Outpatient Appointment Scheduling with Waiting Time Limits

Rachel R. Chen

UC Davis

Joint work with Han Zhu (DUFE), Min Zhang (DUFE) and Zhenzhen Zhang (Tongji)

Introduction

- Long patient waiting times in outpatient clinics remain a serious concern.
  - Mainland China Public Hosp.: average waiting time is 57 minutes
  - HK Public Hosp. Foot & Ankle: average waiting time exceeds 2 hours
  - They negatively impact patients' experiences and perceptions of service quality.

Introduction

- To address this issue, policymakers may impose a waiting time limit (WTL).
  - China Healthcare Commission: WTL = 30 minutes
  - Ministry of Health in Singapore: WTL = 75 minutes
  - The Department of Health in Hong Kong, the Patient's Charter of the United Kingdom Government, …

MINISTRY OF HEALTH
SINGAPORE

MOH and the two clusters want to re-assure the public that we will actively improve the waiting time for SOC consultation. We cannot adopt a one-size-fits-all approach because each medical speciality is different. We recognise the inherent variability and sometimes unpredictability of healthcare services. Notwithstanding, we have established a target that at least 50% of our patients would be seen within 30 minutes of their appointment time, and that 95% of our patients would not have to wait beyond 75 minutes of their appointment times. To achieve, this, we need the co-operation of everyone involved.

Introduction

- When a patient's waiting time limit is exceeded, clinics may need to use flexible or reserved resources to serve the patient —— patient diversion.
- The impact of WTL on the clinic
  - Diverting patients may incur additional costs.
  - Patient diversion disrupts its operations.
- Multiple inherent uncertainties further complicate the analysis.
- It remains unclear how to effectively schedule appointments in the presence of waiting time limits.

Research Questions: We study outpatient appointment scheduling with waiting time limits, in the presence of uncertain service times, patient no-shows, and unpunctual arrivals.

Literature

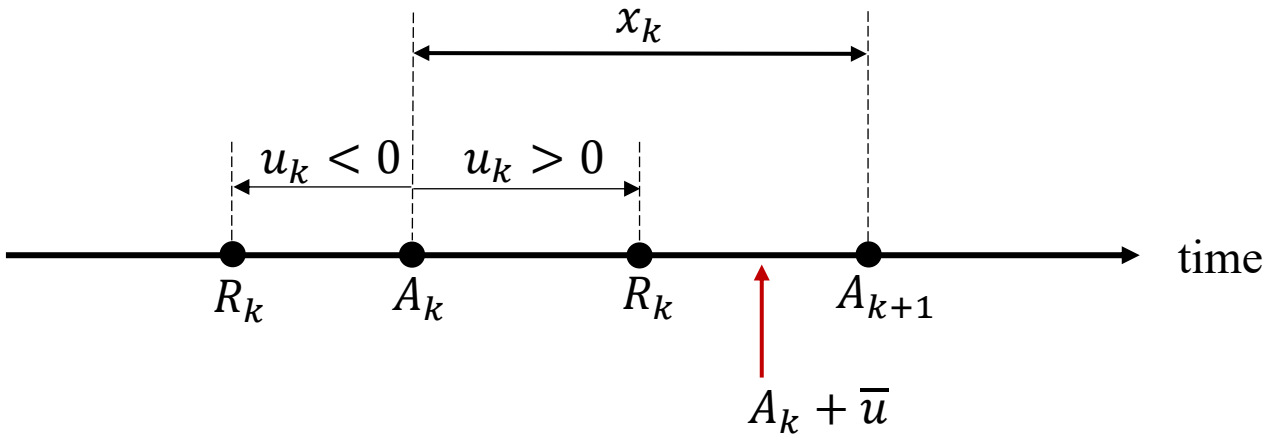
- Appointment scheduling with stochastic service times and no-shows
  - Numerous studies show that the optimal schedule has a dome-shaped pattern. (Hassin and Mendel 2008, Klassen and Yoogalingam 2008, Robinson and Chen 2010, Cayirli and Yang 2014, Zhou and Yue 2019)
  - However, we find the dome-shaped pattern does not necessarily hold in the presence of WTLs.
- Appointment scheduling with waiting time limits
  - hard constraints (Huang et al. 2015, Wen et al. 2020, Zhou et al. 2021, Babashov et al. 2023)
  - soft constraints (Qi 2017, Pan et al. 2020, Wang et al. 2024)
  - Our work considers multiple uncertainties and patient diversion.
- Customer reneging (Jouini et al. 2011, Huh et al. 2013, Lu et al. 2022)
  - Imposing waiting time limits introduces penalties or higher costs.
- Appointment scheduling with patient unpunctuality (Deceuninck et al. 2018, Jiang et al. 2019, Wu and Zhou 2022)
  - None of these studies take into consideration waiting time limits.

Our work differs from prior research by simultaneously considering waiting time limits, uncertain service times, patient no-shows, and unpunctuality.

Model Setup

We study the case with a single doctor serving N patients.

- A_k: the appointment time of patient k (k = 1,…, N).
- R_k: the actual arrival time of patient k (k = 1,…, N).
- $u_k = R_k - A_k \in [\,\underline{u}, \overline{u}\,]$: the unpunctual time of patient k (k = 1,…, N).
- $x_k = A_{k+1} - A_k$: the job allowance (k = 1,…, N-1).
- decision variables: $\boldsymbol{x} = (x_1, \ldots, x_{N-1})$

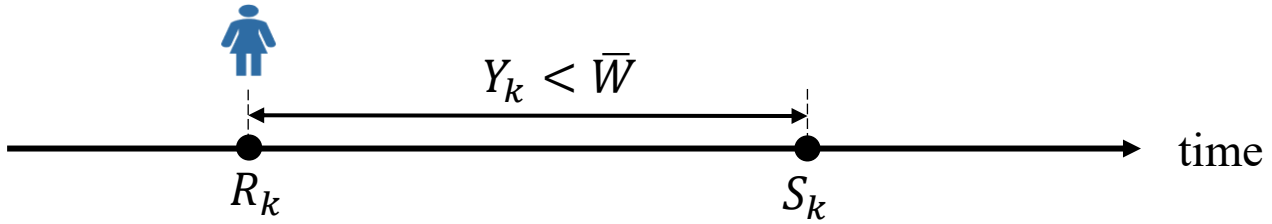


we model no-show patients as “ghosts”, who show up at their latest possible arrival times with zero service time.

Model Setup

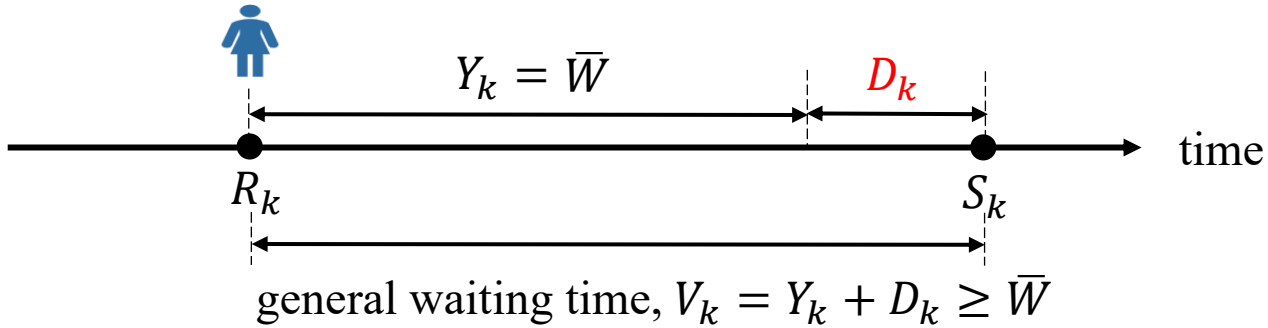
- $\bar{W}$: waiting time limit.
- S_k: general service starting time of patient k when $\bar{W} = \infty$.
- Y_k: actual waiting time of patient k, measured from her arrival time.
- D_k: virtual waiting time of patient k.

(1) Patients who are seen by the current doctor.



Patient diversion

(2) Patients who are unseen by the current doctor.



- $t_k = \mathbb{I}\{V_k \geq \bar{W}\} \in \{0,1\}$ identifies whether patient k is diverted.

Mathematical Formulation

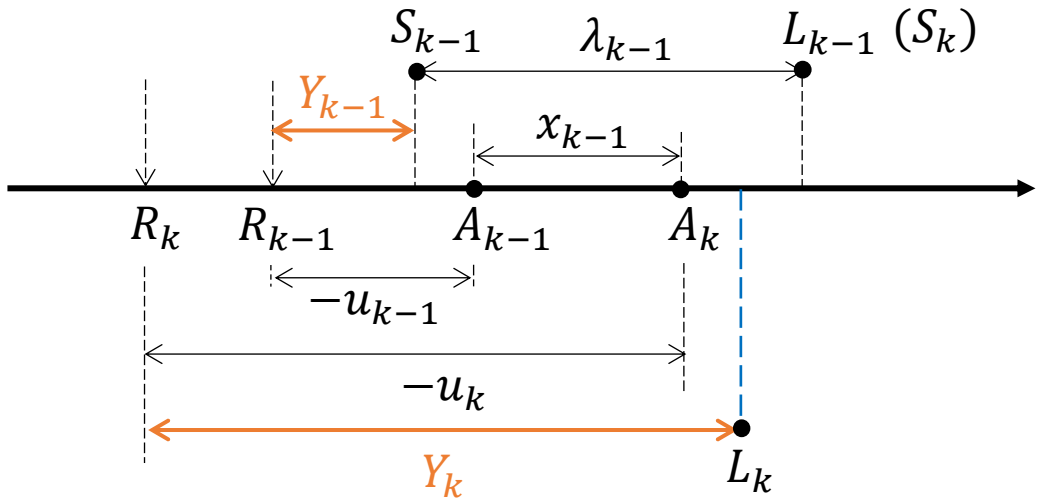
- Waiting time limits and patient unpunctuality lead to numerous scenarios of system dynamics.
- A scenario:

Patient k is diverted.

$$R_k \to R_{k-1} \to A_{k-1} \to A_k \;\&\; V_{k-1} < \bar{W} \;\&\; V_k \geq \bar{W}$$

Patient k arrives earlier than patient k-1, and both arrive early.

Patient k-1 receives service.



$$D_k = V_k - Y_k.$$

$$Y_k = -u_k + u_{k-1} - x_{k-1} + Y_{k-1} + \lambda_{k-1} + D_{k-1} - D_k.$$

$$I_k = 0.$$

Mathematical Formulation

Due to patient unpunctuality, the arrival sequence of any two adjacent patients is uncertain, potentially matching or differing from their appointed order.

- Out-of-order case

26 scenarios

  - Out-of-order arrivals of two adjacent patients

$$(1)\ R_k \to R_{k-1} \to A_{k-1} \to A_k \qquad (4)\ A_{k-1} \to R_k \to R_{k-1} \to A_k$$
$$(2)\ R_k \to A_{k-1} \to R_{k-1} \to A_k \qquad (5)\ A_{k-1} \to R_k \to A_k \to R_{k-1}$$
$$(3)\ R_k \to A_{k-1} \to A_k \to R_{k-1} \qquad (6)\ A_{k-1} \to A_k \to R_k \to R_{k-1}$$

18

  - Patients may exit the queue due to waiting time limits.

$$(1)\ V_{k-1} < \bar{W}\ \&\ V_k < \bar{W}\ \ (2)\ V_{k-1} < \bar{W}\ \&\ V_k \geq \bar{W}\ \ (3)\ V_{k-1} \geq \bar{W}\ \&\ V_k \geq \bar{W}$$

- In-order case

  - $R_k \leq L_{k-1}$: $V_{k-1} < (\geq)\bar{W}\ \&\ V_k < (\geq)\bar{W}$ (4 scenarios)
  - $L_{k-1} < R_k \leq L_{k-1} + D_{k-1}$: (1) $V_{k-1} \geq \bar{W}\ \&\ V_k < \bar{W}$ (2) $V_{k-1} \geq \bar{W}\ \&\ V_k \geq \bar{W}$ (2 scenarios)
  - $R_k > L_{k-1} + D_{k-1}$: (1)$V_{k-1} < \bar{W}\ \&\ V_k < \bar{W}$ (2) $V_{k-1} \geq \bar{W}\ \&\ V_k < \bar{W}$ (2 scenarios)

8

Mathematical Formulation

- Proposition 1 unifies expressions of system dynamics of 26 scenarios.

Proposition 1. In both out-of-order and in-order cases, the k^{th} patient's general waiting time V_k and the doctor's idle time I_k are given by

$$V_k = \begin{cases} (-u_1)^+, k = 1; \\ (V_{k-1} + \lambda_{k-1} - x_{k-1} + u_{k-1} - u_k)^+, \forall k \geq 2. \end{cases} \quad (1)$$

$$I_k = \begin{cases} (u_1)^+, k = 1; \\ (-V_{k-1} - \lambda_{k-1} + x_{k-1} - u_{k-1} + u_k)^+, \forall k \geq 2. \end{cases} \quad (2)$$

Mathematical Formulation

$$\min \mathbb{E}[\sum_{k=1}^{N}(\alpha W_k + \theta t_k + \beta I_k) + \gamma O]$$

① ② ③ ④

①patients' waiting cost + ② penalty cost + ③ idle time cost + ④overtime cost.

$$s.t. \quad V_k = \begin{cases} (-u_1)^+, k = 1; \\ (V_{k-1} + \lambda_{k-1} - x_{k-1} + u_{k-1} - u_k)^+, \forall k = 2, \dots, N \end{cases}$$

(1) general waiting time

$$I_k = \begin{cases} (u_1)^+, k = 1; \\ (-V_{k-1} - \lambda_{k-1} + x_{k-1} - u_{k-1} + u_k)^+, \forall k = 2, \dots, N \end{cases}$$

(2) idle time

$$t_k = \mathbb{I}(V_k \geq \bar{W}), \forall k = 1, \dots, N$$

(3) identify whether patient k is diverted

$$\lambda_k = \xi_k(1 - t_k), \forall k = 1, \dots, N$$

(4) actual service time

$$Y_k = \min(V_k, \bar{W}), \forall k = 1, \dots, N$$

(5) actual waiting time

$$W_k = (Y_k - (-u_k)^+)^+, \forall k = 1, \dots, N$$

(6) the clinic-concerned waiting time

$$O = \left(\sum_{k=1}^{N-1} x_k + u_N + V_N + \lambda_N - \mathrm{T}\right)^+$$

(7) overtime

$$x_k \in X, V_k, Y_k, I_k, W_k, O \geq 0, t_k \in \{0,1\}, \forall k = 1, \dots, N$$

Mathematical Formulation

- Sample Average Approximation (SAA) + linearization operations: a deterministic mixed-integer linear program (DMILP).

$$\min \frac{1}{S}\sum_i^S\left[\sum_{k=1}^N\left(\alpha W_k^i + \theta t_k^i + \beta I_k^i\right) + \gamma O^i\right]$$

$$s.t. \quad V_1^i = \left(-u_1^i\right)^+, \forall i = 1, \dots, S$$

$$I_1^i = \left(u_1^i\right)^+, \forall i = 1, \dots, S$$

$$V_k^i - I_k^i = V_{k-1}^i + \lambda_{k-1}^i - x_{k-1} + u_{k-1}^i - u_k^i, \forall k = 2, \dots, N, \forall i = 1, \dots, S$$

$$V_k^i \le M_{1,k}^i v_k^i, \forall k = 2, \dots, N, \forall i = 1, \dots, S$$

$$I_k^i \le M_{2,k}^i \mu_k^i, \forall k = 2, \dots, N, \forall i = 1, \dots, S$$

$$v_k^i + \mu_k^i \le 1, \forall k = 2, \dots, N, \forall i = 1, \dots, S$$

$$v_k^i, \mu_k^i \in \{0,1\}, \forall k = 2, \dots, N, \forall i = 1, \dots, S$$

$$Y_k^i \le \bar{W}, \forall k = 1, \dots, N, \forall i = 1, \dots, S$$

$$Y_k^i \le V_k^i, \forall k = 1, \dots, N, \forall i = 1, \dots, S$$

$$Y_k^i \ge \bar{W} t_k^i, \forall k = 1, \dots, N, \forall i = 1, \dots, S$$

$$Y_k^i \ge V_k^i - M_{2,k}^i t_k^i, \forall k = 1, \dots, N, \forall i = 1, \dots, S$$

$$W_k^i \ge Y_k^i - \left(-u_k^i\right)^+, \forall k = 1, \dots, N, \forall i = 1, \dots, S$$

$$O^i \ge V_N^i + \lambda_N^i + \sum_{k=1}^{N-1} x_k + u_N^i - T, \forall i = 1, \dots, S$$

$$\lambda_k^i = \xi_k^i\left(1 - t_k^i\right), \forall k = 1, \dots, N, \forall i = 1, \dots, S$$

$$x_k = \Delta \sum_{p=0}^P 2^p y_{k,p}, \forall k = 1, \dots, N-1$$

$$1 \le \sum_{p=0}^P 2^p y_{k,p} \le \bar{y}, \forall k = 1, \dots, N-1$$

$$y_{k,p} \in \{0,1\}, \forall k = 1, \dots, N-1, p = 0,1, \dots, P$$

$$V_k^i, Y_k^i, I_k^i, W_k^i, O^i \ge 0, t_k^i \in \{0,1\}, \forall k = 1, \dots, N, \forall i = 1, \dots, S$$

The Tailored Integer L-shaped Method

- The standard integer L-shaped method
  - The second stage problem: S subproblems are mixed integer nonlinear programs.
  - Directly solving subproblems individually by off-the-shelf solvers is time-consuming.

- The tailored integer L-shaped method
  - Proposition 2 shows that the subproblem Q(x, i) has good properties, which allow us to deduce its optimal value without using any optimization solver.

Proposition 2. Given a first-stage solution $\boldsymbol{x}$ and scenario i, the unique feasible solution $(W_k^{i^*}, t_k^{i^*}, I_k^{i^*}, O^{i^*})$, $k = 1,\ldots, N$ satisfying (42) – (50) is the optimal solution to subproblem (SP), so that $Q^*(\boldsymbol{x}, i) = \sum_{k=1}^{N}\left(\alpha W_k^{i^*} + \theta t_k^{i^*} + \beta I_k^{i^*}\right) + \gamma O^{i*}$.

Numerical Analysis —— The Tailored ILSM

Table 2 Performance of the tailored ILSM

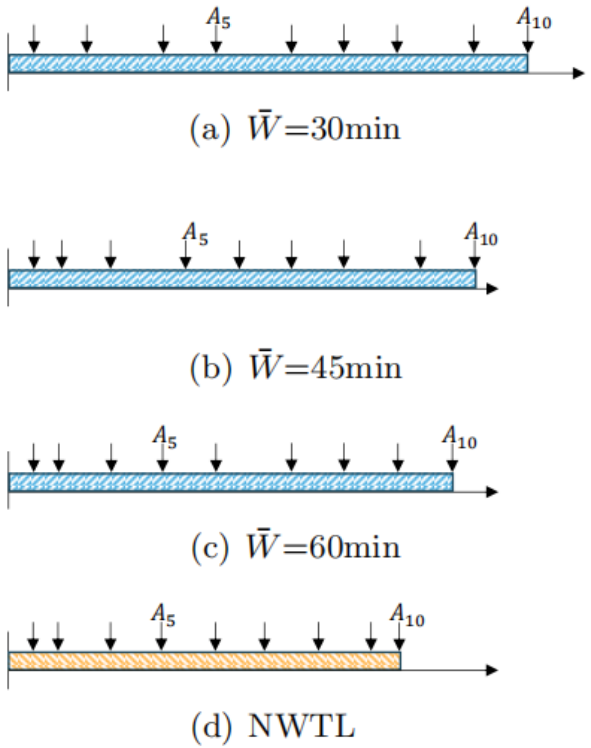
$p_{ns}=0.1$			RunningTime(s)			Gap(%)			$p_{ns}=0.2$			RunningTime(s)			Gap(%)		
T	N	S	Std-ILSM	SAA	ILSM	Std-ILSM	SAA	ILSM	T	N	S	Std-ILSM	SAA	ILSM	Std-ILSM	SAA	ILSM
60	4	1500	26	23	8	0	0	0	60	4	1500	26	17	8	0	0	0
		2000	28	39	10	0	0	0			2000	32	29	11	0	0	0
		2500	34	55	12	0	0	0			2500	40	45	13	0	0	0
		3000	46	60	16	0	0	0			3000	49	97	19	0	0	0
	5	1500	47	34	15	0	0	0		5	1500	53	32	17	0	0	0
		2000	68	57	22	0	0	0			2000	65	42	24	0	0	0
		2500	88	88	28	0	0	0			2500	81	64	28	0	0	0
		3000	102	119	34	0	0	0			3000	112	84	37	0	0	0
	6	1500	85	48	29	0	0	0		6	1500	83	40	31	0	0	0
		2000	122	77	39	0	0	0			2000	120	64	44	0	0	0
		2500	137	104	49	0	0	0			2500	140	152	53	0	0	0
		3000	162	222	58	0	0	0			3000	159	330	59	0	0	0
120	8	1500	277	257	126	0	0	0	120	8	1500	1530	1015	595	0	0	0
		2000	393	320	190	0	0	0			2000	1901	2685	701	0	0	0
		2500	479	670	202	0	0	0			2500	2375	3518	870	0	0	0
		3000	642	1158	283	0	0	0			3000	2750	7200	1056	0	2.99	0
	10	1500	2415	1657	940	0	0	0		10	1500	4831	3229	1780	0	0	0
		2000	3196	4235	1230	0	0	0			2000	6279	7200	2257	0	2.87	0
		2500	4715	6610	1736	0	0	0			2500	7200	7200	2945	1.97	2.90	0
		3000	5206	7200	2031	0	1.93	0			3000	7200	7200	3661	2.23	3.15	0
	12	1500	7200	6239	2860	2.95	0	0		12	1500	7200	6091	4237	3.50	0	0
		2000	7200	7200	4056	4.23	3.81	0			2000	7200	7200	5945	4.70	3.63	0
		2500	7200	7200	5825	3.92	4.37	0			2500	7200	7200	6830	4.88	3.97	0
		3000	7200	7200	6860	5.44	4.88	0			3000	7200	7200	7200	5.28	3.87	2.32

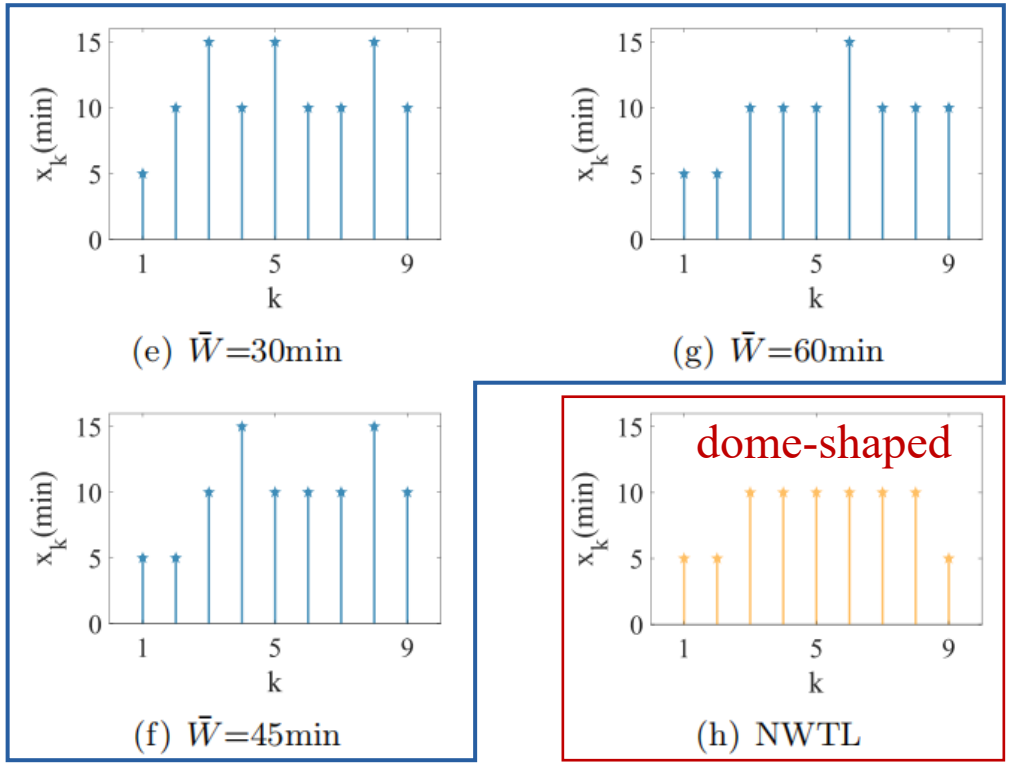
The tailored ILSM outperforms both benchmarks in all instances.

Numerical Analysis —— The Optimal Schedule

- The optimal schedule in the presence of waiting time limits **has larger job allowances** than that without waiting time limits.
- The well-known **dome-shaped** pattern is **no longer guaranteed** under waiting time limits.

larger job allowances





not dome-shaped

Figure 4 **Optimal schedules of the WTL systems under different $\bar{W}$.**

Numerical Analysis ——The Impact of Waiting Time Limits

- Waiting time limits help reduce variation in patient waiting times across different positions in the schedule, thereby **enhancing fairness** in the schedule.

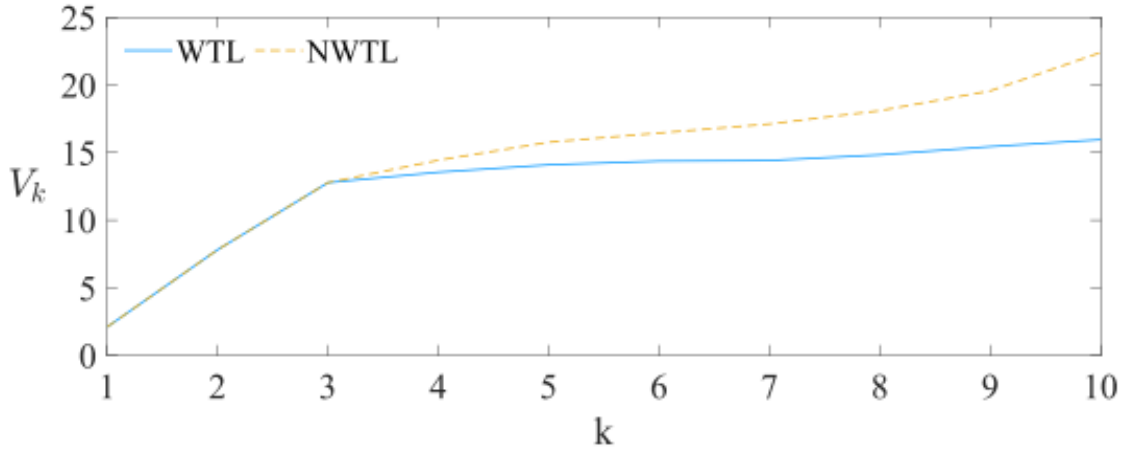
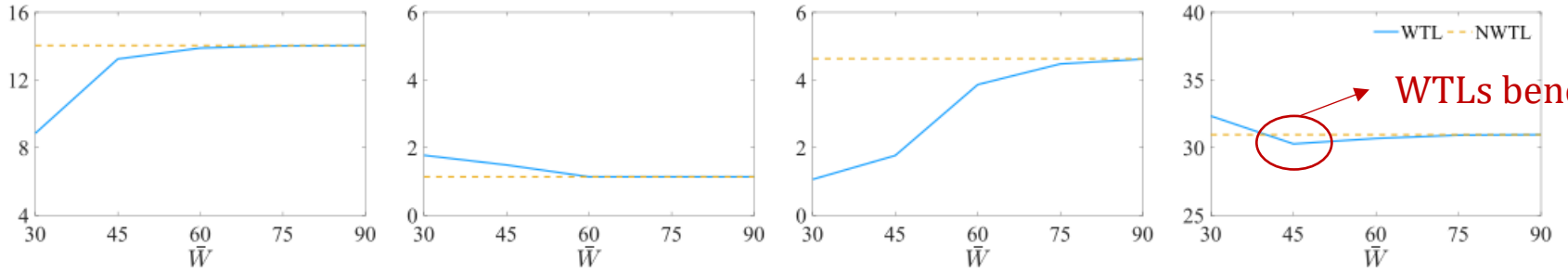


Figure 5 **Average general waiting times for patients of different positions in a line**

Numerical Analysis ——The Impact of Waiting Time Limits

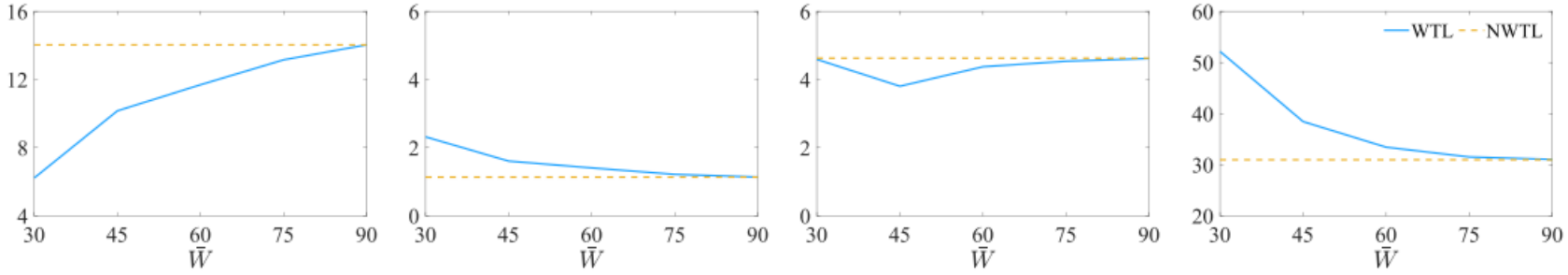
- Patient waiting time ↓
- Doctor utilization ↓ (idle time ↑; overtime ↓).
- When θ is relatively low, waiting time limits may benefit the clinic.



(a) Average waiting time (b) Idle time (c) Overtime (d) Total cost

WTLs benefit the clinic

Figure 6 Performance of both WTL and NWTL systems when unit penalty cost is low ($\theta = 20$).



(a) Average waiting time (b) Idle time (c) Overtime (d) Total cost

Figure 7 Performance of both WTL and NWTL systems when unit penalty cost is high ($\theta = 75$).

Numerical Analysis ——The Impact of Waiting Time Limits

- In the presence of waiting time limits, **the total cost** of the system is **minimized** when patients tend to **arrive slightly late** on average.

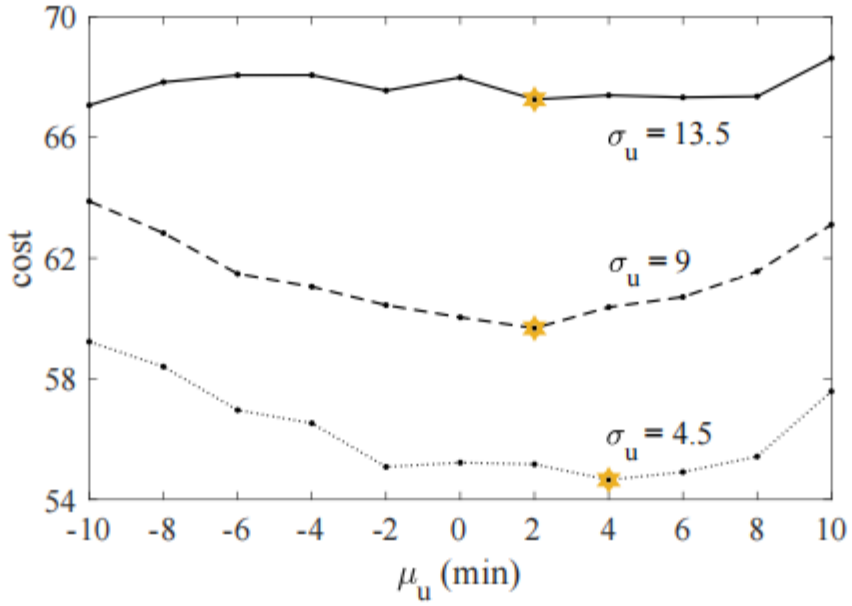


Figure 8 Comparison of Total Costs under Different Unpunctuality Parameters.

Numerical Analysis —— Different Regulatory Contexts with $\bar{W}$ and θ

- Configurations of waiting time limits ($\bar{W}$) and the unit penalty cost (θ) across different regulatory contexts.

  - Clinic self-regulation
    - The clinic proactively sets $\bar{W}$;
    - θ: the clinic's additional resource cost for each diverted patient (exogenous).

  - Planner-imposed waiting time limit
    - A social planner sets $\bar{W}$;
    - θ: the clinic's additional resource cost for each diverted patient (exogenous).

  - Social planner joint regulation
    - A social planner simultaneously decides on both $\bar{W}$ and θ_2(a fine);
    - θ_1: the clinic's additional resource cost for each diverted patient (exogenous);
    - $\theta = \theta_1 + \theta_2$.

(1) The clinic has incentives to misreport its true cost of serving each diverted patient.
(2) The social planner can combine waiting time limits with fines to improve social welfare.

Conclusion

- We study appointment scheduling with waiting time limits, in the presence of uncertain service times, patient no-shows, and unpunctuality.
- We introduce the concept of virtual waiting time, which helps us unify the modeling of system dynamics of 26 scenarios into one stochastic program. (Proposition 1)
- We develop a tailored integer L-shaped method. (Proposition 2)

Conclusion

- We unravel the impact of waiting time limits on **job allowances**, **the optimal schedule**, **fairness** in patient waiting, **doctor utilization**, and **clinic costs**. In the presence of waiting time limits,

  (1) the optimal schedule has larger job allowances;
  (2) the optimal schedule does not necessarily exhibit the dome-shaped pattern;
  (3) the presence of waiting time limits improves fairness in the schedule;
  (4) the doctor utilization is reduced;
  (5) when the penalty cost is relatively low, waiting time limits may benefit the clinic;
  (6) the total cost of the system is minimized when patients tend to arrive slightly late on average.

- Our results offer valuable insights for clinics and policymakers.