Capacity Allocation Optimization Method for Itinerary Completion in a Destination Outpatient Clinic

<table>
<thead>
<tr>
<th>Journal</th>
<th>IIE Transactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>Draft</td>
</tr>
<tr>
<td>Manuscript Type</td>
<td>Regular Paper</td>
</tr>
<tr>
<td>Focus Issue</td>
<td>Operations Engineering</td>
</tr>
</tbody>
</table>
Capacity Allocation Optimization Method for Itinerary Completion in a Destination Outpatient Clinic

Abstract: Destination medical centers are comprised of an integrated suite of specialist services that provide care for a heterogeneous population of patients with different time sensitivities to completing diagnosis and treatment planning, which is called an itinerary of care. In collaboration with one such organization, we develop an earmarked capacity plan that ensures patient itineraries complete before priority-specific deadlines. We develop methods for designing a time-varying capacity allocation scheme for optimizing itinerary completion delays based on patient type and priority. The stochastic optimization that drives this scheme has a large state space, as well as an objective and constraints that are probabilistic and non-convex. We develop tractable solution approaches by decomposing the problem into two stages and then transforming non-linear stochastic patient flow network models into a tractable deterministic optimizations and applying new linearizing approximations. The methodology is demonstrated through an application to improving on-time service completion for national and international patients at a destination medical center. This is a significant factor in these high priority patients’ satisfaction with their visits. A case study applying our methods at the Mayo Clinic demonstrates a potential of up to 14% improvement in on-time completion.

Keywords: patient access management, queueing networks with deadlines, matrix geometric methods, History: New Submission.

1 Introduction
As health care moves toward more consolidation, the resulting outpatient healthcare networks will serve a diverse patient population in terms of conditions treated but also in terms of geography and medical urgency. The Mayo Clinic is an example of such a destination clinic, where patients travel from all over the world to receive diagnosis and treatment from an integrated suite of specialty services. This growing business model leads to new considerations for access to care, as different patients have different time sensitivities (deadlines) to completing all the steps in their diagnosis and treatment process. The care process of diagnosing and developing a treatment plan, often involving several different specialist services, is typically called the patient’s itinerary. Managing itinerary completion deadlines that are differentiated by patient type creates a need for a more sophisticated approach to managing capacity that transcends the first-come-first-served approach that is the prevailing standard in the outpatient environment.

Two examples from a partner health system illustrate this point. First, patients who present with a suspicion of breast cancer need to go through diagnostic testing, surgery consultation, and surgery quickly. These cancer patients will be much more time sensitive compared to patients with non-malignant cysts or receiving routine breast exams. Second, at a destination clinic patients who travel long distances to receive care need to complete their itinerary faster because they cannot easily travel back and forth to the clinic. Appointment delays can cause extended stays, requiring them to stay on or near the medical campus, forcing patients to take more time off of work, disrupting their daily lives, taking them away from family and friend support structures, and incurring financial costs. For the destination clinic model to be successful, it is imperative to design operational mechanisms that allow time-sensitive patients to flow unimpeded through the...
integrated network of services required to diagnose and treat their conditions while maintaining sufficient access for less time-sensitive patients. To do so, we develop an optimization method that controls only the first appointment(s) of a patient’s itinerary, which can be booked in advance. These appointment(s) that are pre-booked represent the beginning of an itinerary and are called *initial appointment(s)*. Appointments that are generated after consulting with clinicians and performing diagnostics during the initial appointment(s) are called subsequent visits, which are modeled as a stochastic process generated by the initial appointment(s).

We demonstrate how our optimization can be used to control the mix and timing of initial appointments to maximize the number of time-sensitive patients that can meet their *itinerary completion* deadlines at a destination hospital. The optimization controls the number of new initial appointments across a network of services required during a patient’s itinerary. This metric of itinerary completion within a queueing network differentiates our work from other literature on outpatient scheduling, since previous approaches typically seek to minimize delays to start treatment rather than time to complete treatment. As opposed to the single appointment model typically studied in outpatient scheduling, the queueing network dynamics where patients have multiple, sequential appointments, each of which must be completed before the patient can be given a diagnosis and treatment plan makes itinerary completion important in our context. Delays at any step in the itinerary, which depend on the number of initial visits scheduled in each service on each day, can delay the entire itinerary.

We contribute to the literature by endogenizing the itinerary completion metric as a function of the capacity allocated across the queueing network. By doing so, the objective of maximizing itinerary completion becomes non-linear in the capacity decision variable. This non-linearity, combined with the large state space of the queueing network problem, makes traditional optimization methods intractable. As will be demonstrated, capturing the full distribution on the discrete-time itinerary completion times in a queueing network with time-varying arrivals and capacities is challenging in itself. In this work, we contribute to the literature by developing a phase-type representation of these sojourn times and presenting an alternate, compact form of the traditional phase-type generator that ensures computational tractability. To handle the non-linearity, we develop an heuristic decomposition method and accompanying linearizing approximations that enable the application of linear programming for determining optimal capacities at each service by day of the week.

Through our case study at the Mayo Clinic we develop several encouraging managerial insights. First, we show that significant benefits can be gained by controlling only a small fraction (< 25%) of all the patients that will be using shared network resources. This is important because a pilot often involves a single service or small subset of services that are willing to take the risks incurred by an early adopter. Thus our results indicate that such a pilot will likely be able to demonstrate success to the larger organization even without initial buy-in from other services that share the pilot service’s network resources. Second, we show that constant staffing over time enables
optimization to achieve greater benefit than time-varying staffing. This provides support for
a conjecture proposed by our collaborators at Mayo Clinic that, over the course of the week,
staffing across the service network should be level. Finally, we demonstrate that moving time-
sensitive patients earlier in the week is beneficial, but an unforeseen result is that Mondays
should typically be avoided because pushing low priority patients to the end of the week results
in increased demand for follow-up appointments on Monday.

In Sec. 2, we review the literature. Sec. 3 develops a capacity allocation optimization and
presents a decomposition method that provides a tractable solution. In Sec. 4, we perform a case
study based on data from the Mayo Clinic’s breast cancer services. Sec. 5 concludes the paper.

2 Patient Flow and Optimization in Stochastic Networks

Much of the outpatient scheduling literature focuses on a single clinic in which patients arrive
for a single appointment and waiting times are captured using queueing models (see the survey
paper Cayirli and Veral (2003)). Deglise-Hawkinson et al. (2016) focuses on capacity planning in
an integrated care environment, assuming that all subsequent appointments can be completed in
overtime if necessary. This work focuses on time to obtain an initial appointment. In our model
of integrated care, we are concerned instead with the time to complete an entire itinerary that
involves multiple visits across a network of specialist services, where the itinerary unfolds over
multiple days, endogenizing delays along a patient’s itinerary, rather than assuming subsequent
visits can be completed in overtime. Under the assumption of unlimited overtime, the question
of itinerary completion is moot as all subsequent appointments could potentially be scheduled
on the same day as the original appointment. However, we know from our collaboration with
the Mayo Clinic that itinerary completion is indeed a significant problem.

Other approaches focus on patient scheduling with patient no-shows, doctor availability, or
other capacity and workload related features (see Cayirli and Veral (2003)). A number of recent
papers have considered priority scheduling with respect to dynamic scheduling problems with
patients of different priority levels that may consume single or multiple resources with holding
costs for delays or overtime costs (see Patrick et al. (2008), Erdelyi and Topaloglu (2010), Goc-
gun and Ghate (2012)). Some papers have considered more complex treatment itineraries with
multiple appointments over multiple days with scheduling time windows, focusing on dynamic
scheduling (see Sauré et al. (2012), Gocgun and Puterman (2013)).

In our setting patient itineraries are not known in advance, so we focus on static capacity
allocation rather than dynamic scheduling of individual appointments. Specifically, we optimize
capacities for different priority patients on clinicians’ schedules in advance of observing appoint-
ment requests, with the objective of improving the expected rate at which itineraries will be
completed before a deadline. From a modeling perspective, this work differs from the works
mentioned in that only a patient’s initial visit is controllable with downstream appointments
following a stochastic trajectory through a network of services over a time. We develop models
for downstream appointment delays not considered in previous work. Finally, rather than consid-
ering holding times or scheduling windows to start a job, we consider priority specific deadlines
to complete a job (which may differ depending on the day of arrival); e.g., high priority patients
must complete their itinerary by Friday afternoon regardless of their day of arrival.

The literature on inpatient scheduling also considers networks of services like the one in
our setting. Approaches to model complex care trajectories in these networks of services have
included simulation (see Hancock and Walter (1983), Lim et al. (1975)) and stochastic modeling
(see Connors (1970)) and optimization Gallivan and Utley (2005), Chow et al. (2011), Adan et al.
(2009), Bekker (2011), Helm and Van Oyen (2014). These works focus on scheduling to reduce
workload variability and probability of exceeding capacity. Blocking within the network (during
the patient’s treatment episode) and treatment delays have not been considered. Specifically,
the time to complete treatment is assumed to be exogenous to the scheduling policy, whereas we
endogenize this metric into our network model.

Finally, there have been numerous results on sojourn times in queueing networks, which
we refer to here as itinerary completion. These results are surveyed by Boxma and Daduna
(1990) and Disney and Konig (1985). According to these surveys, for a quasi-overtake free path
in networks with Poisson arrivals, exponential service times, multiple server nodes, multiple
customer classes, sojourn time has product form. Results also exist for general service times
Reiman (1984) and BCMP networks Kamed Ha and Zhang (1995). However, it is well-established
that few results exist for non-overtake-free networks Boxma and Daduna (1990). Our system
does not have a BCMP network structure, arrivals are not time stationary, and capacity is also
not time stationary. These features, to the best of our knowledge, have not been considered in
the literature. Hence we require new approximation methods for determining the distribution
on sojourn times in our context.

The modeling contributions of this work include (1) developing new phase-type blocking mod-
eels for patient sojourn times in queueing networks for which service time is not continuous, but is
characterized by the number of shifts/days interrupted by system “downtime” until service can
occur, (2) developing a compact representation of the maximum of n phase-type random variables
leveraging the special structure of our flow problem, (3) a decomposition method to eliminate
non-linearities in sojourn times, (4) new approximations that linearize blocking calculations in
queueing networks and can be used in an LP framework.

In the next section, we develop and analyze our analytical framework for statistically charac-
terizing patient flow trajectories and using them to build a workload profile for the entire network
of clinical services as a function of the capacity plan.

3 Priority Capacity Allocation Model for Patient Care Networks
Let D be the set of patient types and U the set of services offered in the care network. Let
θ^H_{d,k} be the decision variable for the number of appointment-slots reserved for high priority (e.g.,
national/international) type k (e.g., diagnosis k) initial appointments on day d. Similarly let
θ^L_{d,k} represent the low priority (e.g., local/regional) patients. We assume that these slots for new

For ScholarOne support, you may contact 434-964-4100 or 888-503-1050 (US-based numbers).
patients can always be filled due to high demand and long waiting lists (often on the order of 6-12 weeks) for the modeled services. Finally, let $C_{u,d}$ be the aggregate maximum number of appointment slots available in service $u$ on day $d$ (based on available staff and care resources) and let $\theta_k^\ell$ be the weekly throughput requirements for type $k$ patients of priority $\ell \in \{H, L\}$. Program 1 formulates an optimization model that minimizes the expected number of high priority patients that fail to complete their itinerary before Friday, which is a key service metric for our partner health system, since limiting weekend waiting is important to patient satisfaction.

**PROGRAM 1:**

$$\min \sum_{k \in \mathcal{D}} \sum_{d=1}^5 \Theta^H_{k,d} \mathbb{P}_{C,\Theta}(\text{Length of type k itinerary starting on day } d \geq 6 - d) \quad (1)$$

$$s.t. \sum_{d=1}^5 \Theta^\ell_{k,d} \geq \theta_k^\ell \quad \forall k \in \mathcal{D}, \ell \in \{H, L\} \quad (2)$$

Eq. 1 minimizes the expected number of deadlines missed. Note that because itinerary delays have been endogenized, $\mathbb{P}(\text{Length of type k itinerary starting on day } d \geq 6 - d)$ is a function of decision variables $\Theta$ and the exogenous capacity limit matrix $C = \{C_{u,d}\}$. This leads to a non-linear (and in fact non-convex) objective. Eq. 2 maintains throughput targets for both high and low priority patients, ensuring that both patient types are still allocated the same number of appointment slots as they have historically used. This approach keeps the model from negatively impacting new patient access of low priority (less time-sensitive) patients. As will be shown, the size and complexity of the problem, combined with the discrete nature of the appointment reservation system (appointments that exceed capacity overflow to the next day) make this program unsuitable for conventional optimization techniques.

Recall from the discussion of Sec. 1 that our industry partner has analyzed the problem and found two key contributors to itinerary delays were (1) overutilization of key services, and (2) no staffing on the weekends. We leverage this insight to heuristically decompose the problem defined in Program 1 into a two stage problem, where each stage addresses one of the key problems. The first stage adjusts capacity allocation to improve the timing of new patient admissions in a manner that minimizes internal blocking across all services in the network due to overutilization of resources caused by subsequent itinerary visits generated after the initial appointment.

Based on the first stage optimal schedule, $\Theta^\ast$, we are able to obtain off-line the blocking probabilities, $\beta_{u,d}$, for each service $u$ and day $d$. The second stage takes the blocking probabilities, $\mathcal{B} = \{\beta_{u,d}\}$, calculated from the first stage as exogenous inputs. We then develop a phase-type model to capture sojourn times of an individual patient moving through the network as their itinerary of care evolves. Because of the discrete time appointment system, appointments in the itinerary that cannot be serviced on a particular day must try again to get an appointment on the following day. That is, for each visit in the itinerary, the patient is able to either obtain their next visit in service $u$ on day $d$ (w.p. $1 - \beta_{u,d}$) or is blocked (w.p. $\beta_{u,d}$) and must try again on the following day. Our phase-type model captures this blocking mechanism and can be

For ScholarOne support, you may contact 434-964-4100 or 888-503-1050 (US-based numbers).
used to compute the sojourn time through the congested network with time-dependent blocking probabilities $B$. Due to the underlying structure of the model, we find that there are many different configurations of capacity allocation that can achieve (very nearly) the same level of blocking. Leveraging this fact, we design a second stage optimization that solves the second key problem with itinerary completion: weekend delays. Specifically, we reoptimize the capacity allocation to maximize the expected number of patients that can complete their itinerary before the weekend under the condition that blocking is no greater than in the optimal solution to stage 1. By doing so, this allows us to calculate the sojourn time off-line (independent of the decision variable) because the blocking dynamics in all feasible solutions will be the same as in the first stage problem; essentially decoupling the sojourn time distribution from the capacity decision variables. This removes the non-linearity caused by endogenizing the sojourn times into the network capacity model and allows for tractable solutions even for very large networks.

3.1 Model for Patient Flow in an Integrated Care Network

In this section, we develop an offered load model of patient flow to characterize the load on each of the clinical services over time. We develop an aggregate workload model by first building a model of the care pathways for each type of patient (i.e., probabilities of whether or not each service is required at a given time after the initial appointment). Here, patient type refers to the patient’s primary reason for visiting the organization (i.e., primary diagnosis) and time-sensitivity (which has an effective on the quantity and type of services required), though it can be general. We then superimpose capacity constraints on the workload process to approximate the blocking in each service by day of the week.

3.1.1 System Design and Modeling Assumptions.

The system model presented here functions by reserving a certain number of appointments of each type to each service by day of week with a repeating weekly schedule. This is to illustrate the application and is not a modeling limitation. In practice, physicians prefer repeating schedules to accommodate their varied activities, such as research and operating room time. Another mild assumption is that patient care paths are independent of one another. This is true except in the case of blocking. To capture blocking effects, we later superimpose capacities on our offered load model – similar to the modified offered load approach proposed by Massey and Whitt (1994). In several instances, we employ the normal approximation for the number of appointments requested on a given day. This approximation can be justified by the central limit theorem, but also has been validated for similar healthcare applications (see Connors (1970), Isken (2002)).

3.1.2 Patient Flow Stochastic Location Process.

In this section we develop a stochastic location process to model patient flow through the network of clinical services. This process forms the basis of the stochastic-arrival-location model of clinical service workload. The vector state space for the stochastic location process is $S^0 = \{[x_1, x_2, \ldots, x_M] : x_i \in \mathbb{Z}^+ \forall i\}$, where $x_i$ is the number of appointments required in service $i$. This is a deviation from the stochastic location models employed in the research mentioned previously, where the state space was a scalar representing the location of the patient.
The full state space is then \( \mathcal{S} = \mathcal{S}^0 \cup \{ \Delta \} \), where \( \Delta \) represents that the patient is at home. In the outpatient setting there are two major differences that require this modified formulation. First, a patient can have more than one appointment at a given clinical service in a day. Second, a patient could have appointments at multiple clinical services on a single day, requiring the vector state space. The stochastic location process is an \( \mathcal{S} \)-valued function, \( L_{s,k}(t) \) for \( t \in \mathbb{R} \), where \( s \) is the time that the patient started treatment and \( k \in \mathcal{D} \) is the patient type. \( L_{s,k}(t) \) takes outcomes in \( \Sigma_s \), the set of right-continuous functions with left limits that map \( \mathbb{R} \) to \( \mathcal{S} \). \( \Sigma_s \) must contain the care paths of all patients who start new treatment at the clinic at time \( s \), so we require for any function \( \sigma_s \in \Sigma_s \), that \( \sigma_s(t) = \Delta \ \forall t < s \) and \( \sigma_s(s) \in \mathcal{S}^0 \). The function space \( \Sigma \) is the collection of all \( \Sigma_s \). We define a probability measure on \( \Sigma \) that is associated with the stochastic location process, \( \mathbb{P}_s : \Sigma \rightarrow [0, 1] \), that assigns 0 probability to \( \Sigma_t \) for \( t \neq s \) and for \( \Gamma \subseteq \Sigma_s \), it assigns the probability associated with the set of location functions \( \Gamma \).

The process \( L_{s,k}(t) \) will be generated from data to capture historical pathways and the correlations that they contain. As an example, consider the set \( \Gamma_t^{u,2} \equiv \{ \sigma : \sigma \in \Sigma_s \text{ for } s < t, e_u \sigma(t) = 2 \} \) where \( e_u \) is the unit vector with all 0’s and a 1 in the \( u \)th column. In words, \( \Gamma_t^{u,2} \) is the set of all location processes in which the patient requires two appointments in clinical service \( u \) at time \( t \). Therefore, \( \mathbb{P}_s(\Gamma_t^{u,2}) \) is the probability that a patient who initiates a new treatment at time \( s \) requires two appointments in clinical service \( u \) at time \( t \). In our application, time \( t \) is in days making it feasible to forecast and manage appointments. For notational convenience we let \( \mathbb{P}(L_{s,k}(t) \cdot e_u = m) = p_{s,k,u}(m, t - s) \) be the probability that type \( k \) patient starting treatment on day \( s \) requires \( m \) appointments on day \( t \).

### 3.1.3 The Clinical Service Workload Process

To characterize the workload in terms of number of appointments by day of the planning horizon, we combine the stochastic location process from Sec. 3.1.2 with a controlled arrival/admission stream, \( \Theta \). In our model, we make the reasonable assumption that \( \Theta \) is deterministic, since high demand and long waiting lists in the services modeled at our partner health system ensure that capacity reservations can almost always be filled. We consider a repeating planning horizon of 5 days to match the Mayo Clinic’s work week. For any given capacity allocation, \( \Theta \), the demand for clinical service \( u \) on day \( d_1 \) of week \( t \), \( D_{u,d_1}^t \), is given by Eq. 3 and the cyclostationary steady state demand is likewise given by Eq. 4, as will be explained.

\[
D_{u,d_1}^t(\Theta) = \sum_{d_2=1}^5 \sum_{k \in \mathcal{D}} \sum_{j=1}^{\Theta_{k,d_2}} \sum_{n=0}^{t} e_u \cdot L_{d_2+5n,k}^{j,n}(d_1 + 5t) \tag{3}
\]

\[
D_{u,d_1}^\infty = \lim_{t \to \infty} D_{u,d_1}^t. \tag{4}
\]

\( L_{s,k}^{j,n}(\cdot) \) is the \((j,n)\)th i.i.d. instance of the location process \( L_{s,k}(\cdot) \), representing a patient who was the \( j \)th patient of type \( k \) that was scheduled on week \( n \). The first sum in Eq. 3 is over the days of the planning horizon, \( d_2 \), the second is over all patient types, \( k \), the third is over the number, \( j \), of patients of type \( k \) that were scheduled for their first appointment on day \( d_2 \), and the fourth
is over weeks, \( n \), from 0 to \( t \).

A key to our first stage optimization is that the moments of the process can be calculated analytically and linearly. Let \( M_u \) denote the maximum number of appointments an individual patient could require in clinic \( u \) on a given day, which can be determined using historical data. It is possible to develop a one-to-one mapping from the vector function space to a larger scalar function space that admits a \( J_1 \) Skorohod topology and removes measure theoretic concerns surrounding arrival-location models. Therefore, the mean workload in service \( u \) on day \( d_1 \) follows from the monotone convergence theorem and Eq.’s 3 and 4, where:

\[
\mu_{u,d_1}(\Theta) = E \left[ \sum_{d_2=1}^5 \sum_{k \in D} \sum_{j=1}^c \lim_{t \to \infty} \sum_{n=0}^t e_u \cdot L_{d_2+5n,k}^{j,\infty}(d_1+5t) \right] \\
= \sum_{d_2=1}^5 \sum_{k \in D} \Theta_{k,d_2} \cdot \sum_{n=0}^\infty \sum_{m=1}^{M_u} m \cdot p_{d_2+5n,k,u}(m, d_1 - d_2 + 5(t-n)). \tag{5}
\]

Eq. 5 is linear in the decision variable \( \Theta \). The variance is also linear in the decision variable as shown in the following theorem, which is an extension of Helm and Van Oyen (2014) to vector location functions. For notational convenience let \( \hat{p}_{d_1,d_2,u}^{n,k,m,t} = p_{d_2+5n,k,u}(m, d_1 - d_2 + 5(t-n)) \). It suffices to compute the finite and infinite horizon variance as follows with proof given in Online Appendix A:

**Theorem 3.1.** For deterministic arrival streams, the variance in number of appointments requested (offered load) for resource \( u \) on day \( d_1 \) is given by

\[
\sigma^2_{u,d_1}(\Theta) = \sum_{d_2=1}^5 \sum_{k \in D} \Theta_{k,d_2} \lim_{t \to \infty} \sum_{n=0}^t \sum_{m=1}^{M_u} m^2 \cdot \hat{p}_{d_1,d_2,u}^{n,k,m,t}(1 - \hat{p}_{d_1,d_2,u}^{n,k,m,t}) - \sum_{q>m} 2m \cdot q \cdot \hat{p}_{d_1,d_2,u}^{n,k,m,t} \cdot \hat{p}_{d_1,d_2,u}^{n,k,q,t}. \tag{6}
\]

### 3.1.4 Validation of the Workload Process Model using Historical Data.

Stochastic location models have been proven effective in accurately modeling and forecasting workloads in a hospital inpatient setting (see Helm and Van Oyen (2014)). In this section, we show that the same holds for modeling appointments in an outpatient setting at the Mayo Clinic. The validation experiment was designed by first dividing Mayo Clinic’s data on breast cancer patients into a training set and a test set. The training set, consisting of two years of appointments, was used to empirically estimate the stochastic location process’ distribution by calculating the fraction of patients that required \( m \) appointments in service \( u \), \( d \) days after their initial appointment for all \( u \in U, \ m = 1, \ldots, M_u \), and days \( d = 1, 2, \ldots \) for each patient type. The test set was the next 6 months of data. We combined the actual number of new breast cancer patient arrivals (from historical data) with the stochastic location process to predict the number of appointments across the suite of services used by breast cancer patients. We then calculated the absolute error between the model and the actual number of appointments and averaged the absolute error across all the days of the 6 month testing horizon. The average absolute error between predicted and actual workloads across the 5 key breast cancer services varied between...
0.49 appointments for a smaller service to 2.4 appointments for a larger service, with 4 of the 5 services having errors less than 1 appointment per day on average.

To test the variance calculation we also compared the 75% quantile predicted by the model with the actual 75% quantile in the historical data. We do so because the quantiles are used later to capture the blocking metric, so if the model can accurately predict the quantiles it is capturing the variability in the way we wish to use it. The quantiles for the analytical model were calculated by approximating the workload by a Normal random variable (RV) and calculating the quantile using the formula $\mu_{u,d}(\Theta) + 0.6745 \cdot \sqrt{\sigma_{u,d}^2(\Theta)}$ for the 75% workload quantile in service $u$ on day $d$, where $\mu_{u,d}(\Theta)$ and $\sigma_{u,d}^2(\Theta)$ are calculated using Eq. 5 and 6 respectively. At the 75% quantile the average absolute errors ranged from 0.2 appointments to 1.4 appointments with an average across the 5 services of 0.82. These errors are small enough for the level of accuracy needed for this decision support system. Additionally, this analysis should assuage concerns about using an offered load model in a capacitated system, since both the mean and quantile approximation closely match the true workload in the true capacitated system. This analysis also validates the Normal workload approximation that we later use in our optimization models.

3.2 Stage 1: Workload and Blocking Optimization

The workload smoothing stage develops a capacity allocation that minimizes blocking (i.e., the event where a patient can’t get an appointment on the day they request) over all clinical services across the week. To accomplish this, we develop analytical methods for calculating the blocking at each service based on the workload model from Sec. 3.1. The development and linearization of the blocking calculation are presented in Sec. 3.2.1. The framework is then incorporated into a linear program that minimizes blocking in Sec. 3.2.2.

3.2.1 Linearized Blocking Calculations for Optimization Models.

Blocking in queueing networks is non-linear and typically difficult to compute. To overcome this, we approximate the workload in each clinical resource by a Normal distribution (see 3.1.1 for reasoning) with mean and variance parameters given by Eq’s 5 and 6. The Normal workload approximation has been used effectively in many healthcare settings (see for example Isken (2002), Connors (1970)). The expected number of blockages and the blocking probability for Normal workloads is non-linear in the patient schedule. To approximate the expected overflow linearly, let $M' = \{1, 2, \ldots, N\}$ be an index that creates an $N$-point discrete grid $M$ approximating the real line with respect to the workload distribution to be approximated. The one-to-one function $m(i) : M' \rightarrow M$ maps the integer values of $M'$ to the grid values $M$. Table 1 and Fig. 1 show an example of a grid mapping with constant probability intervals.

<table>
<thead>
<tr>
<th>$i \in M'$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m(i) \in M$</td>
<td>$-\infty$</td>
<td>-0.4</td>
<td>-0.2</td>
<td>-0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.9</td>
<td>1.2</td>
<td>1.5</td>
<td>1.8</td>
<td>2.2</td>
<td>2.6</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

Table 1: Sample grid mapping from the integers to the grid $M = \{-0.6, -0.4, -0.2, -0.1, 0, 0.1, 0.2, 0.4, \ldots, 3.1\}$

Fig. 1 (a) shows a grid design where each interval contains an equal amount of probability density. Fig. 1 (b) shows how the grid is used to approximate expected overflow. On the $m(i)$
interval, the realized overflow is approximated for each grid point as \((\mu_{u,d}(\Theta) + m(i)\sigma_{u,d}(\Theta) - C_{u,d})^+\). The realized value for gridpoints 1, 2, 3, and 4 is represented by the 4 lines on the graph in Fig. 1 (b), where the capacity is given by the flat dashed line. The positive difference between the gridpoint realization line and the capacity line is the amount of overflow associated with that gridpoint. The realized overflow value is multiplied by the probability mass contained within interval, \(\Phi(m(i)) - \Phi(m(i-1))\) with \(m(0) \equiv -\infty\), where \(\Phi(\cdot)\) is the Normal cdf. Thus the expected overflow is approximated by \(E[\text{Overflow}] \approx \sum_{i=2}^{N} (\Phi(m(i)) - \Phi(m(i-1))) \cdot (\mu(\Theta) + m(i)\sigma(\Theta) - C)^+\). In the limit, as \(N \to \infty\) this approximation will converge to the exact overflow value.

All that remains is to design a set of constraints in the linear program that will set the overflow variable for the workload realization at gridpoint \(i\) in service \(u\) on day \(d\), \(\delta_{u,d,i}\), to the correct value. The following constraint set, when combined with an objective function that minimizes expected overflow (and hence \(\delta_{u,d,i}\)) will achieve this result.

\[
\mu_{u,d}(\Theta) + m(i) \cdot \sigma_{u,d}(\Theta) - C_{u,d} \leq \delta_{u,d,i} \quad \text{for } i \in M' = \{1, 2, 3, \ldots, N\}. \tag{7}
\]

In the above approximation, \(\sigma(\Theta)\) is still non-convex in the decision variable \(\Theta\) because it is the square root of the variance. However, from Eq. 6 the variance, \(\sigma^2(\Theta)\), of the workload can be calculated linearly in \(\Theta\). We propose to approximate the square root of \(\sigma^2(\Theta)\) with a new approach based on Newton’s method. Letting \(\hat{\sigma}\) be an initial guess for the standard deviation, the one-step Newton’s method is given by \(\sqrt{\hat{\sigma}^2(\Theta)} \approx \frac{1}{2} \left( \frac{\sigma^2(\Theta)}{\hat{\sigma}} + \hat{\sigma} \right)\). In our application, a high level of accuracy can be achieved if \(\hat{\sigma}\) is set to the standard deviation of the historical workload of the current system because the specific mix of patients does not greatly affect the workload variability given that the average throughput remains constant (which is what we want). To validate this claim, Table 2 demonstrates the quality the approximation for the four most common breast cancer services (using separate training and test sets) considering two different types of schedules: (1) Likely Case: The workload is statistically identical across all days the week, (2) Worst Case: The entire patient load is scheduled on the first day of the week, in an effort to

For ScholarOne support, you may contact 434-964-4100 or 888-503-1050 (US-based numbers).
bound the worst case error. The worst case (2) will almost certainly not be optimal so the error here is less important. Based on numerical experiments, the optimal schedule is much more likely to resemble the likely case (1), a balanced schedule, so the fact that the approximation errors are negligible indicates that this approximation is a good choice for our problem.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Actual</td>
<td>Approx</td>
<td>Actual</td>
<td>Approx</td>
</tr>
<tr>
<td>All</td>
<td>Likely</td>
<td>5.50</td>
<td>5.50</td>
<td>4.04</td>
<td>4.04</td>
</tr>
<tr>
<td>Mon</td>
<td>Worst</td>
<td>5.59</td>
<td>5.59</td>
<td>5.95</td>
<td>6.47</td>
</tr>
<tr>
<td>Tue</td>
<td>Worst</td>
<td>5.44</td>
<td>5.44</td>
<td>3.68</td>
<td>3.70</td>
</tr>
<tr>
<td>Wed</td>
<td>Worst</td>
<td>5.37</td>
<td>5.38</td>
<td>3.44</td>
<td>3.50</td>
</tr>
<tr>
<td>Thu</td>
<td>Worst</td>
<td>5.56</td>
<td>5.56</td>
<td>3.20</td>
<td>3.29</td>
</tr>
<tr>
<td>Fri</td>
<td>Worst</td>
<td>5.51</td>
<td>5.51</td>
<td>3.25</td>
<td>3.31</td>
</tr>
</tbody>
</table>

Table 2: Comparison of actual standard deviation with Newton’s method-based approximation for a likely case and the worst case by day of week

Intuition suggests that the worst case (1) will produce the largest possible deviations in variance from the current system; the variance will be much higher on Monday and lower the other days of the week. This represents the case in which the approximation should perform the worst, as the initial guess will be farthest away from the actual standard deviation. Even in this unrealistic scenario, the worst error (still less than 9%) occurs in the diagnostic clinic. The next largest error is less than 3%. Outside of the diagnostic clinic, there is essentially no error. For this application, our Newton’s method-based approximation is sufficiently accurate.

3.2.2 Stage 1: Workload Smoothing Optimization Model.
We present one particular formulation of the outpatient workload smoothing optimization that stabilizes the workloads to improve system level access for outpatients. In this linear programming formulation, time is discretized into days with a planning horizon of 1, . . . , 5 to correspond to a weekly (business week) schedule. We use the notation of PROGRAM 1. $\mu_{u,d}(\cdot)$ and $\sigma^2_{u,d}(\cdot)$ are defined by Eq.’s 5 and 6 and $\hat{\theta}_{k,d}$ is the maximum number of type $k$, priority $\ell$ patients allowed to be scheduled on day $d$. The stochastic location process probabilities in these equations are calculated off-line and enter the LP as data.
\[
\min_{\Theta, \delta, i, u, d} \sum_{u \in U} \sum_{d=1}^{5} w_{u,d} \sum_{i \in M'} [\Phi(m(i + 1)) - \Phi(m(i))] \delta_{u,d,i} \\
\text{s.t.} \\
\mu_{u,d}(\Theta^L, \Theta^N) + m(i) \cdot \frac{1}{2} \left( \frac{\sigma_{u,d}^2(\Theta^L, \Theta^N)}{\hat{\sigma}_{u,d}} + \hat{\sigma}_{u,d} \right) - C_{u,d} \leq \delta_{u,d,i} \quad \forall u \in U, i \in M', d = 1, \ldots, 5 \\
\sum_{d=1}^{5} \Theta_{k,d}^{\ell} \geq 0 \\
\Theta_{k,d}^{\ell} \leq \tilde{\theta}_{k,d}^{\ell} \\
\Theta_{k,d}^{\ell} \in \mathbb{R}^+, \delta_{u,d,i} \geq 0 \\
\forall \ell \in \{H, L\}, k \in U, d = 1, \ldots, 5, u \in U, i \in M'.
\] (8)

The objective, Eq. 8, will drive the system to minimize the weighted approximation of the expected number of patient overflows across the week over all the clinical services. In fact, \( \sum_{i \in M'} [\Phi(m(i + 1)) - \Phi(m(i))] \delta_{u,d,i} \) is a simple Riemann integral approximation, which converges to the true expected value as the intervals go to zero. Weights, \( w_{u,d} \), are included because a blockage in one service may be more critical than a blockage in another. Constraints 9 calculate the amount of workload overflow at grid level \( m(i) \) for \( i \in M' \) as defined in Sec. 3.2.1. If the workload level associated with grid level \( m(i) \) is below capacity, the variable \( \delta_{u,d,i} \) will be set to 0, otherwise it will be forced to equal the amount of overflow at grid level \( m(i) \). Since the objective minimizes a weighted sum of the \( \delta_{u,d,i} \)'s, the \( \delta_{u,d,i} \)'s will be set as small as the constraint allows. Here we approximate the standard deviation with the Newton’s method-based approach detailed in Section 3.2.1. Eq. 10 constrains the optimal weekly volume served to be equal to the current weekly volume. Eq. 11 ensures that an upper bound on the number of patients of each type arriving for treatment on a given day is respected.

### 3.3 Stage 2: Phase-Type Models for Itinerary Completion

In Sec. 3.1 we developed stochastic models to capture the aggregate level system behavior. In this section, we develop and analyze new stochastic models for patient sojourn time through the network. Recall that itinerary completion, which is the objective of the second stage optimization, is defined as a patient completing an entire treatment path before a deadline. We develop a phase-type approach to characterize the patient’s sojourn time from treatment initiation to completion. Phase-type approximations were proposed by Casale (2010), which motivates our use of a phase-type approach in our discrete-time queueing network. However, the author does not consider time-varying capacities and arrival rates, as is necessary in our context. The author further notes that in queueing network models the phase-type generator matrix is often prohibitively large and grows combinatorially with the number of queues and jobs in the network. Hence, their method for passage times can only be applied to small queueing network models. One contribution of our
work is to identify an alternate representation for the generator matrix that significantly reduces its size and allows us to apply our method to much larger queueing networks.

3.3.1 Phase-Type Model for Critical Path Flow Times.
To calculate treatment length, including delays caused by blocking, for a treatment that begins on day \( d \), we break the treatment into appointments along the patient’s critical path. The critical path is defined as the set of appointments that, if delayed, will delay the entire treatment cycle. This prevents the model from allowing an auxiliary appointment, such as a dental appointment made by a breast cancer patient, for reasons of convenience and unrelated to cancer, from affecting itinerary completion for the main purpose of the visit to the destination medical center. In reality, auxiliary appointments are added while a patient is on-site, but if these appointments are not available the patient will often not stay over the weekend to get one the next week due to the inconvenience. Details of how to determine the critical path are discussed in Sec. 4.1.

The critical path is defined by the tuple \( \mathcal{C} = (\mathcal{R}, \mathcal{P}) \), where \( \mathcal{R} \) are the services that lie along the critical path and \( \mathcal{P} \) are the precedence relations. For example, \( \mathcal{R} = \{u_1, u_2, \ldots, u_n\} \) and \( \mathcal{P} = \{(u_i, c) : u_i \in \mathcal{R}, c = 1, 2, \ldots, N\} \), where the tuple \((u_i, c)\) represents the scenario where an appointment at service \( u_i \) is required in the \( c^{th} \) step (in the sense of order of precedence) of treatment. That is if \( a < b \), then \((u_i, b)\) can only occur after \((u_j, a)\) have been completed for all \( j \), but they can occur in the same step if \( a = b \) (in either order).

While we eventually incorporate multiple appointments that can be done in parallel at each precedence level, for purposes of exposition we begin with a base model that assumes each precedence level has only one appointment. Let \( \beta_{u,d} \) be the matrix of blocking probabilities in clinical service \((u)\) on day \((d)\), which is the blocking associated with the optimal schedule from the Stage 1 workload smoothing model. \( \beta_{u,d} \) can be calculated off-line from \( \Theta^* \), the optimal schedule of Sec. 3.2.2. If \( C_{u,d} \) is the capacity limit in service \( u \) on day \( d \), and \( \mu_{u,d}(\Theta^*) \) and \( \sigma_{u,d}(\Theta^*) \) are the mean and standard deviation of the workload under schedule \( \Theta^* \), then the blocking probability is given by the standard normal excess probability

\[
\beta_{u,d} = 1 - \Phi \left( \frac{C_{u,d} - \mu_{u,d}(\Theta^*)}{\sigma_{u,d}(\Theta^*)} \right). \tag{13}
\]

The total time to complete the treatment segment is denoted by the phase-type RV \( \delta_{\mathcal{C},d}(\mathbf{B}) \), which captures the sojourn time through a congested network with blocking matrix \( \mathbf{B} \) and indexes the precedence so that day 1 begins on the arrival day of week, \( d \). Letting \( K \) be the completion deadline (in absolute time), the probability that a patient will not complete their itinerary by time \( K \) given they were admitted on day \( d \in \{0, 1, \ldots, K\} \) is given by \( \mathbb{P}(\delta_{\mathcal{C},d}(\mathbf{B}) > K - d) \). Sec. 3.3.2-3.3.6 below develop a method to compute the itinerary completion probability.

3.3.2 Phase-type Base Model.
In this section, a base model is developed in which all items along the critical path must be completed. The next two sections incorporate the possibility that not all patients need appointments at all of the services, and each precedence level may have multiple tasks. To calculate the
CDF for the sojourn time \( \delta_{d,d}(B) \) for the base model, consider that on each day \( d \) the patient is either able to get the appointment at service \( u \) (w.p. \( 1 - \beta_{u,d} \)) or is blocked from getting the appointment (w.p. \( \beta_{u,d} \)). The time to obtain an appointment at service \( u_i \), given an initial attempt to schedule it on day \( d \), therefore follows a discrete phase-type distribution characterized by a Markov Chain (MC) with a transition probability matrix (tpm)

\[
\begin{bmatrix}
T_u & T_0^u \\
0 & 1
\end{bmatrix}
\]  

(14)

where \( T_u \cdot 1 + T_0^u = 1 \) and \( 1 \) is the vector of 1’s. In particular, the phase-type RV is the time until absorption in state 6 of the discrete time MC defined by the Eq. 14. The absorbing state 6 refers to the completion of the patient’s itinerary. The phase-type distribution is completely characterized by \( T_u \), which is called the generator matrix. The CDF and pmf of \( \delta_{u,d}(B) \) is given by (see Latouche and Ramaswami (1999))

\[
F(x) = 1 - e_d(T_u)^x \cdot 1
\]

(15)

\[
f(x) = e_d(T_u)^{x-1} \cdot T_0^u,
\]

(16)

where \( e_d \) is the unit column vector with 1 in the \( d \)th position and 0 elsewhere. The intuition behind Eq. 15 is that the probability that, given an initial appointment on day \( d \), the patient has not completed their itinerary by day \( x \), which is \( e_d(T_u)^x \cdot 1 \), where \( e_d \) is the starting state. Eq. 16 is the associated pdf. The total itinerary delay is the sum of the correlated delays for each service along the critical path. We can show that this sum preserves the phase-type distribution.

The tpm for the total length of the critical path is given by the block diagonal matrix

\[
\begin{bmatrix}
T_c & T_0^c \\
0 & 1
\end{bmatrix} = \begin{bmatrix}
T_{u_1}^1 & T_{u_1}^2 & \cdots & T_{u_1}^n \\
0 & T_{u_2}^1 & \cdots & T_{u_2}^n \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & T_{u_n}^1 \\
0 & 0 & \cdots & 1
\end{bmatrix}
\]

(17)

where \( T_{u_1}^1 \) and \( T_{u_n}^0 \) are defined as in Eq. 14. To see that this matrix represents the desired tpm of a MC whose time to absorption represents the length of a patient’s itinerary, consider the phase transition diagram (Fig. 2) of the chain described by Eq. 17. From Fig. 2, when a patient starts task 1 of their critical path on day \( d \), they are either blocked from receiving an appointment (w.p. \( \beta_{u_1,d} \)) or they complete their task (w.p. \( 1 - \beta_{u_1,d} \)). If the patient is blocked, they transition to the next day, \((d+1) \mod 5\) (i.e., Monday transitions to Tuesday and so forth, and Friday transitions back to Monday), but remain within task 1. Otherwise, they transition...
Figure 2: State transition diagram of a MC whose time to absorption represents the length of a patient’s care path. The state is a tuple (task, day), where $u_i$ is the task and the days are from 1, $\ldots$, 5, (weekdays)
to task 2. When the patient is on their final task and they are not blocked they transition to the absorbing state, (0, 0), indicating treatment is completed. The pmf and CDF of the total treatment completion time are given by:

$$F(x) = 1 - e_d(T_c)^x \cdot 1$$  \hspace{1cm} (18)$$
$$f(x) = e_d(T_c)^{x-1} \cdot T_c^0.$$  \hspace{1cm} (19)$$

Remark 3.1. With the phase-type structure it is also possible to capture precedence relationships that require task $j$ to begin no earlier than $n = 1, \ldots, N$ days after task $i$ is completed (where $N$ is the length of the planning horizon). This is useful if, for example, task $i$ is a medical test that will require $n$ days to process. This delayed precedence can be captured by including mandatory transitions w.p. 1 for the required waiting time until the patient can transition to $j$.

3.3.3 Adding Probabilistic Resource Needs to the Phase-type Itinerary Completion Model.

From the data, patients may require each task along their critical path with probability less than one. This section develops a richer phase-type model in which each visit along the critical path can be skipped (i.e., it was not needed for the given patient). An example of the appointment requirement probabilities is given in Table 3. In this case, all patients have a visit to the breast diagnostic clinic on their first day, while 17% visit medical oncology. On their second day of treatment 24% of patients visit the breast diagnostic clinic and 4% visit medical oncology, etc.

Let $\nu_{k,u_i}(d)$ be the probability that a patient of type $k$ will require task $i$ (requiring service $u_i$) in precedence level $d$ along their critical path. Fig. 3 illustrates the modification to the phase-type MC to incorporate this feature. When a patient completes any task $j$, they will move to the next task only if the task is required, w.p. $\nu_{k,u_j+1}(d)$. If task $j+1$ is not required (w.p. $1 - \nu_{k,u_j+1}(d)$), then the MC can jump to task $j+2$. If both tasks $j+1$ and $j+2$ are not required (w.p. $(1 - \nu_{k,u_j+1}(d))(1 - \nu_{k,u_j+2}(d))$, then the MC can jump to task $j+3$ and so forth.

The new transition probability matrix then has the upper triangular form (Eq. 20).
Now let \( \eta \) for \( j = 2 \), subsequent and the start of the itinerary terms of number of days after the initial appointment of the patient’s itinerary, with day zero being the day of arrival subsequently and the start of the itinerary.

Table 3: Sample of a stochastic location process empirical distribution for breast cancer patients, identified in terms of number of days after the initial appointment of the patient’s itinerary, with day zero being the day of critical path on service \( i \). For a type \( k \) patient, let \( \kappa_k \) be the probability vector in which \( \kappa_k(i) \) is the probability that the patient begins their critical path on service \( i \). \( \kappa_k = \left[ \nu_{k,u_1}(1), \nu_{k,u_2}(2)(1-\nu_{k,u_1}(1)), \ldots, \nu_{k,u_n}(n) \prod_{j=1}^{n-1}(1-\nu_{k,u_j}(j)) \right]^t. \)

For ScholarOne support, you may contact 434-964-4100 or 888-503-1050 (US-based numbers).
scheduled for their initial appointment on day \(d\). This distribution is given by
\[
\eta_{k,d} = [\nu_{k,u_1}(1)e_d, \nu_{k,u_2}(2)(1 - \nu_{k,u_1}(1))e_d, \ldots, \nu_{k,u_n}(n) \prod_{j=1}^{n-1} (1 - \nu_{k,u_j}(j))e_d, 0]' \tag{23}
\]
The \((N \cdot n + 1) \times 1\) vector \(\eta_{k,d}\) combines the initial service the patient enters \((\kappa_k)\) with the day the patient enters that service \((e_d)\), to start the patient’s path in the correct location at the correct time. Now, for a patient of type \(k\), we can calculate the discrete time phase-type distribution of the probability that itinerary completes on or before stage/day \(x\) (CDF denoted by \(F_k\)) and the probability that a patient leaves the system at time \(x\) (pmf denoted by \(f_k\)) as follows:
\[
F_{k,d}(x) = 1 - \eta_{k,d}(T_c)^x \cdot 1 \tag{24}
\]
\[
f_{k,d}(x) = \eta_{k,d}(T_c)^{x-1} \cdot T_c^0. \tag{25}
\]

3.3.4 Appointments in Parallel: Maximum of Phase-type Distributions.
The third and final feature of the outpatient service model is that the critical path may contain non-identical appointments that can occur in parallel. That is, in precedence level \(d\), there can be \(n\) parallel appointments with the time to obtain each appointment denoted by \(X_1, X_2, \ldots, X_n\). From Sec. 3.3.2, each \(X_i\) follows a phase-type distribution characterized by generator matrix \(T_i\). The time to complete all appointments is given by \(X = \max\{X_1, X_2, \ldots, X_n\}\). To analyze the distribution of \(X\) it is necessary to present the following definition (from Davio (1981)).

Definition 3.1. The Kronecker Product of matrices \(A\) and \(B\), \(A \otimes B\), is an operation on two matrices, such that if \(A\) is an \(m \times n\) matrix and \(B\) is a \(p \times q\) then the \(mp \times nq\) matrix is:
\[
A \otimes B = \begin{bmatrix}
a_{11}B & \ldots & a_{1n}B \\
\vdots & \ddots & \vdots \\
a_{m1}B & \ldots & a_{mn}B
\end{bmatrix}. \tag{26}
\]

Lemma 3.1 shows that the maximum of \(n\) phase-type distributions is a phase-type distribution. The proof follows by showing that the Davio (1981) result for two phase-type distributions can be applied recursively (see Appendix A). While it is possible to derive a closed form solution for the generator matrix for the maximum, it is notationally complex and so we present the more simple recursive formula for the generator matrix, from which it is easier to gain intuition.

Lemma 3.1. Let \(X_1, X_2, \ldots, X_n\) be phase-type distributed RVs with generating matrices \(T_1, \ldots, T_n\) and the normalizing absorbing probability vectors \(T_1^0, \ldots, T_n^0\). Let \(X^{(n)} = \max\{X_1, X_2, \ldots, X_n\}\). Then \(X^{(n)}\) is phase-type distributed and the generator matrix for \(X^{(n)}\) is upper triangular and has block form defined by the recursive relationship
\[
T_{X^{(n)}} = \begin{bmatrix}
T_{X^{(n-1)}} \otimes T_n & T_{X^{(n-1)}} \otimes T_n^0 & T_{X^{(n-1)}} \otimes T_n^0 \\
0 & T_{X^{(n-1)}} & 0 \\
0 & 0 & T_n
\end{bmatrix}, \tag{27}
\]
where \(T_{X^{(1)}} = T_1\).
From Lemma 3.1 the maximum of \( n \) phase-type distributions has block upper triangular form. To obtain the closed form solution for the maximum of \( n \) phase-type distributions, first consider the result that the Kronecker product is distributive among blocks of a block matrix. That is,
\[
\begin{bmatrix}
A & B \\
D & C
\end{bmatrix} \otimes N = \begin{bmatrix}
A \otimes N & B \otimes N \\
D \otimes N & C \otimes N
\end{bmatrix}.
\] (28)

The closed form solution comes from combining the result from Eq. 28 and Lemma 3.1.

3.3.5 A Tractable Representation of the “Max” Phase-type Generator.

A drawback of the general representation of the generator matrix for the maximum of \( L \) phase-type distributions given by Eq. 27 is the size of the matrix. If each matrix in the maximum is size \( N \times N \), then the generator matrix of the maximum would be of size \((N+1)^L - 1 \times (N+1)^L - 1\). The exponential growth of the size of the matrix in the number of phases, \( L \), is a concern. In our case study, each individual generator matrix is \( 5 \times 5 \) (5 critical path services), so the generator matrix is \( 7,775 \times 7,775 \), with 60,450,625 entries. Adding just one more service would generate a matrix with 2,176,689,025 entries.

Since representations for phase-type distributions are in most cases not unique, we exploit the special structure of our phase-type model to obtain a representation, \( V \), that is significantly more compact. We do so by eliminating the possibility of certain transitions and thereby reducing the state space. First, in the general form of the phase-type distribution, transitions can occur from any state to any other state. Our MC, however, never transitions more than one day ahead (i.e., transitions occur only from Day 1 to Day 2, or Day 2 to Day 3, etc.). Arranging the states properly in the MC representation, allows us to create a sparse block matrix structure that provides for efficient computation of powers. Secondly, for the general form, each individual phase-type RV is allowed to start in any state, independent of the other RVs in the maximum. Thus it is possible that the first RV of the maximum starts on day 1, while the second starts on day 3. In the our model, however, all parallel tasks begin the MC on the same day. This allows us to eliminate states such as \((1,3)\), where the one parallel appointment is attempting to get scheduled on day 1 while the second is attempting to get scheduled on day 3. The compact matrix, \( V \), has the following form (with explanation to follow):

\[
V = 
\begin{bmatrix}
\text{Day 1} & \text{Day 2} & \text{Day 3} & \text{Day 4} & \text{Day 5} \\
\text{Day 1} & 0 & V_{1,2} & 0 & 0 \\
\text{Day 2} & 0 & 0 & V_{2,3} & 0 \\
\text{Day 3} & 0 & 0 & 0 & V_{3,4} \\
\text{Day 4} & 0 & 0 & 0 & V_{4,5} \\
\text{Day 5} & V_{5,1} & 0 & 0 & 0 \\
\end{bmatrix}.
\] (29)

Define \( i \oplus 1 \) as \( 1 \) if \( i + 1 = N + 1 \) (\( N \) being the length of the planning horizon) and as \( i + 1 \) otherwise (similar to mod). The blocks of zeroes in Eq. 29 are a result of the fact that on day \( i \), the only possible next state for each appointment is day \( i \oplus 1 \) or that the appointment has been completed. Each block, \( V_{i,i+1} \), represents the transitions on day \( i \). These transitions capture how...
many of the $L$ services the patient was able to obtain an appointment at on day $i$. The $2^L$ states represented by $V_{i,i\oplus 1}$ take the form $(a_1, a_2, \ldots, a_L)$, where $a_j = 1$ means that the patient has completed their appointment at service $j$ and $a_j = 0$ means that the patient has not yet gotten an appointment at service $j$. As an example, trying to get an appointment at services 1 and 3 on day $i$, having already completed an appointment at service 2 yields one possible transition:

$$P((0, 1, 0) \rightarrow (0, 1, 1)) = P(\text{Get Appt at Svc 3})P(\text{Not Get Appt at Svc 1}) = \beta_{1,i}(1 - \beta_{3,i}).$$

For each day $(V_{i,i\oplus 1})$, we only consider the remaining services for which the patient is able to get appointments. The events describing the daily transition for getting an appointment in service $u_j$ on day $i$, can be captured by the simplified $2 \times 2$ t.p.m.

$$A_{u_j,i} = \begin{bmatrix} \beta_{u_j,i} & 1 - \beta_{u_j,i} \\ 0 & 1 \end{bmatrix}. \quad (30)$$

$\beta_{u_j,i}$ is the probability that the patient couldn’t get an appointment in service $u_j$ on day $i$. If blocked, the patient is directed to the following day (e.g., $i \oplus 1$) and attempts to get the appointment again. Otherwise, the patient has finished the task and enters the absorbing state (i.e., task complete) for that task. Thus, $A_{u_j,i}$ describes the daily transition for a single service.

The transition probabilities, $V_{i,i\oplus 1}$, for the possible outcomes of trying to get appointments in $L$ different services (i.e., success or failure on day $i$) can be calculated by combining the single service t.p.m.’s using the Kronecker Product, $\bigotimes_{j=1}^L A_{u_j,i}$.

**Theorem 3.2.** Suppose there are $L$ phase-type distributed RVs, $X_{u_1}, \ldots, X_{u_L}$, with generator matrices, $T_{u_1}, \ldots, T_{u_L}$, following the structure in Eq. 14. Let $A_{u_j,i}$ be the compact representation of the single service t.p.m.’s for day $i$ given by Eq. 30. Then Eq. 29 is a generator for $\max_j \{X_{u_j}\}$, where

$$\begin{bmatrix} V_{i,i\oplus 1} & 0 \\ 0 & V_0^{\ominus 1} \end{bmatrix} = \bigotimes_{j=1}^L A_{u_j,i}. \quad (31)$$

**Corollary 3.1.** Let $U(n)$ be the computational complexity of multiplying $n \times n$ upper triangular matrices together. Let $N$ be the length of the planning horizon for the phase-type distribution described by Eq. 29. The computational complexity of calculating the CDF of the itinerary completion phase-type distribution for the maximum of $L$ phase-type distributions using the compact approach is given by (with proof in Online Appendix A)

$$O_{\text{compact}}(F(x)) = N \cdot x \cdot U(2^L). \quad (32)$$

First note that the computational complexity grows linearly in the length of the planning horizon, which is far slower than the traditional representation. As an illustration of the importance of the above decomposition method, we compare the computational complexity of the compact representation ($O_{\text{compact}}$) with the standard representation ($O_{\text{full}}$) of the phase-type distribution.
Let $M(n)$ be the computational complexity of regular matrix multiplication. Then
\[
O_{\text{compact}}(F(x)) = N \cdot x \cdot U(2^L) \quad (33)
\]
\[
O_{\text{full}}(F(x)) = x \cdot M((N + 1)^L) \quad (34)
\]

We compact the state space and compute the matrix power by multiplying smaller sub-matrices instead of the entire matrix. With the 5 day planning horizon and 5 services that could be potentially visited in parallel, the compact representation only requires multiplication of matrices of size $31 \times 31$ that have 961 entries, whereas the general representation requires multiplication of matrices of size $7,775 \times 7,775$ matrix with 60,450,625 entries. With 6 services, the contrast becomes even more stark: 3,969 entries versus 2,176,689,025 entries.

### 3.3.6 Phase-type Model for Itinerary Completion.

We have developed a phase-type model of blocking that considers (1) completing a sequence of appointments with precedence constraints, (2) some tasks may be completed in parallel, and (3) the idea that some appointments along the critical path may not be required by all patients. This section combines all of these features into one phase-type model to calculate the probability of on-time treatment completion given an initial appointment on day $d$. Let $\mathcal{R}_d \subseteq \mathcal{R}$ be a cluster of services that are required in precedence level $d$. Let $V(\mathcal{R}_d)$ be the compact form generator for the phase-type model of completion time per Eq. 29 from Sec. 3.3.4. The phase-type model that captures itinerary completion is obtained by replacing the generator, $T_{\text{ud}}^1$ for each individual service in Eq. 20 with $V^i(\mathcal{R}_d)$. The transition from precedence level $d$ to $d + i$, $V^i(\mathcal{R}_d)$, (or to the absorbing state $V^0(\mathcal{R}_d)$) is calculated the same way as $T_{\text{ud}}^i$ (or $T_{\text{ud}}^0$) from Eq. 21 by subtracting the row sum of each row of the matrix from 1 and multiplying by the probability the cluster of services in level $d + i$ is needed, $P(1 \{\mathcal{R}_{d+i}\})$. It is too large to display here. Because this generator matrix has block form with a significant number of zero blocks, the computation of the power of the phase-type generator matrix for the entire path, $V(\mathcal{C})$, can be calculated by multiplying the smaller blocks of the total matrix together using an algorithm similar to the one proposed in the proof of Corollary 3.1.

Let $\eta_{k,d}$ be the initial distribution for patients of type $k$ beginning their itinerary on day $d$ of the planning horizon (Eq. 23). Realizing that $\delta_{\mathcal{C},d}(\mathcal{B})$ is the RV for the time it takes to complete an itinerary, the probability that a type $k$ patient’s itinerary completes before the end of the work week given that they were admitted on day $d \in \{1, \ldots, 5\}$ is then given by
\[
P(1_{k,d}\{\text{Itinerary Complete}\} = 1) = P_{k,d}(\delta_{\mathcal{C},d}(\mathcal{B}) \leq 6 - d) = 1 - \eta_{k,d}(V(\mathcal{C}))^{6-d} \cdot 1. \quad (35)
\]

### 3.3.7 Validation of Itinerary Completion Model for Mayo Clinic Breast Cancer Patients.

To validate whether this phase-type approach modeled our system well, we parameterized the model using historical data under the current scheduling system at Mayo Clinic based. The parameterization process using training data is described in greater detail in Sec. 4. The blocking rates were estimated based on the actual schedule that Mayo used over the 6 month testing data.

For ScholarOne support, you may contact 434-964-4100 or 888-503-1050 (US-based numbers).
Over that time period, the phase-type model predicted itinerary completion rate at 74% while the actual over that time was 72%. The reason our model slightly overpredicts itinerary completion is because we only modeled the critical path, and not all the services the patient will visit. This level of accuracy was deemed sufficient for our application at the Mayo Clinic and adding additional services to increase the accuracy was not considered necessary. In Sec. 3.4 we develop an optimization model that maximizes itinerary completion for high priority patients.

3.4 Stage 2: Itinerary Completion Optimization.

For the second stage optimization, we begin with the workload dynamics of the system resulting from the first stage workload and blocking optimization, and minimize the expected number of time-sensitive (high priority) patients who fail to complete their itinerary by the deadline. The probability of on-time treatment completion for type \( k \) patients who have their initial appointment on day \( d \) can be calculated from the phase-type model CDF, \( F_{k,d}(\cdot) \), of Sec. 3.3.6. By constraining the blocking probability, \( \beta_{u,d} \) from Eq. 13 in the priority scheduling optimization to respect the service levels under \( \Theta^* \), we ensure that the itinerary completion model is properly parameterized with similar flow dynamics to \( \Theta^* \). This is important because \( F_{k,d}(t) \), the itinerary completion probability, is dependent on the blocking probabilities from the solution of stage 1. Thus, as long as the blocking is not much worse in the second stage optimization, \( F_{k,d}(t) \) will be a good estimate of the probability of itinerary completion within \( t \) days. We enforce the blocking rates by finding the standard deviation multiplier, \( b_{u,d} = \Phi^{-1}(1-\beta_{u,d}) \), that forces the workload quantile \( (\mu_{u,d}(\Theta) + b_{u,d}\sigma_{u,d}(\Theta)) \) to maintain the same blocking probabilities as stage 1, \( \beta_{u,d} \) (see Eq. 37). The stage 2 optimization is as follows.

\[
\min_{\Theta^H, \Theta^L} \sum_{k \in \mathcal{D}} \sum_{d=1}^{5} \Theta^H_{k,d}(1 - F_k(6 - d)) 
\]

\( s.t. \)

\[
\mu_{u,d}(\Theta^H, \Theta^L) + \beta_{u,d} \cdot \frac{1}{2} \left( \frac{\sigma^2_{u,d}(\Theta^H, \Theta^L)}{\hat{\sigma}_{u,d}} + \hat{\sigma}_{u,d} \right) \leq C_{u,d} + \epsilon \quad \forall u \in \mathcal{U}, d = 1, \ldots, 5 \]  

\[
\sum_{d=1}^{5} \Theta^L_{k,d} \geq \theta^L_k \quad \forall \ell \in \{H, L\}, k \in \mathcal{D} \]  

\[
\Theta^L_{k,d}, \Theta^H_{k,d} \in \mathbb{R}^+ . \]  

The objective, Eq. 36, minimizes the expected number of incomplete itineraries across priority patient types. Itinerary completion of a type \( k \) patient beginning their itinerary on day \( d \) is modeled as a Bernoulli RV with probability \( F_{k,d}(6 - d) = \mathbb{P}_{k,d}(\delta_{e,d}(B) \leq 6 - d) \) from Eq. 35. Eq. 37 is the constraint that enforces the service level factor, \( \beta_{u,d} \), for any given schedule \( \Theta^L \) and \( \Theta^H \) using our Newton’s method approximation of the square root function from Section 3.2.1. The Eq.’s in 37 ensure that the solution yields a flow that is characterized by \( F_{k,d}(\cdot) \), the minimum blocking solution of the stage 1 workload smoothing model. We add \( \epsilon \) to ensure that there are sufficiently many solutions for our optimization engine (CPLEX) to solve the problem easily and not be hampered by much if any precision errors. We enforce weekly throughput targets.
4 Analysis and Case Study of Itinerary Completion Improvement

In this section, we present an application of our optimization method to improve itinerary completion for national/international breast cancer patients at the Mayo Clinic. In principle the methods can be applied to an ensemble of services offered by the Mayo Clinic, but our industry partner indicated interest in addressing one suite of services as an initial pilot.

4.1 Data and Model Parametrization

We begin by describing the patient flow process of a breast cancer suspect at a destination clinic and highlight two key problems that must be addressed by a capacity allocation plan: (1) appointment delays caused by high utilization and variable workloads, (2) operational delays caused by the fact that clinical services are closed on the weekends, which drives the need to complete itineraries by Friday afternoon for patients that are not within commuting distance of the clinic. Figure 4 shows a simplified diagram of breast cancer patient flow for 5 key breast cancer services (out of a total of 77 services that breast cancer patients may use) at the Mayo Clinic. New suspected breast cancer patients typically arrive to the clinic to begin their itinerary in the breast diagnostic clinic (BDC). Based on the results from the BDC, subsequent appointments may be made with other specialty services in the integrated network. Typically, patients will then be scheduled for a surgical consult or a medical oncology consult based on the appropriate treatment course. After these consults, the patient may be referred to other services to complete the treatment plan such as plastic surgery (if surgery is the chosen option) or radiation oncology. Further, if the initial consult determines that the best treatment option lies with a different specialist service, the patient may be referred there. For example, if the surgical consult determines that surgery is not appropriate, the patient may be referred to medical oncology.

![Figure 4: Simplified example of offered load flow model for breast cancer patients.](image)

A factor that complicates the modeling of patient itineraries is the fact that a patient’s full itinerary (quantity, timing, and services required) is not known at the time of the patient’s initial visit; new appointments are made dynamically when more information becomes available about the patient’s condition. This workload uncertainty is a significant contributor to two problems. First, frequent overutilization of key services causing itinerary delays. Second, a breast cancer patient who cannot complete all the steps of their itinerary by Friday (e.g., they completed BDC and the surgery consult, but have not yet had a radiation oncology appointment), must incur two days of non-value added waiting over the weekend. Waiting over the weekend has also been found to play an important role in decreased patient satisfaction for these time-sensitive patients.

In discussion with our industry partner, we determined that, while downstream visits cannot
be planned in advance, the clinic does have control over the start time of the initial visit of the itinerary – in the breast cancer itinerary above this is typically an appointment at the BDC. We leverage this control to design a capacity plan that reserves appointment slots by patient type (time-sensitive vs. less time-sensitive) to mitigate both problems described above. By optimizing the number of appointments reserved for different types of patients over the days of the week, we are able to reduce the overutilization in key breast cancer services. Moreover, this reduces the number of visits of time-sensitive patients on high utilization days (where itinerary delays are likely) as well as the number of itineraries that span weekends (where delays are inevitable).

The first challenge in parameterizing the itinerary completion model was to determine how to categorize the clinical resources. Key considerations include: (1) Are the definitions of resources amenable to measurement of patient workload (e.g., number of appointments, OR time, etc.)? (2) Is it possible to quantify the resource’s capacity limit? (3) Is the resource a bottleneck; i.e., is the resource capacity constrained? To account for these, we categorize resources by clinical service (e.g., general surgery, radiation oncology) and, where appropriate, by the particular type of appointment at the clinical service (e.g., surgical consult, registered nurse, diagnostic imaging). The workload and staffing capacity limits at these services were quantified in terms of their number of appointments per day, as explained in more detail below.

Patient type was the tuple of diagnosis and geo-code (e.g., breast cancer and national). Not only did patients from different regions exhibit different care paths and resource usage, but also the national/international patients were time-sensitive high priority patients because of travel constraints. We parameterized the breast cancer stochastic location processes using patient flow data from 2006 - 2011. The data contained, for each patient, the type and timing of each appointment in their itinerary. From this, we calculated the probabilities of requiring service $u$, $d$ days after the start of the itinerary.

The total workload at each resource was measured as the number of appointments at the resource; not all resources are assumed to have equal service times/hours of operation. Appointment lengths, however, are generally standard within each service. Long appointments are accounted for by assigning them as a group of multiple shorter appointments. To calculate the service’s utilization, we divide the total appointments by the FTEs allocated to the service to get the ratio of appointments per FTE. A service’s capacity was determined by multiplying the number of appointments per staff member by the number of staff members on duty, which was obtained from a separate dataset that logs staffing on each day.

An additional complication was that the clinical services that we studied did not exclusively serve breast cancer patients. In our case study, we only control the breast cancer patient arrival stream so we model initial visits for other diagnoses as exogenous demand. The arrival distribution for the exogenous demand is approximated as a truncated normal RV with the mean and variance parameters estimated from historical data.

We implemented the critical path model developed above specifically for breast cancer re-
sources that: (1) are commonly used meaning that a significant proportion of patients from a
given patient type visited at least once, and (2) frequently caused itinerary completion failures, as
logged in a separate database. For the breast cancer patients in our proof-of-concept case study,
we identified 5 services forming the critical path: Breast Diagnostic Clinic (BDC), Medical On-
cology, Radiation Oncology, General Surgery, and Plastic Surgery. These services were further
validated as being critical to the breast cancer itinerary in discussions with medical professionals.

4.2 Case Study Results

In our analysis we present several possibilities based on different system configurations that were
of interest to the Mayo Clinic. We compare a system with time-varying staffing by day of the
week (current practice) with a system that has constant staffing on each day. We show that our
model can achieve significant gains by only controlling the capacity for new breast cancer patient
initial appointments, which accounted for less than 25% of the total volume in any one service.

To illustrate how the two-stage method works, we first present the intermediate (stage 1)
results, then we present results after both stages are run.

Stage 1: Blocking Optimization. First, we demonstrate the aggregate results for each of
the five critical path services under the current system staffing. Then we drill down into the
bottleneck resource for the breast cancer itinerary, which is the physician consult in the BDC,
and analyze the system under the current staffing and under stable staffing with different levels
of control on the percent of appointments that can be booked by the service in advance.

To perform the experiment, we parameterized the model as described above and then solved
the stage 1 workload smoothing optimization that minimizes blockages throughout the week.
The optimization determines the number of initial appointment(s) (that are booked in advance
by day of week and comprise the start of a patient’s itinerary) to assign to each service on
each day of the week, while working within the staffing constraints. Note that each service
also has to serve subsequent appointments that are generated after a patient’s initial (typically
diagnostic) appointment(s). We call this non-controllable (NC) workload, which is accounted for
via the stochastic location process. The other type of NC workload is the exogenous demand,
i.e, non-breast cancer patients and patients that are referred from other services.

Fig. 5 shows the average utilization of the five critical path resources under the original
(current) staffing and the result after the Stage 1 optimization. The primary gains are in the
services that are most heavily utilized by breast cancer patients, BDC and General Surgery, where
the utilization is significantly smoothed, reducing the chances of midweek blocking that can cause
itinerary completion failures. Note that both scenarios serve the same number of patients, so the
lower utilizations from the optimized appointment allocation result from smoothing the workload
across the days of the week (relative to the maximum capacity on each day).

Because we are studying breast cancer patients, the BDC which is where most breast cancer
patients begin their itinerary, provides a rich environment for illustrating the insights from this
intermediate stage. Thus we demonstrate further capabilities of our model by drilling down

For ScholarOne support, you may contact 434-964-4100 or 888-503-1050 (US-based numbers).
into the activities at the BDC, where the key appointments were (1) Physician Visit (MD) and (2) Diagnostic Procedure (PR). To do so, we perform the following experiment. We begin by parameterizing the stage 1 workload smoothing model (as we did above with the five services) considering the key services provided within the BDC: MD and PR. We then ran the workload smoothing optimization under the case where we control the number of new breast cancer appointments (Fig. 6 (b), representing about 25% of the total volume in MD and PR). We next ran the optimization controlling all new itinerary start appointments within the BDC (Fig. 6 (c)), while still accounting for the subsequent appointments that are generated randomly after the initial appointment(s) of the itinerary and the exogenous demand referred from other services. Comparing Fig. 6 (b) and (c) demonstrates that a partial implementation controlling only one service can be nearly as effective as a full implementation, which requires more extensive buy-in. We compare these two optimization results with the workload generated from the current appointment allocation shown in Fig. 6 (a). Fig. 6 shows the results in terms of the average utilization of the physicians and diagnostic imaging machines by day of week. Here we consider stable staffing, by which we mean that the service has the same staffing each day of the week, which was a proposal being considered by Mayo and presents a challenge of matching the appointment reservation plan to a new staffing mix. The varying staffing scenario (i.e., the number of FTEs changes from day to day), produces similar results.

From Fig. 6, the physician visit (MD) is clearly the bottleneck resource at the BDC. In fact, the mean physician’s workload at the BDC is over capacity on three out of the five weekdays (Fig 6 (a)). The optimal schedule (Fig. 6 (b)) smooths the physician workload - only slightly exceeding capacity on Thursday. Note that there is only a marginal benefit when going from controlling only new breast cancer initial appointments (Fig. 6 (b)), accounting for 25% of total demand, to controlling all patient initial appointments at the diagnostic clinic (Fig. 6 (c)). While there is more day of week variation in the mean workloads for the diagnostic procedure (PR) service, this is acceptable to achieve the gains in the physician workload because the average utilization of PR is well below 100%.

Having established that physician appointments are the bottleneck, we drill down again into

For ScholarOne support, you may contact 434-964-4100 or 888-503-1050 (US-based numbers).
Figure 6: Workloads for the physician (MD) appointment type and the diagnostic procedure (PR) appointment type for the (a) original schedule, (b) Stage 1 optimal schedule controlling only new breast cancer patients, (c) Stage 1 optimal schedule controlling all patient schedules. MD capacity is the bottleneck

The physician workload for a more detailed analysis of the study in Fig. 6. Fig. 7 shows the non-controllable (NC) demand (exogenous demand and subsequent visits that are generated dynamically during a patient’s itinerary), represented by the black bar, and controllable (C) demand, represented by the gray bar, in terms of average number of appointments needed by day of week. The solid line shows the appointment capacity based on the staffed FTEs. In this analysis, we consider two different scenarios. The first scenario considers a constant staffing level (i.e., number of appointments that can be served) by day of week (solid line in Fig. 7 (a) and (b)), where we expect to see more relative improvement. The other scenario represents the current practice where the staffing differs by day of week (solid line Fig. 7 (c) and (d)). Comparing Fig. 7 (a) and (b) with Fig. 7 (c) and (d) demonstrates that the optimization can be more effective under constant staffing than under time-varying staffing.

Figure 7: Physician appointments at the Breast Diagnostic Clinic for controllable (C) and non-controllable (NC) patients. (a) current schedule with constant staffing, (b) optimal schedule with partial control and constant staffing, (c) current allocation with varying staffing, (d) optimal allocating with partial control and varying staffing

The key insight from Fig. 7 is the following: by controlling the schedules of less than 25% of the patient population, significant gains can be made to reduce congestion in critical clinical services. This is true whether the capacity is constant or varies over time, though it is more pronounced in the constant staffing scenario. Note, all of this is done only by controlling the
timing and mix of new initial visits for breast cancer itineraries, the follow-on appointments that
determine the full workload are stochastic. For constant staffing, Thursday is still overutilized
due to the extremely high level of non-controllable workload because new breast cancer initial
appointments on Monday, Tuesday or Wednesday often generate follow-up workload on Thursday.

**Stage 2: Itinerary Optimization.** Returning to the original network model with the five
breast cancer services, we calculated (1) blocking probabilities to parameterize the national/international
patients’ phase-type care paths, (2) blocking quantiles to ensure that blocking probabilities are
bounded above by the first stage solution (see Eq. 37), and (3) the phase-type care paths with
blocking incorporated. For (2) it was necessary to determine an appropriate \( \epsilon \) in Eq. 37 so
that the optimization would solve efficiently and accurately. In this case study, the optimization
solved quickly for very small values of \( \epsilon \) (e.g., \( \epsilon = 0.34 \), which allows workload to exceed the
optimal workload by no more than 0.34 patients), so the final (2nd stage) optimized network
workloads deviated less than 1% in each service on each day from the workloads generated by the
optimal schedule from the workload smoothing stage. In this study, appointment reservations
were allowed to be fractional to represent the average number of appointments for that day of
week, allowing for small fluctuations from week to week. This was determined to be appropriate
in the context of the study, however the integer model also solves quickly and can be employed if
desired. The difference between the integer and continuous solutions is small because workloads
do not change much with small changes in capacity allocation. For model validation, we input
the current appointment scheduling practice and calculated an itinerary completion rate of 74%
using our model. The actual itinerary completion rate from the data was 72%, indicating that,
despite the complexities of both the workload and sojourn time models, we are capturing the
reality of itinerary completion with good accuracy.

In this study, we consider three scenarios. The first scenario (Fig. 8 (a)) implements the
current appointment allocation for high priority (national/international) and low priority (lo-
cal/regional) patients, in which priority is determined by itinerary completion time-sensitivity
due to travel constraints. The primary metric is for national/international patients to be able to
complete their treatment by Friday so that they do not have a weekend delay or become forced
to travel long distances multiple times to complete their itinerary. The second scenario optimizes
the appointment allocation under the current (time-varying) staffing levels (Fig. 8 (b)). The
third scenario optimizes the appointment allocation under a fixed staffing level (Fig. 8 (c)).

Fig. 8 shows the model decision variables, the number of new breast cancer initial appoint-
ments, for all three scenarios just described. Note these graphs include only new breast cancer
initial appointments, exogeneous demand and follow-up appointments are not displayed. The
appointment workloads across the five critical path breast cancer services are also not displayed.
For reference, the workloads resulting from this stage 2 optimization are nearly identical to the
workloads from the stage 1 workload smoothing optimization (e.g., Fig. 5) due to the workload
constraints in the stage 2 optimization.

For ScholarOne support, you may contact 434-964-4100 or 888-503-1050 (US-based numbers).
Figure 8: Comparing the current first appointment day schedule with the stage 2 optimal schedule for national/international patients versus local/regional patients.

In both optimizations, the national/international demand is pushed closer to the beginning of the week, while maintaining the optimal workload/blocking profile so that mid-week blockages don’t delay itinerary completion. This gives the national/international patients the best chance to complete their itinerary before the weekend. For the optimal appointment allocations (Fig. 8 (b) and (c)), the number of new treatments that start on Monday is low because there will be a significant number of follow-up appointments on Mondays, being driven by the fact that most of the local/regional patients are being shifted to Thursday and Friday. Because most national/international itineraries can be completed in 4 days or less after optimizing the workloads across the five services (stage 1), starting the majority of “priority” patients on Tuesday still allows most to complete their itinerary by Friday.

The optimization using the time-varying (current) staffing capacities can achieve 85% itinerary completion. However, if staffing is stabilized (constant over the week instead of varying from day to day), the optimization can achieve 88% itinerary completion. Thus, depending on the level of change an organization can enact, it is possible to see an improvement of 11% to 14% in itinerary completion, which was considered a significant improvement by our industry collaborator.

Several managerial insights can be gained from this case study. The first is a simple rule of thumb that can help guide organization change: placing national/international patients near the beginning of the week but not many on Mondays (to allow for local/regional patient prior week overflow) can significantly improve itinerary completion. Second, the optimization is most effective when overall staffing is constant over the course of the week as time-varying staffing makes it more difficult to accommodate the stochastic downstream flows from previous days. Third, priority appointment reservation can improve patient service/throughput without adding capacity. Finally, the fact that these benefits were achieved while controlling a small percent of the total patient initial appointments (< 25% at each service) demonstrates that a pilot involving buy-in from only a subset of the departments in a care network would be able to demonstrate immediate benefits and prove the effectiveness of this method, thus reducing barriers to implementation.
5 Conclusions
In this paper, we develop an optimization approach to a queueing network model for priority appointment allocation in a network of healthcare services with patient classes with different time-sensitivities. We applied our new approach to meet itinerary completion deadlines for national/international patients at the Mayo Clinic. Using decomposition, linearizing approximations, and phase-type modeling, we transform the non-linear stochastic optimization into two-stages of tractable LP models. The blocking optimization stage stabilizes the environment through which patients flow. Within this stabilized environment, the second stage allows for further differentiation of patients by priority. The theoretical models were tested on breast cancer patients at the Mayo Clinic, indicating valuable improvements in itinerary completion.

References


For ScholarOne support, you may contact 434-964-4100 or 888-503-1050 (US-based numbers).


A Appendix

Proof of Theorem 3.1

Proof. Each pair of location functions, \((L_{j_1,n_1}^{1,k_1}(t_1), L_{j_2,n_2}^{2,k_2}(t_2))\) where \((j_1, n_1, k_1, s_1) \neq (j_2, n_2, k_2, s_2)\) represents two different patients. Therefore the two location processes are independent, which follows from our assumption that the care paths of two different patients are independent. To see that each pairing in Eq. 3 does indeed represent two different patients, note that each patient’s stochastic process is uniquely indexed by the patient type, \(k\), the week in which they are admitted, \(n\), the day of the week they were admitted, \(s\), and their admission number on the day they are admitted, \(j\). Therefore the processes \(L_{j_1,n_1}^{1,k_1}(t_1)\) and \(L_{j_2,n_2}^{2,k_2}(t_2)\) are independent and their covariance term is necessarily zero. The variance of the number of appointments can be calculated by

\[
\sigma_{d_1,u}^2(\Theta) = \text{Var} \left[ \sum_{d_2=1}^{5} \sum_{k \in \mathbb{D}} \sum_{j=1}^{5} \sum_{n=0}^{t} \lim_{t \to \infty} \sum_{m=0}^{t} \epsilon_u \cdot L_{d_2+5n,k}(d_1 + 5t) \right] = \sum_{d_2=1}^{5} \sum_{k \in \mathbb{D}} \sum_{j=1}^{5} \sum_{n=0}^{t} \lim_{t \to \infty} \sum_{m=0}^{t} \text{Var} \left[ \epsilon_u \cdot L_{d_2+5n,k}(d_1 + 5t) \right] = \sum_{d_2=1}^{5} \sum_{k \in \mathbb{D}} \lim_{t \to \infty} \sum_{n=0}^{t} \sum_{m=1}^{M_u} \left[ m^2 \cdot \hat{p}_{n,k,m,t}^{d_1,d_2,u} \left( 1 - \hat{p}_{n,k,m,t}^{d_1,d_2,u} \right) - \sum_{q>m} 2m \cdot q \cdot \hat{p}_{n,k,m,t}^{d_1,d_2,u} \cdot \hat{p}_{n,k,q,t}^{d_1,d_2,u} \right].
\]

(40)

The first equality follows from the monotone convergence theorem and from the independence of the stochastic location processes within the sum, allowing us to take the variance on the inside of the sum. The second equality follows from the variance of the stochastic location process:

\[
\text{Var} [\epsilon_u \cdot L_{s,k}(t)] = \mathbb{E} \left[ (\epsilon_u \cdot L_{s,k}(t))^2 \right] - \mathbb{E} [\epsilon_u \cdot L_{s,k,u}(t)]^2 = \sum_{m=0}^{M_u} m^2 p_{s,k,u}(m, t-s) - \left( \sum_{m=0}^{M_u} m p_{s,k,u}(m, t-s) \right)^2 = \sum_{m=0}^{M_u} m^2 p_{s,k,u}(m, t-s) - \sum_{m=0}^{M_u} m^2 p_{s,k,u}(m, t-s)^2 - \sum_{m=0}^{M_u} \sum_{n>m} 2mn \cdot p_{s,k,u}(m, t-s) \cdot p_{s,k,u}(n, t-s).
\]

The final equality of the variance calculation follows by applying the multinomial formula. \(\square\)

Proof of Lemma 3.1

Proof. To show the result, we begin with the maximum of two phase-type distributions and then apply the relationship recursively to obtain the general result. The maximum of two phase-type distributions can be calculated by

For ScholarOne support, you may contact 434-964-4100 or 888-503-1050 (US-based numbers).
The transitions are then defined by state, because this is the absorbing state for the chain max of 2 phase-type distributions; the general case follows from the same arguments.

For the sake of expositional clarity we present a proof for the case of the maximum of 2 phase-type distributions; the general case follows from the same arguments.

Proof of Theorem 3.2

Proof. This can be shown via equivalence of the compact MC with the MC for the general solution in Eq. 27. For the sake of expositional clarity we present a proof for the case of the maximum of 2 phase-type distributions; the general case follows from the same arguments.
Consider 2 phase-type distributions, \( X_{u_1} \) and \( X_{u_2} \) with generator matrices \( T_{u_1} \) and \( T_{u_2} \) respectively. The state space for the general solution is given by the \( \{(i, j) : i, j \in \{1, 2, \ldots, 6\}\} \), where 6 denotes service completed. Due to the structure of \( T_{u_1} \) and \( T_{u_2} \), the only possible transitions and their probabilities are given by

\[
\begin{align*}
\mathbb{P}(i,j) \to (i+1,j+1) &= \beta_{i,u_1} \beta_{j,u_2}, \\
\mathbb{P}(i,j) \to (i+1,6) &= \beta_{i,u_1} (1-\beta_{j,u_2}), \\
\mathbb{P}(i,j) \to (1,1) &= \beta_{i,u_1} \beta_{j,u_2}, \\
\mathbb{P}(i,j) \to (6,1) &= \beta_{i,u_1} (1-\beta_{j,u_2}), \\
\mathbb{P}(i,j) \to (1,6) &= \beta_{i,u_1}, \\
\mathbb{P}(i,j) \to (6,6) &= 1-\beta_{i,u_1}, \\
\mathbb{P}(i,j) \to (6,6) &= 1-\beta_{j,u_2},
\end{align*}
\]

(47)

where state \((6,6)\) is the absorbing state. For this chain, the difference between \( i \) and \( j \) will remain constant until one of them enters the absorbing state. Considering this and the fact that the initial distribution in our application \( (e_d) \) starts the search for both appointments on the same day of the week further reduces the state space that needs be modeled to the state space \( \{(i,i) : i \in \{1, 2, \ldots, 5\}\} \cup \{(i,6) : i \in \{1, 2, \ldots, 6\}\} \cup \{(6,i) : i \in \{1, 2, \ldots, 5\}\} \) because there is no path to any of the states where \( i \neq j \) for \( i, j < 6 \).

Another way to compute the \( V_{i,j} \) blocks of Eq. 29 for the example of \((i,j) = (1,2)\):

\[
V_{1,2} = \begin{bmatrix}
\text{State} & (2,2) & (2,6) & (6,2) \\
(1,1) & \beta_{1,u_1} \beta_{2,u_2} & \beta_{1,u_1} (1-\beta_{2,u_2}) & (1-\beta_{1,u_1}) \beta_{2,u_2} \\
(1,6) & 0 & \beta_{1,u_1} & 0 \\
(6,1) & 0 & 0 & \beta_{2,u_2}
\end{bmatrix},
\]

(49)

Comparing Eq. 49 with the transition probabilities from the general phase-type representation Eq.’s 47 - 48, each non-zero transition is accounted for in the compact representation.

**Proof of Corollary 3.1**

**Proof.** We prove the result by construction, presenting an algorithm for computing \( F(x) \) that achieves the required complexity. Consider a matrix that has block form

\[
V = \begin{bmatrix}
0 & V_{1,2} & 0 & 0 & 0 \\
0 & 0 & V_{2,3} & 0 & 0 \\
0 & 0 & 0 & \ddots & 0 \\
0 & 0 & 0 & 0 & V_{N-1,N} \\
V_{N,1} & 0 & 0 & 0 & 0
\end{bmatrix}.
\]

(50)

Each time \( V \) is taken to the \( x^{th} \) power, each block \( V_{i,i+1} \) is shifted up by \( x-1 \) blocks (wrapping around to the bottom when the block reaches the top most block in the matrix) and multiplied sequentially on the left by the non-zero blocks to the left of it (again wrapping around to the right as necessary). This is best demonstrated by a simple example.
\[
V^2 = \begin{bmatrix}
0 & 0 & v_{1,2}v_{2,3} & 0 & 0 \\
0 & 0 & 0 & \ddots & 0 \\
v_{N-1,N}v_{N,1} & 0 & 0 & 0 & 0 \\
0 & v_{N,1}v_{1,2} & 0 & 0 & 0 \\
\end{bmatrix}, \tag{51}
\]

\[
V^3 = \begin{bmatrix}
0 & 0 & 0 & \ddots & 0 \\
v_{N-2,N-1}v_{N-1,N}v_{N,1} & 0 & 0 & 0 & 0 \\
0 & v_{N-1,N}v_{N,1}v_{1,2} & 0 & 0 & 0 \\
0 & 0 & v_{N,1}v_{1,2}v_{2,3} & 0 & 0 \\
\end{bmatrix}, \tag{52}
\]

where the diagonal dots, \( \ddots \), in Eq. 52 represent a continuation of the pattern from the columns to the left of it where non-zero blocks will appear along the diagonal.

Thus the calculation of \( V^x \) decomposes into the multiplication of \( x \) matrices for each non-zero block and shifting that block’s position up by \( x \) blocks. Because each matrix \( V_{d,d+1} \) is a portion of the Kronecker product of \( L \) matrices of size \( 2 \times 2 \), the size of \( V_{d,d+1} \) is \( (2^L - 1) \times (2^L - 1) \).

To calculate the non-zero block of each column of \( V^x \) it requires \( x - 1 \) multiplications of the \( (2^L - 1) \times (2^L - 1) \) matrices, which is of order \( x \cdot U(2^L) \). Since there are \( N - 1 \) columns, this procedure must be repeated \( N - 1 \) times, leading to the desired complexity on the order of \( N \cdot x \cdot U(2^L) \).