < \alpha_0 + \alpha S$, (i) if patient j is not deteriorated, rearranging the sequence as $[k, j, i]$ reduces the maximum completion times of immediate and deteriorated delayed patients without worsening the solution; (ii) if patient j is deteriorated, we have

$$\begin{aligned} C_j &= (1+\beta)S_j + \alpha_0 + \alpha S - \beta S \\ &> (1+\beta)S + \alpha_0 + \alpha S - \beta S \\ &> S + \alpha_0 + \alpha S. \end{aligned} \tag{27}$$

Since $C_j \leq S + p^{sd}$, we have $p^{sd} > \alpha_0 + \alpha S$, leading to contradiction and proving the desired result.

(d) Consider an optimal schedule q_1 where unstable delayed patients do not have priority over stable delayed patients if $p^{sd} \geq \alpha_0 + \alpha S$ and there exist some stable delayed patients served before some unstable delayed patients. Without loss of generality, suppose there is one stable delayed patient i served before one unstable delayed patient j at time t. Let q_2 represent the schedule obtained from q_1 by interchanging the positions of patient i & j.

If no immediate patients (stable and unstable) are served between patient i & j, consider three cases. (1) When $t > S$, since $\frac{\alpha_0 + \alpha S}{\beta} < \frac{p^{sd}}{0}$, according to Proposition 1, $C_i(q_2) \leq C_j(q_1)$. Meanwhile, the number of deteriorated delayed number is no larger than that in q_1, so q_2 is at least as good as q_1. (2) When $t \leq S - p^{sd}$, both patients i & j are not deteriorated in q_1. Since $S_i(q_2) = C_j(q_2) = t + \alpha_0 + \alpha t < t + \alpha_0 + \alpha S < t + p^{sd} < S$, both patient i & j are not deteriorated in q_2. According to Proposition 1, $C_i(q_2) \leq C_j(q_1)$. Meanwhile, the number of deteriorated delayed patients in q_2 is no larger than that in q_1, so q_2 is at least as good as q_1. (3) When $S - p^{sd} < t \leq S$, $S_j(q_1) = t + p^{sd} > S$, and so patient j is deteriorated in q_1. We have

$$C_j(q_1) = t + p^{sd} + \alpha_0 + \alpha S - \beta S + \beta(t + p^{sd}) = t + p^{sd} + \max\{\alpha_0 + \alpha S - \beta S + \beta(t + p^{sd}),\ \alpha_0 + \alpha(t + p^{sd})\}.$$

In q_2, since $t \le S$, there is at most one deteriorated delayed patient. We have

$$C_i(q_2) = t + \alpha_0 + \alpha t + p^{sd} < t + p^{sd} + \max\{\alpha_0 + \alpha S - \beta S + \beta(t + p^{sd}),\ \alpha_0 + \alpha(t + p^{sd})\} < C_j(q_1),$$

making the maximum completion times of delayed and immediate patients shorter than those of q_1. Meanwhile, stable delayed patient i may not deteriorate in q_2, making the number of deteriorated delayed patients in q_2 fewer than or equal to that in q_1. As a result, q_2 is at least as good as q_1.

If there are immediate patients (stable and unstable) served between patient i & j, consider three cases. (1)When $t > S$, since $p^{sd}/0 \to +\infty$, according to Proposition 1, interchanging positions of stable patient i and patients after i decreases the completion time without worsening the solution, forming the the case where no immediate patients are served between patient i and j. (2) Consider when $t \le S$ and unstable delayed patient j is not deteriorated in q_1. Without loss of generality, suppose the total service time of immediate patients served between patients i & j is Δ in q_1 and Δ' in q_2. Since $p^{sd} \ge \alpha_0 + \alpha S$, we have $C_j(q_2) \le C_i(q_1)$, which indicates $\Delta' \le \Delta$. Thus, $S_i(q_2) \le S_j(q_1) \le S$, and stable delayed patient i is not deteriorated in q_2. Besides, since $C_j(q_1) = (t + p^{sd} + \Delta)(1 + \alpha) + \alpha_0 > \alpha_0 + (1 + \alpha)t + \Delta' + p^{sd} = C_i(q_2)$, we have $C^i_{max}(q_2) \le C^i_{max}(q_1)$ and $C^d_{max}(q_2) \le C^d_{max}(q_1)$, and thus q_2 is at least as good as q_1. (3) Consider when $t \le S$ and unstable delayed patient j is deteriorated in q_1. Without loss of generality, suppose the total service time of immediate patients served between patients i & j is Δ in q_1 and Δ' in q_2. Since $p^{sd} \ge \alpha_0 + \alpha S$, we have $C_j(q_2) \le C_i(q_1)$, which indicates $\Delta' \le \Delta$. Thus, the condition of stable delayed patient i is uncertain in q_2, i.e., the number of deteriorated delayed patients in q_2 is no more than that in q_1. Next we show that q_2 is feasible. Note that

$$\begin{aligned} C_j(q_1) &= t + p^{sd} + \Delta + \alpha_0 + \alpha S - \beta S + \beta(t + p^{sd} + \Delta) \\ &= t + p^{sd} + \Delta + \max\{\alpha_0 + \alpha S - \beta S + \beta(t + p^{sd} + \Delta),\ \alpha_0 + \alpha(t + p^{sd} + \Delta)\}, \end{aligned}$$

and $C_i(q_2) = t + \alpha_0 + \alpha t + \Delta' + p^{sd} < C_j(q_1)$. Thus, patients after j in q_1, i.e., patients after i in q_2, get treatments in q_2 earlier than q_1, making q_2 a feasible schedule. Since the number of deteriorated delayed patients in q_2 is no more than that in q_1 and q_2 is feasible, q_2 is at least as good as q_1.

From the analysis above, we can see that unstable delayed patients have priority over stable delayed patients if $p^{sd} \ge \alpha_0 + \alpha S$.

(e) Among patients with ST $> S$, if unstable ones (both immediate and delayed) have no priority over stable delayed patients, there must be at least one stable delayed patients served before unstable patients. Without loss of generality, suppose stable delayed patient i served before unstable immediate patient j in the optimal schedule q_1. According to Theorem 2(a), no stable immediate patients are served between patients i & j. Let us obtain q_2 by swapping the positions of patients i & j. Since $\frac{p^{sd}}{0} > \frac{\beta_0 + \beta S}{\beta}$, according to Proposition 1, $C_i(q_2) < C_j(q_1)$. Since patients before patients i and j remain the same in q_1 and q_2, patients

after i & j receive treatments earlier in q_2 than q_1 and $C_i(q_2) < C_j(q_1) < \bar{D}^i < \bar{D}^d$, q_2 is feasible. Besides, the number of deteriorated delayed patients in q_2 is no more than that in q_1, q_2 is at least as good as q_1. Similar argument shows that unstable delayed patients have priority over stable delayed patients.

(f) When $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$ and let us focus on patients with FT $\leq \bar{D}^i$. If $p^{sd} \geq \alpha_0 + \alpha S$, consider an optimal schedule q_1 where unstable delayed patients with FT $\leq \bar{D}^i$ do not have priority over unstable immediate patients and there exists at least one unstable immediate patient i served before at least one unstable delayed patient j. Let q_2 represent the schedule obtained from q_1 by interchanging patient i & j. We first show that there are no stable immediate patients served between patients i & j. Without loss of generality, suppose stable immediate patient k gets service between patients i & j in q_1. Swapping the positions of patients k & j, if the type of patient j remains the same as before, according to Proposition 1, the completion time of k after interchanging is no longer than $C_j(q_1) \leq \bar{D}^i$. Besides, the completion time of other patients is no longer than that of q_1. This schedule is feasible. Since the number of deteriorated delayed patients after interchanging is at most the same as that in q_1. This schedule is at least as good as q_1.

If, however, the type of patient j becomes different, patient j must be deteriorated in q_1 while not deteriorated after swapping the positions of k and j, which means, $S - p^{si} < S_k(q_1) \leq S$. We use C'_k to denote the completion time of patient k after interchanging. Then,

$$\begin{aligned} C_j(q_1) &= (S_k(q_1) + p^{si})(1+\beta) + \alpha_0 + \alpha S - \beta S, \\ C'_k &= (1+\alpha) S_k(q_1) + \alpha_0 + p^{si}. \end{aligned}$$

It follows that

$$\begin{aligned} C_j(q_1) - C'_k &= (\beta - \alpha) S_k(q_1) + \beta p^{si} + \alpha S - \beta S \\ &> (\beta - \alpha)(S - p^{si}) + \beta p^{si} + \alpha S - \beta S \\ &> \alpha p^{si} \\ &> 0. \end{aligned}$$

These prove that no stable immediate patients are served between i & j.

Next consider two cases. (1) When the types of unstable delayed patient i remains the same in schedules q_1 and q_2, according to Proposition 1, we have $C_i(q_2) \leq C_j(q_1) \leq \bar{D}^i$. Thus, we have $C^i_{max} \leq \bar{D}^i$ and $C^d_{max}(q_2) < C^d_{max}(q_1)$, q_2 is feasible. Besides, since the number of deteriorated delayed patients in q_2 is no more than that in q_1 , q_2 is at least as good as q_1. (2) When the types of unstable delayed patient j are different between q_1 and q_2, then patient j must be deteriorated in q_1 while not in q_2, which means, $\frac{S-\beta_0}{1+\beta} < t \leq S$. Then,

$$C_j(q_1) = [\beta_0 + (1+\beta)t](1+\beta) + \alpha_0 + \alpha S - \beta S, \tag{28}$$

$$C_i(q_2) = [\alpha_0 + (1+\alpha)t](1+\beta) + \beta_0. \tag{29}$$

Recall that $t > \frac{S-\beta_0}{1+\beta}$. It follows that Equation (28) − Equation (29)

$$\begin{aligned}
&= (\beta-\alpha)(1+\beta)t + \beta_0\beta - \alpha_0\beta + \alpha S - \beta S \\
&> (\beta-\alpha)(S-\beta_0) + \beta_0\beta - \alpha_0\beta + \alpha S - \beta S \\
&> 0.
\end{aligned}$$

In q_2, since the number of deteriorated delayed patients is smaller than that in q_1 and q_2 is feasible, contradicting the optimality of q_1 and proving the desired result.

If $p^{sd} < \alpha_0 + \alpha S$, among patients with FT $\leq S + p^{sd}$, consider an optimal schedule q_1 where unstable delayed patients do not have priority over unstable immediate patients and there must be at least one unstable immediate patient served before unstable delayed patient. Without loss of generality, suppose unstable immediate patient i gets service before unstable delayed patient j at time t. Let q_2 represent the schedule obtained from q_1 by interchanging patients i and j. A similar argument as above shows that no stable immediate patients are served between patients i & j. When the types of unstable delayed patient j in q_1 & q_2 remain the same, according to Proposition 1, $C_j(q_1) \geq C_i(q_2)$. Since $C^i_{max}(q_2) \leq \bar{D}^i$ and $C^d_{max}(q_2) \leq C^d_{max}(q_1)$, q_2 is feasible. Besides, the number of deteriorated delayed patients in q_2 is no more than that in q_1. q_2 is at least as good as q_1. When the type of unstable delayed patient j is different in q_1 & q_2, unstable delayed patient j is deteriorated in q_1 while not deteriorated in q_2, i.e., $\frac{S-\beta_0}{1+\beta} < t \leq S$. We have

$$\begin{aligned}
C_j(q_1) &= \alpha_0 + (\alpha-\beta)S + (1+\beta)C_i(q_1) = \alpha_0 + \beta_0 + \beta\beta_0 + (\alpha-\beta)S + (1+\beta)^2 t, \\
C_i(q_2) &= \beta_0 + (1+\beta)C_j(q_2) = \alpha_0 + \beta_0 + \alpha_0\beta + (1+\alpha)(1+\beta)t.
\end{aligned}$$

Then,

$$\begin{aligned}
C_j(q_1) - C_i(q_2) &= \beta(\beta_0-\alpha_0) + (\alpha-\beta)S + (1+\beta)(\beta-\alpha)t \\
&> \beta(\beta_0-\alpha_0) + (\alpha-\beta)S + (\beta-\alpha)(S-\beta_0) \\
&> \alpha\beta_0 - \alpha_0\beta \\
&> 0.
\end{aligned}$$

Since the number of deteriorated delayed patients in q_2 is smaller than that in q_1 and q_2 is feasible, contradicting the optimality of q_1 and proving the desired result.

If $p^{sd} < \alpha_0 + \alpha S$, among patients with ST $> S$ and FT $\leq \bar{D}^i$, consider an optimal schedule q_3 where unstable deteriorated delayed patients have no priority over unstable immediate ones. There must be at least one unstable immediate patient served before unstable delayed patients. Without loss of generality, suppose unstable immediate patient i gets service before unstable delayed patient j at time t. Let q_4 represent the schedule obtained from q_3 by swapping the positions of i and j. A similar argument as above shows

that no stable immediate patients are served between patient i and j. Since $\frac{\alpha_0+\alpha S}{\beta} < \frac{\beta_0+\beta S}{\beta}$, according to Proposition 1, we have $C_j(q_3) \geq C_i(q_4)$. Since $C^i_{max}(q_4) \leq \bar{D}^i$ and $C^d_{max}(q_4) \leq C^d_{max}(q_3)$, q_4 is feasible. Besides, since the number of deteriorated delayed patients in q_4 is no more than that in q_3, q_4 is at least as good as q_3, proving the desired result.

(g) If $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$ and $\bar{D}^i \leq \min\{C+\alpha_0\beta-\alpha\beta_0, S+(1+\beta)p^{sd}\}$, consider an optimal schedule q_1, where the unstable immediate patients are not served consecutively from time 0. Since $\bar{D}^i \leq S+(1+\beta)p^{sd}$, unstable immediate patients have priority over stable delayed patients. Also from part (a), we know unstable immediate patients have priority than stable immediate patients. Thus, only unstable delayed patients can be served before unstable immediate ones in q_1. Without loss of generality, suppose unstable delayed patient i gets service before unstable immediate patients j at time t. Since $C_j(q_1) \leq C+\alpha_0\beta-\alpha\beta_0$, we have

$$C_i(q_1) = S_j(q_1) = (C_j(q_1)-\beta_0)/(1+\beta) \leq \alpha_0 + (1+\alpha)\frac{S-\beta_0}{1+\beta} < S.$$

Thus,

$$t = S_i(q_1) = \frac{C_i(q_1)-\alpha_0}{1+\alpha} \leq \frac{\alpha_0+(1+\alpha)\frac{S-\beta_0}{1+\beta}-\alpha_0}{1+\alpha} \leq \frac{S-\beta_0}{1+\beta}.$$

Let q_2 represents the schedule obtained from q_1 by interchanging patients i & j. Since $t \leq \frac{S-\beta_0}{1+\beta}$, we have

$$S_i(q_2) = C_j(q_2) \leq \beta_0 + (1+\beta)t \leq S.$$

Patient i is not deteriorated in q_2, which means, the type of patient i in q_2 remains the same as in q_1. Since $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$, according to Proposition 1, $C_j(q_1) \geq C_i(q_2)$. Since the number of deteriorated delayed patients in q_2 is no more than that in q_1 , and the maximum completion times of deteriorated delayed & immediate patients in q_2 are no more than those in q_1, q_2 is at least as good as q_1, proving the desire result.

When $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$ and $\bar{D}^i > \min\{C+\alpha_0\beta-\alpha\beta_0, S+(1+\beta)p^{sd}\}$, according to the proof shown above, unstable immediate patients with FT $< \min\{C+\alpha_0\beta-\alpha\beta_0, S+(1+\beta)p^{sd}\}$ have priority over unstable delayed patients. For unstable immediate patients with FT $\geq C+\alpha_0\beta-\alpha\beta_0$, consider an optimal schedule q_3, where unstable immediate patients with FT $\geq C+\alpha_0\beta-\alpha\beta_0$ are not served consecutively and therefore there exists at least one patient (unstable delayed or stable delayed) served among unstable immediate patients. We consider both types below. Without loss of generality, suppose unstable delayed patient j gets service between unstable immediate patients i & k and unstable immediate patient i gets service at time t. $S_j(q_3) = C_i(q_3) \geq C+\alpha_0\beta-\alpha\beta_0 > S$, thus unstable delayed patient j is deteriorated in q_3. Let q_4 represents the schedule obtained from q_3 by interchanging the positions of i & j. When $t > S$, $S_j(q_4) = t > S$, unstable delayed patient j is deteriorated in q_4. Since $\frac{\alpha_0+\alpha S}{\beta} < \frac{\beta_0+\beta S}{\beta}$, according to Proposition 1, $C_i(q_4) \leq C_j(q_3)$. Since $C^i_{max}(q_4) \leq C^i_{max}(q_3)$ and $C^d_{max}(q_4) \leq C^d_{max}(q_3)$, q_4 is feasible. Meanwhile, the

number of deteriorated delayed patients in q_4 is no more than that in q_3, q_4 is at least as good as q_3. When $\alpha_0 + (1+\alpha)\frac{S-\beta_0}{1+\beta} \le t \le S$, $S_j(q_4) = t \le S$, unstable delayed patient j is not deteriorated in q_4. We have

$$C_j(q_3) = \alpha_0 + (\alpha-\beta)S + (1+\beta)C_i(q_3) = \alpha_0 + \beta_0 + \beta\beta_0 + (\alpha-\beta)S + (1+\beta)^2 t, \tag{30}$$

$$C_i(q_4) = \beta_0 + (1+\beta)C_j(q_4) = \alpha_0 + \beta_0 + \alpha_0\beta + (1+\alpha)(1+\beta)t. \tag{31}$$

Equation (30) - Equation (31) =

$$\begin{aligned}
&= \beta(\beta_0 - \alpha_0) + (\alpha-\beta)S + (1+\beta)(\beta-\alpha)t \\
&\ge \beta(\beta_0 - \alpha_0) + (\alpha-\beta)S + (1+\beta)(\beta-\alpha)[\alpha_0 + (1+\alpha)\frac{S-\beta_0}{1+\beta}] \\
&\ge (\beta-\alpha)\alpha S + \alpha(\beta_0 - \alpha_0) + (\beta-\alpha)(\alpha_0\beta - \alpha\beta_0) \\
&\ge 0.
\end{aligned}$$

Since $C^i_{max}(q_4) \le C^i_{max}(q_3)$ and $C^d_{max}(q_4) \le C^d_{max}(q_3)$, q_4 is feasible. Besides, the number of deteriorated delayed patients in q_4 is smaller than that in q_3, contradicting the optimality of q_3 and proving the desired result.

Finally, we consider the case when there exist stable delayed patients served between unstable immediate patients with FT $\ge C + \alpha_0\beta - \beta_0\alpha$. Without loss of generality, in an optimal schedule q_5, suppose stable delayed patient j is served between unstable immediate patients i & k. Let q_6 represent the schedule obtained from q_5 by swapping the positions of j & k. Since $S_j(q_5) = C_i(q_5) \ge C + \alpha_0\beta - \beta_0\alpha > S$, patient j is deteriorated in both q_5 and q_6. According to Proposition 1, we have $C_k(q_5) \ge C_j(q_6)$. Since $C^i_{max}(q_6) < C^i_{max}(q_5)$ and $C^d_{max}(q_6) \le \bar{D}^i \le \bar{D}^d$, q_6 is feasible. Thus, q_6 is at least as good as q_5, completing the proof.

□

Proof of Proposition 2

1. **When $\frac{\alpha_0}{\alpha} \le \frac{\beta_0}{\beta}$, among patients with wait times $< \tau$, immediate patients, if any, should be served consecutively before D^i and as many delayed patients as possible should be served before immediate patients. The same priority result also holds for patients with wait times $\ge \tau$.**

We first show that the immediate patients in question should be served consecutively. Consider an optimal schedule q_1, among patients with wait times $< \tau$, immediate patients if any are not served consecutively and therefore there exists at least one delayed patient served among immediate patients. Without loss of generality, suppose delayed patient j gets service between immediate patients i & k and immediate patient i gets service at time t. We use q_2 to denote the schedule obtained from q_1 by interchanging patients i & j. When the types of patient j are the same in schedules q_1 and q_2, according to Proposition 1, we have $C_j(q_1) \ge C_i(q_2)$. In schedule q_2, since $C^i_{max}(q_2) < C^i_{max}(q_1)$ and $C^d_{max}(q_2) < C^d_{max}(q_1)$, q_2 is feasible. Besides, the number of deteriorated delayed patients in q_2 is no larger than that in q_1, q_2 is at least as good

as optimal schedule q_1. When the types of patient j are different in schedules q_1 & q_2, then patient j must be deteriorated in schedule q_1 while not in schedule q_2. Thus, we have $\beta_0 + (1+\beta)t > S$ and $t \leq S$, i.e., $\frac{S-\beta_0}{1+\beta} < t \leq S$. In addition,

$$C_j(q_1) = [\beta_0 + (1+\beta)t](1+\beta) + \alpha_0 + \alpha S - \beta S, \tag{32}$$

$$C_i(q_2) = [\alpha_0 + (1+\alpha)t](1+\beta) + \beta_0. \tag{33}$$

Equation (32) $-$ Equation (33)

$$\begin{aligned} &= (\beta-\alpha)(1+\beta)t + \beta_0\beta - \alpha_0\beta + \alpha S - \beta S \\ &> (\beta-\alpha)(S-\beta_0) + \beta_0\beta - \alpha_0\beta + \alpha S - \beta S \\ &> \beta_0\alpha - \alpha_0\beta \\ &> 0. \end{aligned} \tag{34}$$

Since $C^i_{max}(q_2) < C^i_{max}(q_1)$ and $C^d_{max}(q_2) < C^d_{max}(q_1)$, q_2 is feasible. Besides, the number of deteriorated delayed patients of q_2 is smaller that of q_1, q_2 is better than optimal schedule q_1.

When $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$, among patients with wait times $< \tau$, since immediate patients should be served consecutively, we could treat immediate patients as one "big" immediate patient f. Without loss of generality, suppose patient f gets service at time t. We have

$$C_f = \frac{\beta_0}{\beta}[(1+\beta)^{N_i} - 1] + (1+\beta)^{N_i}t, \tag{35}$$

$$p_f = \frac{\beta_0}{\beta}[(1+\beta)^{N_i} - 1] + [(1+\beta)^{N_i} - 1]t. \tag{36}$$

Thus, the I2D ratio of patient f equals $\frac{\beta_0}{\beta}$. Among delayed patients with wait times $< \tau$, consider there are n_1 delayed patients served before immediate patients in an optimal schedule q. Suppose that there could be at most n_2 $(> n_1)$ delayed patients served before immediate patients, without making $C^i_{max} > \min\{\bar{D}^i, \tau\}$. Without loss of generality, suppose $n_2 = n_1 + 1$. Denote the first delayed patient served after immediate patients in q as patient j. Let q^* be obtained by swapping the positions of patients j & f. When the type of patient j remains the same between q and q^*, according to Proposition 1, we have $C_f(q^*) \leq C_j(q)$. Since $C^i_{max}(q^*) \leq \min\{\bar{D}^i, \tau\}$ and $C^d_{max}(q^*) < C^d_{max}(q)$, q^* is feasible. Besides, the number of deteriorated delayed patients in q^* is no more than that in q, q^* is at least as good as q.

If the type of patient j becomes different between q and q^*, then patient j is not deteriorated in q^* while deteriorated in q, i.e., $\frac{S-\beta^f_0}{1+\beta^f} < t \leq S$. Denote $\frac{\beta_0}{\beta}[(1+\beta)^{N_i} - 1]$ as β^f_0 and $(1+\beta)^{N^i} - 1$ as β^f, we have

$$C_f(q^*) = \alpha_0 + \beta^f_0 + \alpha_0\beta^f + (1+\alpha)(1+\beta^f)t, \tag{37}$$

$$C_j(q) = \alpha_0 + (\alpha-\beta)S + (1+\beta)\beta^f_0 + (1+\beta)(1+\beta^f)t. \tag{38}$$

Equation (38) - Equation (37)

$$\begin{aligned}
&= \beta\beta_0^f - \alpha_0\beta^f + (\alpha-\beta)S + (1+\beta^f)(\beta-\alpha)t \\
&> \beta\beta_0^f - \alpha_0\beta^f + (\alpha-\beta)S + (1+\beta^f)(\beta-\alpha)\frac{S-\beta_0^f}{1+\beta^f} \\
&> \beta_0^f\alpha - \alpha_0\beta^f \\
&> [(1+\beta)^n - 1](\frac{\beta_0}{\beta}\alpha - \alpha_0) \\
&> 0.
\end{aligned}$$

Since $C^i_{max}(q^*) \le \min\{\bar{D}^i, \tau\}$ and $C^d_{max}(q^*) < C^d_{max}(q)$, q^* is feasible. Besides, the number of deteriorated delayed patients in q^* is smaller than that in q, contradicting to the optimality of q, proving the desired result.

Similar argument shows that the same priority holds for patients with wait times $\ge \tau$.

2. **When $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$, among patients with wait times $< \tau$, then immediate patients, if any, completing treatment before $C + \alpha_0\beta - \beta_0\alpha$ are served consecutively from time 0 and no delayed patients are served before them. Immediate patients who complete treatment after $C + \alpha_0\beta - \beta_0\alpha$, if any, are served consecutively and as many delayed patients as possible should be served before immediate patients. The same priority results also hold for patients with wait times $\ge \tau$.**

We first show that the immediate patients in question should be served consecutively. Among patients with wait times $< \tau$, consider an optimal schedule q_1 where immediate patients with service finish times $\le C + \alpha_0\beta - \beta_0\alpha$, if any, are not served consecutively from time 0 and there must be at least one delayed patient served between immediate patients whose wait times $\le C + \alpha_0\beta - \beta_0\alpha$. Without loss of generality, suppose delayed patient j gets service between immediate patients i & k at time t ($t < \tau$). Since $C_k(q_1) \le C + \alpha_0\beta - \beta_0\alpha$, we have $C_j(q_1) = S_k(q_1) \le \alpha_0 + (1+\alpha)\frac{S-\beta_0}{1+\beta} < S$. Then,

$$t = \frac{C_j(q_1) - \alpha_0}{1+\alpha} \le \frac{\alpha_0 + (1+\alpha)\frac{S-\beta_0}{1+\beta} - \alpha_0}{1+\alpha} \le \frac{S-\beta_0}{1+\beta}.$$

Let q_2 represent the schedule obtained from q_1 by interchanging positions of j and k. Since $t \le \frac{S-\beta_0}{1+\beta}$, we have

$$S_j(q_2) = C_k(q_2) \le \beta_0 + (1+\beta)t \le S.$$

Thus, patient j is not deteriorated in q_2, i.e., the type of patient j in q_2 remains the same as that in q_1. Since $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$, according to Proposition 1, $C_k(q_1) \ge C_j(q_2)$. q_2 is feasible. Since $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$, according to Proposition 1, $C_k(q_1) \ge C_j(q_2)$. Since the maximum completion times of deteriorated delayed and immediate patients in q_2 are no more than those in q_1, and the number of deteriorated delayed patients in q_2 is no more than that in q_1, q_2 is at least as good as q_1.

Similar argument shows that delayed patients will only be served after all immediate patients with service finish times $\leq C+\alpha_0\beta-\beta_0\alpha$.

Next, we show that among patients with wait times $<\tau$, immediate patients with service finish times $>C+\alpha_0\beta-\beta_0\alpha$, if any, should be served consecutively. Consider an optimal schedule q_3 where these immediate patients are not served consecutively and therefore there exists at least one delayed patient served between immediate patients. Without loss of generality, suppose delayed patient j gets service between immediate patients i & k at time t. Let q_4 be the schedule obtained from q_3 by interchanging patients i & j. When $t>S$, if any, since $\frac{\beta_0+\beta S}{1+\beta}>\frac{\alpha_0+\alpha S}{1+\beta}$, according to Proposition 1, we have $C_j(q_3)\geq C_i(q_4)$. Since the maximum completion times of deteriorated delayed and immediate patients in q_4 are smaller than those in q_3, and the number of deteriorated delayed patients in q_4 is no more than that in q_3, q_4 is at least as good as q_3. If $\alpha_0+(1+\alpha)\frac{S-\beta_0}{1+\beta}<t\leq S$, we have

$$C_j(q_3)=\alpha_0+(\alpha-\beta)S+(1+\beta)C_i(q_3)=\alpha_0+\beta_0+\beta\beta_0+(\alpha-\beta)S+(1+\beta)^2t, \tag{39}$$

$$C_i(q_4)=\beta_0+(1+\beta)C_j(q_4)=\alpha_0+\beta_0+\alpha_0\beta+(1+\alpha)(1+\beta)t. \tag{40}$$

Equation (39) - Equation (40)

$$\begin{aligned}
&=\beta(\beta_0-\alpha_0)+(\alpha-\beta)S+(1+\beta)(\beta-\alpha)t\\
&>\beta(\beta_0-\alpha_0)+(\alpha-\beta)S+(1+\beta)(\beta-\alpha)[\alpha_0+(1+\alpha)\frac{S-\beta_0}{1+\beta}]\\
&>(\beta-\alpha)\alpha S+\alpha(\beta_0-\alpha_0)+(\beta-\alpha)(\alpha_0\beta-\alpha\beta_0)\\
&>0.
\end{aligned}$$

Since the maximum completion of deteriorated delayed and immediate patients are no more than those in q_3, q_4 is feasible. Besides, the number of deteriorated delayed patients in q_4 is smaller than that in q_3, contradicting the optimality of q_3. Thus, immediate patients with FT $>C+\alpha_0\beta-\beta_0\alpha$ are served consecutively.

Finally, we show that among patients with wait times $<\tau$, as many delayed patients should be served before immediate patients with FT $>C+\alpha_0\beta-\beta_0\alpha$.

Among patients with wait times $<\tau$, since immediate patients with FT $>C+\alpha_0\beta-\beta_0\alpha$ are served consecutively, we could treat immediate patients with FT $>C+\alpha_0\beta-\beta_0\alpha$ as one "big" immediate patient f. Without loss of generality, suppose in an optimal schedule q, there are m ($m\geq 1$) immediate patients with FT $>C+\alpha_0\beta-\beta_0\alpha$ and patient f gets service at time t.

$$C_f=\frac{\beta_0}{\beta}[(1+\beta)^m-1]+(1+\beta)^m t \tag{41}$$

$$p_f=\frac{\beta_0}{\beta}[(1+\beta)^m-1]+[(1+\beta)^m-1]t \tag{42}$$

Thus, the I2D ratio of patient f equals $\frac{\beta_0}{\beta}$.

Among patients with wait times $<\tau$, suppose there are n_1 delayed patients served before immediate patients

with FT $> C + \alpha_0\beta - \beta_0\alpha$ in optimal schedule q, but there could be at most $n_2(n_2 > n_1)$ delayed patients having priority than immediate patients with FT $> C + \alpha_0\beta - \beta_0\alpha$. Without loss of generality, suppose $n_2 = n_1 + 1$. Denote the first delayed patient served after immediate patients with FT $> C + \alpha_0\beta - \beta_0\alpha$ as patient j. Let q^* be obtained by swapping the positions of patient j & f in q. When the type of patient j remains the same in q and q^*, according to Proposition 1, we have $C_f(q^*) \leq C_j(q)$. Since $C^i_{max}(q^*) \leq \bar{D}^i$ and $C^d_{max}(q^*) < C^d_{max}(q)$, q^* is feasible. Besides, the number of deteriorated delayed patients in q^* is no more than that in q, q^* is at least as good as q. If the type of patient j becomes different between q and q^*, then patient j is not deteriorated in q^*. Denote $\frac{\beta_0}{\beta}[(1+\beta)^m - 1]$ as β_0^f and $(1+\beta)^m - 1$ as β^f, we have

$$C_f(q^*) = \alpha_0 + \beta_0^f + \alpha_0\beta^f + (1+\alpha)(1+\beta^f)t, \tag{43}$$

$$C_j(q) = \alpha_0 + (\alpha - \beta)S + (1+\beta)\beta_0^f + (1+\beta)(1+\beta^f)t. \tag{44}$$

Recall that immediate patients' finish times $> C + \alpha_0\beta - \beta_0\alpha$, we have $t > \alpha_0 + (1+\alpha)\frac{S-\beta_0}{1+\beta}$. Equation (44) - Equation (43)

$$\begin{aligned}
&= \beta\beta_0^f - \alpha_0\beta^f + (\alpha - \beta)S + (1+\beta^f)(\beta - \alpha)t \\
&> \beta\beta_0^f - \alpha_0\beta^f + (\alpha - \beta)S + (1+\beta^f)(\beta - \alpha)[\alpha_0 + (1+\alpha)\frac{S-\beta_0}{1+\beta}] \\
&> (\beta_0 - \alpha_0)[(1+\beta)^m - 1] + (\alpha - \beta)S + (1+\beta)^m(\beta - \alpha)\alpha_0 + (1+\alpha)(\beta - \alpha)(1+\beta)^{m-1}(S - \beta_0) \\
&> (\beta - \alpha)S[(1+\beta)^{m-1}(1+\alpha) - 1] + (\beta_0 - \alpha_0)[(1+\beta)^m - 1] + (\beta - \alpha)(1+\beta)^{m-1}[\alpha_0(1+\beta) - \beta_0(1+\alpha)] \\
&> (\beta - \alpha)S[(1+\beta)^{m-1}(1+\alpha) - 1] + (\beta - \alpha)(1+\beta)^{m-1}(\alpha_0\beta - \beta_0\alpha) + (\beta_0 - \alpha_0)[(1+\beta)^{m-1} - 1 + \alpha(1+\beta)^{m-1}] \\
&> 0,
\end{aligned} \tag{45}$$

which suggests that victims following victim f in q^* get treatment earlier than those in q. Since $C^i_{max}(q^*) \leq \bar{D}^i$ and $C^d_{max}(q^*) \leq C^d_{max}(q)$, q^* is feasible. Besides, the number of deteriorated delayed patients in q^* is smaller than that in q, q^* is better than q, proving the desired result.

Among patients with wait times $\geq \tau$ in the optimal schedule, similar argument can be used to prove the results.

□

Appendix B: Algorithms for the Model with both Unstable and Stable Patients

For ease of understanding, the algorithms presented in this section make references to different patient clusters marked in Figure 4 of the main text.

Algorithm 3: Optimal schedule when $p^{sd} \geq \alpha_0 + \alpha S$, $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$, and $\bar{D}^i \leq S + (1+\beta)p^{sd}$

1. Schedule all stable immediate patients backwards from $\bar{D}^i$—denote this cluster as cluster 1;
2. Schedule all unstable immediate patients backwards from ST(cluster 1)—denote this cluster as cluster 2;
3. Insert as many unstable delayed patients onwards from 0 to ST(cluster 2) as possible—denote the number as n^d;

if $n^d < N^d$ **then**

4. Move clusters 1 & 2 backward to ensure no idle time in the system. Schedule $N^d - n^d$ unstable delayed patients forwards from FT(cluster 1);
5. Schedule N^{sd} stable delayed patients to the end;

else

6. Move cluster 2 backwards to ensure no idle time in the system;
7. Insert N^{sd} stable delayed patients forwards from FT(cluster 2). If necessary, move cluster 1 backwards to ensure no idle time.

end

Algorithm 4: Optimal schedule when $p^{sd} \geq \alpha_0 + \alpha S$, $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$, and $\bar{D}^i > S + (1+\beta)p^{sd}$

1. Schedule N^{si} stable immediate patients from $\bar{D}^i$ backwards—denote this cluster as cluster 1;
2. Schedule all unstable immediate patients from ST(cluster 1) backwards—denote this cluster as cluster 2;
3. Insert as many unstable delayed patients onwards from time 0 to ST(cluster 2) as possible—denote this cluster as cluster 3 and this number as n^d;

if $n^d < N^d$ **then**
- 4. If necessary, move clusters 1 & 2 backwards to ensure no idle time in the system. Schedule $N^d - n^d$ unstable delayed patients forwards from FT(cluster 1);
- 5. Schedule N^{sd} stable delayed patients at the end;

else
- 6. Let n^{sd}_{max} represent the maximum number of stable delayed patients served between [FT(cluster 3), $\min\{S + p^{sd}$, ST(cluster 1)$\}$]. Set $n^{sd} = n^{sd}_{max}$;
- **while** $n^{sd} \geq 0$ **do**
  - 7. Schedule n^{sd} stable delayed patients backwards from $\min\{S + p^{sd}$, ST(cluster 1)$\}$—denote this cluster as cluster 4;
  - 8. Insert N^i unstable immediate patients onwards from FT(cluster 3). If necessary, move cluster 4 backwards to ensure no idle time before the last unstable immediate patients;
  - 9. If necessary, move clusters 4 & 1 backwards to ensure no idle time in the system;
  - 10. Schedule $N^{sd} - n^{sd}$ stable delayed patients to the end;
  - **if** *the resulting schedule is feasible* **then** record this schedule. Done;
  - **else** $n^{sd} = n^{sd} - 1$;
- **end**

end

Algorithm 5: Optimal schedule when $p^{sd} \geq \alpha_0 + \alpha S$, $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$, and $\bar{D}^i \leq \min\{S + (1+\beta)p^{sd}, C + \alpha_0\beta - \beta_0\alpha\}$

1. Schedule N^{si} stable immediate patients backwards from $\bar{D}^i$—denote this cluster as cluster 1;
2. Schedule N^i unstable immediate patients forwards from time 0—denote this cluster as cluster 2;
3. Insert as many unstable delayed patients between FT(cluster 2) and ST(cluster 1) as possible—denote this cluster as cluster 3 and the number of unstable delayed patients in cluster 3 as n^d;

if $n^d < N^d$ **then**
- 4. If necessary, move cluster 1 backwards to ensure no idle time in the system;
- 5. Schedule $N^d - n^d$ unstable delayed patients forwards from FT(cluster 1);
- 6. Schedule N^{sd} stable delayed patients to the end;

else
- 7. Insert N^{sd} stable delayed patients forwards from FT(cluster 3). If necessary, move cluster 1 backwards to time 0 to ensure no idle time.

end

Algorithm 6: Optimal schedule when $p^{sd} \geq \alpha_0 + \alpha S$, $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$, and $\bar{D}^i > \min\{S + (1+\beta)p^{sd}, C + \alpha_0\beta - \beta_0\alpha\}$

Initialization: Denote the number of unstable immediate patients with FT $\leq C + \alpha_0\beta - \beta_0\alpha$ as n^i and the number of unstable immediate patients with FT $> C + \alpha_0\beta - \beta_0\alpha$ should be $N^i - n^i$. Set $n^i = 0$;

1. Schedule N^{sd} stable immediate patients backwards from $\bar{D}^i$—denote this cluster as cluster 1;

while $n^i \leq N^i$ **do**

2. Schedule n^i unstable immediate patients forwards from time 0—denote this cluster as cluster 2[1];

3. Schedule $N^i - n^i$ unstable immediate patients backwards from ST(cluster 1)—denote this cluster as cluster 2[2];

4. Insert as many unstable delayed patients onwards between [FT(cluster 2[1]), ST(cluster 2[2])]—denote this cluster as cluster 3 and the number of patients in cluster 3 as n^d;

if $n^d < N^d$ **then**

5. If necessary, move clusters 2[2] & 1 backwards to ensure no idle time in the system;

6. Schedule $N^d - n^d$ unstable delayed patients forwards from FT(cluster 1);

7. Schedule N^{sd} stable delayed patients to the end;

else

8. Denote the maximum number of stable delayed patients served in [FT(cluster 3), $S + p^{sd}$] as n^{sd}_{max} and the number of stable delayed patients served before $S + p^{sd}$ as n^{sd}. Let $n^{sd} = n^{sd}_{max}$;

while $n^{sd} \geq 0$ **do**

9. Schedule n^{sd} stable delayed patients forwards from FT(cluster 3)—denote this cluster as cluster 4;

10. Schedule $N^i - n^i$ unstable immediate patients onwards from FT(cluster 4);

11. Schedule N^{si} stable immediate patients onwards from the finish time of the last unstable immediate patients;

12. Schedule $N^{sd} - n^{sd}$ stable delayed patients to the end;

if *The resulting schedule is feasible* **then** record this schedule;

else $n^{sd} = n^{sd} - 1$;

end

end

13. $n^i = n^i + 1$;

end

14. Compare all feasible schedule and get the best solution.

Algorithm 7: Optimal schedule when $p^{sd} < \alpha_0 + \alpha S$, $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$, and $\bar{D}^i \leq S + (1+\beta)p^{sd}$

Initialization: Denote n^{sd} as the number of stable delayed patients served without deterioration in the optimal schedule, and n^{sd}_{max} as the largest number of stable delayed patients that can be served within $S + p^{sd}$ units of time, we have $n^{sd} \leq n^{sd}_{max}$. Set $n^{sd} = n^{sd}_{max}$;

while $n^{sd} \geq 0$ **do**

1. Schedule $N^{sd} - n^{sd}$ stable delayed patients from $\bar{D}^d$ backwards—denote this cluster as cluster 4[1];

if $\bar{D}^i > S + p^{sd}$ **then**

2. Schedule N^{si} stable immediate patients backwards from $\min\{\text{ST(cluster 4[1])}, \bar{D}^i\}$—denote this cluster as cluster 1;

3. Schedule n^{sd} stable delayed patients backwards from $\min\{\text{ST(cluster 1)}, S + p^{sd}\}$—denote this cluster as cluster 4[2];

else

4. Schedule n^{sd} stable delayed patients backwards from $\min\{\text{ST(cluster 4[1])}, S + p^{sd}\}$—denote this cluster as cluster 4[2];

5. Schedule N^{si} stable immediate patients backwards from $\min\{\text{ST(cluster 4[2])}, \bar{D}^i\}$—denote this cluster as cluster 1;

end

6. Schedule all unstable immediate patients backwards from $\min\{$ST(cluster 1), ST(cluster 4[2])—denote this cluster as cluster 2;

7. Insert unstable delayed patients from time 0 to $\bar{D}^d$, keeping the order of other patients unchanged. If necessary, move the next cluster backward to ensure no idle time;

8. **if** *the resulting schedule is feasible* **then** evaluate the schedule;

9. $n^{sd} = n^{sd} - 1$;

end

10. Compare all feasible schedules and get the best solution.

Algorithm 8: Optimal schedule when $p^{sd} < \alpha_0 + \alpha S$, $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$, and $\bar{D}^i > S + (1+\beta)p^{sd}$

Initialization: Denote n^{sd} as the number of stable delayed patients served without deterioration and n^{sd}_{max} as the largest number of stable delayed patients that can be served within $S + p^{sd}$ units of time. Set $n^{sd} = n^{sd}_{max}$;

while $n^{sd} \geq 0$ **do**

1. Schedule $N^{sd} - n^{sd}$ stable delayed patients backwards from $\bar{D}^d$—denote this cluster as cluster 4[1];
2. Schedule stable immediate patients from $\min\{\text{ST(cluster 4[1])}, \bar{D}^i\}$—denote this cluster as cluster 1;

if *ST(cluster 1)* $\leq S + (1+\beta)p^{sd}$ **then**

3. Schedule n^{sd} stable delayed patients backwards from $\min\{S + p^{sd}, \text{ST(cluster 1)}\}$—denote this cluster as cluster 4[2];
4. Schedule all unstable immediate patients from ST(cluster 4[2]) backwards—denote this cluster as cluster 2;
5. Inset unstable delayed patients from time 0 onwards, keeping the order of other patients unchanged. If necessary, move clusters backwards to ensure no idle time;

if *the resulting schedule is feasible* **then** evaluate the schedule;

else

if $n^{sd} = 0$ **then**

6. Schedule N^i unstable immediate patients backwards from ST(cluster 1);
7. Insert N^d unstable delayed patients forwards from time 0. If necessary, move the next cluster backwards to ensure no idle time;

else

8. Denote n^{sd} stable delayed patients without deterioration as cluster 4[2] and n^i as the number of unstable immediate patients served before cluster 4[2]. Set $n^i = 0$;

while $n^i \leq N^i$ **do**

9. Schedule $N^i - n^i$ unstable immediate patients backwards from ST(cluster 1)—denote this cluster as cluster 2[1];
10. Schedule n^{sd} stable delayed patients, i.e., cluster 4[2] backwards from $\min\{S + p^{sd}, \text{ST(cluster 2[1])}\}$;
11. Schedule n^i unstable immediate patients backwards from ST(cluster 4[2])—denote this cluster as cluster 2[2];
12. Insert N^d unstable delayed patients forwards from time 0. If necessary, move the next cluster backwards to ensure no idle time;
13. **if** *The resulting schedule is feasible* **then** evaluate the schedule;
14. $n^i = n^i + 1$;

15. $n^{sd} = n^{sd} - 1$;

16. Compare all feasible schedules and get the best one.

Algorithm 9: Optimal schedule when $p^{sd} < \alpha_0 + \alpha S$, $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$ and $\bar{D}^i \leq \min\{S + (1+\beta)p^{sd}, C + \alpha_0\beta - \beta_0\alpha\}$

Initialization: Denote n^{sd} as the number of stable delayed patients served without deterioration and n^{sd}_{max} as the largest number of stable delayed patients that can be served within $S + p^{sd}$ units of time. Set $n^{sd} = n^{sd}_{max}$;

1. Schedule N^i unstable immediate patients forwards from time 0—denote this cluster as cluster 2;

while $n^{sd} \geq 0$ **do**

2. Schedule $N^{sd} - n^{sd}$ stable delayed patients backwards from $\bar{D}^d$—denote this cluster as cluster 4[1];

if $\bar{D}^i \geq S + p^{sd}$ **then**

3. Schedule N^{si} stable immediate patients backwards from $\min\{\text{ST(cluster 4[1])}, \bar{D}^i\}$—denote this cluster as cluster 1;

4. Schedule n^{sd} stable delayed patients backwards from $\min\{\text{ST(cluster 1)}, S + p^{sd}\}$—denote this cluster as cluster 4[2];

else

5.Schedule n^{sd} stable delayed patients from $\min\{\text{ST(cluster 4[1])}, S + p^{sd}\}$—denote this cluster as cluster 4[2];

6. Schedule N^{si} stable immediate patients from $\min\{\text{ST(cluster 4[2])}, \bar{D}^i\}$ backwards—denote this cluster as cluster 1;

end

7. Insert unstable delayed patients forwards from FT(cluster 2), keeping the order of other patients unchanged. If necessary, move the next cluster backward to ensure no idle time;

8. **if** *the resulting schedule is feasible* **then** record the schedule;

9. $n^{sd} = n^{sd} - 1$;

end

10. Compare all feasible schedules and get the best solution.

Algorithm 10: Optimal schedule when $p^{sd} < \alpha_0 + \alpha S$, $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$, and $\bar{D}^i > \min\{S + (1+\beta)p^{sd}, C + \alpha_0\beta - \beta_0\alpha\}$

Initialization: Let m_1,m_2 and m_3 be the numbers of unstable immediate patients with FT $\leq \min\{C + \alpha_0\beta - \beta_0\alpha, S + (1+\beta)p^{sd}\}$, FT $> C + \alpha_0\beta - \beta_0\alpha$, and FT between $\min\{C + \alpha_0\beta - \beta_0\alpha, S + (1+\beta)p^{sd}\}$ and $C + \alpha_0\beta - \beta_0\alpha$, respectively. Denote n^{sd} as the number of stable delayed patients served without deterioration in the optimal schedule and n^{sd}_{max} as the largest number of stable delayed patients that can be served within $S + p^{sd}$ units of time. Set $n^{sd} = n^{sd}_{max}$ and $m_1 = m_3 = 0$;

while $n^{sd} \geq 0$ **do**

1. Schedule $N^{sd} - n^{sd}$ stable delayed patients backwards from $\bar{D}^d$—denoted as cluster 4[1];
2. Schedule N^{si} stable immediate patients backwards from $\min\{$ST(cluster 4[1]), $\bar{D}^i\}$—denoted as cluster 1;

while $m_1 \leq N^i$ **do**

3. Schedule m_1 unstable immediate patients forwards from time 0—denoted as cluster 2[1];

if $S + (1+\beta)p^{sd} < C + \alpha_0\beta - \beta_0\alpha$ **then**

while $m_3 \leq 1$ **do**

4. Schedule $N^i - m_1 - m_3$ unstable immediate patients backwards from ST(cluster 1)—denote this as cluster 2[2];
5. Schedule m_3 unstable immediate patients backwards from $\min\{C + \alpha_0\beta - \beta_0\alpha,$ ST(cluster 2[2])$\}$—denote this cluster as cluster 2[3];
6. Schedule n^{sd} stable delayed patients backwards from ST(cluster 2[3])—denote this cluster as cluster 4[2];
7. Let n^d_{max} and n^d_{min} represent the maximum and minimum number of unstable delayed patients served before cluster 4[2]. Set $n^d = n^d_{min}$;

while $n^d \leq n^d_{max}$ **do**

8. Insert n^d unstable delayed patients forwards from FT(cluster 2[1]). If necessary, move the next cluster backwards to ensure no idle time before cluster 2[3];
9. Insert $N^d - n^d$ unstable delayed patients forwards from FT(cluster 2[3]). If necessary, move the next cluster backwards to ensure no idle time in the system;

if *the resulting schedule is feasible* **then** evaluate the schedule;

10. $n_d = n_d + 1$;

11. $m_3 = m_3 + 1$;

else

12. Schedule $N^i - m_1$ unstable immediate patients backwards from ST(cluster 1)—denote this cluster as cluster 2[2];
13. Schedule n^{sd} stable delayed patients backwards from $\min\{S + p^{sd},$ ST(cluster 2[2])$\}$—denote this cluster as cluster 4[2];
14. Insert unstable delayed patients forwards from FT(cluster 2[1]). If necessary, move the next cluster backwards to ensure no idle time in the system;
15. **if** *the resulting schedule is feasible* **then** evaluate the schedule;

16. $m_1 = m_1 + 1$;

17. $n^{sd} = n^{sd} - 1$;

18. Compare all feasible schedules and get the best solution.

Appendix C: Explanation of Figure 4(f)

Figure 4(f) concerns the parameter setting where $p^{sd} < \alpha_0 + \alpha S, \frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$ and $\bar{D}^i > S + (1+\beta)p^{sd}$. In this setting, with stable delayed patients, the strict priority order between unstable delayed and unstable immediate patients before $\bar{D}^i$ as pressed for by the condition $\frac{\alpha_0}{\alpha} \leq \frac{\beta_0}{\beta}$ does not hold. What we can show is that in the early portion and the late portion of the schedule before $\bar{D}^i$, unstable delayed patients have higher priority than unstable immediate ones, but this priority order may not prevail throughout. The intuition is explained belo···

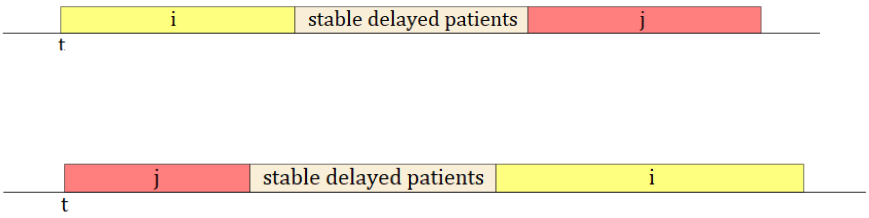


Figure 7 Illustration for Swapping Patients i and j

Suppose that the unstable delayed patient i gets service at time t, and then a cluster of n stable delayed patients get served, followed by the unstable immediate patient j. Let us consider that $t < S$, so patient i has not deteriorated. Let C be the completion time for these patients and it follows that

$$C = \alpha_0(1+\beta) + (1+\alpha)(1+\beta)t + np^{sd}(1+\beta) + \beta_0.$$

Let us swap the positions of unstable delayed patient i and unstable immediate patient j. We use C' to denote the completion time of these patients after this swap. If patient i is not deteriorated, then

$$C' = \beta_0(1+\alpha) + (1+\alpha)(1+\beta)t + np^{sd}(1+\alpha) + \alpha_0.$$

We have

$$C' - C = \beta_0\alpha - \alpha_0\beta + np^{sd}(\alpha - \beta).$$

If, however, patient i is deteriorated, then

$$C' = \beta_0(1+\beta) + (1+\beta)^2 t + np^{sd}(1+\beta) + \alpha_0 + \alpha S - \beta S.$$

We have

$$C' - C = \beta(\beta_0 - \alpha_0) + (1+\beta)(\beta - \alpha)t + (\alpha - \beta)S.$$

Thus, regardless of the deterioration status of patient i after the swap, the sign of $C' - C$ is undetermined. Though prioritizing delayed patients could decrease the number of deteriorated cases, it may increase the completion time of all patients, leading to an infeasible solution. Therefore, with stable delayed patients, a strict priority between unstable immediate patients and unstable delayed patients does not exist in the optimal schedule. This is a sharp contrast to Theorem 1 (a), where there are no stable delayed patients.

Appendix D: Algorithms for the Model with a Mandated Rest Period

Algorithm 11: Optimal schedule with a mandated rest period for providers when $\frac{\alpha_0}{\alpha} > \frac{\beta_0}{\beta}$

Initialization: $m_1 = 0;\ m_1^s = 0;\ n_1 = 0;\ m_2^s = 0;$

while $m_1 \leq N^i$ **do**
  if $\tau \leq C + \alpha_0\beta - \beta_0\alpha$ **then** $m_1^s = m_1$;
  while $n_1 \leq N^d$ **do**
    while $m_1^s \leq m_1$ **do**
      1. Schedule m_1^s immediate patients forwards forwards from time 0—denote this cluster as cluster 1;
      2. Schedule n_1 delayed patients forwards from FT(cluster 1)—denote this cluster as cluster 2;
      3. Schedule $m_1 - m_1^s$ immediate patients forwards from FT(cluster 2)—denote this cluster as cluster 3;
      while $m_2^s \leq N^i - m_1$ **do**
        if $m_1 - m_1^s \neq 0$ **then**
          4. Set $m_2^s = N^i - m_1$ and schedule m_2^s immediate patients backwards from $\bar{D}^i$—denote this cluster as cluster 4;
          5. Insert as many delayed patients forwards from FT(cluster 3)$+r$ as possible. If necessary, move the next cluster backwards to ensure no idle time in the schedule;
          if *Feasible* **then** 6. Record this schedule;
        else if $m_1 - m_1^s = 0$ **then**
          7. Schedule m_2^s immediate patients forwards from FT(cluster 2)$+r$—denote this cluster as cluster 4;
          8. Schedule $m_2 - m_2^s$ immediate patients backwards from $\bar{D}^i$—denote this as cluster 5;
          9. Insert as many delayed patients forwards from FT(cluster 4) as possible. If necessary, move the next cluster backwards to ensure no idle time in the schedule;
          if *Feasible* **then** 10. Record this schedule;
        11. $m_2^s = m_2^s + 1$;
      end
      12. $m_1^s = m_1^s + 1$;
    end
    13. $n_1 = n_1 + 1$;
  end
  14. $m_1 = m_1 + 1$;
end
15. Compare all feasible schedules and find the optimal one.