Online Supplement for "An Integrated Approach to Improving Itinerary Completion in Coordinated Care Networks"

This document serves as the online supplement for the main paper "An Integrated Approach to Improving Itinerary Completion in Coordinated Care Networks." Section A details the transition dynamics in the network. Section B provides additional details for the case study including the simulation platform and alternate template designs and scheduling rules. Section C specifies the characterization for the itinerary completion time (ICT) in the general network setting. In the interest of space, detailed proofs for all theorems and lemmas in the main paper are relegated to the technical companion (Anonymous 2024).

A. Additional Details for Patient Flow Models

A.1. Setting without Parallel Appointment

Denote the total number of target patients blocked at the end of day $t+1$ as $M_u^B(t+1)$. Recall that $N_u(t+1)$ is the total number of requests and $B_u(t+1)$ is the blocking probability. Conditioning on $N_u(t+1)=n$, $\Lambda_u^e(d(t+1))=\lambda_e$, and $\big(n-C_{u,d(t+1)}\big)^+=b$, we have that $M_u^B(t+1)$ follows the hypergeometric distribution

$$\mathbb{P}\big(M_u^B(t+1)=m|b,\lambda_e\big)=\frac{c(n-\lambda_e,m)\cdot c(\lambda_e,b-m)}{c(n,b)}, \tag{19}$$

where $c(a,b)$ is the binomial coefficient. Note that M_u^B does not follow a binomial distribution because, in total, there are exactly b patients blocked. Thus, we are sampling b patients *without* replacement from the joint pool of target patients and exogenous patients. In contrast, the binomial distribution assumes each patient has an independent Bernoulli trial, i.e., sampling *with* replacement. In other words, the blocking events are correlated among patients. Since b is random depending on $N_u(t)$ via (3), the number of blocked target patients is essentially a *doubly-stochastic* random variable.

We assume there are K types of patients and we use the superscript $k=1,\ldots,K$ to denote the patient count in type k. Once we have $M_u^B(t+1)$, under the random ordering assumption, the joint distribution of $\{M_{u,s}^{B,k}(t+1),\forall k,s\}$ (across all stages and patient types) follows the multinomial distribution with parameters $M_u^B(t+1)$ and probabilities $\left\{\frac{N_{u,s}^k(t+1)}{\sum_{k,s}N_{u,s}^k(t+1)},\forall k,s\right\}$, and $M_{u,s}^{NB,k}(t+1)=N_{u,s}^k(t+1)-M_{u,s}^{B,k}(t+1)$ for each u,k,s.

A.2. Setting with Multiple Classes and Parallel Appointments

Similar as the above section, we consider K types of patients. To specify the patient flow model with potentially parallel appointments in each stage, we denote $\mathcal{R}_{k,s}=\{u_1,u_2,\ldots,u_n\}$ as the set of resources

(stations to visit) that are required to complete stage s for a type k patient. We further define the following vector tracking the appointment completion status for resource group $\mathcal{R}_{k,s}$ as

$$\mathcal{B}_{\mathcal{R}_{k,s}} = (a_{u_1}, a_{u_2}, \ldots, a_{u_n}),$$

where for each $u_j \in \mathcal{R}_{k,s}$, $a_{u_j} = 0$ indicates that the appointment at resource u_j has not yet been completed and $a_{u_j} = 1$ indicates that it has been completed. Given $N_{k,s}$ total possible resources that can be used by a type k patient at stage s, each resource group is one possible combination of the $N_{k,s}$ resources. That is, $\mathcal{R}_{k,s} \in \mathcal{P}(\{u_1, \ldots, u_{N_{k,s}}\})$, where $\mathcal{P}$ is the power set.

For illustration purposes, we explain the patient flow model by considering the case where each stage contains only *two* stations for the patient to visit; the model framework can be generalized easily. In this scenario, the blocking status vector $\mathcal{B}_{\mathcal{R}_{k,s}}$ can take four possible values: $(0,0)$, $(1,0)$, $(0,1)$, $(1,1)$, where $(1,1)$ represents that the patient finished all appointments required in the current stage and is ready to move to the next stage on her care path.

Now, we define the following patient count that differentiates not only by k, s but also by the blocking status. That is, we denote $M^{k,\mathcal{B}}_{\mathcal{R},s}(t)$ that counts the total number of type k patients whose blocking status is $\mathcal{B} = (a_1, a_2)$ in stage s, at the end of day t. We drop the index k, s from $\mathcal{B}$ and $\mathcal{R}$ for notational simplicity. The total number of target patients requesting an appointment from station u on day $t+1$ follows:

$$N^k_{u,s}(t+1) = \sum_{\mathcal{R}: u \in \mathcal{R}} \sum_{\mathcal{B}: a_u = 0} M^{k,\mathcal{B}}_{\mathcal{R},s}(t) + \tilde{M}^k_{\mathcal{R},s}(t) + \Lambda_u(t+1).$$

Here, the first double-summation represents all the patients who are in stage s and need to visit station u, yet have not finished their appointments at u. The second term represents the patients who have finished all appointments in stage $s-1$ by the end of day t and now need to visit station u in stage s. That is,

$$\tilde{M}^k_{\mathcal{R},s}(t) = \sum_{\mathcal{R}_{s-1}} Mult\big(M^{k,(1,1)}_{\mathcal{R},s-1}(t), p_{\mathcal{R}_{s-1},\mathcal{R}_s}\big),$$

where each term in the summation denotes the number of patients, out of those $M^{k,(1,1)}_{\mathcal{R},s-1}(t)$ who completed all appointments in stage $s-1$, that request appointments from resource group $\mathcal{R}_s$ in stage s with probability $p_{\mathcal{R}_{s-1},\mathcal{R}_s}$, where this patient count follows a multinomial distribution.

Transitions in the stochastic system. Once we get $N^k_{u,s}(t+1)$, we can then define $N_u(t+1)$ and calculate $B_u(t+1)$ in (3). To obtain $M^{k,\mathcal{B}}_{\mathcal{R},s}(t+1)$, we first calculate the following intermediate variables: $N^{b,k,\mathcal{B}}_{\mathcal{R},s}(t+1;u)$, which counts the number of patients, out of all patients in the same category determined by $(k, s, \mathcal{B}, \mathcal{R})$, that are blocked at station u on day $t+1$. We then $M^B_u(t+1)$ from the hypergeometric distribution that samples blocked patients *without* replacement from the joint

pool of the target patients and exogenous patients. Then, the joint distribution of $N_{\mathcal{R},s}^{b,k,\mathcal{B}}(t+1;u)$'s follow a multinomial distribution with parameters $M_u^B(t+1)$ and the proportions of patients from the corresponding category of $(k,s,\mathcal{B},\mathcal{R})$. With these intermediate variables, we can characterize the transitions in the patient counts to capture the new blocking status.

Mean-field model. Once we have the stochastic patient flow model, the corresponding mean-field model can be written by taking the expectation of the random quantities, similar to what we show in Section 3.2.1. In particular, in the mean-field model, the transitions from $m_{\mathcal{R},s}^{k,\mathcal{B}}(t)$'s to $m_{\mathcal{R},s}^{k,\mathcal{B}}(t+1)$'s (the deterministic counterparts for $M_{\mathcal{R},s}^{k,\mathcal{B}}(\cdot)$) are much simplified. For example, for a patient in a starting status $(0,0)$, the transition probabilities into status $(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$ are simply $\beta_{u_1}(t+1)\beta_{u_2}(t+1)$, $\big(1-\beta_{u_1}(t+1)\big)\beta_{u_2}(t+1)$, $\beta_{u_1}(t+1)\big(1-\beta_{u_2}(t+1)\big)$, and $\big(1-\beta_{u_1}(t+1)\big)\big(1-\beta_{u_2}(t+1)\big)$, respectively, where $\beta_u(t+1)$ is the blocking probability on day $t+1$ in the mean-field model. The stability condition (9) ensures that we have an equilibrium solution $\beta=\{\beta_{u,d}\}$, which can be solved numerically from the mean-field model.

B. Additional Results for the Case Study

B.1. Algorithm Setup and Constraints

In the case study, we impose the following constraints according to our healthcare partner:

$$\sum_{d=1}^{D}\Theta_{k,d}\geq\theta_k,\quad\forall k\in\mathfrak{K}, \tag{20}$$

$$\sum_{k\in\mathfrak{K}}\Theta_{k,d}\leq C_{1,d},\quad\forall d=1,\ldots,D, \tag{21}$$

$$\mathbb{P}_\infty(N_{u,d}\geq C_{u,d}\mid\Theta)\leq\gamma_{u,d},\quad\forall u\in\mathfrak{U},d=1,\ldots,D. \tag{22}$$

Constraints (20) are throughput constraints where θ_k denotes the weekly throughput requirement for type k patients. Constraints (21) ensure that the root appointment allocation will not exceed the capacity of station 1 (BDC). Constraints (22) are service-level constraints ensuring that blocking at any given service will not be too high and negatively impact both target and exogenous patients. We estimate the target service level, $\gamma_{u,d}$, from historical data.

For the iterative algorithm proposed in Section 4, we obtain the initial blocking probabilities from a pre-processing step that employs a workload smoothing optimization. We set the tolerance term $\epsilon=0.03$ in (17) to solve the optimization efficiently.

B.2. Simulation Platform

We describe the flow of events in our discrete-even simulation. At the beginning of each day, we first determine the number of new patients. Patients getting root appointments arrive according to the template $\Theta=\{\Theta_{k,d}\}$ in the deterministic arrival setting. Table 5 shows the historical template. To account for non-integer Θ, we generate arrivals according to a random variable $\lfloor\Theta_{k,d}\rfloor+\text{Bern}(\Theta_{k,d}-$

$\lfloor\Theta_{k,d}\rfloor$), where Bern($p$) is the Bernoulli random variable with parameter p, s.t. $(1+p)\lfloor\Theta_{k,d}\rfloor = \Theta_{k,d}$. For each new patient of type k starting on day d, we generate a priori a realization of her care path by considering $S \cdot U$ independent Bernoulli trials, one for each station in each stage, where each trial is given by Bern($p^k_{u,s}$). We generate exogenous arrivals from the truncated Gaussian distribution. Each station maintains a pool of appointments, which records the unique ID (generated at admission time) of each patient that currently has an appointment at station u to be finished. We assume a constant service time for each appointment. Jointly estimated with the capacity of each station, we set the service time to be 9.4, 3.2, 4.1, 6.7, and 8.0 minutes for stations 1 through 5, respectively.

Random Ordering. To implement in the simulation the random ordering rule (same blocking probability for each patient as in (3)), we randomly sample from the list of all patients (including target and exogenous) requesting an appointment and admit patients up to the capacity $C_{u,d}$. The remaining target patients are blocked, staying in the pool of appointments for the next day; exogenous patients are lost. If a patient is not blocked, they join the pool of appointments for the next station(s) of their pre-generated care path on the following day.

FCFS Discipline. Unlike random ordering, the FCFS (first-come-first-served) discipline assumes that each service station fulfills the appointments according to the time they were reserved each day. We make two assumptions in the simulation: 1) a target patient reserves all the appointments in the next stage at the moment she finishes all the appointments in her current stage; 2) a blocked target patient reserves the retrial appointment immediately on a future day after knowing she would be blocked at the beginning of the current day.

In addition to following a strict FCFS policy, we consider the combination of FCFS with alternate scheduling rules. We process all appointments on the current day using a FCFS order to determine the blocked patients, and then readjust appointments on future days using the following rules:

- **BJF rule.** Under the BJF (Blocked Jobs First) rule, retrial appointments for previously blocked patients are prioritized.
- **SJF rule.** Under the SJF (Shortest-remaining-time Jobs First) rule, appointments from patients with fewer remaining stages are prioritized.
- **CJF rule.** Under the CJF (Closest-to-deadline Jobs First) rule, appointments from patients with projected completion time closer to the coming Friday (deadline) are prioritized.

Note that the BJF rule is essentially equivalent to the pure FCFS discipline since the blocked patients reserve the retrial appointments at the beginning of the day, so they are prioritized over other non-blocked patients (whose appointments have not yet begun processing).

Priority of Exogenous Patients. Under the random ordering setting, exogenous patients join the queue with target patients and have the same chance of being blocked. Under the FCFS scheduling disciplines (including the combination with other rules), we use the following way to schedule

exogenous patients amongst the target patients. Assume that a given station has N_{tar} target patient appointments scheduled, and N_{exo} exogenous patient appointments scheduled. We use a factor α to control the number of exogenous patients to be prioritized today, i.e., $\lfloor \alpha N_{\text{exo}} \rfloor$ of the exogenous appointments should be prioritized to be processed before other patients, and the rest will be randomly mixed with the target patients for processing or blocking. To implement this in the simulation, we first put the $\lfloor \alpha N_{\text{exo}} \rfloor$ exogenous patient appointments to the current service order list (to give them the highest priority). We then append the N_{tar} target patient appointments to the order list. For the remaining $N_{\text{exo}} - \lfloor \alpha N_{\text{exo}} \rfloor$ exogenous patient appointments, we generate an index from the discrete uniform distribution with range $\{0, 1, \ldots, N_{\text{exo}}\}$ to indicate which target patient appointment the exogenous patient appointment is inserted ahead of. For example, if the generated index is 3, the exogenous patient appointment will be placed between the second and third of the N_{tar} target patient appointments. This uniform mixing results in the target patients and the $(1-\alpha)$ proportion of exogenous patients being blocked with the same probability.

Care Path Estimation. We use visit appointment data (for appointments after the initial visit) to estimate care paths for target patients, grouped by geocode into four types. Based on input from our healthcare partner, we only include periods of low blocking to prevent estimation bias. To determine low blocking periods we compare the daily workload to the capacity – estimated as the 95th percentile of the daily workload over the entire dataset. We then estimate the probability of service needs by patient type for each subsequent day, based on the empirical portion of patients attending a service on that day within their itinerary. The sample size ranges between 1,000 to 1,600 patients (20,000 to 30,000 total visits) for each type except for the lower-volume international patients. We also performed a sensitivity analysis using aggregated routing probabilities across all patient types; see Table 2. The resulting solution deviates little from the solution using the path estimates differentiated by patient type, indicating our model's robustness against potential estimation errors.

Performance Metrics. On each day of the simulation, we record the total number of patients blocked, the blocking probability (fraction of patients blocked), and the number of exogenous patients blocked. For each patient, we record her ICT and whether or not her itinerary completes by the target deadline. Table 5 shows the historical template. Table 6 shows the workload under the historical template using the random-ordering setting. Table 7 shows the workload under the historical template using the FCFS-BJF rule with $\alpha = 0.4$. Comparing the two tables, we can see that the average workload of each station from the FCFS-BJF rule is close to that in the random-ordering setting. Detailed comparison including variance comparison and additional scheduling rules are in Section B.5.

B.3. Comparison with the Exact Solution

In this section, we conduct experiments in small-scale systems and perform an exhaustive search to obtain the "true" optimal solutions, which are compared with the solutions from our iterative

	International	National	Regional	Local	Total
Mon	2%	36%	27%	35%	10.91
Tue	2%	45%	22%	32%	10.02
Wed	1%	43%	26%	30%	13.29
Thu	2%	38%	29%	32%	13.87
Fri	3%	36%	24%	36%	8.23
Total	2%	40%	26%	33%	56.33

Table 5 The average number of root appointments allocated by day of the week from historical data, and the proportion from each patient type.

Total workload	BDC	Med Onco	Rad Onco	Gen Surg	Plas Surg
Mon	48.9	124.0	92.1	59.3	54.7
Tue	44.0	145.2	95.3	64.3	60.0
Wed	53.2	141.4	102.8	68.5	54.7
Thu	57.1	124.1	108.2	64.9	53.4
Fri	37.4	102.2	56.6	36.8	38.5
% of BC	66%	2%	1%	25%	8%
Capacity	51	152	118	72	60

Table 6 Estimated total average workload, capacity, and proportion of workload contributed by breast cancer (BC) patients for each station in the five-station network under random ordering. In the baseline, we use the constant capacity (estimated from the empirical average).

Total workload	BDC	Med Onco	Rad Onco	Gen Surg	Plas Surg
Mon	48.8	124.1	92.1	59.4	54.9
Tue	44.0	145.1	95.3	64.2	60.0
Wed	53.2	141.3	102.8	68.5	54.7
Thu	57.4	124.1	108.2	65.0	53.3
Fri	38.2	102.3	56.5	36.9	38.5
% of BC	66%	2%	1%	25%	8%
Capacity	51	152	118	72	60

Table 7 Estimated total average workload, capacity, and proportion of workload contributed by breast cancer (BC) patients for each station in the five-station network under the FCFS-BJF rule and $\alpha = 0.4$.

algorithm. We show that the resulting templates and itinerary completion rates are close, providing evidence that our algorithm can find near-optimal solutions. Below we explain the experimental setting and our design of the exhaustive search and then present the performance comparison.

Experimental setting. We consider a small system consisting of two service stations: BDC and general surgery. To reduce the problem size, we use 4 days per week. We assume the capacity of BDC station is 50 appointments per day and 70 for the general surgery station, corresponding to the average capacity in the baseline (see Table 6). Patients are divided into two types: a normal type and a prioritized type, with a total of 20 and 30 root appointments per week, respectively.

Brute-force search. The search consists of two rounds, with the first round being a coarse grid search and the second round being a refined search. To start, for each of the two types of patients, we divide the total number of root appointments to be allocated into 8 partitions (coarse grid) and enumerate all combinations of allocation for 4 days. We do a Cartesian product of the two patient types to get a total of 27,225 possible templates. We calculate the completion rates for these templates via simulation and identify a few candidates that achieve the following: 1) at least 80% completion

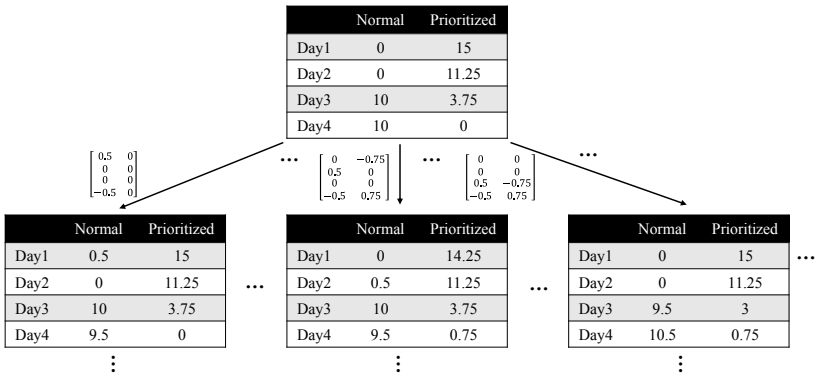
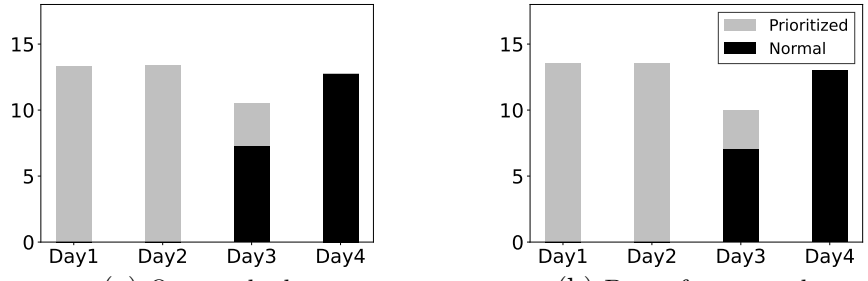


Figure 9 **Constructing the search space using the tree-search method with one of the six candidate templates.**

rate for the prioritized patients and 2) maximum blocking probability being less than 0.25. This yields six candidate templates that we use as initial points to refine in the second round.

In the second round, we employ a perturbation method that starts from six candidate templates and constructs the search space with a refined grid using a tree-search method as illustrated in Figure 9. For each of the six templates selected as an initial solution (the starting "node" in the tree), we do a breadth-first search to collect 5000 templates, which results in analyzing 30,000 templates in total. For each node, its child nodes are acquired by adding the perturbations. Each perturbation moves 2.5% of the total capacity from one day to another.

Solution comparison. Figure 10 compares the optimal template solved from our iterative algorithm versus the one solved using the brute-force search. Table 8 compares the completion rates under the two templates. Not only is the template from our algorithm close to the one from the exhaustive search, but both templates also produce nearly identical completion performance.



(a) Our method (b) Brute-force search

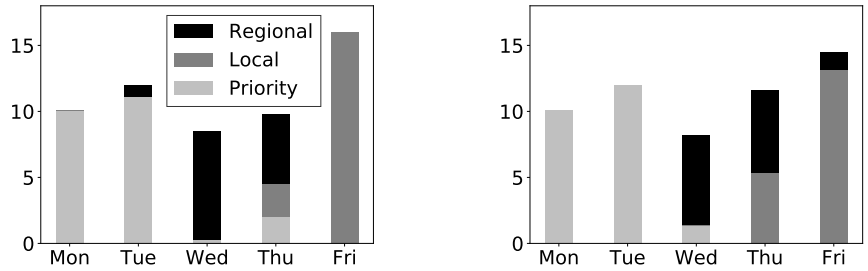
Figure 10 **Templates solved from our method and from brute-force search under the 2-station network setting.**

	Prioritized	Normal
Our method	$83.2\% \pm 0.4\%$	$21.5\% \pm 0.4\%$
Brute-force search	$83.4\% \pm 0.2\%$	$23.2\% \pm 0.2\%$

Table 8 Completion rates in the small-scale, 2-station network system.

B.4. Network Optimal and Network-agnostic Templates

Figure 11 shows the templates solved from the network-agnostic and full-network optimization problems in Section 5.3.3.



(a) Network optimal (b) Network-agnostic

Figure 11 **Comparison of network optimal and network-agnostic templates when a 75% completion rate for regional patients is set as a constraint for optimizing the completion rate of the priority patients.**

B.5. Comparison of Alternate Scheduling Rules

The template solved from our iterative algorithm assumes the random ordering as described in Section 2.1. In this section we demonstrate via numerical experiments that (i) the workload under random ordering is close to that from other scheduling rules, and (ii) the optimal templates solved using random ordering result in relatively robust performance under other scheduling rules. We compare random ordering with the following rules: FCFS-BJF, FCFS-SJF, and FCFS-CJF; all three are introduced in Section B.2 (FCFS-BJF is equivalent to pure FCFS as explained there).

Workload comparison. Table 9 and Table 10 compare the mean and standard deviations of workload of the BDC station under the different scheduling rules using the baseline 5-station network. The role of α has been introduced in the Online Supplement, Section B.2. We implement the optimal template (solved using our iterative algorithm) in the simulation using different scheduling rules and calculate the corresponding workload. We can see that the mean and standard deviation varies little across these scheduling rules. The workloads for other stations in the network are similar.

	Random	BJF, $\alpha=0.4$	BJF, $\alpha=0.8$	SJF, $\alpha=0.4$	SJF, $\alpha=0.8$	CJF, $\alpha=0.4$	CJF, $\alpha=0.8$
Mon	46.4	46.4	46.4	46.4	46.6	46.3	46.4
Tue	49.2	49.2	49.2	49.2	49.3	49.2	49.2
Wed	45.2	45.3	45.3	45.2	45.3	45.3	45.4
Thu	49.9	49.9	49.9	49.7	49.7	49.9	49.9
Fri	45.2	45.3	45.7	45.4	45.7	45.3	45.5

Table 9 Comparison of the mean of BDC workload with different scheduling rules under the optimal template.

Robustness of the optimal template. In Figure 12, we compare itinerary completion performance by evaluating the following templates in our simulation under the FCFS-BJF rule: the optimal-template (solved from our iterative algorithm), the historical template, the front-loaded template,

	Random	BJF, $\alpha = 0.4$	BJF, $\alpha = 0.8$	SJF, $\alpha = 0.4$	SJF, $\alpha = 0.8$	CJF, $\alpha = 0.4$	CJF, $\alpha = 0.8$
Mon	3.1	3.1	3.2	3.1	3.2	3.1	3.1
Tue	3.3	3.3	3.3	3.3	3.3	3.3	3.3
Wed	3.8	3.8	3.9	3.8	3.9	3.9	3.9
Thu	3.5	3.5	3.6	3.5	3.6	3.5	3.6
Fri	3.3	3.4	3.5	3.4	3.5	3.4	3.5

Table 10 Comparison of the standard deviation of BDC workload with different scheduling rules under the optimal template.

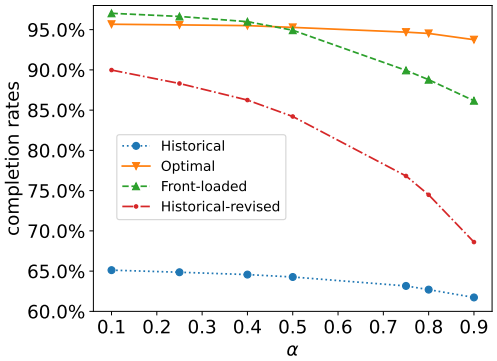


Figure 12 **Completion rates of target patients under BJF and different α values.**

and the historical revised template. We also vary the ratio α, which determines the percentage of exogenous patient appointments to be prioritized, from 0.1 to 0.9 to demonstrate the robustness of the optimal template under different resulting exogenous workload.

We can observe from the figure that the optimal template achieves the best completion rates when $\alpha \geq 0.5$, and the second best (with less 1% difference from the best) for $\alpha < 0.5$; the front-loaded template outperforms the optimal template marginally when $\alpha < 0.5$. This is because when α is small, fewer exogenous patients get priority over the target patients, so the high level of blocking caused by the front-loaded template primarily affects the exogenous patients rather than the target patients. Since exogenous patients leave the system after being blocked (no retrial), this dynamic reduces the effective workload and improves the overall completion rates by unfairly favoring target patients. Moreover, the performance of the front-loaded template degrades significantly as α increases and exogenous patients get higher priority. In contrast, our algorithm accounts for the workload and service-level requirement for both target and exogenous patients. Hence, the resulting optimal template achieves a smoothed workload and low blocking probability for both the target and exogenous patients. Correspondingly, the itinerary completion rates of target patients are far less affected as α increases, maintaining a desirable rate of around 95%.

First-order effect to completion rate improvement. Table 11 compares the target patients' completion rates using the historical, historical-revised, and optimal templates under different scheduling rules. We can observe that, under a given template, different scheduling rules may lead to very different completion rates. For example, the SJF or CJF rules help to further improve the completion

	Historical		Historical-revised		Optimal	
Random ordering	61.8%±0.3%	0.14	85.2%±0.2%	0.71	95.4%±0.2%	0.05
BJF, $\alpha = 0.4$	64.6%±0.3%	0.11	86.3%±0.3%	0.62	95.5%±0.2%	0.05
BJF, $\alpha = 0.8$	62.7%±0.3%	0.06	74.5%±0.6%	0.23	94.5%±0.3%	0.04
SJF, $\alpha = 0.4$	65.1%±0.3%	0.11	78.4%±0.4%	0.62	93.8%±0.2%	0.05
SJF, $\alpha = 0.8$	63.6%±0.2%	0.06	73.9%±0.8%	0.23	92.0%±0.3%	0.04
CJF, $\alpha = 0.4$	65.7%±0.4%	0.11	95.1%±0.2%	0.62	96.2%±0.2%	0.05
CJF, $\alpha = 0.8$	65.2%±0.4%	0.06	85.8%±0.5%	0.23	96.1%±0.2%	0.05

Table 11 Completion rates of the target patients ("mean ± 95%CI) and blocking probabilities of exogenous patients (the highest among stations/days) under different scheduling rules and templates.

rates. SJF clears existing jobs faster and is known to reduce the overall workload in the scheduling literature (Silberschatz et al. 2018, Smith 1956). CJF prioritizes patients who are closer to violating their completion deadline, which is aligned with the objective of maximizing the fraction of target patients that complete their itineraries by the deadline. In both cases, however, the improvement is marginal compared to the impact of optimized template itself. For example, under the historical template, the best completion rate of target patients is 65.7% regardless of what scheduling rule is used. In contrast, by employing the optimal template, the completion rate of target patients is at least 92.0% even under the worst-performing scheduling rules.

We note one exception, where the CJF rule with $\alpha = 0.4$ achieves a significantly higher (95.1%) completion rate than other combinations of scheduling rules for the historical-revised template. The reason is similar to our explanation above: the high completion rate for target patients is achieved by sacrificing exogenous patients, who experience a blocking probability as high as 62%. When exogenous patients get higher priority (CJF with $\alpha = 0.8$), their blocking probability reduces to 23% but the target patient completion rates deteriorate significantly to 85%. In contrast, our optimization framework accounts for the blocking of both exogenous and target patients, achieving near zero blocking for exogenous patients while providing high completion rates for target patients.

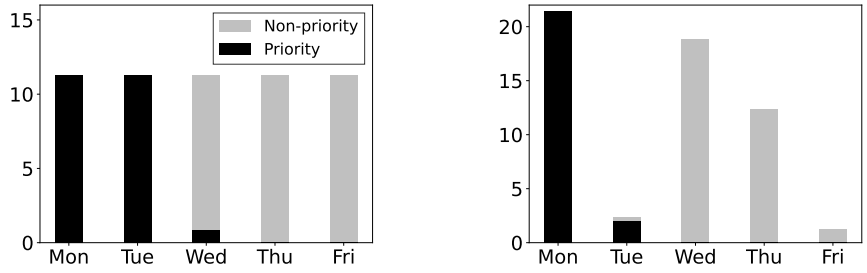
To summarize, the numerical results show that our random ordering assumption is a good approximation to the system dynamics and generates representative workload distributions. Moreover, this approximation is sufficient for tactical planning of capacity allocation – the root appointment templates solved using this approximation provide satisfactory itinerary completion rates across different scheduling disciplines. These results imply that, while using scheduling rules can improve the completion rates further, proper capacity allocation in the tactical phase is the *first-order* issue. While smarter scheduling rules can bring marginal improvement under a given template, scheduling alone cannot bridge the gap between optimized and non-optimized templates.

B.6. Comparison with Other Heuristic Templates

Based on the insights we summarize in Section 5.3, we can design an improved front-loaded heuristic template by setting the daily allowed number of root appointments to be more evenly spread out

across days of the week, i.e., smoothing the workload and reducing blocking. To do so, we set the daily number of root appointments to be the total number of root appointments divided by 5. We then front-load the root appointments of target patients to give them the priority. This evenly-distributed template is illustrated in Figure 13a, and it achieves a very high completion rate of 96.0% for the target patients in the baseline setting. This is not surprising since the template is quite close to the optimal template solved from our algorithm. However, this does not preclude the necessity of developing an algorithm, since (a) the capacities of each service stations are the same on each day in our baseline case study setting, and (b) the heuristic template does not address more nuanced system requirements such as when we have different levels of priority or need to account for the completion rates for non-priority patients.

For point (b), as we have emphasized in Sections 5.3 and 5.4, our framework is flexible and can incorporate various objectives and constraints beyond completion rates for target patients, e.g., using different weights to differentiate target patients and incorporating desirable completion rates for non-priority patients as constraints. Heuristic designs such as the evenly-distributed template cannot easily account for these features. To demonstrate point (a), we consider a setting with time-varying capacity in addition to the workload. For example, for the BDC station, we change its capacity to 75, 25, 64, 51, and 41 on Monday through Friday, respectively, with Tuesday having the least amount of capacity. In this scenario, the evenly-distributed template causes a high blocking probability of 0.64 on Tuesdays, which reduces the completion rates to less than 90% for national patients as demonstrated in Table 12 (recall that national patients constitute the majority of priority patients). Meanwhile, our algorithm can respond to this capacity change naturally. The optimized template solved from our algorithm under this time-varying capacity is shown in Figure 13b. The optimized template reduces the root appointments for Tuesdays to account for the low capacity and maintains high completion rates, with the average completion rates for national patients being over 97%.



(a) The evenly-distributed template (b) The optimized template

Figure 13 The evenly-distributed and optimized templates for time-varying capacities of BDC station.

	International	National	Regional	Local
Evenly-distributed	$95.4 \pm 0.9\%$	$89.1 \pm 0.3\%$	$50.3 \pm 0.9\%$	$39.4 \pm 0.2\%$
Optimized	$94.8 \pm 0.9\%$	$97.8 \pm 0.1\%$	$81.6 \pm 0.6\%$	$51.4 \pm 0.6\%$

Table 12 Comparison of completion rates for the evenly-distributed template and the optimized template under the uneven BDC capacities setting.

C. Comprehensive Framework for Phase-type ICT Characterization

We consider the transition from one stage (with multiple appointments) to the next stage (with multiple appointments). We first define a set, $\mathcal{R} = \{u_1, u_2, \ldots, u_n\}$, that contains the group of resources that are required to complete a particular stage of treatment. We define the state space of the phase-type distribution for completing all the appointments in group $\mathcal{R}$ as

$$\mathcal{S}_{\mathcal{R}} = \{(a_{u_1}, a_{u_2}, \ldots, a_{u_n}, d)\}, \tag{23}$$

where for each $u_j \in \mathcal{R}$, $a_{u_j} = 0$ indicates that the appointment at resource u_j has not yet been completed on day d and $a_{u_j} = 1$ indicates that it has been completed. Given U total possible resources that can be used by the patient, each group, $\mathcal{R}$, is one possible combination of the U resources. That is, $\mathcal{R} \in \mathcal{P}(\{u_1, \ldots, u_U\})$, where $\mathcal{P}$ is the power set, or the set of all possible subsets of $\{u_1, \ldots, u_U\}$.

C.1. Phase-type network sojourn time with parallel appointments and deterministic patient path

We first consider the case where each group of appointments must be finished before moving to the next stage, i.e., no probabilistic resource requirements. For ease of exposition, we begin by illustrating the phase-type generator for a simpler setting with S stages, where each stage contains only *two* stations for the patient to visit. That is, let $s = 1, 2, \ldots, S$ denote one of the S stages, $u_{s,j} \in \{u_1, \ldots, u_U\}$ $(j = 1, 2,\ s = 1, \ldots, S)$ denote the j^{th} station the patient needs to visit in stage s, and $\mathcal{R}_s = \{u_{s_1}, u_{s_2}\}$ for the two stations to visit. Then, the state for stage s is given by

$$\mathcal{S}_{\mathcal{R}_s} = (a_{u_{s,1}}, a_{u_{s,2}}, d)$$

where $d = 1, \ldots, 5$ represents the day of week. We specify the transition matrix for the general setting at the end of this subsection.

Block corresponding to transitions within one stage. We first characterize $V^1_{\mathcal{R}_s}$, which characterizes the transitions within stage s; i.e., at least one of the appointments from the stations in $\mathcal{R}_s = \{u_{s,1}, u_{s,2}\}$ has not yet been obtained. As mentioned, for exposition, we consider two resources are needed in each stage $s = 1, 2, \ldots, m$. Note that there can be at most three combinations for $(a_{u_{s,1}}, a_{u_{s,2}})$: $(0,0)$, $(0,1)$, and $(1,0)$, denoting that neither appointment was able to be completed, the appointment in $u_{s,2}$ was able to be completed but not $u_{s,1}$, and the appointment in $u_{s,1}$ was able to be completed

but not $u_{s,2}$, respectively. The transition from day d to day $d+1$ corresponding to sample paths where the patient has not completed both appointments in stage s is thus given by:

$$\mathbf{V}^1_{\mathcal{R}_s}(d,d+1) = \begin{bmatrix} & (0,0,d+1) & (0,1,d+1) & (1,0,d+1) \\ (0,0,d) & \beta_{u_{s,1},d}\cdot\beta_{u_{s,2},d} & \beta_{u_{s,1},d}(1-\beta_{u_{s,2},d}) & (1-\beta_{u_{s,1},d})\beta_{u_{s,2},d} \\ (0,1,d) & 0 & \beta_{u_{s,1},d} & 0 \\ (1,0,d) & 0 & 0 & \beta_{u_{s,2},d} \end{bmatrix}, \tag{24}$$

Specifically, $\mathbf{V}^1_{\mathcal{R}_s}(d,d+1)$ is the matrix that represents the time to complete both jobs $u_{s,1}$ and $u_{s,2}$ in stage s, which is the maximum of two phase-type distributions. As a result, block $V^1_{\mathcal{R}_s}$ can be written as

$$\mathbf{V}^1_{\mathcal{R}_s} = \begin{bmatrix} & \textit{Day } 1 & \textit{Day } 2 & \textit{Day } 3 & \textit{Day } 4 & \textit{Day } 5 \\ \textit{Day } 1 & \mathbf{0} & \mathbf{V}^1_{\mathcal{R}_s}(1,2) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \textit{Day } 2 & \mathbf{0} & \mathbf{0} & \mathbf{V}^1_{\mathcal{R}_s}(2,3) & \mathbf{0} & \mathbf{0} \\ \textit{Day } 3 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}^1_{\mathcal{R}_s}(3,4) & \mathbf{0} \\ \textit{Day } 4 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}^1_{\mathcal{R}_s}(4,5) \\ \textit{Day } 5 & \mathbf{V}^1_{\mathcal{R}_s}(5,1) & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{25}$$

Note that the process will stay in block $\mathbf{V}^1_{\mathcal{R}_s}$ until all the appointments in group $\mathcal{R}_s$ have been successfully completed. The transitions to the next stage, $s+1$, where all appointments in stage s have been completed are given below.

Block corresponding to transitions from one stage to another stage. Let $\mathbf{V}^2_{\mathcal{R}_s\to\mathcal{R}_{s+1}}(d,d+1)$ represent the transition from the group of resources, $\mathcal{R}_s$ on day d to the group of resources $\mathcal{R}_{s+1}$ on day $d+1$. For illustration, we also let $\mathcal{R}_{s+1}=\{u_{s+1,1},u_{s+1,2}\}$, indicating that two resources are required in stage $s+1$. Hence specifying the transition from states $(a_{u_{s,1}},a_{u_{s,2}},d)$ to states $(a_{u_{s+1,1}},a_{u_{s+1,2}},d+1)$ gives us

$$\mathbf{V}^2_{\mathcal{R}_s\to\mathcal{R}_{s+1}}(d,d+1) = \begin{bmatrix} & (0,0,d+1) & (0,1,d+1) & (1,0,d+1) \\ (0,0,d) & (1-\beta_{u_{s,1},d})(1-\beta_{u_{s,2},d}) & 0 & 0 \\ (0,1,d) & (1-\beta_{u_{s,1},d}) & 0 & 0 \\ (1,0,d) & (1-\beta_{u_{s,2},d}) & 0 & 0 \end{bmatrix}, \tag{26}$$

Note, when the process initially enters $s+1$ on day $d+1$, none of the new appointments generated for this stage have been completed yet. Hence it is only possible to transition to state $(0,0,d+1)$ (indicating that none of the new appointments for stage $s+1$ have yet been completed) among all the states for $(a_{u_{s+1,1}},a_{u_{s+1,2}},d+1)$. As a result, block $\mathbf{V}^2_{\mathcal{R}_s\to\mathcal{R}_{s+1}}$ can be written as

$$\mathbf{V}^2_{\mathcal{R}_s\to\mathcal{R}_{s+1}} = \begin{bmatrix} & \textit{Day } 1 & \textit{Day } 2 & \textit{Day } 3 & \textit{Day } 4 & \textit{Day } 5 \\ \textit{Day } 1 & \mathbf{0} & \mathbf{V}^2_{\mathcal{R}_s\to\mathcal{R}_{s+1}}(1,2) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \textit{Day } 2 & \mathbf{0} & \mathbf{0} & \mathbf{V}^2_{\mathcal{R}_s\to\mathcal{R}_{s+1}}(2,3) & \mathbf{0} & \mathbf{0} \\ \textit{Day } 3 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}^2_{\mathcal{R}_s\to\mathcal{R}_{s+1}}(3,4) & \mathbf{0} \\ \textit{Day } 4 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}^2_{\mathcal{R}_s\to\mathcal{R}_{s+1}}(4,5) \\ \textit{Day } 5 & \mathbf{V}^2_{\mathcal{R}_s\to\mathcal{R}_{s+1}}(5,1) & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{27}$$

Full transition matrix. Now we are ready to specify the full transition matrix.

$$\left[\begin{array}{c|c} \mathbf{T}_{\mathcal{C}} & \mathbf{T}^0_{\mathcal{C}} \\ \hline \mathbf{0} & 1 \end{array}\right] = \left[\begin{array}{ccccc|c} \mathbf{v}^1_{\mathcal{R}_1} & \mathbf{v}^2_{\mathcal{R}_1\to\mathcal{R}_2} & \mathbf{0} & \mathbf{0}\ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{v}^1_{\mathcal{R}_2} & \mathbf{v}^2_{\mathcal{R}_2\to\mathcal{R}_3} & \mathbf{0}\ldots & \mathbf{0} & \mathbf{0} \\ & \ddots & \ddots & & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\ldots & \mathbf{v}^1_{\mathcal{R}_\mathbf{S}} & \mathbf{v}^0_{\mathcal{R}_\mathbf{S}} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\ldots & \mathbf{0} & 1 \end{array}\right] \tag{28}$$

Here, $\mathbf{V}^0_{\mathcal{R}_\mathbf{S}}$ is defined similarly to $\mathbf{V}^2_{\mathcal{R}_\mathbf{s}\to\mathcal{R}_{\mathbf{s}+1}}$ $(s<S)$, except for each block $\mathbf{V}^0_{\mathcal{R}_\mathbf{S}}(d,d+1)$, in which all transitions are to the absorbing state.

Extension to general numbers of stations in each stage. Consider the process is in set of resources $\mathcal{R}_s$ in stage s and transitions to resources $\mathcal{R}_{s+1}$ in stage $s+1$. Then we can define the transitions as follows:

$$\left[\begin{array}{c|c} \mathbf{V}^1_{\mathcal{R}_s}(d,d+1) & \mathbf{V}^2_{\mathcal{R}_s}(d,d+1) \\ \hline \mathbf{0} & 1 \end{array}\right] = \bigotimes_{u_j\in\mathcal{R}_s} A_{u_j,d}, \tag{29}$$

where $\bigotimes$ denotes the Kronecker product. (29) gives us $\mathbf{V}^1_{\mathcal{R}_s}(d,d+1)$ to specify $\mathbf{V}^1_{\mathcal{R}_s}$ and also

$$\mathbf{V}^2_{\mathcal{R}_s\to\mathcal{R}_{s+1}}(d,d+1) = \left[\mathbf{V}^2_{\mathcal{R}_s}(d,d+1)\ 0_{|\mathcal{S}_{\mathcal{R}_s}|\times|\mathcal{S}_{\mathcal{R}_{s+1}}|-1}\right]. \tag{30}$$

Here, $0_{|\mathcal{S}_{\mathcal{R}_s}|\times|\mathcal{S}_{\mathcal{R}_{s+1}}|}$ in (30) is a $|\mathcal{S}_{\mathcal{R}_s}|\times|\mathcal{S}_{\mathcal{R}_{s+1}}|$ matrix of zeros to account for the fact that patients only transition to their next stage in the state where none of the appointments have been scheduled yet. $A_{u_j,d}$ is the following 2×2 matrix:

$$A_{u_j,d} = \begin{bmatrix} \beta_{u_j,d} & 1-\beta_{u_j,d} \\ 0 & 1 \end{bmatrix}. \tag{31}$$

C.2. Phase-type network sojourn time with parallel appointments and stochastic patient path

In this section, we consider the case where only a subset of all possible resource groups may be required in a given stage. Consider U total possible resources, where the possible resource groups form the powerset $\mathcal{P}(\{u_1,\ldots,u_U\})$. Again, for the purposes of exposition, we focus on the simplified setting, where in each stage $s=1,\ldots,S$, there are two possible resources that *may* require appointments, $u_{s,1},u_{s,2}$.

In the simplified setting, the possible groups of resources for stage s are given by $\mathcal{R}_s\in\mathcal{P}(\{u_{s,1},u_{s,2}\}) = \{\{u_{s,1},u_{s,2}\},\{u_{s,1}\},\{u_{s,2}\},\emptyset\}$. In other words, for stage s, there are four possible outcomes: visiting both stations, visiting $u_{s,1}$, visiting $u_{s,2}$, or skip this stage, with probabilities $\mathbb{P}(\{u_{s,1},u_{s,2}\}) = \nu_{u_{s,1}}\nu_{u_{s,2}}$, $\mathbb{P}(\{u_{s,1},\}) = \nu_{u_{s,1}}(1-\nu_{u_{s,2}})$, and $\mathbb{P}(\{u_{s,2},\}) = (1-\nu_{u_{s,1}})\nu_{u_{s,2}}$, and $\mathbb{P}(\emptyset) = (1-\nu_{u_{s,1}})(1-\nu_{u_{s,2}})$, respectively. In our phase-type generator matrix specified below, we omit the last one, the null set $\emptyset$, since it contains no resources. Let $\mathcal{R}_{s,k}$, $k=1,\ldots,3$ be the set of appointments that need to be completed prior to exiting stage s, representing each of the three non-empty outcomes.

To capture all possible resource groups, we first need to enlarge the state space to be

$$\big(S_{\mathcal{R}_{s,1}}, S_{\mathcal{R}_{s,2}}, S_{\mathcal{R}_{s,3}}, d\big),$$

where $S_{\mathcal{R}_{s,k}}$ is the state space defined in (23) above for a given resource group $\mathcal{R}_{s,k}$, i.e., the tuple of a_u's for all $u \in \mathcal{R}_{s,k}$. For example, $S_{\mathcal{R}_{s,1}} = \big\{(a_{u_{s,1}}, a_{u_{s,2}})\big\}$ since appointments are required in both resources for group $\mathcal{R}_{s,1}$. Similarly, $S_{\mathcal{R}_{s,2}} = \big\{a_{u_{s,1}}\big\} = \{0\}$ since only $u_{s,1}$ is required in resource group $\mathcal{R}_{s,2}$; we only need $a_{u_{s,1}} = 0$ to track whether stage s is finished or not when there is a single resource.

Block corresponding to transitions within one stage. The transitions within a single stage are defined by the phase-type block matrix $U^1_{\mathcal{R}}$, which is the probabilistic analogue to $V^1_{\mathcal{R}_s}$ defined in the previous subsection. In the simplified setting, we have

$$\mathbf{U}^1_{\mathcal{R}} = \begin{bmatrix} \mathbf{V}^1_{\mathcal{R}_{s,1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^1_{\mathcal{R}_{s,2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}^1_{\mathcal{R}_{s,3}} \end{bmatrix}, \tag{32}$$

Here, each block represents the phase-type transitions for attempting to complete all the appointments in the corresponding resource group. Block $V^1_{\mathcal{R}_{s,k}}$ captures the transitions prior to completing all the appointments in the resource group $\mathcal{R}_{s,k}$ and can be specified as the general form of (25) with $\mathbf{V}^1_{\mathcal{R}_{s,k}}(d, d+1)$ given by (29). $\mathbf{U}^1_{\mathcal{R}}$ has a diagonal structure because once a process enters a given block corresponding to group $\mathcal{R}_{s,k}$, it will stay in that block until all the appointments for resources in the group have been scheduled. Then the process will leave the block and transition to stage $s+1$ with probability defined by the blocks that follow. Note that $\mathcal{R}_{s,k}$'s are not necessarily all the same size, as the size of each block is determined by the number of resources in that block.

Block corresponding to transitions from stage s to stage $s+1$. In the setting with probabilistic resource requirements, transition from stage s to $s+1$ indicates that the patient has completed all appointments required in stage s. We specify the block for transitions out of stage s (i.e. completion of all required appointments), $U^2_{\mathcal{R}}$. Again, for illustration, we focus on the simplified setting where there are only two stations that may need to be visited in stage s and stage $s+1$, i.e., each stage has three possible resource groups (excluding the empty set) denoted as $\mathcal{R}_{s,k}$ and $\mathcal{R}_{s+1,k}$ $(k = 1, 2, 3)$, respectively. In this simplified setting, the block is given by

$$\mathbf{U}^2_{\mathcal{R}} = \left[\begin{array}{c|ccc} & \mathcal{R}_{s+1,1} & \mathcal{R}_{s+1,2} & \mathcal{R}_{s+1,3} \\ \hline \mathcal{R}_{s,1} & p_{s+1,1} \cdot \mathbf{V}^2_{\mathcal{R}_{s,1}\to\mathcal{R}_{s+1,1}} & p_{s+1,2} \cdot \mathbf{V}^2_{\mathcal{R}_{s,1}\to\mathcal{R}_{s+1,2}} & p_{s+1,3} \cdot \mathbf{V}^2_{\mathcal{R}_{s,1}\to\mathcal{R}_{s+1,3}} \\ \mathcal{R}_{s,2} & p_{s+1,1} \cdot \mathbf{V}^2_{\mathcal{R}_{s,2}\to\mathcal{R}_{s+1,1}} & p_{s+1,2} \cdot \mathbf{V}^2_{\mathcal{R}_{s,2}\to\mathcal{R}_{s+1,2}} & p_{s+1,3} \cdot \mathbf{V}^2_{\mathcal{R}_{s,2}\to\mathcal{R}_{s+1,3}} \\ \mathcal{R}_{s,3} & p_{s+1,1} \cdot \mathbf{V}^2_{\mathcal{R}_{s,3}\to\mathcal{R}_{s+1,1}} & p_{s+1,2} \cdot \mathbf{V}^2_{\mathcal{R}_{s,3}\to\mathcal{R}_{s+1,2}} & p_{s+1,3} \cdot \mathbf{V}^2_{\mathcal{R}_{s,3}\to\mathcal{R}_{s+1,3}} \end{array}\right]. \tag{33}$$

While (33) may seem complex, it is easily interpreted. First, we have added labels for the rows and columns. The row labels, $\mathcal{R}_{s,k}$, indicate which resource set the patient has just completed in stage s. The column labels, $\mathcal{R}_{s+1,k}$, represent which resource group the patient will require in stage $s+1$ of treatment. Note these labels are for exposition and do not represent the matrix states, nor do they directly indicate the size of the blocks, which is fully defined by the $\mathbf{V}$'s. Inside the matrix, $p_{s+1,k}$ is the probability of requiring resource group $\mathcal{R}_{s+1,k}$ in stage $s+1$. Thus, each block entry (row $\mathcal{R}_{s,k}$ to column $\mathcal{R}_{s+1,k}$) represents the probability of finishing the remaining appointments of resource group $\mathcal{R}_{s,k}$ and transitioning to resource group $\mathcal{R}_{s+1,k}$, denoted by $\mathbf{V}^2_{\mathcal{R}_{s,k}\to\mathcal{R}_{s+1,k}}$; this event occurs with probability $p_{s+1,k}$. Recall from before that, for a given k, the matrix blocks $\mathbf{V}^2_{\mathcal{R}_{s,k}\to\mathcal{R}_{s+1,k}}$ only differ from each other by the number of zero blocks required to expand the state space to the appropriate size for resource group $\mathcal{R}_{s+1,k}$, which may be different for each k.

Block corresponding to transitions from stage s to stage $s+d$. The transitions from stage s to $s+d$ are defined by matrix $\mathbf{U}_s^{d+1}$, which has the exact same form as (33) except that we replace resource group $\mathcal{R}_{s+1,k}$ with resource group $\mathcal{R}_{s+d,k}$ and resource group probabilities $p_{s+1,k}$ with $p_{s+d,k}$, where

$$p_{s+d,1} = \nu_{u_{s+d,1}}\nu_{u_{s+d,2}}\Pi_{q=s+1}^{s+d-1}(1-\nu_{u_{q,1}})(1-\nu_{u_{q,2}})$$

and $p_{s+d,2},\ p_{s+d,3}$ can be defined similarly by replacing the first two terms with $\nu_{u_{s+d,1}}(1-\nu_{u_{s+d,2}})$ and $(1-\nu_{u_{s+d,1}})\nu_{u_{s+d,2}}$, respectively.

Full transition matrix. Now we are ready to specify the full transition matrix with at total of S possible stages.

$$\left[\begin{array}{c|c}\mathbf{T}_b & \mathbf{T}_b^0\\ \hline \mathbf{0} & 1\end{array}\right] = \left[\begin{array}{ccccccc|c}\mathbf{U}_1^1 & \mathbf{U}_1^2 & \mathbf{U}_1^3 & \mathbf{U}_1^4 & \dots & \mathbf{U}_1^{\mathbf{S}} & & \mathbf{U}_1^0\\ \mathbf{0} & \mathbf{U}_2^1 & \mathbf{U}_2^2 & \mathbf{U}_2^3 & \dots & \mathbf{U}_2^{\mathbf{S}-1} & & \mathbf{U}_2^0\\ & \ddots & \ddots & & & & & \vdots\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{U}_{\mathbf{S}}^1 & & \mathbf{U}_{\mathbf{S}}^0\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & & 1\end{array}\right], \tag{34}$$

$$\mathbf{U}_s^0 = \Pi_{q=s+1}^{S-s}(1-\nu_{u_{q,1}})(1-\nu_{u_{q,2}})\cdot\begin{bmatrix}\mathbf{V}^0_{R_{s,1}}\\ \mathbf{V}^0_{R_{s,2}}\\ \mathbf{V}^0_{R_{s,3}}\end{bmatrix} \tag{35}$$

Note that the last column, $\mathbf{U}_s^0$, captures transitions to the absorbing state, where each block $\mathbf{V}^0_{R_{s,k}}$ is defined similarly as $\mathbf{V}^0_{R_S}$ in (28).