A Capacity Allocation Planning Model for Integrated Care and Access Management: Online Appendices

Jivan Deglise-Hawkinson[1] • **Jonathan E. Helm**[2] • **Todd Huschka**[3] • **David L. Kaufman**[4] • **Mark P. Van Oyen**[5*]

1 jivan@umich.edu; Revenue Management – Operations Research, American Airlines, Fort Worth, TX
2 helmj@indiana.edu; Operations & Decision Technologies, Indiana University, Bloomington, IN
3 huschka.todd@mayo.edu; Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN
4 davidlk@umich.edu; Management Studies, University of Michigan–Dearborn, Dearborn, MI
5 vanoyen@umich.edu; Industrial & Operations Engineering, University of Michigan, Ann Arbor, MI
* Corresponding author: vanoyen@umich.edu

A Proofs

Proof of Theorem 1: Our methodology can be developed for a finite planning horizon; however, we instead focus on an infinite horizon cyclo-stationary model. To this end, we augment the time index with a 'w' (for weeks) as follows: In this case, T is the number of days in one cycle (e.g., $T = 5$ days at our partner institution), and stationarity may be reached by taking the limit $w \to \infty$. Let

$$W_t^k = \sum_{k_1 \in \mathcal{K}'} \sum_{\tau \in \mathcal{C}(k_1)} \sum_{t_1=1}^{T} \lim_{w \to \infty} \sum_{j=0}^{w} \sum_{i=1}^{\alpha_{t_1,j}^{k_1,\tau}} L_{t_1+jT,i}^{k_1,\tau}(t + wT - (t_1 + jT)) \cdot \mathbf{e}_k \cdot s_k, \tag{31}$$

where $L_{t_1+jT,i}^{k_1,\tau}(t + wT - (t_1 + jT))$ is the i^{th} i.i.d instance of the $L_{t_1+jT}^{k_1,\tau}(t + wT - (t_1 + jT))$ random variable. For notational convenience, if $t+wT-(t_1+jT)$ is negative then $L_{t_1+jT}^{k_1,\tau}(t+wT-(t_1+jT))$ will be the vector of 0's. Notice that the second inner sum in Equation (31) considers $\alpha_{t_1,j}^{k_1,\tau}$, all type (k_1, τ) patients that were scheduled for their root appointment on day t_1 of week $j \leq w$. For each one of those patients, a given realization of $L_{t_1+jT}^{k_1,\tau}(t + wT - (t_1 + jT))$ is multiplied by $\mathbf{e}_k \cdot s_k$, which will determine the workload each one of those patients will impose in specialty k, $t + wT - (t_1 + jT)$ days later (hence, on day t of week w in this cyclo-stationary model). The second inner sum is over all weeks j from 0 to w, and the third inner sum is over all days t_1 included in those weeks. Therefore, this captures all type (k_1, τ) patients that were scheduled for a root appointment earlier than day t of planning horizon w, and the sum of their resource requirements in specialty k on day t of week w. The last two outer sums consider all patient types $\tau \in \mathcal{C}(k_1), k_1 \in \mathcal{K}'$.

Since the second innermost sum in Equation (31) is non-decreasing in w, we may apply the Monotone Convergence Theorem to interchange the expectation and the limit. Then, since the number of patients

scheduled, $\alpha_{t_1,j}^{k_1,\tau}$, is independent of the stochastic location process, $L_{t_1+jT,i}^{k_1,\tau}(\cdot)$, we can apply Wald's equation to express the random sum $\sum_{i=1}^{\alpha_{t_1,j}^{k_1,\tau}}$ in Equation (31) in terms of $\mathbb{E}[\alpha_{t_1}^{k_1,\tau}]$, yielding:

$$\begin{aligned}\mathbb{E}[W_t^k] &= \sum_{k_1\in\mathcal{K}'}\sum_{\tau\in\mathcal{C}(k_1)}\sum_{t_1=1}^{T}\cdot\lim_{w\to\infty}\sum_{j=0}^{w}\mathbb{E}[\alpha_{t_1}^{k_1,\tau}]\,\mathbb{E}\left[L_{t_1+jT,i}^{k_1,\tau}(t+wT-(t_1+jT))\cdot\mathbf{e}_k\cdot s_k\right]\\ &= \sum_{k_1\in\mathcal{K}'}\sum_{\tau\in\mathcal{C}(k_1)}\sum_{t_1=1}^{T}\mathbb{E}[\alpha_{t_1}^{k_1,\tau}]\cdot\sum_{j=0}^{\infty}\sum_{m=1}^{M_k}m\cdot r_{t_1}^{k_1,\tau,k}\left(m,t-t_1+jT\right)\cdot s_k.\end{aligned}$$

□

Proof of Theorem 2: Here again we begin by augmenting the time index with the week index, 'j', to capture the fact that the system being modeled is cyclo-stationary with a weekly pattern. Following the general idea of the proof in Theorem 1, we can formulate the steady state variance of the specialty k workload random variable on day t of our planning horizon as:

$$\mathrm{Var}[W_t^k] = \mathrm{Var}\left[\sum_{k_1\in\mathcal{K}'}\sum_{\tau\in\mathcal{C}(k_1)}\sum_{t_1=1}^{T}\sum_{j=0}^{\infty}\sum_{i=1}^{\alpha_{t_1,j}^{k_1,\tau}}L_{t_1+jT,i}^{k_1,\tau}(t+wT-(t_1+jT))\cdot\mathbf{e}_k\cdot s_k\right].$$

As in the proof of Theorem 3.1, the Monotone Convergence Theorem can once again be used to interchange the expectation and the infinite sum. Also, note that $(L_{t_1+jT,i}^{k_1,\tau}(t+wT-(t_1+jT))\cdot\mathbf{e}_k)_i$ is a sequence of i.i.d. random variables which are also independent of $\alpha_{t_1,j}^{k_1,\tau}$; Wald's equation applies. Moreover, the variance of $L_{t_1+jT,i}^{k_1,\tau}(t+wT-(t_1+jT))\cdot\mathbf{e}_k$ is the same for every i. Then, the variance of this random sum can be expressed as follows:

$$\begin{aligned}&\mathrm{Var}\left[\sum_{i=1}^{\alpha_{t_1,j}^{k_1,\tau}}L_{t_1+jT,i}^{k_1,\tau}(t+wT-(t_1+jT))\cdot\mathbf{e}_k\cdot s_k\right]\\ &\quad=\mathbb{E}[\alpha_{t_1,j}^{k_1,\tau}]\cdot\mathrm{Var}[L_{t_1+jT,i}^{k_1,\tau}(t+wT-(t_1+jT))\cdot\mathbf{e}_k\cdot s_k]\\ &\qquad+\mathbb{E}[L_{t_1+jT,i}^{k_1,\tau}(t+wT-(t_1+jT))\cdot\mathbf{e}_k\cdot s_k]^2\cdot\mathrm{Var}[\alpha_{t_1,j}^{k_1,\tau}].\end{aligned}$$

By using the same arguments as the proof of the previous theorem, Equation (8) follows. □

Proof of Proposition 1: For any nonnegative discrete random variable β, it is somewhat well known that $\mathbb{E}[\beta]=\sum_{l=0}^{\infty}\mathbb{P}(\beta>l)$. In particular, $\beta_{t\oplus1}^{k,\tau}$ is nonnegative, so:

$$\begin{aligned}\mathbb{E}[\beta_{t\oplus1}^{k,\tau}] &= \sum_{l=0}^{\infty}\mathbb{P}\left(\beta_{t\oplus1}^{k,\tau}>l\right)\\ &= \sum_{l=0}^{\infty}\sum_{i\in\mathcal{I}}y_t^{k,\tau}(i,l)\cdot\Psi(i),\end{aligned}$$

where we are using our previously defined binary variables $y_t^{k,\tau}(i,l)$ (see constraints (11) and (12)). Similarly,

$$\mathbb{E}[(\beta_{t\oplus1}^{k,\tau})^2]=\sum_{l=0}^{\infty}(2l+1)\cdot\mathbb{P}\left(\beta_{t\oplus1}^{k,\tau}>l\right),$$

which can be rewritten as:

$$\mathbb{E}[(\beta_{t\oplus 1}^{k,\tau})^2] = \sum_{l=0}^{\infty} \left[(2l+1) \cdot \sum_{i \in \mathcal{I}} y_t^{k,\tau}(i,l) \cdot \Psi(i) \right].$$

Then,

$$\begin{aligned}
\tilde{\beta}_{t\oplus 1}^{k,\tau} &= \sum_{l=0}^{\infty} \left[(2l+1) \cdot \sum_{i \in \mathcal{I}} y_t^{k,\tau}(i,l) \cdot \Psi(i) \right] - \left[\sum_{l=0}^{\infty} \sum_{i \in \mathcal{I}} y_t^{k,\tau}(i,l) \cdot \Psi(i) \right]^2 \\
&= \sum_{l=0}^{\infty} \left[(2l+1) \cdot \sum_{i \in \mathcal{I}} y_t^{k,\tau}(i,l) \cdot \Psi(i) \right] - \sum_{(l_1,l_2) \in (\mathbb{Z}^+)^2} \sum_{(i_1,i_2) \in \mathcal{I}^2} y_t^{k,\tau}(i_1,l_1) \cdot y_t^{k,\tau}(i_2,l_2) \cdot \Psi(i_1) \cdot \Psi(i_2).
\end{aligned}$$

Finally, replacing the binary products $y_t^{k,\tau}(i_1,l_1) \cdot y_t^{k,\tau}(i_2,l_2)$ by a binary variable $z_t^{k,\tau}(i_1,i_2,l_1,l_2)$ satisfying Equations (15)–(17), we have expressed the variance of the carryover random variable linearly in our decision variables (see expressions (11)–(14)). □

Proof of Theorem 3: The proof of the linearization result is by induction on t. We first show the result for a finite horizon problem. Then, augmenting the time index with 'w' and taking the limit $w \to \infty$ gives, for a stable system, the result for the cyclo-stationary model. For $t = 1$, $\beta_1^{k,\tau}$ is assumed to be a fixed input value, independent of Θ. Then, the $D_1^{k,\tau}(i)$ are independent of (and hence trivially linear in) Θ since $\mathbb{E}[D^{k,\tau}]$ and $\mathrm{Var}[D^{k,\tau}]$ are independent of Θ. For the induction hypothesis, assume that the $D_t^{k,\tau}(i)$ can be expressed linearly in Θ. Then, by the construction in Equations (11)–(14), the induction hypothesis implies that the $\beta_{t+1}^{k,\tau}(i)$ can be expressed linearly in Θ. Since expectation is a linear operator, it follows that $\overline{\beta}_{t+1}^{k,\tau}$ can also be expressed linearly in Θ. Also, by Proposition 4.1, $\tilde{\beta}_{t+1}^{k,\tau}$ can be expressed linearly in Θ. In order to show that the $D_{t+1}^{k,\tau}(i)$ can be expressed linearly in Θ, under Equation (10) it is sufficient to show that $\mathbb{E}[D_{t+1}^{k,\tau}]$ and $\mathrm{Var}[D_{t+1}^{k,\tau}]$ can be expressed linearly in Θ. For the mean, $\mathbb{E}[D_{t+1}^{k,\tau}] = \mathbb{E}[X_{t+1} + \beta_{t+1}^{k,\tau}] = \mathbb{E}[X_{t+1}] + \overline{\beta}_{t+1}^{k,\tau}$, which can also be expressed linearly in Θ since $\mathbb{E}[X_{t+1}]$ is independent of Θ and we have already shown the linearity of $\overline{\beta}_{t+1}^{k,\tau}$. For the variance, $\mathrm{Var}[D_{t+1}^{k,\tau}] = \mathrm{Var}[X_{t+1} + \beta_{t+1}^{k,\tau}]$. Note that X_{t+1} is independent of $\beta_{t+1}^{k,\tau}$, and so $\mathrm{Var}[D_{t+1}^{k,\tau}] = \mathrm{Var}[X_{t+1}] + \tilde{\beta}_{t+1}^{k,\tau}$. We have already shown the linearity of $\tilde{\beta}_{t+1}^{k,\tau}$, and $\mathrm{Var}[X_{t+1}]$ is independent of Θ. Hence, both $\mathbb{E}[D_{t+1}^{k,\tau}]$ and $\mathrm{Var}[D_{t+1}^{k,\tau}]$ can be expressed linearly in Θ, which implies, by Equation (9), that the $D_{t+1}^{k,\tau}(i)$ can be expressed linearly in Θ. □

B Linear Approximation of Carryover Variance

Using Proposition 4.1 to exactly linearize the carryover variance requires many binary variables, which results in a computational challenge for today's commercial solvers. For the case studies, we instead use an approximation. Towards understanding the approximation, it is useful to consider two extremes: $\mathbb{P}(D_t^{k,\tau} < \Theta_t^{k,\tau}) \approx 1$ and $\mathbb{P}(D_t^{k,\tau} \geq \Theta_t^{k,\tau}) \approx 1$. If, $\mathbb{P}(D_t^{k,\tau} < \Theta_t^{k,\tau}) \approx 1$, then $\mathrm{Var}[\beta_{t\oplus 1}^{k,\tau}] = \mathrm{Var}[(D_t^{k,\tau} -$

$\Theta_t^{k,\tau})^+] \approx \text{Var}[0] = 0$. On the other hand, if $\mathbb{P}(D_t^{k,\tau} \geq \Theta_t^{k,\tau}) \approx 1$, then $\text{Var}[\beta_{t\oplus 1}^{k,\tau}] = \text{Var}[(D_t^{k,\tau} - \Theta_t^{k,\tau})^+]$ $\approx \text{Var}[D_t^{k,\tau} - \Theta_t^{k,\tau}] = \text{Var}[D_t^{k,\tau}]$. The approximation of $\text{Var}[\beta_{t\oplus 1}^{k,\tau}]$, again denoted $\tilde{\beta}_{t\oplus 1}^{k,\tau}$, linearly weights these extreme cases as follows:

$$\tilde{\beta}_{t\oplus 1}^{k,\tau} := 0 \cdot \mathbb{P}(D_t^{k,\tau} < \Theta_t^{k,\tau}) + \text{Var}[D_t^{k,\tau}]\,\mathbb{P}(D_t^{k,\tau} \geq \Theta_t^{k,\tau}) = \left(\text{Var}[X_t^{k,\tau}] + \tilde{\beta}_t^{k,\tau}\right)\mathbb{P}(D_t^{k,\tau} \geq \Theta_t^{k,\tau}).$$

To achieve this, we define $v_t^{k,\tau}(i)$ to be equal to the DIP variance $(\text{Var}[X_t^{k,\tau}] + \tilde{\beta}_t^{k,\tau})$ on day t if and only if $y_t^{k,\tau}(i,0)$ equals 1 (or equivalently, $D_t^{k,\tau}(i) \geq \Theta_t^{k,\tau}$). The following constraints will assure that this definition is met:

$$v_t^{k,\tau}(i) \leq M \cdot y_t^{k,\tau}(i,0), \tag{32}$$

$$v_t^{k,\tau}(i) \geq \left(\text{Var}[X_t^{k,\tau}] + \tilde{\beta}_t^{k,\tau}\right) - M \cdot \left(1 - y_t^{k,\tau}(i,0)\right), \tag{33}$$

$$v_t^{k,\tau}(i) \leq \left(\text{Var}[X_t^{k,\tau}] + \tilde{\beta}_t^{k,\tau}\right). \tag{34}$$

Then,

$$\tilde{\beta}_{t\oplus 1}^{k,\tau} = \sum_{i\in\mathcal{I}} v_t^{k,\tau}(i) \cdot \Psi(i). \tag{35}$$

We do something similar for $\text{Var}[\alpha_t^{k,\tau}]$. By Equation (3), $\alpha_t^{k,\tau} + \beta_{t\oplus 1}^{k,\tau} = D_t^{k,\tau}$. In general, $\alpha_t^{k,\tau}$ and $\beta_{t\oplus 1}^{k,\tau}$ are not independent. Still, for our linear approximation, we define $\tilde{\alpha}_t^{k,\tau}$ such that $\tilde{\alpha}_t^{k,\tau} + \tilde{\beta}_{t\oplus 1}^{k,\tau}$ equals $\text{Var}[D_t^{k,\tau}]$. The part of the DIP variance that is not allocated to $\beta_{t\oplus 1}^{k,\tau}$ gets allocated to $\alpha_t^{k,\tau}$: $\tilde{\alpha}_t^{k,\tau} = \text{Var}[D_t^{k,\tau}]\,\mathbb{P}(D_t^{k,\tau} < \Theta_t^{k,\tau})$; so:

$$\tilde{\alpha}_t^{k,\tau} = \sum_{i\in\mathcal{I}} \left(\text{Var}[X_t^{k,\tau}] + \tilde{\beta}_{t\oplus 1}^{k,\tau} - v_t^{k,\tau}(i)\right) \cdot \Psi(i). \tag{36}$$

Figure 13 illustrates the idea.

C Notation and Mixed Integer Program

The purpose of this deterministic MIP model is to create a template (Θ), that decides how much capacity is to be reserved for different patient classes on a given day in a given service (i.e., department). Patients are categorized by the service (k) they need and their class (τ). This model will create a cyclo-stationary template – the template does not change from one cycle to the next.

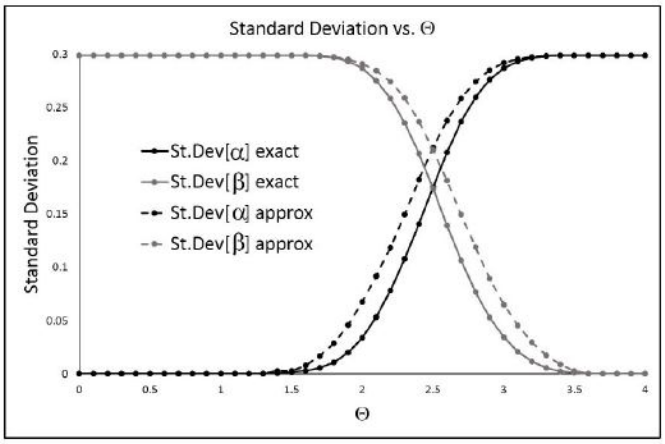

Figure 13: Illustration of the linear approximation of variance. In this example, D is Normally distributed with mean 2.5 and standard deviation 0.3. (Our approximations are validated for a practical setting in the case study presented in section 5.2.) We are interested in the variances of β and α where $\beta = [D - \Theta]^+$ and $\alpha = D - \beta$ (recall Equations (2) and (3)). The exact standard deviations of β and α were calculated using Monte-Carlo simulation, and are plotted vs. Θ. In particular, if Θ is low relative to $\mathbb{E}[D]$, then $\text{Var}[\beta] \approx V[D]$ and $\text{Var}[\alpha] \approx 0$. For the linear variance approximation, $\text{Var}[\beta] = \text{Var}[D]\,\mathbb{P}(D \geq \Theta)$ and $\text{Var}[\alpha] = \text{Var}[D]\,\mathbb{P}(D < \Theta)$. The resulting approximate standard deviations are greater than or equal to the exact values. In general, when Θ is somewhat far away from $\mathbb{E}[D]$, the linear approximations match the exact values, but when Θ is near $\mathbb{E}[D]$, the approximate standard deviations are a bit larger. This means that, under the linear approximation, constraints on the performance metrics are actually a bit conservative.

C.1 Indices

- i index for DIP grid, $i = 1, \ldots, I$.
- j index for number of arrivals per day, $j = 0, \ldots, J$.
- k index for specialty, $k = 1, \ldots, K$.
- τ index for patient class (e.g., Urgent or Non-Urgent), $\tau = 1, \ldots, Y$.
- t index for days, $t = 1, \ldots, T$.
- w index for number of weeks in an itinerary of care, $w = 1, 2, \ldots, Z$.
- h index for number of slots required by a patient, $h = 1, \ldots, H_k$ for department k.
- n number of service level constraints on the access delay, $n = 1, 2, \ldots, N$.

C.2 Parameters/Fixed Input Data

C_t^k	total usable capacity (in hours) on day t in service k.
$\overline{X}_t^{k,\tau}$	$\mathbb{E}[X_t^{(k,\tau)}]$: expected value of the exogenous demand for type (k,τ) on day t.
$\tilde{X}_t^{k,\tau}$	$\text{Var}[X_t^{(k,\tau)}]$: variance of the exogenous demand for type (k,τ) on day t.
$\hat{D}_t^{k,\tau}$	type (k,τ) standard deviation of the DIP on day t under the current system – the guess for the one-step Newton's method approximation in Equation (10).
$m(i)$	number of standard deviations the grid point associated with index i is away from the mean of a given random variable.
M	sufficiently large integer chosen by the user.
$\hat{W}_t^k$	standard deviation of the current system workload for service k – the guess in Equation (28).
$\Psi(i)$	the probability mass for grid point i.
$f_t^{k,\tau}(j)$	the probability that there are j exogenous type (k,τ) arrivals on day t.
$r_{t_1}^{k_1,\tau,k}(m,t-t_1)$	probability that a class τ root appointment in department k_1 on day t_1 will result in m downstream appointment slots t days later in department k.
s_k	length of a time slot in department k in hours.
H_k	maximum number of time slots a patient may need in department k on a given day.

C.3 User Inputs for Possible Constraints

$TFAV_n^{k,\tau}$	n^{th} waiting time target (in days) for patient of type (k,τ).
$p_n^{k,\tau}$	bound set on the expected percentage of type (k,τ) patients exceeding $TFAV_n^{k,\tau}$ days of access delay.
B_k	limit on the non-urgent mean access delay in specialty k.
i^*	grid point that guarantees the desired maximum violation probability of workload exceeding a department's daily capacity (recall Equation (30)).
O_{bound}^k	mean overtime bound set in hours for specialty k.

C.4 Decision Variables

$\Theta_t^{k,\tau}$	number of type (k,τ) reserved slots on day t. (nonnegative, integer)
$\beta_t^{k,\tau}(i)$	conditional overflow demand for type (k,τ) from day $t-1$ to day t of the planning horizon, conditioned on realized DIP at $m(i)$ standard deviations above the mean.
$\epsilon_t^{k,\tau}(i)$	type (k,τ) helper variable that assures that $D_t^{k,\tau}(i)$ is nonnegative at grid level i.
$b_t^{k,\tau}(i)$	binary variable indicating whether the $m(i)$ standard deviations above the mean will cause the DIP to be non-negative.
$y_t^{k,\tau}(i,l)$	binary indicator variable equal to 1 when $D_t^{k,\tau}(i) - \Theta_t^{k,\tau} \geq l$ and 0 otherwise.
$v_t^{k,\tau}(i)$	continuous variable equal to the type (k,τ) total DIP variance on day t when $D_t^{k,\tau}(i) - \Theta_t^{k,\tau} \geq 0$, and equal to 0 otherwise.
$O_t^k(i)$	number of workload hours that are processed in overtime at grid level i for service k on day t.
$\delta_{t,n}^{k,\tau}(i)$	conditional total number of type (k,τ) slots remaining over the next $TFAV_n^{k,\tau}$ days after all prior demand is scheduled, on day t, conditioned on realized DIP at $m(i)$ standard deviations above the mean.
$x_{t,n}^{k,\tau}(i)$	binary helper variable that equals 1 when $\left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t\oplus l}^{k,\tau}\right) - \beta_t^{k,\tau}(i) \geq 0$ and equals 0 otherwise.
$\gamma_{t,n}^{k,\tau}(i,j)$	fraction of patients requesting an appointment of type (k,τ) that exceed $TFAV_n^{k,\tau}$ days of waiting for the appointment, given that there are j type (k,τ) requests on day t, and there are $\delta_t^{k,\tau}(i)$ remaining type (k,τ) slots in the template after all prior demand is scheduled.

C.5 Convenience Variables – Can Be Substituted Out Prior to Optimization

$D_t^{k,\tau}(i)$	conditional type (k,τ) DIP level, conditioned on realized DIP at $m(i)$ standard deviations above the mean.
$\overline{D}_t^{k,\tau}$	type (k,τ) mean DIP on day t.
$\tilde{D}_t^{k,\tau}$	type (k,τ) DIP variance on day t.
$\overline{W}_t^k$	expected workload induced on service k on day t.
$\tilde{W}_t^k$	workload variance in service k on day t.
$\bar{\beta}_t^{k,\tau}$	expected carryover demand of type (k,τ) patients to day t.
$\bar{\alpha}_t^{k,\tau}$	type (k,τ) expected fulfilled/scheduled demand on day t.
$\tilde{\beta}_t^{k,\tau}$	variance for type (k,τ) carryover demand from day $t-1$ to t.
$\tilde{\alpha}_t^{k,\tau}$	variance of fulfilled demand for type (k,τ) on day t.
$G_{t,n}^{k,\tau}$	expected percentage of type (k,τ) patients on day t that will exceed $TFAV_n^{k,\tau}$ days of access delay.
$\bar{O}_t^k$	expected overtime hours of service k on day t.

The objective of this MIP formulation is to minimize the mean access delay of urgent patients in a given department k. (Note that the optimization could also be used to minimize the mean access delay of urgent patients across all departments by adding a sum over k in the numerator and denominator.) The objective function is expression (4):

$$\min \frac{\sum_{t=1}^{T} \overline{\beta}_t^{k,\text{Urgent}}}{\sum_{t=1}^{T} \overline{X}_t^{k,\text{Urgent}}}.$$

The user is able to enter constraints on (i) the mean access delay of non-urgent patients for every

department k (constraints (37)), (ii) the percentage of type (k,τ) patients exceeding $TFAV_n^{k,\tau}$ days of access delay (constraints (38)), (iii) the maximum violation probability of the workload exceeding a department's daily capacity (constraints (39)), and (iv) the maximum mean number of hours performed in overtime in a given department (constraints (40)):

$$\sum_{t=1}^{T} \overline{\beta}_t^{k,\text{Non-Urgent}} \leq B_k \cdot \left(\sum_{t=1}^{T} \overline{X}_t^{k,\text{Non-Urgent}} \right), \forall k; \tag{37}$$

$$G_{t,n}^{k,\tau} \leq p_n^{k,\tau}, \forall k, \forall \tau, \forall t, \forall n; \tag{38}$$

$$\overline{W}_t^k + \frac{1}{2} m(i^*) \cdot \left(\frac{\tilde{W}_t^k}{\hat{W}_t^k} + \hat{W}_t^k \right) \leq C_t^k, \forall k, \forall t; \tag{39}$$

$$\overline{O}_t^k \leq O_{\text{bound}}^k, \forall k, \forall t. \tag{40}$$

Constraints (41)-(45) allow us to precisely define $D_t^{k,\tau}(i)$, and ensure that it is nonnegative:

$$D_t^{k,\tau}(i) = \overline{D}_t^{k,\tau} + \frac{1}{2} m(i) \cdot \left(\frac{\tilde{D}_t^{k,\tau}}{\hat{D}_t^{k,\tau}} + \hat{D}_t^{k,\tau} \right) + \epsilon_t^{k,\tau}(i), \forall k, \forall \tau, \forall t, \forall i; \tag{41}$$

$$\epsilon_t^{k,\tau}(i) \leq -\left(\overline{D}_t^{k,\tau} + \frac{1}{2} m(i) \cdot \left(\frac{\tilde{D}_t^{k,\tau}}{\hat{D}_t^{k,\tau}} + \hat{D}_t^{k,\tau} \right) \right) + M \cdot b_t^{k,\tau}(i), \forall k, \forall \tau, \forall t, \forall i; \tag{42}$$

$$\epsilon_t^{k,\tau}(i) \leq M \cdot (1 - b_t^{k,\tau}(i)), \forall k, \forall \tau, \forall t, \forall i; \tag{43}$$

$$-M \cdot \left(1 - b_t^{k,\tau}(i)\right) \leq \overline{D}_t^{k,\tau} + \frac{1}{2} m(i) \cdot \left(\frac{\tilde{D}_t^{k,\tau}}{\hat{D}_t^{k,\tau}} + \hat{D}_t^{k,\tau} \right), \forall k, \forall \tau, \forall t, \forall i; \tag{44}$$

$$\overline{D}_t^{k,\tau} + \frac{1}{2} m(i) \cdot \left(\frac{\tilde{D}_t^{k,\tau}}{\hat{D}_t^{k,\tau}} + \hat{D}_t^{k,\tau} \right) \leq M \cdot b_t^{k,\tau}(i), \forall k, \forall \tau, \forall t, \forall i. \tag{45}$$

Constraints (46) and (47) define the helper binary variable $y_t^{k,\tau}(i,l)$ to be equal to 1 when $D_t^{k,\tau}(i) - \Theta_t^{k,\tau} \geq l$ and 0 otherwise. We use these binary variables to calculate the $\beta_{t\oplus 1}^{k,\tau}(i)$ (constraints (48) and (49)), which allow us to calculate the mean carryover demand, $\overline{\beta}_{t\oplus 1}^{k,\tau}$ (Equation (50)). Finally (Equation (51)), the mean number of type (k,τ) patients scheduled on a given day t can be calculated using the mean carryover demand and the mean exogenous demand.

$$-M \cdot \left(1 - y_t^{k,\tau}(i,l)\right) \leq D_t^{k,\tau}(i) - \Theta_t^{k,\tau} - l, \forall k, \forall \tau, \forall t, \forall i, \forall l; \tag{46}$$

$$D_t^{k,\tau}(i) - \Theta_t^{k,\tau} - l \leq M \cdot y_t^{k,\tau}(i,l), \forall k, \forall \tau, \forall t, \forall i, \forall l; \tag{47}$$

$$D_t^{k,\tau}(i) - \Theta_t^{k,\tau} \leq \beta_{t\oplus 1}^{k,\tau}(i), \forall k, \forall \tau, \forall t, \forall i; \tag{48}$$

$$\beta_{t\oplus 1}^{k,\tau}(i) \leq D_t^{k,\tau}(i) - \Theta_t^{k,\tau} + M \cdot \left(1 - y_t^{k,\tau}(i,0)\right), \forall k, \forall \tau, \forall t, \forall i; \tag{49}$$

$$\overline{\beta}_{t\oplus 1}^{k,\tau} = \sum_{i \in \mathcal{I}} \beta_{t\oplus 1}^{k,\tau}(i) \Psi(i), \forall k, \forall \tau, \forall t; \tag{50}$$

$$\overline{\alpha}_t^{k,\tau} = \overline{D}_t^{k,\tau} - \overline{\beta}_{t\oplus 1}^{k,\tau} = \overline{X}_t^{k,\tau} + \overline{\beta}_t^{k,\tau} - \overline{\beta}_{t\oplus 1}^{k,\tau}. \tag{51}$$

The optimization uses the helper variables $v_t^{k,\tau}(i)$ that equal the type (k,τ) total DIP variance on day t when $D_t^{k,\tau}(i) - \Theta_t^{k,\tau} \geq 0$, and equal 0 otherwise. This is assured by constraints (52)-(54). The variances of the carryover and fulfilled demands are calculated (Equations (55) and (56)) using $v_t^{k,\tau}(i)$ and the approximation detailed in expressions (32)–(36).

$$v_t^{k,\tau}(i) \leq M \cdot y_t^{k,\tau}(i,0), \forall k, \forall \tau, \forall t, \forall i; \tag{52}$$

$$v_t^{k,\tau}(i) \geq \tilde{D}_t^{k,\tau} - M \times \left(1 - y_t^{k,\tau}(i,0)\right), \forall k, \forall \tau, \forall t, \forall i; \tag{53}$$

$$v_t^{k,\tau}(i) \leq \tilde{D}_t^{k,\tau}, \forall k, \forall \tau, \forall t, \forall i; \tag{54}$$

$$\tilde{\beta}_{t\oplus 1}^{k,\tau} = \sum_{i\in\mathcal{I}} v_t^{k,\tau}(i) \cdot \Psi(i), \forall k, \forall \tau, \forall t; \tag{55}$$

$$\tilde{\alpha}_t^{k,\tau} = \sum_{i\in\mathcal{I}} \left(\tilde{X}_t^{k,\tau} + \tilde{\beta}_t^{k,\tau} - v_t^{k,\tau}(i)\right) \cdot \Psi(i), \forall k, \forall \tau, \forall t. \tag{56}$$

In Equations (57) and (58), we calculate the expected value and variance, respectively, of the total workload on day t for department k. We define Z as the maximum number of weeks an itinerary of care can be for any patient type. (Based on the one year dataset, patients had less than a 0.1% probability of exceeding a three week (15 day) itinerary. Hence, to avoid summing w from 0 to ∞ in our MIP, we truncated this sum at $Z = 2$.) Recall that $\mathbf{e}_k$ is a column vector with all 0's except a 1 in the k^{th} row.

$$\overline{W}_t^k = \sum_{k_1=1}^{K}\sum_{\tau=1}^{Y}\sum_{t_1=1}^{T} \overline{\alpha}_{t_1}^{k_1,\tau} \cdot \sum_{w=0}^{Z}\sum_{h=1}^{H_k} h \cdot r_{t_1}^{k_1,\tau,k}(h, t - t_1 + wT) \cdot \mathbf{e}_k \cdot s_k, \forall k, \forall t; \tag{57}$$

$$\begin{aligned}\tilde{W}_t^k = \sum_{k_1=1}^{K}\sum_{\tau=1}^{Y}\sum_{t_1=1}^{T}\sum_{w=0}^{Z} \Bigg[& \tilde{\alpha}_{t_1}^{k_1,\tau} \left(\sum_{h=0}^{H_k} h \cdot r_{t_1}^{k_1,\tau,k}(h, t - t_1 + wT) \cdot \mathbf{e}_k \cdot s_k\right)^2 + \overline{\alpha}_{t_1}^{k_1,\tau} \\ & \cdot \sum_{h=0}^{H_k} \left(h^2 \cdot s_k^2 \cdot r_{t_1}^{k_1,\tau,k}(h, t - t_1 + wT) \cdot \mathbf{e}_k \left(1 - r_{t_1}^{k_1,\tau,k}(h, t - t_1 + wT) \cdot \mathbf{e}_k\right) \right. \\ & \left. - \sum_{h<q\leq H_k} 2hq \cdot s_k^2 \cdot r_{t_1}^{k_1,\tau,k}(h, t - t_1 + wT) \cdot \mathbf{e}_k \cdot r_{t_1}^{k_1,\tau,k}(q, t - t_1 + wT) \cdot \mathbf{e}_k \right) \Bigg], \forall k, \forall t. \end{aligned} \tag{58}$$

Now, we are able to calculate the amount of workload (in hours) that has to be performed in overtime at grid level i (constraints (59)). Using the Normal distribution assumption on the offered workload, we then compute (Equation (60)) the mean overtime (in hours) on day t in department k.

$$O_t^k(i) \geq \overline{W}_t^k + \frac{1}{2} m(i) \cdot \left(\frac{\tilde{W}_t^k}{\hat{W}_t^k} + \hat{W}_t^k\right) - C_t^k, \forall k, \forall t, \forall i; \tag{59}$$

$$\overline{O}_t^k = \sum_{i\in\mathcal{I}} O_t^k(i) \cdot \Psi(i), \forall k, \forall t. \tag{60}$$

The remaining constraints of the MIP (constraints (61)-(67)) formulate $G_{t,n}^{k,\tau}$, the percentage of type (k,τ) patients arriving on day t that will exceed $TFAV_n^{k,\tau}$ days of waiting for a root appointment.

$$\delta_{t,n}^{k,\tau}(i) \geq \left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t\oplus l}^{k,\tau} \right) - \beta_t^{k,\tau}(i), \forall k, \forall \tau, \forall t, \forall i, \forall n; \tag{61}$$

$$\delta_{t,n}^{k,\tau}(i) \leq \left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t\oplus l}^{k,\tau} \right) - \beta_t^{k,\tau}(i) + M \cdot (1 - x_{t,n}^{k,\tau}(i)), \forall k, \forall \tau, \forall t, \forall i, \forall n; \tag{62}$$

$$\delta_{t,n}^{k,\tau}(i) \leq M \cdot x_{t,n}^{k,\tau}(i), \forall k, \forall \tau, \forall t, \forall i, \forall n; \tag{63}$$

$$\gamma_{t,n}^{k,\tau}(i,j) \geq \left(1 - \frac{\delta_{t,n}^{k,\tau}(i)}{j} \right), \forall k, \forall \tau, \forall t, \forall i, \forall j, \forall n; \tag{64}$$

$$-M \cdot (1 - x_{t,n}^{k,\tau}(i)) \leq \left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t\oplus l}^{k,\tau} \right) - \beta_t^{k,\tau}(i), \forall k, \forall \tau, \forall t, \forall i, \forall n; \tag{65}$$

$$M \cdot x_{t,n}^{k,\tau}(i) \geq \left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t\oplus l}^{k,\tau} \right) - \beta_t^{k,\tau}(i), \forall k, \forall \tau, \forall t, \forall i, \forall n; \tag{66}$$

$$G_{t,n}^{k,\tau} = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \gamma_{t,n}^{k,\tau}(i,j) \cdot \Psi(i) \cdot f_t^{k,\tau}(j), \forall k, \forall \tau, \forall t, \forall n. \tag{67}$$

C.6 Discrete Event Simulation Model for Patient Scheduling

In this section, we present our discrete event simulation that was developed and analyzed using Visual C++. The simulation serves several purposes towards building up our APT decision support framework: (1) In Online Appendix D, we develop an iterative method that integrates the simulation with the template optimization to refine estimates of the DIP distributions that the optimization is built upon; (2) In section 5.2, we use the simulation to validate the analytical approximations of the stochastic parameters and metrics in the optimization; (3) In sections 5.3 and 5.4 we conclude with a case study in which we use the optimization to design improved templates for our partner health system and employ the simulation to calculate the impact of those templates on the competing metrics mentioned above. We first present the simulation dynamics and then discuss model verification and validation.

We describe the timeline for the simulation with the help of Figure 14, which provides a pictorial representation of the sequence of events that occur during the simulation. To begin, the simulation inputs include: (1) a template (Θ) denoting the number of root appointments reserved for each patient class, in each service, for each day of the time horizon to be simulated, (2) the daily resource capacities over the simulation time horizon (C_t^k), (3) the stochastic location functions for downstream appointments based on historical data, (4) the empirical exogenous demand distribution for each $X_t^{k,\tau}$ based

on historical data, and (5) the historical internal referral workload mean and variance from patients starting their itineraries outside of the three departments we consider (GI, GIM, and Neurology). The methods and data we use to calculate (2)–(5) for our partner health system will be described in our case study, in section 5.

The simulation time-step is one day. In Figure 14, the simulation events and timeline are denoted by the text in the boxes with dashed borders (**S1** is the first step/event, **S2** occurs second, etc.). At the beginning of each day, new root appointment requests are generated for each patient type (Step **S1** in Figure 14). Each of these patients is assigned to the first available slot reserved for their type in our template (Step **S2**). In Figure 14, availability of root appointments is indicated by the term A/B, where A is the number of patient slots already scheduled for a root appointment on that day and B is the total number of slots available in that service on the given day. At this point, access delay is calculated for each patient by subtracting the arrival day from the day that they are able to first be accommodated in the scheduling template. After all urgent patients have been assigned a root appointment, the simulation generates realizations of itineraries (i.e., all the downstream appointments generated as a result of the root appointment) for all the patients whose root appointment occurs on the current day of the simulation (Step **S3**). These subsequent visits and internal referrals are scheduled into the remaining slots not reserved for root appointments, with the appointments that exceed the service's total capacity being served through overtime. Hence there is no template capacity listed for the internal referral slots; rather, the number listed in the figure is the total number of internal referrals already scheduled. It is important to note that patients can be scheduled for multiple time slots within the same service. After all appointment requests (root appointments and internal referrals) are accommodated, the simulation clock is updated to the next day (Step **S4**) and the process repeats.

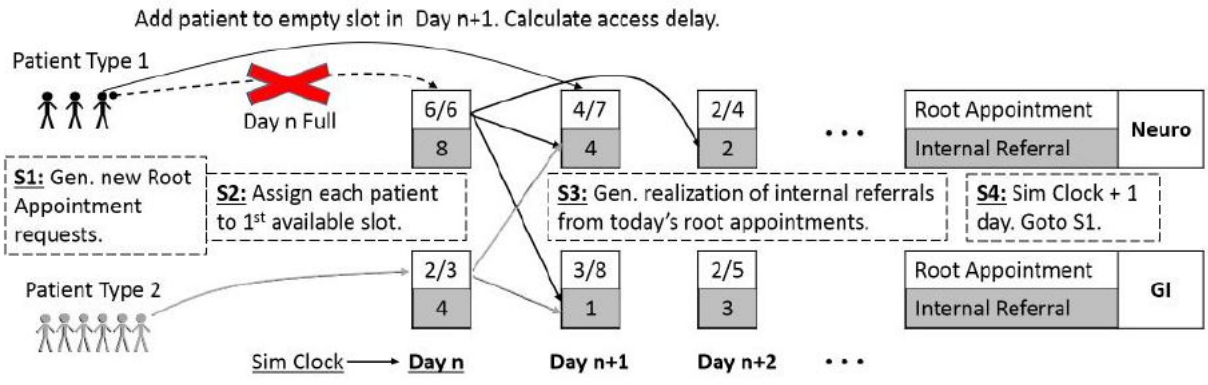

Figure 14: Simulation time-line indicating four steps in each day.

For initial validation, we employed strategies suggested in Sargent (2005) using both black box and

white box testing approaches. We also generated output from each of the steps listed above and tested them to ensure that patient generation, itinerary generation, and workload calculations were occurring as specified. Using Welch's method (Law and Kelton 2000), we determined that a warm-up period of 500 days was sufficient. For computational efficiency, we used the batch means method with 5,000 batches of 50 weeks each. This batch length and number of batches was determined by testing numerous different scheduling templates and ensuring that four key metrics (DIP mean, DIP variance, carryover mean, carryover variance) from which the other metrics are calculated had sufficiently tight 95% confidence intervals. For example, for the DIP mean the 95% confidence interval for each template tested was smaller than 1 appointment slot.

In the following sections, we first present our model for characterizing demand in progress, which is used to capture the key features of importance to our industry partner: access delay and workload requirements (and by corollary overtime and utilization).

D DIP Distribution Adjustment Combining Optimization and Simulation

While representing demand in progress using a Normal distribution enables tractable optimization and is often a good approximation (Figure 3a) this is not always the case (Figure 3b). However, we find that an iterative adjustment to the DIP probability masses that integrates the simulation with the optimization can significantly improve the accuracy of the analytical probability mass approximation. In this section, we introduce an algorithm that is used to test and adjust the DIP probability mass approximation and produces accurate results in few iterations for test cases from our case study (section 5.2).

The algorithm starts with the assumption of a Normally distributed DIP distribution. An optimal template is generated using APT. This template is then simulated to obtain the true probability masses of the DIP distribution. If the approximate probability masses from the optimization are close enough to the simulated probability masses, then the template was designed using an accurate representation of the true DIP probability masses. If not, we update the DIP probability mass estimates using the simulated probability masses and rerun the optimization, repeating the procedure above.

In this algorithm we define $\Psi(i)[n]$ as the probability mass associated with grid point i (i.e., the probability mass of the DIP random variable that lies in between $\mu + m(i) \cdot \sigma$ and $\mu + m(i+1) \cdot \sigma$) during the n^{th} iteration of the algorithm. This is the probability mass that is used as a parameter input to the optimization during the n^{th} iteration. (Since the probability masses may now vary by department (k), patient class (τ), and day (t), we really have $\Psi_t^{k,\tau}(i)[n]$; but, we suppress the subscript and superscripts.)

Likewise $\Psi^S(i)$ is the actual probability mass associated with grid point i as determined by the discrete event simulation (section C.6)

0.	n=0. Set ϵ approximation error tolerance.
	Initialize $\Psi(i)[0]$ with the probability masses of the Standard Normal.
1.	Input $\Psi(i)[n]$ into the APT optimization and obtain optimal template Θ^*
2.	Simulate Θ^* and obtain simulated grid point probability masses $\Psi^S(i)$.
3.	IF $\lvert\Psi^S(i) - \Psi(i)[n]\rvert < \epsilon$ THEN terminate algorithm ELSE GOTO 4.
4.	$\Psi(i)[n+1] = \Psi^S(i)$
5.	$n = n + 1$. GOTO 1.

Table 8: Algorithm for obtaining accurate DIP distribution approximations.