

Online Supplement

Dialysis Clinic Patient Flow and Work Design

Vincent W. Slaugh and Andre A. Cire

Appendix EC.1: Proofs

Proof of Theorem 1. Statement (a). The completion (2) imply that the tasks for the i -th patient, $i \in \mathcal{P}$, have the following contiguous schedule, obtained by subtracting $p^\downarrow$, $p^\uparrow$, and d from (2):

- Start put-on time $s_i := (i - 1)p^\downarrow + [b(i) - 1] \cdot d + [b(i) - 1] \cdot p^\uparrow$;
- Start treatment time $r_i := i \cdot p^\downarrow + [b(i) - 1] \cdot d + [b(i) - 1] \cdot p^\uparrow$; and
- Start take-off time $f_i := i \cdot p^\downarrow + b(i) \cdot d + [b(i) - 1] \cdot p^\uparrow$.

Assume without loss of generality that the number of patients n is a multiple of the number of stations C , i.e., $n/C = m$ for m integer; otherwise, this can be adjusted by adding artificial patients. Partition the patient indices $\mathcal{P}$ in m blocks, $\{(k - 1)C + 1, \dots, kC\}$, $k = 1, \dots, m$. Note that the i -th patient is in the k -th block if and only if $b(i) = k$. The blocks are illustrated in Figure 3, each composed of a set of C patients that are in treatment simultaneously.

For $i = 1$, we have $b(i) = 1$, $s_1 = 0$. Consider the following two cases:

Within blocks: Pick all patients i such that $b(i) = k$ for any $k \in \{1, \dots, m\}$, and let $j := (k - 1)C + 1$ be the index of the first patient within this block. By construction, $s_l = s_{l-1} + p^\downarrow$ for $l = j + 1, \dots, j + C - 1$, i.e., put-on tasks are performed consecutively and with no PCT idle time between them. Moreover, from Assumption 2, it follows that $d \geq (C - 1) \cdot p^\downarrow$; that is, when the PCT finishes the last put-on of all patients in the block, the j -th patient (the first in that block) is still in treatment. The PCT must then idle until the treatment of the j -th patient ends, which occurs at time f_j . Then, analogously as the put-on times, we have by construction that $f_l - f_{l-1} = p^\downarrow \geq p^\uparrow$ for $l = j + 1, \dots, j + C - 1$, where the last inequality follows from Assumption (1). That is, take-off tasks do not overlap, although the PCT may be idle for $p^\downarrow - p^\uparrow$ periods.

Finally, the PCT is continuously occupied by put-ons over $[s_j, r_{j+C-1}]$. From r_{j+C-1} until f_j , the PCT may be idle, but all C stations remain occupied. During the take-off phase, any gap between consecutive take-offs has length $p^\downarrow - p^\uparrow$; although a station is then available, this gap is shorter than the duration $p^\downarrow$ of a put-on because $p^\uparrow > 0$. Hence, no additional patient can feasibly begin

service before the final take-off ends at $f_{j+C-1} + p^\uparrow$. If another block remains, its first put-on begins immediately at that time.

Between blocks: Fix any block $k < m$. The completion time of the last patient in the k -th block matches with the start time of the first patient in the $(k+1)$ -th block; that is, $z_P(kC) - s_{kC+1}$ equals

$$\begin{aligned} & \left[kC \cdot p^\downarrow + b(kC) \cdot d + b(kC) \cdot p^\uparrow \right] - \left[kC \cdot p^\downarrow + [b(kC+1) - 1] \cdot d + [b(kC+1) - 1] \cdot p^\uparrow \right] \\ &= [b(kC) - b(kC+1) + 1] \cdot (d + p^\uparrow) = 0, \end{aligned}$$

where the last equality follows because $b(kC+1) - b(kC) = 1$. In other words, the blocks do not overlap in time and are consecutive. Thus, the schedule remains feasible.

Statement (b). Let $\tau(i) := (i-1) \bmod C$ for each patient $i \in \mathcal{P}$. The completion times (3) imply that the tasks for the i -th patient, $i \in \mathcal{P}$, have the following contiguous schedule:

- Start put-on time: $s_i := [b(i) + \tau(i) - 1] \cdot p^\downarrow + (b(i) - 1) \cdot d + [b(i) + \tau(i) - 1] \cdot p^\uparrow$;
- Start treatment time: $r_i := [b(i) + \tau(i)] \cdot p^\downarrow + (b(i) - 1) \cdot d + [b(i) + \tau(i) - 1] \cdot p^\uparrow$;
- Start take-off time: $f_i := [b(i) + \tau(i)] \cdot p^\downarrow + b(i) \cdot d + [b(i) + \tau(i) - 1] \cdot p^\uparrow$;

This schedule is valid if the interlaced conditions (a) and (b) are observed, and the schedule is feasible. Conceptually, $\tau(i) + 1$ captures the patient scheduling order within the block (see Figure 3). We next show that the interlaced and feasibility conditions are satisfied by this schedule.

Interlaced Conditions. For $1 \leq i \leq C$, we have $b(i) = 1$ and $\tau(i) = i - 1$. Thus,

$$s_i = [1 + i - 1 - 1] \cdot p^\downarrow + (1 - 1) \cdot d + [1 + i - 1 - 1] p^\uparrow = (i - 1) \cdot p^\downarrow + (i - 1) p^\uparrow.$$

This implies that, for any $l \in \{2, \dots, C\}$, $s_l - s_{l-1} = (l - 1) \cdot p^\downarrow + (l - 1) p^\uparrow - (l - 1 - 1) \cdot p^\downarrow - (l - 1 - 1) p^\uparrow = p^\downarrow + p^\uparrow$, and condition (a) is satisfied. For condition (b), first note that the expressions for the start put-on time and completion time for the i -th patient can be rearranged to

$$s_i = \tau(i) \cdot (p^\downarrow + p^\uparrow) + (b(i) - 1) \cdot (p^\downarrow + d + p^\uparrow), \quad (\text{EC.1})$$

$$z_I(i) = \tau(i) \cdot (p^\downarrow + p^\uparrow) + b(i) \cdot (p^\downarrow + d + p^\uparrow). \quad (\text{EC.2})$$

Consider now any $i > C$. Since $b(i) = b(i - C) + 1$ and $\tau(i) = \tau(i - C)$, which follow from the definitions of $b(\cdot)$ and $\tau(\cdot)$, respectively, the completion time of the $(i - C)$ -th patient matches the start put-on of the i -th patient, i.e.,

$$\begin{aligned} z_I(i - C) &= \tau(i - C) \cdot (p^\downarrow + p^\uparrow) + b(i - C) \cdot (p^\downarrow + d + p^\uparrow) \\ &= \tau(i) \cdot (p^\downarrow + p^\uparrow) + (b(i) - 1) \cdot (p^\downarrow + d + p^\uparrow) = s_i, \end{aligned} \quad (\text{EC.3})$$

in other words, we can assume that, in the schedule the i -th patient starts immediately after the $(i - C)$ -th patient ends and on the same station, satisfying condition (b).

Feasibility. It remains to show that the candidate interlaced schedule is feasible. Similar to the proof of Statement (a), assume that the number of patients n is a multiple of the number of stations C , i.e., $n/C = m$ for m integer; artificial patients can be added otherwise. Under this assumption, partition the patient indices $\mathcal{P}$ into the m blocks $\{(k - 1)C + 1, \dots, kC\}$, for $k = 1, \dots, m$.

Assign the first C patients to distinct stations. For every $i > C$, equation (EC.3) gives $s_i = z_I(i - C)$; thus, patient i can be assigned to the station used by patient $i - C$ and begins exactly when that station becomes available. Hence, the station-capacity constraint is satisfied.

It remains to verify the PCT-capacity constraint. For the first block, consecutive put-on start times differ by $p^\downarrow + p^\uparrow$, so the put-on tasks do not overlap. Moreover, the last put-on in the first block ends at r_C , while the first take-off begins at f_1 , and $f_1 - r_C = d - (C - 1)(p^\downarrow + p^\uparrow) \geq 0$ by Assumption 2. Therefore, the first take-off does not overlap any initial put-on.

Next, fix any block $k \in \{2, \dots, m\}$ and let $j := (k - 1)C + 1$ be its first patient. For any $l \in \{j, \dots, j + C - 2\}$, equations (EC.1) and (EC.3) imply $s_{l+1} - s_l = p^\downarrow + p^\uparrow$, $f_{l+1-C} = s_{l+1} - p^\uparrow = s_l + p^\downarrow$, and $f_{l+1-C} + p^\uparrow = s_{l+1}$. Thus, the put-on of patient l , the take-off of patient $l + 1 - C$, and the put-on of patient $l + 1$ occur consecutively and do not overlap.

After the last put-on in any block beginning with patient j , the first take-off of that block satisfies $f_j - r_{j+C-1} = d - (C - 1)(p^\downarrow + p^\uparrow) \geq 0$. If this is not the final block, its take-offs alternate with the put-ons of the next block as above. After the final block, consecutive take-off start times satisfy $f_{l+1} - f_l = p^\downarrow + p^\uparrow \geq p^\uparrow$, so the remaining take-offs also do not overlap. Therefore, the PCT-capacity constraint is satisfied for all patients, and the candidate interlaced schedule is feasible. ■

Statement (c). Let $s^*(i)$ and $z^*(i)$ be the put-on start time and completion time, respectively, of patient $i \in \mathcal{P}$ in an arbitrary optimal solution. We can relabel patients so that $s^*(1) \leq \dots \leq s^*(n)$. Because all patients occupy a station for the same duration $p^\downarrow + d + p^\uparrow$, their completion times follow the same order, and hence $z^* = z^*(n) = s^*(n) + p^\downarrow + d + p^\uparrow$. Moreover, we can assume that $s^*(1) = 0$; otherwise, all start times could be decreased by $s^*(1)$ while preserving feasibility.

First, suppose that $n \leq C$. Before patient n can start, the PCT must have completed the put-on tasks of patients $1, \dots, n - 1$. Consequently, $s^*(n) \geq (n - 1)p^\downarrow$, and therefore $z^* = s^*(n) + p^\downarrow + d + p^\uparrow \geq np^\downarrow + d + p^\uparrow = z_P(n)$. The push schedule attains this bound, proving the result for $n \leq C$.

Now suppose that $n > C$. We first establish a lower bound on the start times in the second block. For every $j \in \{1, \dots, C\}$ such that $C + j \leq n$, we claim that

$$s^*(C + j) \geq p^\downarrow + d + p^\uparrow + (j - 1)(p^\downarrow + p^\uparrow). \quad (\text{EC.4})$$

By the time patient $C + j$ starts, at least j earlier patients must have completed. Because completion times follow the order of the start times, these include patients $1, \dots, j$. Patient 1 completes at time $p^\downarrow + d + p^\uparrow$. After this time and before patient $C + j$ starts, the PCT must complete the take-off tasks of patients $2, \dots, j$, requiring $(j - 1)p^\uparrow$ time. These take-offs cannot start before time $p^\downarrow + d + p^\uparrow$: for each $l \geq 2$, feasibility of the put-on tasks gives $s^*(l) \geq p^\downarrow$, and hence

$$s^*(l) + p^\downarrow + d \geq 2p^\downarrow + d \geq p^\downarrow + d + p^\uparrow,$$

where the second inequality follows from Assumption 1. Moreover, before patient $C + j$ starts, the PCT must complete the put-on tasks of patients $C + 1, \dots, C + j - 1$, requiring another $(j - 1)p^\downarrow$ time. None of these put-ons can begin before time $p^\downarrow + d + p^\uparrow$, because no station is available before patient 1 completes. Since the PCT can perform at most one task at a time, at least $(j - 1)(p^\downarrow + p^\uparrow)$ time must elapse after time $p^\downarrow + d + p^\uparrow$. This shows (EC.4).

Next, patient $i > C$ can only begin the put-on operation when at least $i - C$ preceding patients have completed. Indeed, immediately before patient i begins service, at most $C - 1$ of the preceding $i - 1$ patients can still occupy stations. Since completion times follow the same order as start times, patient $i - C$ must have completed by time $s^*(i)$. Thus,

$$s^*(i) \geq s^*(i - C) + p^\downarrow + d + p^\uparrow, \quad \forall i > C. \quad (\text{EC.5})$$

For any $i > C$, let $j := \tau(i) + 1$ and $k := b(i)$, so that $i = (k - 1)C + j$. Repeated application of (EC.5), stopping at patient $C + j$, gives

$$\begin{aligned} s^*(i) &\geq s^*(C + j) + (k - 2)(p^\downarrow + d + p^\uparrow) \geq (k - 1)(p^\downarrow + d + p^\uparrow) + (j - 1)(p^\downarrow + p^\uparrow) \\ &= (b(i) - 1)(p^\downarrow + d + p^\uparrow) + \tau(i)(p^\downarrow + p^\uparrow). \end{aligned}$$

It follows that

$$\begin{aligned} z^*(i) &= s^*(i) + p^\downarrow + d + p^\uparrow \geq b(i)(p^\downarrow + d + p^\uparrow) + \tau(i)(p^\downarrow + p^\uparrow) \\ &= [b(i) + \tau(i)]p^\downarrow + b(i)d + [b(i) + \tau(i)]p^\uparrow = z_I(i). \end{aligned}$$

In particular, setting $i = n$ yields $z^* \geq z_I(n)$. Statement (b) shows that the interlaced schedule is feasible and attains completion time $z_I(n)$. Hence, $z^* = z_I(n)$ when $n > C$. ■

Proof of Proposition 1. Using the closed-form expressions from Theorem 1,

$$\begin{aligned}
 \Delta^{PI}(n) &= \underbrace{\left[n p^\downarrow + \left(\left\lfloor \frac{n-1}{C} \right\rfloor + 1 \right) (d + p^\uparrow) \right]}_{z_P(n)} \\
 &\quad - \underbrace{\left[\left(\left\lfloor \frac{n-1}{C} \right\rfloor + 1 \right) (p^\downarrow + d + p^\uparrow) + ((n-1) \bmod C) (p^\downarrow + p^\uparrow) \right]}_{z_I(n)} \\
 &= n p^\downarrow - (\lfloor (n-1)/C \rfloor + 1) p^\downarrow - ((n-1) \bmod C) (p^\downarrow + p^\uparrow)
 \end{aligned}$$

Note that $n p^\downarrow = p^\downarrow + (n-1) p^\downarrow$. Let $q := \lfloor (n-1)/C \rfloor$; then, $(n-1) \bmod C = (n-1) - qC$, and

$$\begin{aligned}
 \Delta^{PI}(n) &= n p^\downarrow - (q+1) p^\downarrow - ((n-1) \bmod C) (p^\downarrow + p^\uparrow) \\
 &= (n - q - 1 - ((n-1) \bmod C)) p^\downarrow - ((n-1) \bmod C) p^\uparrow \\
 &= q(C-1) p^\downarrow - ((n-1) \bmod C) p^\uparrow.
 \end{aligned}$$

Since $n \in \{kC+1, \dots, kC+C\}$, we have $k = \lfloor \frac{n-1}{C} \rfloor = q$. Moreover, observe that $(n-1) \bmod C = (n-1) - kC$. Substituting these identities into the expression above yields

$$\Delta^{PI}(n) = (C-1)k p^\downarrow - [(n-1) - kC] p^\uparrow. \quad \blacksquare$$

Proof of Proposition 2. We first construct a feasible push schedule and establish (b). Let $\tilde{\tau}(i) := (i-1) \bmod 2C$, and schedule patient i to begin the put-on operation at

$$s_T(i) := [\tilde{b}(i) - 1] (p^\downarrow + d + p^\uparrow) + \tilde{\tau}(i) p^\downarrow. \quad (\text{EC.6})$$

Adding the put-on, treatment, and take-off durations to (EC.6) gives

$$s_T(i) + p^\downarrow + d + p^\uparrow = \tilde{b}(i) (p^\downarrow + d + p^\uparrow) + \tilde{\tau}(i) p^\downarrow = [\tilde{b}(i) + \tilde{\tau}(i)] p^\downarrow + \tilde{b}(i) d + \tilde{b}(i) p^\uparrow = z_T(i),$$

which is expression (6). We next show that this schedule is feasible. Assign the first $2C$ patients to distinct stations and, for each $i > 2C$, assign patient i to the station used by patient $i - 2C$. Since $s_T(i + 2C) = s_T(i) + p^\downarrow + d + p^\uparrow = z_T(i)$, each patient begins exactly when the preceding patient assigned to the same station completes. Thus, the station-capacity constraint is satisfied.

For two consecutive patients within the same block, their put-on start times differ by $p^\downarrow$. At the transition between consecutive blocks, the difference between the first put-on start in the new

block and the last put-on start in the preceding block is $p^\downarrow + d + p^\uparrow - (2C - 1)p^\downarrow \geq p^\downarrow$, where the inequality follows from the assumption $d \geq (2C - 1)p^\downarrow - p^\uparrow$. Hence, consecutive put-on start times are always separated by at least $p^\downarrow$, and the put-on tasks do not overlap. The take-off of patient i starts at $s_T(i) + p^\downarrow + d$, so consecutive take-off start times have the same differences as consecutive put-on start times. These differences are at least $p^\downarrow \geq p^\uparrow$ by Assumption 1; hence, the take-off tasks also do not overlap. Because the put-on and take-off operations are performed by different technicians, this proves feasibility. Moreover, within each block every patient starts immediately after the preceding put-on, while the first patient in each subsequent block starts when its station becomes available. Therefore, this is a push schedule.

It remains to establish optimality, proving (a). Consider an arbitrary feasible schedule, and relabel patients so that $s^*(1) \leq \dots \leq s^*(n)$. The put-on specialist must complete the put-ons of patients $1, \dots, j - 1$ before patient j begins; hence, $s^*(j) \geq (j - 1)p^\downarrow$, for $j = 1, \dots, 2C$. Moreover, station capacity implies that, for every $i > 2C$, $s^*(i) \geq s^*(i - 2C) + p^\downarrow + d + p^\uparrow$. Indeed, immediately before patient i starts, at most $2C - 1$ of the preceding $i - 1$ patients can still occupy stations. Therefore, at least $i - 2C$ preceding patients must have completed. Since all patients have the same total service duration, completion times follow the same order as start times, so patient $i - 2C$ must have completed by time $s^*(i)$. For any $i \in \mathcal{P}$, let $j := \tilde{\tau}(i) + 1$. Applying these arguments repeatedly,

$$s^*(i) \geq s^*(j) + [\tilde{b}(i) - 1](p^\downarrow + d + p^\uparrow) \geq \tilde{\tau}(i)p^\downarrow + [\tilde{b}(i) - 1](p^\downarrow + d + p^\uparrow) = s_T(i).$$

Since all patients have the same total service duration, their completion times follow the same order as their start times. Consequently, every feasible schedule has makespan at least

$$s^*(n) + p^\downarrow + d + p^\uparrow \geq s_T(n) + p^\downarrow + d + p^\uparrow = z_T(n).$$

The feasible push schedule constructed above attains this bound and therefore optimizes (5). ■

Proof of Proposition 3. Let $a := \lfloor (n - 1)/2 \rfloor$, $q := \lfloor (n - 1)/(2C) \rfloor$, $r := a - Cq$, $e := \mathbf{1}\{n \bmod 2 = 0\}$ (i.e., $e = 1$ if and only if n is even), and $m := \lceil n/2 \rceil$. By Euclidean division, $r \in \{0, \dots, C - 1\}$. Moreover, $n = 2a + 1 + e$, so $m - 1 = a = Cq + r$ and $n - 1 = 2Cq + 2r + e$. Hence, $\tilde{b}(n) = b(m) = q + 1$, $(m - 1) \bmod C = r$, and $(n - 1) \bmod 2C = 2r + e$. Subtracting the completion-time expressions in Proposition 2 and Theorem 1 therefore gives

$$\begin{aligned} z_T(n) - z_I(m) &= (2r + e)p^\downarrow - r(p^\downarrow + p^\uparrow) \\ &= r(p^\downarrow - p^\uparrow) + ep^\downarrow \\ &= \left(\left\lfloor \frac{n-1}{2} \right\rfloor - C \left\lfloor \frac{n-1}{2C} \right\rfloor \right) (p^\downarrow - p^\uparrow) + \mathbf{1}\{n \bmod 2 = 0\} p^\downarrow \geq 0, \end{aligned}$$

where the inequality follows from $r \geq 0$ and Assumption 1. ■

Proof of Proposition 4. Let $q := \lfloor (n-1)/(2C) \rfloor$ and $r := (n-1) \bmod 2C$, so $n = 2Cq + r + 1$. Proposition 2 gives $z_T(n) = (q+r+1)p^\downarrow + (q+1)(d+p^\uparrow)$. If $0 \leq r \leq C-1$, then $\tilde{b}(n-C) = q$ and $(n-C-1) \bmod 2C = C+r$; otherwise, $\tilde{b}(n-C) = q+1$ and the remainder is $r-C$. Substitution into the definition of $\pi_T(n)$ yields

$$n\pi_T(n) = \begin{cases} (2q+2r+1)p^\downarrow + (2q+1)(d+p^\uparrow), & 0 \leq r \leq C-1, \\ (2q+2r-2C+2)p^\downarrow + 2(q+1)(d+p^\uparrow), & C \leq r \leq 2C-1. \end{cases}$$

For the independent configuration, let $i_1 := \lfloor n/2 \rfloor$ and $i_2 := \lceil n/2 \rceil$. If $r \geq 1$, write $i_1 = Cq + t_1$ and $i_2 = Cq + t_2$, where $t_1 = \lfloor (r+1)/2 \rfloor$, $t_2 = \lceil (r+1)/2 \rceil$, and $t_1 + t_2 = r+1$. Theorem 1 then gives $z_I(i_1) + z_I(i_2) = (2q+r+1)(p^\downarrow + p^\uparrow) + 2(q+1)d$. If $r = 0$, the patient allocations to the two independent pods are Cq and $Cq+1$; the total labor time is $(2q+C)(p^\downarrow + p^\uparrow) + (2q+1)d$. Thus,

$$n\pi_I(n) = \begin{cases} (2q+C)(p^\downarrow + p^\uparrow) + (2q+1)d, & r = 0, \\ (2q+r+1)(p^\downarrow + p^\uparrow) + 2(q+1)d, & 1 \leq r \leq 2C-1. \end{cases}$$

Subtracting these expressions gives (7). Its first and third cases are directly nonnegative. In the second case, $1 \leq r \leq C-1$ and Assumption 2 gives $d - r(p^\downarrow - p^\uparrow) \geq d - (C-1)(p^\downarrow - p^\uparrow) \geq 0$. Hence, $\pi_T(n) \leq \pi_I(n)$ in every case. ■

Appendix EC.2: Additional Considerations

Other important relational nuances and operational features that are difficult to capture in prescriptive models are important to consider for a practical clinic scheduling solution. These include:

Patient Preferences and Behavior. The facility administrator may not have the full ability to dictate schedules to patients. Instead, patients may prefer certain treatment times and consider going to a different clinic if a preferred time is not available. Frequently, the earliest treatment times from 5:00-6:00 AM are highly coveted, as those treatment slots allow patients to have more of a full day after treatment. Patients may also have other medical appointments to schedule around or face restrictions from specialized medical transport services that they require to get to the clinic. Furthermore, patients may be averse to changes from their existing schedule.

Infection Risk and Station Restrictions. When a patient vulnerable to Hepatitis B infection occupies a station, for instance, certain patients are restricted from being in close proximity to the infected patient to help mitigate the risk of infection.

Hybrid Tandem Configurations. Although tandem configurations offer significant labor efficiency advantages, full task specialization across a larger pod may increase walking, weaken technicians' sense of ownership over a particular set of four stations, and disrupt continuity and rapport between PCTs and patients. As a compromise, clinics could consider hybrid strategies that combine features of the interlaced and tandem systems. For example, two technicians' shifts could be staggered so each technician completes a small number of put-on or take-off tasks in the other pod. Outside of the beginning and end of the shift, each technician would be dedicated to their own pod. The schedule would be designed so that no put-ons could happen simultaneously across the two pods early in the day, and no take-offs could happen simultaneously later in the day.

Appendix EC.3: Clinic Configuration Details

We detail the rules for a clinic with S pods, each with C stations. For simplicity, we restrict our attention to clinics with 24 or fewer stations — i.e., $S \cdot C \leq 24$ — so that at most two RNs are needed. The clinic serves n^{TOT} patients with pods $1, \dots, (n^{TOT} \bmod S)$ serving $\lfloor n^{TOT}/S \rfloor + 1$ patients and pods $(n^{TOT} \bmod S) + 1, \dots, S$ serving $\lfloor n^{TOT}/S \rfloor$ patients.

We first examine a push or interlaced single-pod strategy $\phi \in \{P, I\}$. Each technician $i = 1, \dots, S$ has a shift with duration σ_i^{PCT} , in terms of paid labor hours. Generalizing the description above, technicians assigned to an odd-numbered pod come in early to start preparatory work. Each technician i includes additional preparatory and closing work equal to A_i^{PCT} periods when i is even and $\tilde{A}_i^{PCT}$ periods when i is odd. Technicians also take an unpaid break of B periods. For single-pod policies, each technician's shift ends when the last take-off completes in the PCT's pod. The shift duration — expressed in the number of periods — for a PCT in a single pod is then

$$\begin{aligned} \sigma_i^{PCT} = & A^{PCT} \cdot \mathbf{1}\{i \bmod 2 = 0\} + \tilde{A}^{PCT} \cdot \mathbf{1}\{i \bmod 2 = 1\} - B \\ & + z_\phi \left(\lfloor n^{TOT}/S \rfloor + \mathbf{1}\{i \leq n^{TOT} \bmod S\} \right). \end{aligned}$$

If $S \cdot C \leq 12$, only one RN is needed. That RN's shift extends a combined total of A^{RN} periods before the first put-on and after the last take-off. It also includes an unpaid break B .

$$\sigma^{RN} = A^{RN} - B + z_\phi \left(\lceil n^{TOT}/S \rceil \right)$$

If $12 < S \cdot C \leq 24$, the clinic requires 2 RNs, whose shifts occur in a staggered manner. Using shift extensions of A_1^{RN} and A_2^{RN} for the two nurses, we have shift lengths:

$$\begin{aligned} \sigma_1^{RN} = & A_1^{RN} + z_\phi \left(\lceil n^{TOT}/S \rceil - \lfloor 12/S \rfloor \right) - B, \\ \sigma_2^{RN} = & z_\phi \left(\lceil n^{TOT}/S \rceil \right) - \left(z_\phi (\lfloor 12/S \rfloor + 1) - p^\uparrow - d - p^\downarrow \right) + A_2^{RN} - B. \end{aligned}$$

If a tandem strategy is used, two pods with $2C$ stations are combined. We assume that there is an even number of pods — i.e., $S/2$ pods with $S \bmod 2 = 0$ — and index them $j = 1, \dots, S/2$. Combined pods $1, \dots, n^{TOT} \bmod (S/2)$ serve $\lfloor n^{TOT}/(S/2) \rfloor + 1$ patients, and pods $(n^{TOT} \bmod (S/2)) + 1, \dots, S/2$ serve $\lfloor n^{TOT}/(S/2) \rfloor$ patients. Each combined pod has two technicians: technician 1 performs additional preparatory work equal to A_1^{PCT} periods, and technician 2 performs additional closing work equal to A_2^{PCT} periods. PCT 1 leaves when a patient's take-off completes so that only four patients are left in the pod. PCT 2 arrives when the fifth patient's put-on begins. The following shift lengths result for each pod j :

$$\begin{aligned}\sigma_{j,1}^{PCT} &= A_1^{PCT} + z_T \left(\lfloor n^{TOT}/(S/2) \rfloor + \mathbf{1}\{j \leq n^{TOT} \bmod (S/2)\} - 4 \right) - B, \\ \sigma_{j,2}^{PCT} &= z_T \left(\lfloor n^{TOT}/(S/2) \rfloor + \mathbf{1}\{j \leq n^{TOT} \bmod (S/2)\} \right) - \left(z_T(5) - p^\uparrow - d - p^\downarrow \right) \\ &\quad + A_2^{PCT} - B.\end{aligned}$$

If $S \cdot C \leq 12$, only one RN is needed, and the RN's shift has duration

$$\sigma^{RN} = A^{RN} - B + z_T \left(n^{TOT} \right)$$

If $12 < S \cdot C \leq 24$, the clinic requires 2 RNs, whose shifts occur in a staggered manner to maximize labor efficiency while meeting staffing ratios mandated by state or corporate policy. Using shift extensions of A_1^{RN} and A_2^{RN} for the two nurses, we have shift lengths:

$$\begin{aligned}\sigma_1^{RN} &= A_1^{RN} + z_T \left(\lceil n^{TOT}/(S/2) \rceil - \lfloor 12/S \rfloor \right) - B, \\ \sigma_2^{RN} &= z_T \left(\lceil n^{TOT}/(S/2) \rceil \right) - \left(z_T(\lfloor 12/(S/2) \rfloor + 1) - p^\uparrow - d - p^\downarrow \right) + A_2^{RN} - B.\end{aligned}$$

We consider one additional scheme in which a clinic with 12 stations operates one combined pod using the tandem strategy and one solo pod using the interlaced strategy. This configuration takes advantage of the required 12:1 RN-to-patient ratio. In this case, we allocate x^* patients to the combined tandem pod and $(n^{TOT} - x^*)$ patients to the solo interlaced pod such that

$$x^* := \arg \min_{x \in \{0, \dots, n^{TOT}\}} \max \{z_T(x), z_I(n^{TOT} - x)\}$$

In this case, the PCT shift durations are as defined above, and the RN's shift duration is

$$\sigma_1^{RN} = A^{RN} - B + \max \{z_T(x^*), z_I(n^{TOT} - x^*)\}$$

The overall labor efficiency is then the sum of all labor performed by PCTs and RNs divided by the total number of patients. For the push and interlaced schemes, this value is

$$\pi^\phi(n^{TOT}) := \frac{\sum_{i=1}^S \sigma_i^{PCT} + \sigma_1^{RN} + \mathbf{1}\{S \cdot C \geq 13\} \sigma_2^{RN}}{n^{TOT}},$$

and the calculation proceeds analogously for the tandem system.

For the numerical results in Section 5.2, two technicians begin their shifts at 3:45 AM to complete clinic-level preparatory work before the first treatment begins at 5:00 AM, and any remaining technicians arrive at 4:30 AM. One RN's shift begins with the first put-on operation, and the second RN begins once there are at least 13 patients in the clinic. Technicians complete their shifts at the end of the last take-off operation according to the completion equations of Section 4; note that some pods may have one more patient than others, depending on whether the number of patients can be evenly divided by the number of pods. One RN departs when there are always 12 or fewer patients left in the clinic, and the other RN departs 20 minutes after the last take-off operation is completed. Thirty minutes of break time is subtracted from each worker's total number of hours.

Appendix EC.4: Additional Numerical Details

We next provide additional details, insights, and sensitivity analysis associated with our testing.

EC.4.1. Formulation and Computational Details

All formulations tested in §6 are based on the following time-step mathematical program that formulates the single-pod scheduling model (1), where we denote $(x)^+ := \max\{0, x\}$:

$$\min_{x, z} \quad z \tag{EC.7a}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{P}} \sum_{j=(t-p^\downarrow-d_i-p^\uparrow+1)^+}^t x_{ij} \leq C, \quad \forall t \in \mathcal{T}, \tag{EC.7b}$$

$$\sum_{i \in \mathcal{P}} \left(\sum_{j=(t-p^\downarrow+1)^+}^t x_{ij} + \sum_{j=(t-p^\downarrow-d_i-p^\uparrow+1)^+}^{t-p^\downarrow-d_i} x_{ij} \right) \leq w_t, \quad \forall t \in \mathcal{T}, \tag{EC.7c}$$

$$\sum_{j=0}^{T-p^\downarrow-d_i-p^\uparrow} x_{ij} = 1, \quad \forall i \in \mathcal{P}, \tag{EC.7d}$$

$$(t+1) \cdot w_t \leq z, \quad \forall t \in \mathcal{T}, \tag{EC.7e}$$

$$x_{it}, w_t \in \{0, 1\}, \quad \forall i \in \mathcal{P}, \forall t \in \mathcal{T}. \tag{EC.7f}$$

Specifically, variables $x_{it} \in \{0, 1\}$ capture if patient $i \in \mathcal{P}$ starts at period $t \in \mathcal{T}$, and binary variables $w_t \in \{0, 1\}$ represent if the PCT is occupied at period $t \in \mathcal{T}$; domains are adjusted accordingly so

that the shift end time does not exceed T . Inequalities (EC.7b) formulate (1b). Inequalities (EC.7c) define variables w_t and also enforce (1c), since $w_t \leq 1$. Finally, constraints (EC.7d) ensure that every patient is assigned a start time, (EC.7e) evaluate the makespan (adjusted by one because w_t captures the periods for which the PCT is occupied), and (EC.7f) set the variable domains.

The distinct systems are implemented using the following constraints, which adapt the definitions in §4.1 to heterogeneous durations:

- For the optimal push flow system, we impose that all take-offs of a block start before the first put-on of the next block (see, e.g., Figure 3). To this end, we introduce variables s_i and π_i that indicate the start time and position of patient $i \in \mathcal{P}$ in the final schedule, respectively, which satisfy the following constraints:

$$s_i = \sum_{t \in \mathcal{T}} t \cdot x_{it}, \quad \forall i \in \mathcal{P}, \quad (\text{EC.8})$$

$$\pi_i < \pi_j \Leftrightarrow s_i \leq s_j, \quad \forall i, j \in \mathcal{P}, i \neq j, \quad (\text{EC.9})$$

$$\pi_i \in \{1, \dots, n\}, \quad \forall i \in \mathcal{P}. \quad (\text{EC.10})$$

Constraint (EC.8) links variables s_i with x_{it} ; constraint (EC.9) enforces that positioning variables must be consistent with start times, and constraint (EC.10) defines the domain of integer variables. Logical conditions as above can be either linearized using standard big-M inequalities or added using specific solver constructs. Finally, to add the desired block constraint for the push schedule, we add the following logical conditions:

$$\pi_i = k, \pi_j = L \Rightarrow s_i + p^\downarrow + d_i + p^\uparrow \leq s_j, \\ \forall L = \{C + 1, 2C + 1, \dots\}, k \in \{L - C, L - C + 1, \dots, L - 1\}, i, j \in \mathcal{P}, i \neq j. \quad (\text{EC.11})$$

Specifically, we recall that each block consists of C patients. The index L captures the first patient of some block b , and the index k captures a patient in the previous block $b - 1$. Thus, (EC.11) enforces that the k -th take-off ends before the put-on of the L -th patient.

- For the Optimal Interlaced flow system, to ensure the strategic pause between put-ons to interlace them with take-offs, we impose that there is not a put-on within $p^\downarrow + p^\uparrow$ of another put-on throughout the horizon:

$$\sum_{j=t}^{t+p^\downarrow+p^\uparrow-1} \sum_{i \in \mathcal{P}} x_{ij} \leq 1, \quad \forall t \in \{t', \dots, T - p^\downarrow - p^\uparrow\},$$

where t' is the minimum time where interlaced constraints should be imposed; for consistency across testing, we fix it to 7:00 AM.

- The Optimal Tandem is implemented analogously, but with extended variables $x_{itp} \in \{0, 1\}$ representing if patient i starts at time t and is allocated to pod p .
- Finally, the break conditions can be incorporated as artificial tasks with the desired durations, and shift start times are imposed by limiting the index of variables x_{it} accordingly.

EC.4.2. Tractable Sequencing Approaches

The single-pod analysis in §6.1 optimally sequences patients under both the interlaced and push flow systems by solving the mathematical programs described in Appendix EC.4.1. Because treatment durations are predetermined by treatment type and are therefore deterministic, we also evaluate the effect of simpler a priori sequencing strategies on makespan. We consider three rules:

- *Shortest-first*: Patients with shorter treatment durations are ranked first.
- *Longest-first*: Patients with longer treatment durations are ranked first.
- *Cycle*: Inspection of the optimal schedules suggested a simple duration-balancing strategy. Rather than grouping all short or all long treatments together, the schedules typically cycled through short, medium, long, and longest treatments. Because the first and last patients in a cycle had shorter and longer treatment durations, respectively, the take-off of the first patient was less likely to conflict with the put-on of the last patient. This left room to interweave subsequent put-ons, thereby reducing station idle time.

Inspired by this intuition, we implement the following cycle strategy, with partition sizes chosen empirically. We first sort patients from shortest to longest treatment duration and label their ranks $1, \dots, n$. We begin the sequence with ranks $1, \lfloor n/3 \rfloor + 1, \lfloor 2n/3 \rfloor + 1$, and n . We then include the unused ranks from $\lfloor n/3 \rfloor$ through $\lfloor 2n/3 \rfloor$ in ascending order, followed by the remaining ranks above this interval in descending order and, finally, those below it in descending order. Ties are broken arbitrarily. For $n = 12$, this procedure yields the sequence 1, 5, 9, 12, 4, 6, 7, 8, 11, 10, 3, 2.

Figure EC.1 depicts the distributions of makespan and optimality gap for the same instances considered in §6.1, with $n = 12$ and $C = 4$. The optimality gap is measured relative to the optimal makespan for each instance. Among the three simple sequencing rules, Cycle performs best under interlacing, with a mean gap of 2.49% (median: 1.84%; standard deviation: 2.36%), compared with 7.54% for Shortest-first (median: 7.56%; standard deviation: 1.90%) and 8.93% for Longest-first (median: 8.48%; standard deviation: 2.27%). The mean shift end time for interlaced with Cycle is approximately 7:05 PM, only 20 minutes later than that of the optimal interlaced schedule and approximately 2 hours and 49 minutes earlier than that of the push schedule using the same Cycle

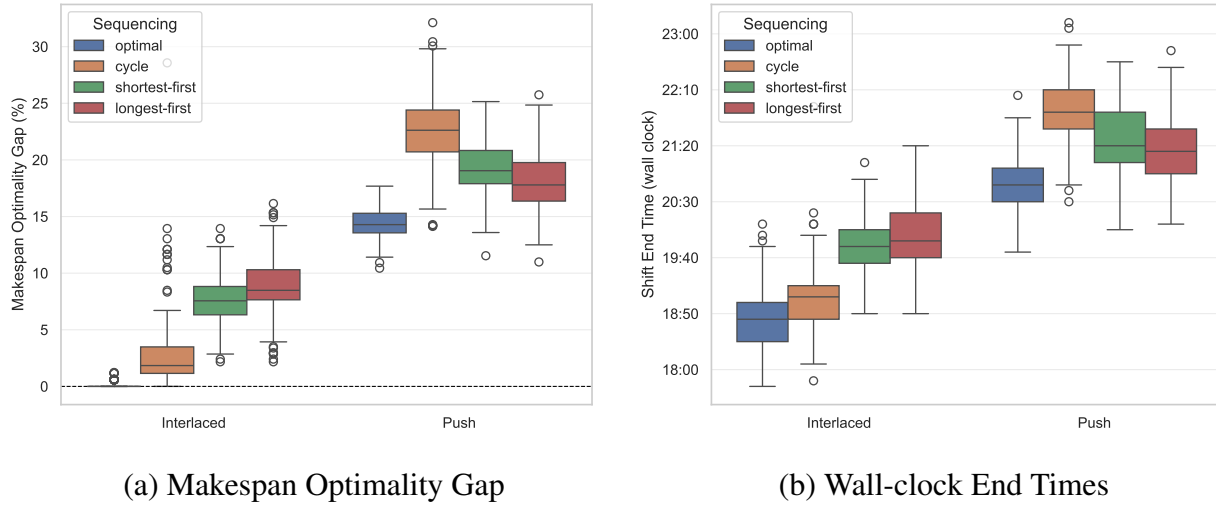


Figure EC.1 (Colored.) Makespan performance of push and interlaced flow systems under optimal, cycle, shortest-first, and longest-first patient sequencing. Each box summarizes 500 instances with 12 patients and four stations.

sequence. These results suggest that, under the tested parameters, clinics can preserve most of the benefits of interlacing relative to a push approach even when leveraging simple sequencing heuristics.

We also find that the sequencing rules are less effective under a push strategy. Our analysis suggests that this occurs because its block structure makes the longest treatment in each block a bottleneck: the next block cannot begin until every patient in the preceding block has completed treatment and take-off. Because Cycle deliberately distributes long treatments throughout the sequence, it can place a patient with a long treatment duration in several blocks, thereby delaying multiple block transitions. Shortest-first similarly places the longest treatments near the end of the day, extending the final completion time. Longest-first performs somewhat better by concentrating longer treatments in earlier blocks. By contrast, interlacing can use dispersed completion times more effectively because each completed take-off creates an opportunity to put on another patient without waiting for the entire block to clear.

EC.4.3. Details of Interlaced Performance

The close relationship between the interlaced and optimal makespan evaluated in §6.1 was due to a frequent alternating pattern of put-on and take-off tasks found in optimal schedules. Figure EC.2a shows the average percentage of instances in which the PCT alternated put-on and take-off tasks. In all tested cases, at least 86% of task transitions followed an interlaced pattern, with over 96% for $n = 12$. The instances that deviated from this pattern, i.e., featuring consecutive put-on or take-off

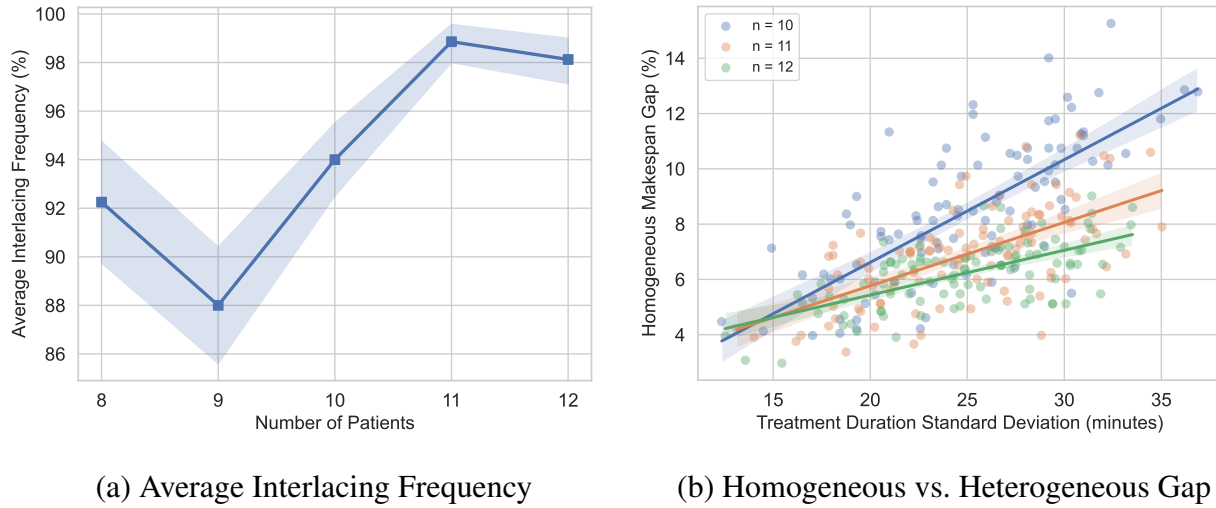


Figure EC.2 (Colored.) Interlacing frequency in optimal schedules, shown for each value of n (Figure EC.2a), and impact of treatment duration heterogeneity on the makespan gap between the optimal solution and the homogeneous-duration benchmark, where all patient durations are set to the instance average (Figure EC.2b).

operations, typically occurred due to relatively short treatment durations, which triggered back-to-back tasks of the same type. In those cases, the maximum time saved was 20 minutes, with a mean of 6.33 minutes and a standard deviation of ≈ 4 minutes.

EC.4.4. Effect of Treatment Duration Variance

We observed that the makespan under heterogeneous treatment durations was significantly smaller than that of schedules assuming constant treatment times. To illustrate this effect, we compute the percentage gap between the optimal makespan and the optimal value obtained by fixing all patient durations to the instance's average. On average, the homogeneous-duration makespan was 7.79% larger than that of the heterogeneous-duration schedule, with a standard deviation of 2.83%. The minimum and maximum observed gaps were 0.63% and 20.0%, respectively. This difference was generally positively correlated with the variance in durations (Pearson's r of 0.37, 0.74, 0.73, 0.63, and 0.63 for $n = 8, 9, 10, 11$, and 12, respectively).

Figure EC.2b illustrates this relationship for $n \in \{10, 11, 12\}$, where each point represents the treatment duration standard deviation by the relative gap, as described above. The line is the best-fit regressor across all instances with the same n . An analysis of optimal schedules indicated that longer treatment durations are more likely to be scheduled with the first group of four patients. Shorter patient durations tended to be assigned to the last group of four patients when there are six or more 3.5-hour patients, and to the middle group of four patients when there are fewer than

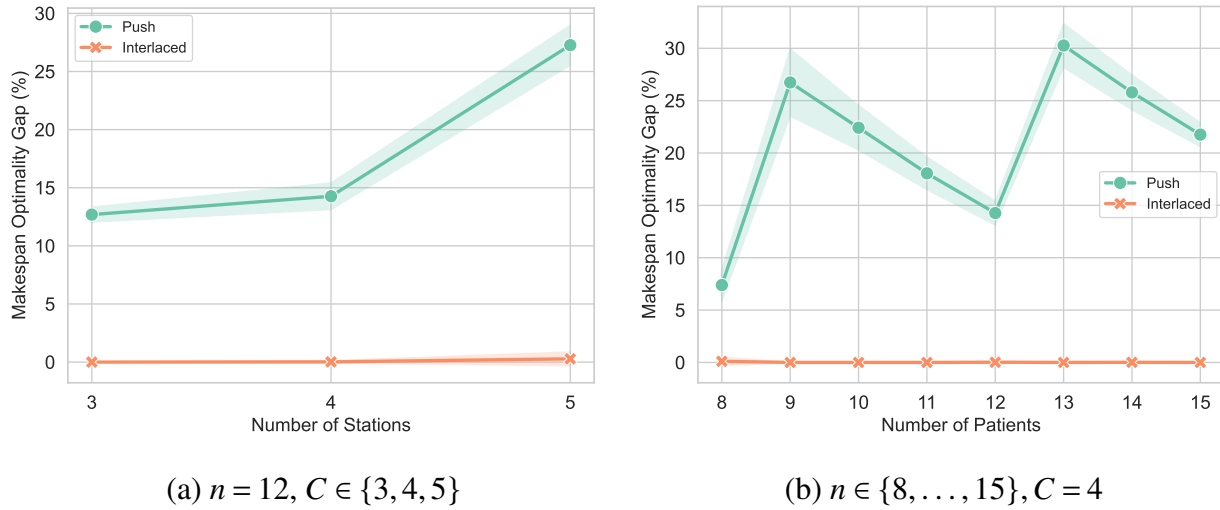


Figure EC.3 (Colored.) Sensitivity of Makespan Optimality Gap for varying numbers of patients n and stations C . “Push” and “Interlaced” correspond to the optimal push and interlaced systems, respectively.

six 3.5-hour patients. This allows the clinic to reduce the technician idle time that is prominent at the beginning and end of the day for the interlaced flow system in Figure 3. Thus, this potential disadvantage of interlacing is largely mitigated when patient durations are heterogeneous.

This finding contrasts with the classic result in queueing theory that increased service-time variability degrades performance (Kingman 1961). In our setting, heterogeneity in treatment durations improves performance by enabling more effective sequencing — placing longer treatments earlier in the day and shorter ones later — to reduce idle time at the boundaries of the interlaced flow. This aligns with prior operations research demonstrating that variability can be beneficial when it is correlated with other operational levers, such as priority rules, skill-based routing, or strategic job sequencing, which can be exploited to improve system efficiency; see, e.g., Armony et al. (2015).

EC.4.5. Additional Sensitivity Analysis on n and C

Figure EC.3 reports the makespan optimality gap of the push and interlaced flow systems relative to the optimal schedule, calculated as in Figure 8. Figure EC.3a fixes the number of patients at $n = 12$ and varies the number of stations $C \in \{3, 4, 5\}$, while Figure EC.3b fixes the number of stations at $C = 4$ and varies the number of patients $n \in \{8, \dots, 15\}$.

The interlaced policy remains remarkably close to optimal across all tested settings, whereas the push policy exhibits substantially larger deviations as a function of both C and n . More precisely, the push gap increases from 12.69% at $C = 3$ to 14.27% at $C = 4$, and then rises sharply to 27.26% at $C = 5$. By contrast, the interlaced gap is essentially zero throughout, taking values

0.00%, 0.02%, and 0.28%, respectively. Figure EC.3b shows a similar pattern when varying the number of patients: the average optimality gap of push is 20.83%, while the corresponding average for interlaced is only 0.02%. The push scheme ranges from 7.39% at $n = 8$ to a peak of 30.26% at $n = 13$, and displays a clear sawtooth pattern as n varies; this behavior is consistent with the homogeneous setting formalized in Proposition 1 and depicted in Figure 4. In contrast, the interlaced policy is exactly optimal for five of the eight patient counts considered and never exceeds 0.11%. These results reinforce that the interlaced policy consistently tracks the optimal makespan across problem sizes, whereas push can become substantially suboptimal, particularly in regimes where the structural effects described in Proposition 1 are most pronounced.