Access and Resource Management for Clinical Care and Clinical Research in Multi-class Stochastic Queueing Networks

by

Jivan Deglise-Favre-Hawkinson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in The University of Michigan
2015

Doctoral Committee:

Professor Mark P. Van Oyen, Chair
Professor Mark S. Daskin
Assistant Professor Mariel S. Lavieri
Professor Blake J. Roessler

This dissertation is dedicated to my friends and mentors, Jonathan Helm and Mark Van Oyen, without whom none of this would have been possible. Their continuous support during this process has been absolutely invaluable and will never be forgotten.

ACKNOWLEDGEMENTS

I would like to thank my parents, Kenneth S. Hawkinson and Bernadette M. Deglise-Favre for all their support and believing in me my whole life. My academic advisor, Mark P. Van Oyen, was by far the best advisor one could hope for. He taught me so much and stood by me through everything. A key player in my academic success was my collaborator and mentor, Blake J. Roessler, whose tireless efforts and valued counsel have truly shaped me both personally and professionally. I would also like to thank my dear friend Olof H. Minto who has taught me so much about life. Thank you to my closest friends in the world Jonathan Helm, Asli Ozen, and Muge Capan for being there for me and helping me through tough times. Words cannot communicate enough how they have helped me throughout these five years. I also acknowledge and thank my godfather Dominique Magrina, and aunt Maryse Magrina for always keeping an eye on me, and supporting me from a distance. And to all my good friends at Michigan who made my time here enjoyable. I would like to thank my roommates Troy Long, Victor Wu, Wendy Shang, and Roman Gayduk, as well as Gregg Schell, Helen Zhang and Fang Dong for being colleagues and friends from start to finish. My crew in all that relates to Michigan sports: Brendan (college-readiness highschool scout) See, Gregory (Michigan man) King, and Marcial (buddy pass) Lapp. Thank you to Julian Pan, a great friend and now colleague, as well as his wife Michelle Pan. And finally, thanks to everyone at Michigan for all the memories.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

Industrial and systems engineering and health services research can come together to make important gains in health care by designing powerful capacity planning and scheduling decision support systems. Targeting the longitudinal integration of clinical care and clinical research studies under shared resources, this work's goal is novel, effective methods to enable better access to care resources, better health outcomes, and cost control. The focus is on fundamental innovations in capacity planning, optimization of system performance metrics, and model-based methods to achieve appointment scheduling for stochastic models of patient flow in networks of clinical care and/or research services. Identifying how to utilize/allocate resources to achieve a care institution's goals is a difficult task due to the heterogeneity, and urgency levels of the patients requiring service as well as the nature of the service processes.

1.1 Current challenges

Based on several collaborations developed with healthcare providers over the course of the last five years, this work targets some key challenges that most large health care institutions with significant research activities are faced with on a daily basis. To appreciate the need for the methods and tools developed in this work, we consider this illustrative example. After repeated visits to the primary care physician, and initial diagnostic tests, there is no clear diagnosis available to Chris. The doctor suspects that the condition might be serious. Chris and the primary care doctor agree that it is wise to be referred to a specialty care provider with the expertise and resources to obtain a diagnosis and treatment plan. Consider some of the questions relevant to the medical center that Chris might be asking:

(Q1) How long will it take me to get an appointment? We will refer to this as the root appointment. Medically, Chris will worry that irreversible damage may be sustained during this wait. The first/root visit is especially important in our methodology. The wait from the time of the visit request until the time of that appointment is referred to as ***access delay***.

(Q2) Chris considers that it will take some random number of visits/consultations and tests in different departments to get a clear diagnosis and treatment, and we refer to this as an *itinerary* of care. Chris may wonder how long it will take to reach the end of that care itinerary, which we call ***itinerary flow time***. Reaching a diagnostic quickly is important to start appropriate treatments as soon as possible if necessary, or relieve the patient's mind if not.

After the results of the diagnostic tests, a research coordinator may contact the specialty physician and try to recruit Chris to be a *participant* in a clinical research trial. Research trials are multi-visit protocols that are resource and time sensitive. Another question will enter Chris' mind:

(Q3) If I were to enroll in this clinical research trial, how long will it take? Clinical trials have rigid protocol requirements that provide an accurate estimation of their durations (from start to the end of the trial). Hence, Chris will be interested to know the ***Time to First Available Visit*** (TFAV) in the clinical trial. Longer waits to start the trial will affect Chris' incentive to accept being part of it.

The medical center needs to be able to answer the above questions. Currently, experienced guesses based on what happened in the past is the only way for medical centers to

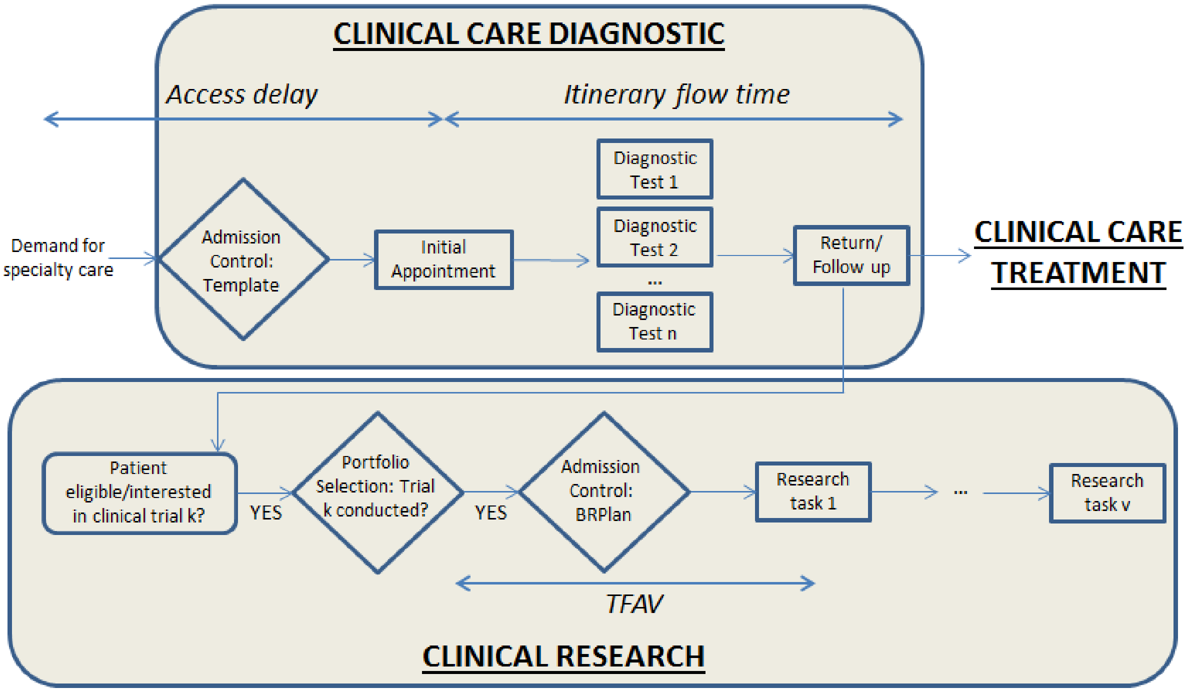


Figure 1.1: Patient flow through clinical care and research appointments under shared resources

estimate Chris' access to the many possible segments of his care/research itineraries. Leveraging forecasting methods to predict patients' resource requirements, this research is able to inform the institution of its predicted utilization levels, and how long patients like Chris would likely have to wait.

Consider the questions that the health care institution might have:

(Q4) How can we achieve differentiated access delays to care and itinerary flow times based on the patient's urgency level or condition? For example, patients with severe spine pain or an urgent problem will need fast diagnosis and access to surgery/treatment. The same resources, however, are used to serve patients with less severe and non-urgent back pain that do not require such fast access to all services. As another example, urology patients that present with a suspicion of cancer need to go through diagnostic testing and surgeon consultation. Those who need surgery are often urgent. These cancer patients will be much more time sensitive compared to patients having non-malignant cysts.

(Q5) When should the specialty physician schedule the follow-up visit? This follow-up visit can only happen once all the diagnostic tests have been performed, in order to provide the patient with a diagnosis. In common practice, it is difficult to predict at the time of the initial/root appointment how long it will take Chris to go through all the required diagnostic services. This length of time will be called the ***diagnostic flow time***.

(Q6) How much capacity should we allocate to clinical care visits vs. clinical research tasks? A tradeoff needs to be considered between the level of service provided to care for patients, and the amount of resources allocated to add research value to the community.

(Q7) Given a commitment to the ongoing care needs of the patient panel/population, which clinical research trials can/should be conducted within the limited resources of our services? Having a larger portfolio of research trials than capacity can handle will create congestion which will affect clinical care patients, and/or create very long access to clinical trial protocols. Hence, it is important to understand how much workload can be handled without hurting clinical care and how this translates to service levels for the trial participants.

(Q8) How should we schedule our human resources performing research tasks to provide efficient coordination between the logistical flows of the research participants and the time and skills required on their visits. Clinical trials usually require specialized skills on each appointment of a multi-visit protocol, and not all the resources have the skill set to perform those visits. To maximize throughput and reduce wait, one does not want physicians/nurses that possess many skills (for care and research) having to allocate their time on relatively simple tasks that physician assistants or technicians could do. *Resource coordination* based on matching a provider to an appropriate skill set to each participant's visit can be a complex, yet crucial element to delivering efficient clinical trials.

While the methods in this work can optimize which mix of clinical care offerings should be offered, we have chosen to focus on treating patient care as obligate and optimizing the mix of clinical research that can be added to it. The proposed capacity planning and scheduling methodology can serve a role as an integrating mechanism for accomplishing both the care and research dimensions of treatment in a manner that maximizes the value to patients and to research communities while managing costs and improving service by optimizing efficiency.

1.2 Capacity allocation and access management in other application fields

Capacity allocation of resources has also been studied in other areas. Manufacturing is one of them, but it is widely recognized that manufacturing models rarely transfer directly to service operations problems. Airline transportation is an area with a rich literature and history of industrial research focused on the operations of servicing transportation needs. As we will see, the problems confronted by airlines are significantly more complex than those in the healthcare operational problems we are targeting. To use differentiated terms for the people served in each systems, let us reserve customers for the airlines and patients for the healthcare application.

Airlines commonly structure the services/seats into differing levels of service (e.g., first class with the most amenities, business class, and economy) with higher fees for the higher levels of service. Contrary to the urgency classes we wish to treat in this dissertation, our urgency classes are not tied to amenities and perks, but focus on how quickly a common service (across patient classes) can be accessed. Of course, even in healthcare, there can be some providers that seek to attract higher paying patients through the promise of boutique services, but that is not the dominant practice.

This relates to the general observation that revenue management is a primary concern in the airlines, which leads to elaborate and complex dynamically changing prices that seek to extract the maximum profit from the customers and even to drive them away from one flight that may be highly utilized to one that may be harder to fill. While there is some movement toward revenue management, attitudes and social values indicate that a revenue management focus with dynamic pricing is inappropriate for the foreseeable future. It is the case that many healthcare institutions struggle to provide sufficient access. New patients tend to bring in significantly more revenue that existing/established ones. In this sense, by creating specific patient classes for new patients, our work does incorporate a dimension of revenue management.

The revenue management approach is motivated by and complicated by the fact that airlines serve in batches, with strong incentives to fly full flights. Of course, utilization is still important in healthcare, and our models include the metrics of utilization (or overutilization) and staff overtime. Clearly, healthcare operations are highly provider-centric, in contrast to major airline considerations such as geography, time zones, various classes of planes, and critical maintenance tasks.

Seat reservations for each patient class (usually determined by the ticket fare, and the patient status with the airline) are made for each aircraft in order to maximize revenue/utilization. However, the concept of customer waiting time (from request to occurrence) does not apply to airlines in the same way that it does for healthcare. Airline customers need to get to their destination on a specific date (or a short time window around it): they either travel for business (need to get someplace at a specific time), or during their vacation time. But ultimately, most customers have a date set in mind, and will choose itineraries/airlines based on pricing options. Hence, a customer usually won't delay her flight date when a flight is sold out; rather, they will pay extra for a higher fare class, choose a different airline/flight, choose a different mode of transportation, or cancel the trip. In healthcare, patients usually try to go through diagnosis and treatment as soon as possible. The demand is carried over until a date with sufficient capacity is available to treat them (acknowledging that patient preference for date and time may play a role, especially for less serious issues).

Interestingly, a good fraction of healthcare practice has patients coming to see a particular clinician/physician; however, in air transport customers seek only the particular geographical itinerary and the date/time. The plane's crew is irrelevant. This is a profound difference. Similarly, many healthcare visits are viewed as an unwelcome and relatively brief intrusion into one's daily schedule. While the particulars of date and time matter, the more brief the disruption of the office visit is, the less concerned the patient is with the date and time of the appointment. This is in marked contrast to the complexities of scheduling airline travel, which can be quite time consuming as a decision process. Most flights require a major commitment of time, money, and effort. Frequently, there are multiple time and date options available from various providers. In a sense, airlines create "booking reservation plans" that they provide for viewing to prospective customers. Thus, customers can see the seats in each class that are available and the pricing. Clearly, a visit to one's primary care provider implies a loyalty to a specific service provider. In some contexts, patients do shop various providers to find the one they like the best (e.g., a knee replacement); however, unlike air travel, healthcare provision attains greater value through the establishment of a relationship with the provider and the development of deeper insights over time. Our models can be employed at the level of an individual physician, or they can be applied to an entire department/service. While few healthcare providers would see a need to post their calendar online, certainly the availability of appointment dates is expected to be part of the conversation as patients set up an appointment. The intended use of our planning models is that the times available to a given type of patient will be limited so as to provide a better experience for the entire set of patients served.

Moreover, in healthcare, patients need various resources in their "itinerary of care", and those needs are stochastic and uncontrollable (they will depend on their condition and its severity, which may be learned over time). In airlines, the “itinerary” of customers going from point A to point B, then B to C, and so forth, usually with C = A is relatively deterministic once the customer buys the ticket. Of course, the vast majority of control over which services to pick lies with the customer (e.g. to get to Boston, I have a choice between a direct flight at some price, or a layover in New York City for another price). Even if we control the pricing to direct the customers to choose specific layovers, or build itineraries that would maximize the effective profit to the airline, this form of control is usually not appropriate in healthcare. Also, healthcare is often more stochastic, particularly in terms of the itinerary of visits that will be generated based on the outcome of a root visit.

After searching the airline literature, we could not find any airline studies that consider details at the passenger level for scheduling purposes. Our belief is that it is too difficult to optimize the system as a whole while considering each passenger individually. In fact, airline transportation problems have developed an elaborate system of decompositions and subproblems to allow solutions that avoid the intractability of treating everything in one model. Note that the airline transportation problem must consider the time of day of each flight, geography and network effects, the pilots, flight attendants, elaborate work rules, training time, rigorous processes of test and maintenance, food services, limited gate capacity, etc.

Because health services rarely have tight restrictions on the time of day that particular services are offered and because they do not require batches of people to be served at the same time, we are able to take a planning approach that models each day in aggregate. We leave time of day scheduling as a complementary piece that others have researched, and we can revisit in the future to take advantage of the daily level planning models we are developing.

It is worth noting the connections to queueing theory. Queueing methods can be used for revenue management (and capacity reservations). Past work in this vein does not directly translate to our healthcare research focusing on access management. However, we have taken the foundations of queueing and built them into MIPs that simultaneously optimize the system while approximating the queueing dynamics.

1.3 Structure of the research

This research seeks to provide differentiated and controlled access to care/research appointments in healthcare networks under shared resources by optimizing capacity allocation/reservation schemes for different patient/appointment types (e.g., care vs. research, urgent vs. less urgent). We are dealing with a large stochastic queueing network with dependencies/precedence between patients and appointments, which makes it difficult to answer the above eight questions in an integrated manner. Chapter II focuses on answering (Q3), (Q7) and (Q8) assuming the set of resources available are devoted only to clinical research visits. Such infrastructures are called "Clinical Research Units" (such as the Michigan Clinical Research Unit (MCRU), where no clinical care visits are performed). Chapter III focuses on (Q1) and (Q4), and Chapter IV will focus on (Q2), (Q4), and (Q5) in a hospital setting where all resources are devoted to clinical care patients (all non-research medical facilities). The conclusion chapter, Chapter V will address (Q6) and leverage the methodologies from the previous three chapters to summarize a high level procedure/algorithm that would answer all questions in a multi-stage optimization approach.

Chapter II creates an operations engineering and management methodology to optimize a complex operational planning and coordination challenge faced in healthcare. We motivate and tailor the method to the needs of sites that perform clinical research trials. The time-sensitive and resource-specific treatment sequences for each of the many trial protocols conducted at a site make it very difficult to capture the dynamics of this unusually complex system. Existing approaches for site planning and participant scheduling exhibit both excessively long and highly variable Time to First Available Visit (TFAV) metrics and high staff overtime costs. We have created a new method, termed CApacity Planning Tool And INformatics (CAPTAIN) that provides decision support to identify the most valuable set of research trials to conduct within available resources and a plan for how to book their participants while constraining (i) the staff overtime costs, and/or (ii) the TFAV by trial. To estimate the site's metrics via a Mixed Integer Program, CAPTAIN combines a participant trajectory forecasting methodology with an efficient visit booking reservation plan to allocate the date for the first visit of every participant's treatment sequence. It also creates a daily nursing staff schedule plan that is optimized together with the booking reservation plan to optimize each nurse's shift assignments efficiently to participants' requirements/needs.

To better service patients, support providers, and obtain greater operational efficiency, Chapter III develops operations models of direct impact to capacity planning and outpatient appointment scheduling. With the increasing popularity of Accountable Care Organizations (ACOs), destination medical programs, and other integrated care options, new challenges have arisen in managing a larger suite of formerly independent/siloed services. In particular, the capacity planning problem for medical services now becomes a coordinated network problem rather than a department by department local optimization. Considering the suite of services as a coordinated whole is critical because, in the integrated care model, patients will often require a sequence of visits to satisfy their diagnosis and care needs for one itinerary of care. Further, this sequence is often not known in advance, leading to decisions that are made at one service having significant and complex impacts on other services in the network.

Within this framework, there are a number of different types of patients with different urgency levels. At a destination hospital, for example, there are local patients versus regional patients (who may travel up to 6 hours to the hospital), and even national/international patients. We focus on managing the mix of patients to provide differentiated access to an initial/root appointment based on level of urgency. We manage this mix by reserving appointment slots across the network for each level of patient urgency and patient specialty type. For any given network capacity reservation scheme, we model the patient flow and access metrics as a queueing network with patients arriving according to a non-homogeneous Poisson process. After arrival, the patient seeks an appointment at the earliest possible time slot. If capacity is available, the patient is scheduled into the slot and if not the patient overflows and attempts to obtain an appointment on the next day. This process repeats until the patient has been scheduled. From a modeling perspective we obtain distributions on the amount of patients scheduled and amount overflowed through a set of recursive stochastic equations. Various access delay metrics can then be calculated from the set of scheduled and overflow random variables in combination with the arrival process.

Having developed results for the underlying stochastic network for a given capacity reservation scheme, the next step is to optimize the scheme relative to the access delay metrics of interest. To do so, we develop a suite of transformations to turn the stochastic metrics into linear, deterministic ones as approximations capturing the essential stochastic and non-linear properties of the system. This allows for the optimization to be solved as a Mixed Integer Program with chance constraints. We parameterize this model with data from a partner that operates a suite of outpatient services. We demonstrate through a case study that we can improve access for urgent patients while still guaranteeing an appropriate service level for other patients types.

In Chapter III, we construct new methodologies to control the access delay to the first appointment of an itinerary of care for each patient class. We are able to provide differentiated access to a care episode based on patients' urgency levels by allocating capacity across the resources. We guarantee a certain service level (e.g., limiting the probability of the workload exceeding capacity, performing limited workload in overtime) for downstream appointments of patient itineraries by assuming their resource needs followed historical patterns. There are limitations to these modeling assumptions: (1) the historical resource requirement needs of patients can be driven by historical blocking (e.g. a patient requires a downstream appointment tomorrow, but due to capacity, the patient had to wait three days historically), (2) the itinerary flow times (from first visit to last visit of an itinerary) are assumed to be fixed to the historical flow times, (3) the downstream demand that exceeds capacity on a given day is assumed to be done in overtime; in practice, this demand may alternatively be accommodated by delaying its service to a later day, and (4) itinerary flow times can not be differentiated by patients' urgency levels. In Chapter IV, we leverage the methods developed in Chapter III to create a model capable of overcoming the limitations mentioned above. Specifically, we will model and control how patients will flow (with delay) through the resources of this capacitated stochastic queueing network. Our main contribution is to provide expedited urgent patient types that flow through downstream itinerary resources by reserving the right amount of capacity for urgent cases while assuring the less urgent patients will not be delayed "too long". By optimizing the tradeoffs between providing differentiated access, itinerary flow times, and matching supply and demand, our Mixed Integer Program provides a resource capacity allocation plan that meets the itinerary flow time targets of a health care institution.

In Chapters II-IV, we assume that the set of resources for clinical care services have no overlap with the set of resources for clinical research. Using clinical care resources to perform clinical research trials can be quite challenging. The issue is really about running two distinct businesses: a clinical care business and a clinical research business. Aside from some operational differences (e.g., clinical research must follow a carefully prescribed protocol and sometimes requires skills/experience and processes not available in clinical care), the financial reality is complex and makes it undesirable to create a flexible operation.

Because clinical care will always trump clinical research (by its very nature, clinical research is a voluntary and elective intervention), dedicated clinical research chairs/beds, equipment, and staff are created to maintain the quality of the research and the scale. Billing insurance companies for clinical research is illegal and puts the business at risk for prosecution and fines if there is a perception that research is being charged as care. If one is willing to take the risk and mix the two businesses, then the back office administrative complexity increases and effort must be expended (even at the provider/caregiver level in documenting their time allocation) to sort out which specific activity can be billed to an insurance company and which needs to be billed to the clinical research sponsor. Further, using care resources to perform research is usually difficult due to the complexity of the research tasks: for example, monitoring the impact of an infusion on the body or the collection, processing, and storage of specimens. However, using clinical research resources for clinical care seems much more feasible in practice, and the combination of methodologies developed in this dissertation could be a first step towards that goal. In the conclusion chapter, Chapter V, we comment on ideas for the integration of clinical care and clinical research under shared resources. We present ideas for how to integrate the methodologies of the three previous chapters, and directions for future research.

CHAPTER II

Access Planning and Resource Coordination for Clinical Research Operations

2.1 Introduction

Clinical research trials are an essential part of discovering and testing new medical devices, procedures and drugs. This paper develops a methodology for appointment visit planning to improve clinical research trial delivery in dedicated performance sites is a gateway to better scientific knowledge, medical discoveries and effective new treatments for many diseases. “These trials, research studies carried out with human volunteers to answer specific questions concerning the effectiveness of a drug, device, treatment or diagnostic method, are designed to advance scientific knowledge and promote discoveries to treat and cure illness and disease, and increase longevity and the quality of life for countless people,” states [50]. While clinical trials are conducted in various settings, academic, medical, and industrial institutions frequently choose to construct clinical research performance sites to provide a shared infrastructure on which to perform them. We will use the term Clinical Research Unit (CRU) to refer to such an organization/site. Operationally, a CRU has intuitive economies of scale and scope by sharing resources across many trials. On the other hand, some clinical trials have a complex operational structure that makes it difficult to plan and schedule them at a shared CRU site. “The conduct of a clinical trial is a complex integration of many activities requiring the coordination of a large number of individuals each with specific expertise," states [25]. This research develops a strategic decision support system called CApacity Planning Tool And INformatics (CAPTAIN) for better resource planning and coordination, faster access to treatment, efficient booking, and trial portfolio selection decision support in CRUs (such as our partnering site, the Michigan Clinical Research Unit or MCRU).

2.1.1 The complexity of CRU operations and insights into their characteristics

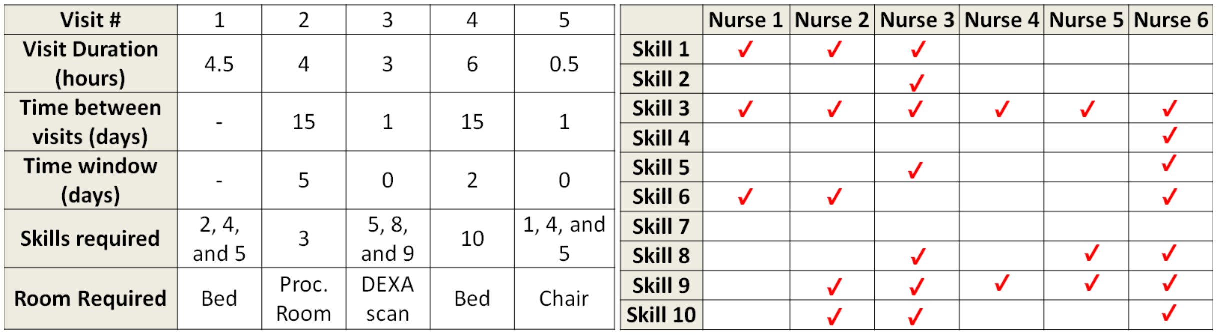
To gain an understanding of a typical CRU, it is useful to consider the perspectives of some key stakeholders, beginning with that of a participant. Participants (who may or may not be accurately called patients) can have many motives to participate in a trial, including the potential to receive a new treatment that is better than existing ones (although this would be speculative). For a participant who is screened and found to be appropriate to proceed in a trial, there will be a protocol specifying the details of how the trial must be conducted. "Clinical trials are detailed plans for medical treatments, for instance a clinical trial protocol can describe at which point of time, in which quantum, and how medications or therapies have to be executed," states [25].

The main operational features of a single clinical trial include: (i) each participant enrolling in a trial has a multi-visit treatment path (often between 1 and 10) with nearly deterministic service times that are specified by the trial's protocol, (ii) participant enrollments are stochastic over time, (iii) the visits of the treatment path are time-sensitive and can range from a few days up to a year (for example, visit 3 of a trial has to happen between 29 and 31 days after the first visit), and (iv) each trial visit requires a number of specific resources (chairs, beds, specific rooms/equipment, and specific skills of the nursing staff). We note that the determinism of service times in feature (i) mentioned above is not a strong modeling assumption because of the detailed protocol instructions. The research trial must have a predetermined number of participants to achieve statistical validity. Either missing a visit or experiencing excessive delay may invalidate that participant, so it is important to ensure that the CRU's care providers and physical resources are available to service the scheduled participants.

Second, individual research investigators and their trial coordinators can be viewed as customers who own the trials and work closely with CRU staff to administer the trials. Trial coordinators will follow a participant recruitment plan (based on judgment and past recruitment efforts for similar trials), possibly with nonstationary enrollment rates, from the initiation of the trial until either a designated number of participants has been reached or, in some cases, until a trial end date has been reached.

Third, operational planners for the CRU try to balance the competing needs for service of the trials/protocols. One difficulty is that the scheduling of all of a participant's visits needs to be done at the enrollment date. There should also be a check that the trial protocol's time and resource requirements will be met for all future visits. Let us elaborate on the complexity introduced by having nurses with varying skills. The left part of Fig. 2.1 provides an example of the specific requirements of a given clinical trial. This trial has 5 visits, and we see that visit 3 lasts 3 hours, needs to happen exactly 1 day after visit 2 (i.e., with a very tight time window), needs skills 5, 8 and 9 from the nursing staff, and uses the DEXA scan room. Based on the skills that each nurse possesses (see an example in the right part of Fig. 2.1), we can see that only nurses 3 and 6 would be able to perform that visit.

Information systems help to track the clinical and operational data needed. However, the complex resource and time coordination needs of the many trials of a CRU cannot be managed well under current planning and scheduling methods. Much manual entry into homegrown spreadsheets is typically required. A few CRUs have developed state of the art systems that identify the key human and physical resources required to fulfill the visits that are on the schedule, and they check for a resource conflict before allowing a centralized CRU



Visit #	1	2	3	4	5
Visit Duration (hours)	4.5	4	3	6	0.5
Time between visits (days)	-	15	1	15	1
Time window (days)	-	5	0	2	0
Skills required	2, 4, and 5	3	5, 8, and 9	10	1, 4, and 5
Room Required	Bed	Proc. Room	DEXA scan	Bed	Chair

	Nurse 1	Nurse 2	Nurse 3	Nurse 4	Nurse 5	Nurse 6
Skill 1	✓	✓	✓			
Skill 2			✓			
Skill 3	✓	✓	✓	✓	✓	✓
Skill 4						✓
Skill 5			✓			✓
Skill 6	✓	✓				✓
Skill 7						
Skill 8			✓		✓	✓
Skill 9		✓	✓	✓	✓	✓
Skill 10		✓	✓			✓

Figure 2.1: Example of the time and resource requirements of a 5 visit clinical trial (left) and the CRU nursing staff's skill sets (right)

scheduler to make an appointment. The methodology of this paper was developed to meet an unfilled need (i) to forecast and control key CRU performance metrics (e.g., avoiding high overutilization costs), (ii) to perform the cost/benefit tradeoffs of accepting/starting a particular new trial, and (iii) to generate a plan for how to keep participant access/wait times to an acceptable limit.

2.1.2 High level insights into CRU functioning

When a new trial is proposed to the CRU, the managerial staff has to make the decision of whether or not the scientific/financial value they will gain from it will outweigh the resources it requires (human as well as room), or whether or not there are even enough resources to accept it. Perhaps more importantly, there is currently no way to accurately determine the negative impact of a new trial imposed on existing trials (i.e., externalities) due to the increased congestion in the system.

CRU performance is dominated by the methods used to schedule the first visit of the treatment sequence required by the trial protocol, because subsequent visit times are dictated in the protocol from the time of the first visit. Failure to meet the required time window and resource requirements for any visit after the first results in a protocol violation that devalues research data and wastes resources. Furthermore, excessive violations will compromise the scientific integrity of the clinical trial if the target enrollment cannot be met. If all visits are properly booked at the time the participant enrolls, we avoid double booked appointments and congestion which may result in protocol violations or overutilization. As a rule, the managerial staff of a CRU would rather delay the start date of a participant's sequence of visits (or perform some visits in overtime) than risk a protocol violation. In current practice and in our planning model, the violation rate metric is approximately 0.

In Fig. 2.2 we illustrate an example of the current practice booking methodology, which we call "First Available Slot" (FAS) dynamic scheduling. We suppose that on day t a trial coordinator contacts the CRU to enroll a participant in a given trial. Suppose that this trial requires 3 visits: the second visit has to be scheduled between 10 and 12 days after the first visit, and the third visit has to be scheduled exactly 16 days after the first visit. The trial coordinator will communicate with the CRU to schedule this participant as soon as possible for his/her first visit. But this can only be done if there will be a nurse available with the required skills (this includes not only the nurses' nominal daily capacities but also the extra capacity that the CRU is willing to spend on overutilization) for all visits of the trial (the same goes for room/equipment availability). In Fig. 2.2, there are no nurses with the right skills available to schedule this participant for his/her first visit on day $t + 1$. On day $t + 2$, there is a nurse available with the required skills for visit 1 of such a trial. However, no nurse with the right skills for visit 2 is available from day $t + 11$ to day $t + 13$ (10 to 12 days after the first visit). There is no choice but to try to reschedule the participant's first visit to day $t + 3$. The same procedure is repeated (as seen in Fig. 2.2) until this participant is finally able to be scheduled on day $t + 5$ with a TFAV of 5 days.

Conceptually, it adds insight to consider the idea that a relatively more "complex" protocol contributes more to the Time to First Available Visit (TFAV) than does a simpler one. Loosely speaking, they cause more congestion to the system, and they also suffer more from

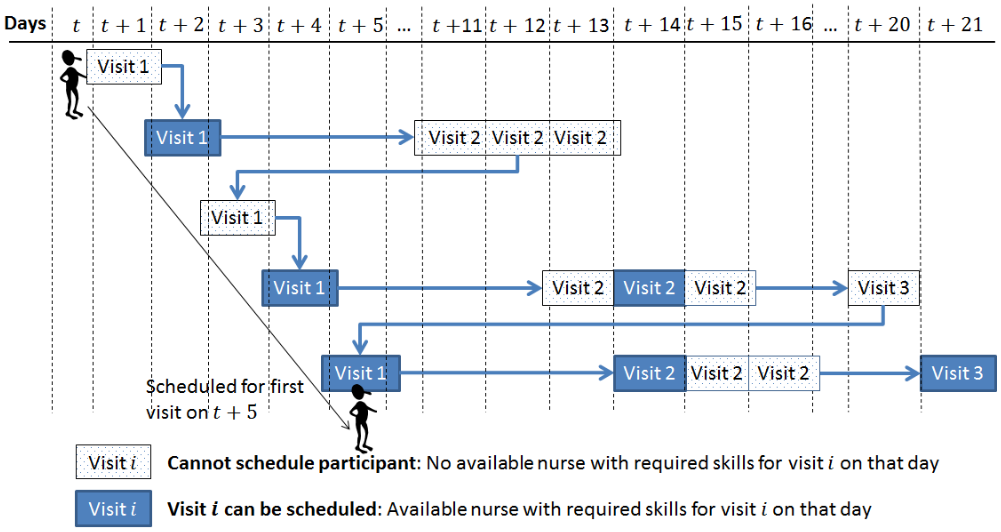

Days
t
$t+1$
$t+2$
$t+3$
$t+4$
$t+5$
...
$t+11$
$t+12$
$t+13$
$t+14$
$t+15$
$t+16$
...
$t+20$
$t+21$
Visit 1
Visit 1
Visit 2
Visit 2
Visit 2
Visit 1
Visit 1
Visit 2
Visit 2
Visit 2
Visit 3
Visit 1
Visit 2
Visit 2
Visit 2
Visit 3
Scheduled for first visit on $t+5$
Visit i
Cannot schedule participant: No available nurse with required skills for visit i on that day
Visit i
Visit i can be scheduled: Available nurse with required skills for visit i on that day

Figure 2.2: Example of the current practice booking methodology, the first available slot

congestion, experiencing relatively poorer access. Our intuition suggests that the complexity of a protocol increases (ceteris paribus) with (i) a larger number of visits in the trial, (ii) smaller visit time windows, (iii) shorter times between visits (i.e., the rate of resource consumption increases) and (iv) larger workloads placed on the CRU. Such "higher density" protocols tend to have a long and more variable TFAV. This effect has been observed in many healthcare settings and is elucidated as follows: if poor capacity planning either utilizes too little capacity (i.e., is inefficient) or permits too many protocols to be conducted then the system will experience significant congestion, thereby delaying the starting times of participants in every trial (increasing TFAV).

The more congested the system, the more difficult it is to schedule multi-visit participants, especially for high density protocols. We use simulation to show this behavior. In the left of Fig. 2.3, we observe the TFAV population average and population standard deviation of 16 trials with protocol characteristics populated by one problem instance of the randomized test suite described in Section 4.1 (see Table 2.1 in Section 4.1). The estimates are good enough to avoid the need for error bars. If we look at the TFAVs of trials in a congested system, we will see highly variable TFAVs (average TFAV for a simple trial vs. a dense trial) as well as high variability in TFAV for participants within the same trial (especially in high density protocols).

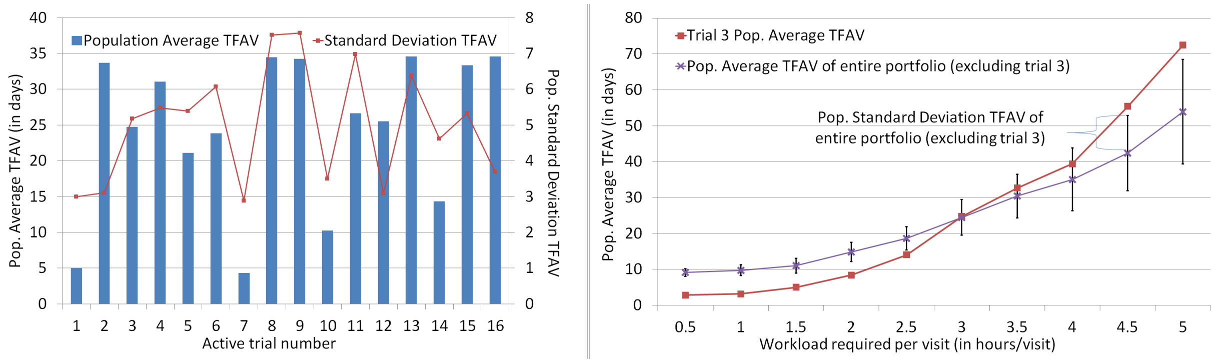


Figure 2.3: Impact of a trial's protocol density on it's own TFAV as well as the rest of the portfolio of trials; Left: Population average and standard deviation of the TFAV under the FAS policy for a given instance of the randomized test suite in Table 2.1; Right: TFAV mean and standard deviation of trial 3 and the same portfolio instance averaged across all 15 other protocols as the mean workload rate of trial 3 varies

In the right of Fig. 2.3, we illustrate the effects on the population average TFAV when the average workload per visit of trial number 3 increases (i.e., the trial's density increases). We used the same problem instance as in the left of Fig. 2.3 while varying the workload per visit of trial 3. Trial 3's enrollment rate constitutes roughly 8% of the total portfolio enrollment rate in the left graph, and the point with a workload rate of 3.0 hours per visit in the right graph corresponds to its value in the base case. First, as expected, the population average TFAV of trial 3 will increase nonlinearly in the average workload per visit. Also, we see not only a nonlinear increase in the population average TFAV over all trials in the portfolio excluding trial 3, but the standard deviation bars on the means are convex increasing in the variability around this TFAV. This shows how the density of one trial not only affects the TFAV for its own participants, but greatly affects the TFAV of participants of all other trials (since this portfolio is conducted under shared resources). This shows how selecting new trials needs to be done carefully while understanding/forecasting the impact that this will have on other trials and the CRU resources.

In clinical research practice, the longer the TFAV, the greater the fraction of participants who choose not to start the trial, thus increasing overall trial duration due to a lower enrollment rate. For example, high TFAV is also a problem for participants that need to start the trial as soon as possible due to their clinical condition (e.g., advanced cancer). Further, "each day it takes to bring [drug] products to the market is revenue lost: therefore, planning is critical to the entire drug development process, including clinical trial operations," [49]. Trial duration is a major problem in current practice.

The above concepts motivate our planning approach to provide improved and controlled access (relative to FAS) for participants overall and especially in high density trials. Under the current booking methodology, the participants enrolled in low density trials will have very good access to care at the expense of the participants enrolled in high density trials. The reason is that under this booking policy, the first available slot is always given to the current enrolling participant even if he/she is enrolled in a low density trial that is easy to book with low access (e.g., a short one visit trial) thus taking critical resource capacity away from the future participants in trial protocols that may be harder to schedule. This motivates our decision to reserve time slots (capacity) for first visits of each participant of each trial (see Section 3.2), while assuring that enough capacity is available for subsequent visits (with our participant flow forecasting methods in Section 3.4).

2.1.3 CRU needs for capacity planning and scheduling tools: A structured survey

To clearly identify the status of common sense approaches in CRUs, MCRU conducted a survey across 80 NIH-sponsored Cancer Centers, with 34 responses. The most difficult challenge was identified as knowing how to "Schedule the number/variety of protocols requested," with more than 95% selecting Strongly Agree or Agree. More than 80% of respondents agreed that a scheduling tool for predicting capacity and overutilization could help improve metrics. CRUs agreed or strongly agreed that “In an effective Clinical Research Unit Planning Tool, I would value the ability to: “predict the impact of a new trial ($> 90\%$), reduce overutilization expenses (80%), and reduce time-to-first-visit (80%). A real time scheduling tool which confirms that the required resources are available for all visits at time of booking was considered very valuable or valuable by 92% of the centers.

2.2 Literature survey

“Many clinical trials fail to deliver because of the lack of a structured, practical, businesslike approach to trial management,” observes [11] (see also [33] for further insight). Rigorous research as well as the software vendor community have left the topic of CRU operations almost untouched. Vendors have targeted the support of individual trials without providing support for managing the CRU as a system of shared resources. Systems like EPIC^{TM} (see [3]) provide basic scheduling and information management, but our research seeks to harness mathematical system modeling and optimization to provide functionality for planning, scheduling, and achievable trial performance. Survey results (provided in Section 1.3) from many leading CRUs indicate that they are experiencing major difficulties and are eager for science-based engineering systems for planning and scheduling. Most of the literature in clinical trial management focuses on how to design trials that are successful and efficient throughout the different stages that lead to drug development/commercialization (see [54, 12]). Some literature treats the recruiting process of participants (see [39]). At the same time, “patient recruitment should not start until the clinics and data center have demonstrated that they are properly staffed and equipped to support this activity,” states [41].

The research models and algorithmic solutions in the literature do not provide a methodology that simultaneously accounts for the following elements that we integrate: (i) timetables for multi-visit care pathways of different durations and resource requirements with specified tolerances for their timing, (ii) a multi-month transient planning horizon, (iii) various room resources, (iv) staff with specific skills/capabilities, (v) variable and uncertain participant enrollments, (vi) forecasting and controlling patient access by type, (vii) trial portfolio selection, (viii) forecasting and controlling overutilization costs, (ix) efficient and practical booking methodologies, and (x) coordinating the nursing staff and the physical resources with participants' treatment needs. Despite a large literature on scheduling and planning (e.g., see [21]), there is little extant work to suit the problem at hand.

[21] uses the term indirect waiting to describe the wait to get an appointment which we have captured at the daily level as TFAV. A prior model by [18] determines the best start times of trials (within a day) and the allocation of individual personnel to clinical tasks using a Mixed Integer Programming model. But this paper does not consider items (ii), (iii) and (vi)-(x). Moreover, the trial protocols they considered do not require more than one day to complete, whereas the sequence of visits to complete a trial can span many weeks in our model. By using historical data to construct a matrix that assigns probabilities for patients' ward locations over time, [28] presents a new method to predict and optimize the daily census at every ward in a hospital. By rearranging the elective/surgical schedules, they can control the offered load on each hospital ward in an equilibrium model that is cyclostationary. Our model differs from theirs in items (i), (ii), (iv), (vi), (vii), (viii), and (x). Relevant literature includes scheduling outpatients to reduce waiting times ([7]), and many others such as [46, 47] and [4]). [45] used approximate dynamic programming (ADP) to consider different patient types that require different levels of access to a single appointment for a diagnostic resource. This ADP approach was extended by [15] who allowed an interval to be specified (earliest date and latest date) for the appointment. Our system differs from

the prior two in items (i)-(iv), (vii), and (x).

2.3 The CAPTAIN methodology

CAPTAIN solves three interrelated problems: (i) selection of the portfolio of candidate clinical trials that can and should be added to maximize the scientific/financial value extracted, (ii) creation of the daily *Booking Reservation Plan* (BRPlan), which is a capacity allocation plan that reserves slots for trial visits, and (iii) construction of a daily nurse schedule (NSPlan) that accounts for every planning visit's requirements for nurse skills.

A key objective of a CRU is to extract the most value from the resources available, so to facilitate item (i), we will use a scalar metric of total value gained and incorporate it as one component in the decision of which trials to conduct. Our goal is to construct an optimized transient/non-homogeneous BRPlan and NSPlan that will estimate and control the critical CRU metrics defined earlier. We use a transient model since the portfolio of active trials in a CRU tend to vary with time: some new trials can be added, while some will end or be interrupted. The length of this planning horizon will be D business days of participant enrollments (e.g., 60 days).

To support implementation, it is important that we be realistic in modeling the way that CRUs would transition from any prior method for scheduling patients to our methodology. We allow for the current portfolio in the CRU (before the incorporation of the CAPTAIN tool), $\mathcal{P}^0$, to have been actively populated with appointments for T days (e.g., 720 business days). This initial condition induces a certain "committed" workload/capacity for each resource (skills and rooms) on each day of our planning horizon. This is generated by the FAS scheduling of patients that enrolled prior to CAPTAIN's period of planning. Those initial workload conditions will be taken as inputs to our math program. Our question is: how should the resource capacity still available be allocated to schedule our new participant enrollments from $\mathcal{P}^0$ and an optimized set of candidate trials while forecasting and controlling the concerned metrics? More information about the transient and implementation aspects of CAPTAIN can be found in Appendix C in Section 2.7.3. Note that all mathematical notations are also described in Appendix A in Section 2.7.1.

2.3.1 Scientific value and the objective function

To help evaluate potential trials to be added, the model accepts an input scientific value rating M_k for every candidate clinical trial $k \in \mathcal{P}^{cand}$. This is an obvious mechanism to obtain the greatest good from the CRU resources. Of course, setting $M_k = 1$ for all trials k is still a very useful model, provided constraints are placed on TFAV to provide equal access to all despite logistical differences between trials. In current practice, they schedule any new participant enrollment to the next available slot such that all future visits can adhere to the protocol requirements while also respecting capacity constraints. As a side effect, this results in a system which offers the best access (low TFAV) to the participants in the low density trials. Complex trial protocols suffer with long TFAV metrics which is a barrier to enrollment, may undermine the participant's health care, and lengthens the duration of the trial. The scientific value input allows the optimization to combine all the relevant metrics in allocating capacity while satisfying constraints on metrics that include TFAV and overutilization limits. Scientific value can have different meanings for different CRUs. While scoring M_k is a challenge outside the scope of this paper, one view is to rate the trials based on the scientific/research knowledge that can be gained from them (this is already done in practice). Scientific value could also incorporate the financial value gained from a trial, especially for industry funded trials. It is not difficult to set up the scoring system to ensure that the scientific goals take priority over finances by weighting the various categories appropriately.

One possible objective of the Mixed Integer Program can be summarized neatly by maximizing the following, but the deep computations all happen in the constraints, which are provided later:

$$\sum_{k \in \mathcal{P}} M_k h_k, \tag{2.1}$$

where the binary decision variable h_k is set to 1 if and only if trial $k \in \mathcal{P}^{cand}$ will be activated, and $\mathcal{P} = \mathcal{P}^{cand} \cup \mathcal{P}^0$. Note that h_k will be a parameter equal to 1 for trials $k \in \mathcal{P}^0$ that are still ongoing at the beginning of our planning horizon. The following sections will introduce and justify the modeling of our system's constraints, and a complete version of the Mixed Integer Program used in the numerical case studies can be found in Appendix D.

2.3.2 The Booking Reservation Plan (BRPlan)

The BRPlan is essentially a patient admission control mechanism that packs the participants' first visit requests into the planning space of physical, human, and time resources via planned appointment time blocks for each type of first visit planned and accounting for all follow-up visits. Our decision variables will set a maximum limit on the number of type k participants that can start their treatment sequence on day d of our planning horizon, thereby controlling the participant flow of enrollments. For each active and candidate trial $k \in \mathcal{P}$, our BRPlan decision variables, $\Theta_{k,d}$, will take values in $\mathbb{Z}^+$, and be defined over the set of days $\{1, ..., D + \overline{TFAV}_k^{max}\}$. The horizon extends $\overline{TFAV}_k^{max}$ days beyond the end of the planning horizon, D, because CAPTAIN schedules the enrolling participants (up to day D) within the population maximum TFAV, $\overline{TFAV}_k^{max}$, which is set by the decision maker. To compute the TFAV access metric and to optimize the BRPlan, a given type k participant enrolling will always be given the first available type k reserved capacity slot within the BRPlan. After the first visit, we account for participant's preference within the time windows of subsequent visits. The first visit for type k is booked to the first available day with sufficient reserved capacity by the BRPlan, so that we can emphasize the effectiveness of the capacity management method of a CRU to guarantee a level of access to the most urgent patients and to do so without dependence on the participants' personal preferences. The BRPlan differs from the current first available slot methodology, because it selects the first available slot reserved for type k participants (which was determined by an informed decision of all protocols' needs derived via optimization).

To quantify the stochastic booking process, we assume that a type k participant enrolling on day t can be scheduled for her first visit no earlier than day $t+1$. We provide the following key definition:

Definition II.1. Let $\alpha_{k,d,t}$ denote the nonnegative integer valued random number of type k first visit participants booked under the BRPlan on day $d \in \{1, ..., D + \overline{TFAV}_k^{max}\}$ and enrolled on day $t \in \{0, ..., D\}$, where $t < d$.

One can compute $\alpha_{k,d,t}$ recursively while considering the type k daily demand/enrollment random variables denoted by $A_{k,t}, t \in \{0, ..., D\}$ taking values in $\mathbb{Z}^+$, and the previously defined BRPlan type k admission limit decision variables, $\Theta_{k,d}$, as:

$$\alpha_{k,d,t} = \min\left\{ \left(\Theta_{k,d} - \sum_{l=0}^{t-1} \alpha_{k,d,l}\right), \left(A_{k,t} - \sum_{l=t+1}^{d-1} \alpha_{k,l,t}\right) \right\}, \tag{2.2}$$

where (i) the left hand side of the minimum is the integer number of type k reservations/slots remaining on day d for first visits after all participants who enrolled before day t have been scheduled, and (ii) the right hand side of the minimum is the random number of type k participants who enrolled on day t and who are not booked for their first visit appointment before day d (or later).

Our goal is to translate this set of stochastic non-linear equations into a set of deterministic linear constraints subject to Θ that can capture the mean of the $\alpha_{k,d,t}$ random variables. This information is needed to accurately estimate (i) how demand will be fulfilled according to a given BRPlan Θ, (ii) the delay to obtain a first visit, and (iii) how the CRU's resources will be utilized. The dynamics of our system are conceptually similar to the idea of blocking in queueing networks, which is difficult to compute except in special cases, and is non-linear in our main decision variable Θ. To overcome this, we consider approximation techniques that can tractably capture the key system dynamics linearly in order to forecast and control our desired metrics in a deterministic Mixed Integer Programming (MIP) optimization.

Although keeping a full information state of prior enrollments allows us to estimate our metrics exactly and linearly in our decision variables, an important mechanism to keep our optimization tractable is to approximate the information state by aggregating the enrollment information for the m or more days prior to current day t, with $m \geq 1$. The random variables $A_{k,t}$ take values in the set $I = \{0, 1, ..., \overline{I}\}$, with $\overline{I}$ being the maximum number of type k enrollments on a given day.

Definition II.2. For all $k \in \mathcal{P}, t \in \{1, ..., D\}, d \in \{t+1, ..., D + \overline{TFAV}_k^{max}\}$, $m \geq 1$, and $j^*(t) = \max\{t - m + 1, 1\}$ we define the set $\mathcal{A}_{k,d,t}(m)$ with elements $\alpha_{k,d,t}^{a_t, a_{t-1}, ..., a_{j^*(t)}, a} \in \mathbb{R}^+$, $a \in \{0, 1, ..., j^*(t) \cdot \overline{I}\}$, and the vector $(a_j)_{j \in \{j^*(t), ..., t\}} \in I$ that represents the conditional expectation of the number of patients that enrolled on day t and are scheduled on day d given a partial history of prior enrollments $(A_{k,t}, ..., A_{k,j^*(t)}, \sum_{i=0}^{j^*(t)-1} A_{k,i}) = (a_t, a_{t-1}, ..., a_{j^*(t)}, a)$, where the oldest history, days $\{0, 1, ..., j^*(t) - 1\}$ are aggregated as a sum.

As a first example, let's assume that t is day 15 and $m = 5$. Then we notice that $j^*(15) = 11$ and that our set $\mathcal{A}_{k,d,15}(5)$ will contain the conditional expectations $\alpha_{k,d,15}^{a_{15}, a_{14}, ..., a_{11}, a}$ of the number of participants enrolled on day 15 and scheduled on day $d > 15$ given the following partial history of enrollment: (a) the aggregated information of total enrollments from days 0 to 10 ($\sum_{i=0}^{10} A_{k,i} = a, \forall a \in \{0, ..., 11\overline{I}\}$), and (b) the daily enrollment information from days 11 to 15 ($A_{k,i} = a_i, \forall a_i \in I, \forall i \in \{11, ..., 15\}$). Now, in the case when $m > t$ (say $m = 20$), we will have $j^*(15) = 1$, and we notice that our set $\mathcal{A}_{k,d,15}(20)$ will contain the conditional expectations $\alpha_{k,d,15}^{a_{15}, a_{14}, ..., a_1, a}$ of the number of participants enrolled on day 15 and scheduled on day $d > 15$ given a full daily history of enrollment from days 0 to 15 (where day 0's enrollment realization will be characterized by a).

Remark II.3. In Definition II.2, we omitted the first day where $t = 0$ due to notational complexity. To address this issue, we will assume $j^*(0) = 0$ and use the convention $\alpha_{k,d,0}^{a_0} := \alpha_{k,d,0}^{a_0,a_{j^*(0)},a}$ to define the elements of the set $\mathcal{A}_{k,d,0}(m)$ similarly to Definition II.2. Hence, in the case $t = 0$, $\alpha_{k,d,0}^{a_0}$ will represent the conditional expectation of the number of patients that enrolled on day 0 and are scheduled on day d given that there were $A_{k,0} = a_0$ type k enrollments on day 0.

In Lemma II.4 and Theorem II.5, we show that a linear MIP can be formulated in our decision variables when we keep a full information state of prior enrollments: $m = \infty$.

Lemma II.4. *For all $k \in \mathcal{P}, t \in \{0, ..., D\}, d \in \{t+1, ..., D + \overline{TFAV}_k^{max}\}$, the elements $\alpha_{k,d,t}^{a_t,a_{t-1},...,a_0} \in \mathcal{A}_{k,d,t}(\infty)$ can be expressed in a MIP as a set of linear constraints subject to our decision variables Θ.*

To prove Lemma II.4 and all the following lemmas and theorems, see Appendix B in Section 2.7.2. We allow general (and time dependent) distributions for type k demand, so the model can use historical/empirical distributions from the given (or from a similar) trial's enrollment processes or use expert opinion to account for different rates of enrollments with time, which may be useful during periods of trial advertisement. With a finite $\overline{I}$, we let $P_{k,t}(a_l), a_l \in I$ with $I = \{0, 1, ..., \overline{I}\} \subset \mathbb{Z}^+$ be the probability that there are a enrollments of type k on day t (which can be drawn from historical and protocol data). Finally, we make the assumption that A_{k,t_1} is independent of A_{k,t_2} for all $t_1 \neq t_2 \in \{0, ..., D\}$ (which is a good assumption in practice even when the arrival process is nonstationary over time due to promotions).

Theorem II.5. *Given any BRPlan matrix Θ, and as $\overline{I} \to +\infty$ in the case of unbounded support for the enrollment distribution, $\overline{\alpha}_{k,d,t}$, the expected value of $\alpha_{k,d,t}$, can be computed linearly in our decision variables Θ as:*

$$\overline{\alpha}_{k,d,t} = \sum_{\alpha_{k,d,t}^{a_t,a_{t-1},\ldots,a_0} \in \mathcal{A}_{k,d,t}(\infty)} \alpha_{k,d,t}^{a_t,a_{t-1},\ldots,a_0} \cdot \prod_{l=0}^{t} P_{k,l}(a_l). \tag{2.3}$$

For every $k \in \mathcal{P}, t \in \{0, 1, ..., D\}$, and d such that $t < d \leq D + \overline{TFAV}_k^{max}$, the set $\mathcal{A}_{k,d,t}(\infty)$ will be composed of $(\bar{I} + 1)^{t+1}$ elements. In our setting, clinical trial daily enrollment/demand rates are rather low (generally less than 1 enrollment per day), which allows us to choose a small $\bar{I}$ while not losing much accuracy. We run into dimensionality issues as the length of our planning horizon gets large (e.g., t is 60 for a 12 business week planning horizon). For tractability, we approximate our booking methodology by reducing the cardinality of the set $\mathcal{A}_{k,d,t}(\infty)$. In all that follows, we use the set defined as $\mathcal{A}_{k,d,t}(m)$ in Definition II.2, with a partial aggregation of the enrollment information from days 0 to $j^*(t) - 1$. The goal is to express $\alpha_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a} \in \mathcal{A}_{k,d,t}(m)$ linearly in our decision variables, but Eq. 2.2 requires taking the conditional expectation of the minimum of two random variables. In the case $m = \infty$ (see Lemma II.4), conditioning on the complete history of enrollments allows the two elements of the minimum of Eq. 2.2 to be deterministic, so the expected value of the minimum will be the minimum of the expected values of each term. In the case $m < t$, this won't hold, and we rely on the following approximation. We define $\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a}$ as an element of the set $\hat{\mathcal{A}}_{k,d,t}(m)$ which approximates $\mathcal{A}_{k,d,t}(m)$, and so we define $\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a}$ to be:

$$\begin{aligned} \min\Bigg\{ &\mathbb{E}\left[A_{k,t} - \sum_{l=t+1}^{d-1} \alpha_{k,l,t} | (A_{k,t}, A_{k,t-1}, ..., A_{k,j^*(t)}, \sum_{i=0}^{j^*(t)-1} A_{k,i}) = (a_t, a_{t-1}, ..., a_{j^*(t)}, a)\right], \\ &\mathbb{E}\left[\Theta_{k,d} - \sum_{l=0}^{t-1} \alpha_{k,d,l} | (A_{k,t}, A_{k,t-1}, ..., A_{k,j^*(t)}, \sum_{i=0}^{j^*(t)-1} A_{k,i}) = (a_t, a_{t-1}, ..., a_{j^*(t)}, a)\right]\Bigg\}. \end{aligned} \tag{2.4}$$

We will assume that the elements $\alpha_{k,d,t}^{a_t,a_{t-1},...,a_{j^*(t)},a}$ of $\mathcal{A}_{k,d,t}(m)$ are close to $\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},...,a_{j^*(t)},a} \in \hat{\mathcal{A}}_{k,d,t}(m)$ for all $k \in \mathcal{P}, \forall t \in 0,...,D, \forall d \in t+1,...,D+\overline{TFAV}_k^{max}$. In other words, we assume equality in the well known Jensen's inequality (i.e. we assume $\mathbb{E}[\min\{X,Y\}|Z] = \min\{\mathbb{E}[X|Z], \mathbb{E}[Y|Z]\}$ for the random variables X, Y and Z). The less variability there is in the random variables $\mathbb{E}[X|Z]$ and $\mathbb{E}[Y|Z]$, the more accuracy we will have in our approximation. Hence, the more enrollment information we condition $\alpha_{k,d,t}$ on (i.e. the larger m is), the closer $\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},...,a_{j^*(t)},a}$ will be to $\alpha_{k,d,t}^{a_t,a_{t-1},...,a_{j^*(t)},a}$.

Proposition II.6. *Given $\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},...,a_{j^*(t)},a} \in \hat{\mathcal{A}}_{k,d,t}(m)$ defined in Eq. 2.4, and any BRPlan matrix Θ, as $\overline{I} \to +\infty$, $\overline{\alpha}_{k,d,t}$, the expected value of $\alpha_{k,d,t}$, can be approximated linearly in our decision variables Θ as:*

$$\overline{\alpha}_{k,d,t} \approx \sum_{\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},...,a_{j^*(t)},a} \in \hat{\mathcal{A}}_{k,d,t}(m)} \hat{\alpha}_{k,d,t}^{a_t,a_{t-1},...,a_{j^*(t)},a} \cdot \left(\prod_{l=j^*(t)}^{t} P_{k,l}(a_l) \right) \cdot \overline{P}_{k,j^*(t)}(a). \tag{2.5}$$

where $\overline{P}_{k,j^*(t)}(a) = \sum_{a_0+a_1+...+a_{j^*(t)-1}=a} \prod_{l=0}^{j^*(t)-1} P_{k,l}(a_l)$ is the probability of having exactly a type k enrollments in the first $j^*(t)$ days (from day 0 to $j^*(t)-1$) of our planning horizon.

Note that as m decreases, $|\hat{\mathcal{A}}_{k,d,t}(m)|$ decreases while trading off some accuracy when computing $\overline{\alpha}_{k,d,t}$ (since we aggregated information on past enrollment realizations). If we let $\mu_{k,d}$ be the random variable designating the number of type k participants scheduled on day d of our planning horizon under the BRPlan policy, we can compute its mean as:

$$\overline{\mu}_{k,d} = \mathbb{E}[\mu_{k,d}] = \sum_{t=0}^{d-1} \overline{\alpha}_{k,d,t}. \tag{2.6}$$

2.3.3 Formulation and control of the Time to First Available Visit

In this subsection, we want to control participants' access to trials. We formulate three TFAV metrics for a given trial k: (i) the average TFAV, denoted $TFAV_k^{avg}$, which is the mean TFAV averaged over all participants of type k that enrolled within our time horizon,

(ii) the TFAV-exception rate, $TFAV_{k,t}^{except}(\omega)$ is the expected value of the percentage of type k participants that enrolled on day t and that had to wait strictly longer than ω days before their initial visit, and (iii) the population maximum TFAV, denoted $\overline{TFAV}_k^{max}$, which captures the maximum expected number of days that a type k participant had to wait within our time horizon. In (ii), the TFAV-exception rate will be defined for each day $t \in \{0, ..., D\}$ by a type k sequence $\big(TFAV_{k,t}^{except}(\omega)\big)_{\omega \in \mathcal{T}}, \mathcal{T} \subset \{0, 1, 2, ..., \overline{TFAV}_k^{max}\}$, where each element is the expected fraction of type k participants enrolling on day t which will exceed ω days of waiting for their first visit (i.e., won't be scheduled before $t + \omega + 1$). By allowing the CRU administrator to control (i)-(iii), we will be able to reduce the access variability between participants of a given trial and shape the access delays across trials. The elements and the cardinality of the set $\mathcal{T}$ are to be custom selected to meet the relative trial priorities and participant treatment needs.

Theorem II.7. *The average type k TFAV can be expressed linearly in our BRPlan variables Θ as follows:*

$$TFAV_k^{avg} = \sum_{t=0}^{D} \sum_{d=t+1}^{D+\overline{TFAV}_k^{max}} \sum_{\alpha_{k,d,t}^{a_t, a_{t-1}, \dots, a_0} \in \mathcal{A}_{k,d,t}(\infty)} (d - t) \cdot \alpha_{k,d,t}^{a_t, a_{t-1}, \dots, a_0} \cdot \sum_{(a_D, a_{D-1}, \dots, a_{t+1}) \in I^{D-t}} \frac{\prod_{l=0}^{D} P_{k,l}(a_l)}{\sum_{i=0}^{D} a_i} \tag{2.7}$$

Next, the customer-average type k TFAV exception rate on day t for a chosen delay ω is computed linearly in Θ as follows:

$$TFAV_{k,t}^{except}(\omega) = \mathbb{E}\left[\sum_{d=t+\omega+1}^{D+\overline{TFAV}_k^{max}} \frac{\alpha_{k,d,t}}{A_{k,t}}\right] \tag{2.8}$$

$$= \sum_{(a_t,...,a_0)\in I^{t+1}} \sum_{d=t+\omega+1}^{D+\overline{TFAV}_k^{max}} \mathbb{E}\left[\frac{\alpha_{k,d,t}}{A_{k,t}}|(A_{k,t}, A_{k,t-1}, ..., A_{k,0}) = (a_t, a_{t-1}, ..., a_0)\right]$$

$$\cdot \prod_{l=0}^{t} P_{k,l}(a_l) \tag{2.9}$$

$$= \sum_{d=t+\omega+1}^{D+\overline{TFAV}_k^{max}} \sum_{\alpha_{k,d,t}^{a_t,a_{t-1},...,a_0}\in\mathcal{A}_{k,d,t}(\infty)} \frac{\alpha_{k,d,t}^{a_t,a_{t-1},...,a_0}}{a_t} \cdot \prod_{l=0}^{t} P_{k,l}(a_l). \tag{2.10}$$

Eq. 2.9 can be derived by conditioning Eq. 2.8's expectation with respect to $(A_t, ..., A_0) = (a_t, ..., a_0)$. Assuming a finite set I (which is a good assumption in practice), we can change the order of summation, and by using the definition of set $\mathcal{A}_{k,d,t}(\infty)$ and its elements $\alpha_{k,d,t}^{a_t,a_{t-1},...,a_0}$, we can derive Eq. 2.10. It will be linear in our decision variables thanks to our result in Lemma 3.1.

Remark II.8. Note that when solving this problem for long planning horizons, we can approximate the average TFAV (resp., TFAV exception rate) formulation in Theorem II.7 (resp. Eq. 2.10) by computing $TFAV_k^{avg}$ (resp., $TFAV_{k,t}^{except}$) as a linear combination of $\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},...,a_{j^*(t)},a}$ terms (see Section 2.3.2 for more details), as follows:

$$TFAV_k^{avg} \approx \sum_{t=0}^{D} \sum_{d=t+1}^{D+\overline{TFAV}_k^{max}} \sum_{\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},a_{j^*(t)},a}\in\hat{\mathcal{A}}_{k,d,t}(m)} (d-t)\cdot\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},a_{j^*(t)},a}$$

$$\cdot \sum_{(a_D,a_{D-1},...,a_{t+1})\in I^{D-t}} \frac{\overline{P}_{k,j^*(t)}(a)\cdot\prod_{l=j^*(t)}^{D} P_{k,l}(a_l)}{a+\sum_{i=j^*(t)}^{D} a_i}, \tag{2.11}$$

$$TFAV_{k,t}^{except}(\omega) \approx \sum_{d=t+\omega+1}^{D+\overline{TFAV}_k^{max}} \sum_{\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},a_{j^*(t)},a}\in\hat{\mathcal{A}}_{k,d,t}(m)} \frac{\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},a_{j^*(t)},a}}{a_t}\cdot\overline{P}_{k,j^*(t)}(a)\cdot\prod_{l=j^*(t)}^{t} P_{k,l}(a_l). \tag{2.12}$$

We can limit $TFAV_k^{avg}$ by constraining it by a specified upper bound on average access wait, denoted $\overline{TFAV}_k^{avg}$. Also, we let $p_{k,t}(\omega)$, with $\omega\in\mathcal{T}$ be the limit placed on the expected percentage of type k participants that enrolled on day t that are permitted to ω days of waiting for their first visit. Note that setting $p_{k,t}(\overline{TFAV}_k^{max}) = 0$ for all $t \in \{0, ..., D\}$ is equivalent to setting a population maximum TFAV limit of $\overline{TFAV}_k^{max}$ on trial k. This assures that type k enrollments on any given day t will be booked for their first visit on a day $d \leq t + \overline{TFAV}_k^{max}$.

2.3.4 The stochastic Participant Resource Requirement Model

The Participant Resource Requirement Model (PRRM) is a stochastic location process model (a particular random field) we developed to tractably capture randomness in participant resource needs over time, and it was inspired by [28]. Let the set of all possible physical locations/resources be $\mathcal{R}$ and the set of all skills of the personnel be $\mathcal{S}$. Then the entire set of resources is indexed by r in $\mathcal{R} \cup \mathcal{S}$. Note that specific skills are required on each visit of a trial (as opposed to specific nurses). Each skill is mapped via the nurse skill sets to one or more staff possessing that skill. This captures the unique skills of each individual staff member. First, we focus on the workloads induced on skills, and we will later assign those workloads to specific nurses.

The protocol information for any trial k includes (i) the number of visits V_k in the trial, (ii) the time window requirements on each visit v, $\Omega_{k,v}$, which represents a finite set of days that visit v can be scheduled after the first visit without provoking a deviation from protocol (patient preference is exercised only within those time windows), (iii) the set of clinical resources needed (for example it could be skill 1, skill 5, and the Procedure Room) needed for each visit v, and (iv) the deterministic workload, $\bar{w}_{k,v,r}$, induced by a visit v on resource r. As an example of item (ii), if $\Omega_{k,v} = \{10, 11, 12\}$ then a type k participant will require resources on a given day between 10 and 12 days after his/her first visit. There may exist some protocols where visit v has no time window requirement, and in this case we let $\Omega_{k,v} = \{1, 2, ..., L^{max}\}$, where L^{max} is the maximum length (in days) of the longest protocol (for a single participant). For first visits, there are no time windows since the participant can start on any day with available capacity, so by convention, we let $\Omega_{k,1} = \{0\}, \forall k \in \mathcal{P}$. We use the protocol information in items (i)-(iv) to create a participant flow model that forecasts the required workload of participants of type k, on and after their first visit. The vector $PRRM_{k,d_1}(d)$ determines, for a randomly sampled participant starting protocol k on day d_1, the number of hours he or she will need each resource $r \in \mathcal{R} \cup \mathcal{S} \cup \mathcal{H}$ on day d, where $\mathcal{H}$ is the "home" resource (i.e., the participant has no appointment in the CRU). Letting e_r be the unit vector with a 1 in the r^{th} column, our goal is to compute $\mathbb{P}\left(PRRM_{k,d_1}(d) \cdot e_r = w\right)$, the probability that a participant starting protocol k on day d_1 will require w hours of resource r on day d. Because participants enrolling in an active trial k on day d_1 have a uniform distribution of preferred days within their visits' time windows, this generates the following probability distribution of the workload induced on resource r (skills and rooms) on day d:

$$\mathbb{P}\left(PRRM_{k,d_1}(d) \cdot e_r = w\right) = \sum_{v=1}^{V_k} \mathbb{1}_{w=\bar{w}_{k,v,r}} \frac{\mathbb{1}_{\{(d-d_1)\in\Omega_{k,v}\}}}{|\Omega_{k,v}|}. \tag{2.13}$$

Remark II.9. From Eq. 2.13, we can compute the expected value of the PRRM, denoted $\overline{PRRM}_{k,d_1,r}(d)$, for all trials k:

$$\overline{PRRM}_{k,d_1,r}(d) := \mathbb{E}\left(PRRM_{k,d_1}(d) \cdot e_r\right) = \sum_{v=1}^{V_k} \frac{\mathbb{1}_{\{(d-d_1)\in\Omega_{k,v}\}}}{|\Omega_{k,v}|} \bar{w}_{k,v,r}. \tag{2.14}$$

In this paper, we assume a uniform distribution within the days of the visit time windows to model patient preferences. Note that in practice, we can generalize this by using patient preference estimated from historical distributions. Also, the required time window of visit v of a protocol is only dictated by the timing of the first visit. A trivial extension of our PRRM formulation would be to consider the time dependence between two subsequent visits (e.g. visit 3 has to occur between 10 and 12 days after visit 1, but also exactly one day after visit 2).

2.3.5 Formulation of the offered daily workload

As is common in queueing problems, computing the workloads in the system is fundamental. The daily offered workload for our planning horizon will depend on (i) $W_r^{init}(d)$, the existing workload on resource r and day d induced by participants that enrolled prior to the start of our horizon, (ii) the portfolio of the selected candidate trials chosen at the beginning of our planning horizon, (iii) the number of daily participant bookings for their first visit by trial, to be optimized by the reserved resource capacity plan, BRPlan, and (iv) the PRRM model of patient flow through subsequent visits. Since we are considering enrollments from days 0 to D, and we limit the trial k population maximum TFAV level to $\overline{TFAV_k^{max}}$, we need to predict the offered workloads up to day $D + L^{max} + \max_k\{\overline{TFAV}_k^{max}\}$ (where L^{max} is the duration in days from the first visit to the last visit of the longest protocol). The following theorem shows how to compute the offered workload on any day d and resource $r \in \mathcal{R} \cup \mathcal{S}$.

Theorem II.10. *Given $\bar{\mu}_{k,d_1}$, the mean offered workload for resource r ($r \in \mathcal{R} \cup \mathcal{S}$) on day d of our planning horizon, with $d = \{1, 2, ..., D + L^{max} + \max_k TFAV_k^{max}\}$, can be expressed linearly as follows:*

$$\bar{W}_r(d) = W_r^{init}(d) + \sum_{k \in \mathcal{P}} \sum_{d_1=1}^{\min\{d, D+\overline{TFAV}_k^{max}\}} \bar{\mu}_{k,d_1} \cdot \overline{PRRM}_{k,d_1,r}(d). \tag{2.15}$$

While Theorem II.10 applies to both the rooms and skills of the personnel, we now focus on the human resources, which we refer to as nurse resources (as non-nursing staff can be treated similarly). Below, index n denotes the n^{th} nurse in set $\mathcal{N}$ of all nurses working in the CRU. In order to achieve the desired targets on the overutilization and TFAV metrics, CAPTAIN has the ability to provide a Nurse Schedule Plan (NSPlan) that (i) plans for each day d, the number of hours $K_n(d)$ each nurse n should work on day d given the clinic's constraints on shift assignments, and (ii) coordinates participants' needs (skills required on each visit) with nurses' skill sets and availability. Feature (i) permits the model to generate a nursing shift assignment plan that takes into account the possible shift lengths of a CRU (for examples, see Case Studies A and B in Section 2.5.2) as well as the CRU's rules on shift assignments (an example is provided when the test suite is described in Section 2.4.1). In the MIP formulation in Appendix D (Section 2.7.4), we omit the nurse staff workload planning constraints, since these must be coded into the MIP in a manner unique to the case study (i.e. rules on shift assignments and shift lengths). Conceptually, the idea is to ensure that $\{(K_n(1), ..., K_n(D + L^{max} + \max_k TFAV_k^{max}) : n \in \mathcal{N}\} \in \Delta$, with the set Δ assuring that all staffing rules are obeyed. The modeling of skill sets allows for the consideration of training/cross-training in addition to hiring/expansion. In (ii), we need to compute decision variables $x_{s,n}(d)$ that determine the number of hours nurse n should be assigned to participants that require skill s on day d of the planning horizon while respecting the allowable shift lengths. Continuity of care is not a concern for the multiple visits of a trial, hence we do not need to assign the same nurse to all the visits of a participant. Note that Theorem II.10 determines the mean workload induced on the nursing skills, but not the nurses themselves (because protocols do not specify which nurse needs to be present for a visit, but rather the skills).

The decision variables $x_{s,n}(d)$ need to ensure that the offered workload on any skill on any day d (expressed in Theorem II.10) has to be assigned only to a nurse n that can perform that skill s, where $\mathcal{N}(s)$ denotes the set of nurses that possess skill s. Observe that such a decision not only affects the workload induced on the skill required by this visit, but also on all other skills that this nurse possesses due to shared capacity. Eq. 2.16 and 2.17 assure that the set of nurses that possess skill s can serve the anticipated workload, and that no nurse n will be assigned workload not in his/her skill set. Let

$$\overline{W}_s(d) = \sum_{n \in \mathcal{N}(s)} x_{s,n}(d), \forall s \in \mathcal{S}, \forall d \in \{1, 2, ..., D + L^{max} + \max_k \overline{TFAV}_k^{max}\}. \tag{2.16}$$

From Eq. 2.16, we can now compute the mean daily workload induced on each nurse:

$$\overline{W}(n, d) = \sum_{s \in \mathcal{S}} x_{s,n}(d), \forall n \in \mathcal{N}, \forall d \in \{1, 2, ..., D + L^{max} + \max_k \overline{TFAV}_k^{max}\}. \tag{2.17}$$

We allow the system administrator to impose a constraint on the maximum allowable, planned mean overutilization for nurse n on day d, denoted $\bar{O}_n(d)$:

$$\overline{W}(n, d) \leq K_n(d) + \overline{O}_n(d), \forall n \in \mathcal{N}, \forall d \in \{1, 2, ..., D + L^{max} + \max_k \overline{TFAV}_k^{max}\}. \tag{2.18}$$

For room resources (e.g., bed, chair, procedure room, exercise room, DEXA scan, etc.), the model is a degenerate case of nurse resources, because there is no need for “room flexibility” or a room “staffing plan.” We model a fixed daily available capacity (of 12 hours in MCRU’s case). From Theorem II.10, the expression for the mean overutilization hours induced each day on any room resource can simply be formulated in Eq. 2.59 (Appendix D in Section 2.7.4).

2.4 Validation approach

Having defined our decision support optimization model, we now validate our computations and approximations. To validate the concept of the BRPlan and our main performance metrics, we begin by assuming that there is no selection of additional trials and that the NSPlan is not optimized (i.e., we continue booking the participants of $\mathcal{P}_0$ in our planning horizon with the same nurse schedule as current practice, but now by following the BRPlan). Section 2.5.1 will test the portfolio selection feature of our problem while Section 2.5.2 will test the improvements made by the BRPlan and NSPlan (which will then be optimized) methodologies over the FAS policy. The validation in this section aims to test the accuracy of our BRPlan approximation and how this affects the forecasting of our metrics: the portfolio selection and nurse daily capacity decisions are not correlated to accuracy, but we do still validate our approach for many portfolio instances and nurse schedules of our randomized test suite (each instance having a fixed portfolio and nurse schedule). Therefore, in this section, our objective is to maximize the mean resource utilization/workload (which is equivalent to minimizing the overall mean TFAV of participants) subject to constraints on our other metrics. For all our following numerical results (validation and sensitivity/tradeoff analyses), we conservatively employ the $m = 1$ approximation, which highlights a worst case scenario for the approximation. Specifically, in the case where $m = 1$, we lose accuracy in computing the mean number of enrolled participants that will be booked on a given day, because we consider the number of enrollments on the current day, and aggregate all the enrollment information of previous days.

2.4.1 Test Suite

Our objective is to perform a performance evaluation that is more scientific than a case study of a specific CRU. Drawing from protocol characteristics from MCRU that are believed to be representative of many sites, we constructed a randomized test suite of protocols, CRU features, and CRU constraints on key metrics, that have characteristics spanning those in practice. We randomized (with uniform distributions) the number of active trials modeled as well as the protocol characteristics of all active and candidate trials. The ranges on the uniformly distributed parameters of our randomized test suite are provided in Table 2.1. In the following experimental studies, we also use a fixed number of rules on shift scheduling: it is required that a given nurse (1) cannot have two 12 hour shifts (the longest possible shift) two days in a row, (2) cannot be on call (a nonzero capacity) more than three days in a row, and (3) is assigned a total planned capacity of 40 hours in a week. Each performance site will have its own set of rules/preferences, and our methods are flexible enough to incorporate such specific requirements (see e.g., [34]).

Parameter	Range
Workload induced on a resource for a given visit (in hours)	1 - 6
Time between successive visits (in business days)	1 - 15
Visit time windows (in ± days)	0 - 10
Number of visits	1 - 10
Enrollment rates (in participants/day)	0.1 - 1
Number of active trials	6 - 16
Number of candidate trials to consider accepting	0 - 8
Number of nurses	6
Number of nurse skill sets	10
Possible shift lengths (in hours/day)	0, 4, 8 or 12
Number of room types	6
Room capacity (in number of hours per day for each type)	12 - 48
Daily overutilization target over all resources (in hours)	0 - 6
Population Average TFAV target $TFAV_k^{avg}$ (in business days)	10 - 25
Mean percentage of participants exceeding a given TFAV target (in %)	0 - 50
Population Maximum TFAV target (in business days)	$TFAV_k^{avg}$ - 40

Table 2.1: Test suite defined by trial protocol features (upper part of the table), CRU features (middle part of the table), and targets/bounds set on the main metrics (lower part of the table)

Note that in practice, candidate trials proposed to the CRU have a fully defined protocol, so the protocol parameters needed for the model will be known prior to the activation of the trial (including the desired number of participants). Estimating enrollments for candidate trials is challenging, but our experience and that of MCRU staff indicates that it can help to use historical enrollment data from trials with similar features. We randomized each nurse skill set as follows. Referring to the table on the right of Fig. 2.1, we use a probability of 0.5 in placing a check mark in each cell to denote that nurse n has skill s, with cells being i.i.d.

2.4.2 Simulation of the BRPlan for validation purposes

For purposes of validation, we wrote a custom C++ simulation and employed the test suite to see how an optimized BRPlan, $\Theta_{k,d}$, would behave in practice. The enrollment rates for all trials are modeled as Poisson processes for simplicity of exposition. The simulation takes in two of the main outputs from CAPTAIN: the BRPlan and the NSPlan. Also, it considers the committed workload (i.e., already scheduled visits) at the beginning of the planning horizon (based on a simulation of the prior history of the system to generate the irrevocable workload from participants that enrolled prior to our planning horizon). After an enrollment of a given type in our planning horizon, the simulation checks for capacity availability to schedule the participant based on the optimized BRPlan, $\Theta_{k,d}$, and the previous enrollments that already occupy some of the slots of the BRPlan. The simulation books each new participant for a first visit on the earliest future day when there is enough resource capacity reserved by the BRPlan for that type of trial. Once this day is identified, the participant is scheduled for the first visit, and all the downstream visits of the protocol are generated according to the PRRM (see Section 2.3.4) and the distributions associated with it (e.g., uniform distributions on the participant preference within the required time windows). The simulation then uses the optimized NSPlan to assign the visit (and workload) to a given nurse with the required skill such that all nurse work rules of the prior subsection are obeyed. This process is repeated on each day of our planning horizon. We run 1,000 sample path realizations of our planning horizon in order to get accurate metric estimates. The presented simulations are accurate to within 1% or better with a 95% confidence interval.

2.4.3 Validation of the BRPlan

Because we approximated the dynamic BRPlan scheduling methodology via a deterministic set of linear equations to be used as constraints in the optimization (see Section 2.3.2), we want to estimate the accuracy of our forecast of the mean daily number of type k participants scheduled under a given BRPlan ($\overline{\mu}_{k,d}$). This section validates our BRPlan approach by comparing these approximations to the sample mean of the simulated model $\bar{\mu}^{sim}_{k,d}$, (which mimics exactly how the BRPlan could be implemented very simply in practice). We will use the Absolute Percentage Error (APE) metric defined as follows:

$$APE_{k,d}^{BRPlan} = |\frac{\overline{\mu}_{k,d} - \bar{\mu}_{k,d}^{sim}}{\bar{\mu}_{k,d}^{sim}}| \cdot 100\%. \tag{2.19}$$

By investigating many problem instances/realizations of the test suite described in Table 2.1, we confirmed the intuition that the accuracy of our approximated analytical BRPlan approach varied from one trial to another within a single simulated CRU instance. The BRPlan reserves capacity for each trial, hence the booking process of each trial is independent of another trial's booking once the $\Theta_{k,d}$ decision variables are all fixed. We found that the APE was most sensitive to two characteristics: (i) the bound we set on the trial's TFAV (mean, max and/or TFAV exception), and (ii) the "density" - a concept hard to quantify but increasing in higher enrollment rates, greater workload induced, shorter time windows, shorter time between visits, and a greater number of visits per participant. In Fig. 2.4, we investigated the sensitivity of the BRPlan APE to the enrollment rate of the trial (Fig. 2.4 left) and the maximum bound set on TFAV for the given trial (Fig. 2.4 right). Note that all other parameters were fixed and were set to a given realization of the randomized test suite of Table 2.1. We averaged our daily APEs over all days of our planning horizon to aid presentation.

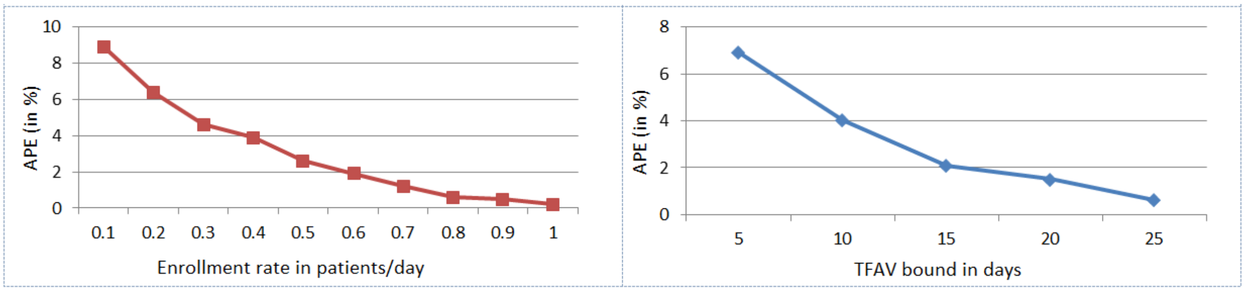


Figure 2.4: BRPlan APE of the number of patients scheduled (as defined in Eq. 2.19) for a given trial as a function of (i) Left: its enrollment rate and (ii) Right: the bound we set on maximum TFAV

While we do not have space to present examples for every conjectured component of trial “density” (see the discussion in Section 2.1.2) our results suggest that our approach works best either for (i) high density trials or for (ii) trials that can tolerate longer waiting times. These two factors are generally related, because higher density trials required longer waits to first visit. Considering the practical desire for high throughput, our BRPlan will usually reflect high utilization which by experimentation we found to correspond to close gaps between the number of patients scheduled and the reserved capacity $\Theta_{k,d}$. In cases approaching heavy traffic, CRUs and/or specific trials will tend to have high enrollment rates and long waits, and in these situations our approximations work very well. For (ii), as the waiting time target gets larger for a trial, the number of days between two successive non zero $\Theta_{k,d}$ values gets larger (ceteris paribus). Hence, the number of participants that enrolled but still haven’t gone through their first visit will also grow larger. By accumulating this unfulfilled demand for first visit (with a fixed arrival rate), the variability around the number of patients that will be scheduled in the reserved capacity will be reduced. Note that when the enrollment interarrival times are deterministic, the APEs are 0 and our analytical approach is exact. In conclusion, our approximated BRPlan methodology tends to work best in high utilization situations and low variability settings such as those with relatively longer TFAV times and more predictable enrollment times.

2.4.4 Validation of the average TFAV

To illustrate the accuracy of the average TFAV estimation method, we will illustrate one problem instance of our randomized test suite from Table 2.1 so we can describe the details driving accuracy. We considered 16 trials that varied a lot based on their characteristics including density. In Fig. 2.5, we compare the analytical average TFAV optimized via the linearization in Eq. 2.45 to the average TFAV computed via the simulation (by computing the TFAV APE value similarly to Eq. 2.19). The APE accuracy of our estimations for each trial is shown in Table 2.2. Just as we suggested in Section 4.3, we see that the accuracy

(APE) varies across trials with different characteristics. As we see later in a test case (see Fig. 2.7 in Section 5.2), the bounds on population average TFAV reveal that the trials with higher waiting time bounds achieved smaller APEs. This aligns with our statement that our methods work best in a high utilization setting.

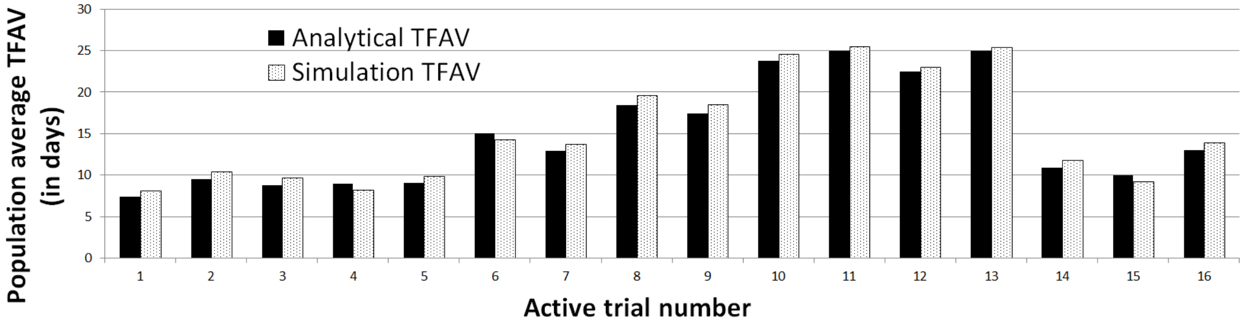


Figure 2.5: Validation of period 1 trials' population average TFAV

Trial:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
APE	7.8	8.3	8.8	8.9	7.4	5.0	5.6	5.8	5.5	3.2	1.7	2.0	1.4	7.1	8.6	5.6

Table 2.2: Absolute Percentage Errors (APEs) in % of the population average TFAV by trial over one instance/realization of the randomized test suite in Table 2.1

2.4.5 Validation of the mean offered workload

We provide validation of our analytical workloads by calculating the daily average APE for each nurse skill ($S_i, i = 1, ..., 10$) and each of the 30 instances of the randomized test suite of Table 2.1 considered. In Table 3, we then provide an average and standard deviation (STD) over those 30 instances of the daily APEs. We note that the forecasting of the total aggregate workload induced on the entire CRU (denoted "Ag. Skills" in Table 2.3) is very accurate with only a 1.5% error on average. Similarly, the error for room resources can be found, but we omit them because rooms are not the capacity constraint and represent a special case of human skills (since rooms have fixed daily capacities).

Workload induced on:	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Ag. Skills
Average APE (%)	2.4	2.5	2.3	2.5	2.8	2.5	2.3	3.4	2.5	3.5	1.5
STD APE (%)	4.6	3.9	2.8	3.3	3.6	3.6	4.0	4.2	4.0	4.7	2.3

Table 2.3: Daily average Absolute Percentage Errors (APEs) (average and standard deviation over 30 instances of the randomized test suite of Table 2.1) of the offered workload forecasts by skill S_i.

2.5 Optimization results

Having formulated our performance metrics linearly in our decision variables, we can constrain them based on managerial inputs: (i) $\bar{O}_r(d)$, the maximum allowable number of overutilization hours for resource r (for nurses and/or rooms) and each day d, (ii) $\overline{TFAV}_k^{avg}$, the upper limit of the population average TFAV for each trial, (iii) $TFAV_{k,t}^{except}(\omega)$, the TFAV exception limiting the percentage of type k participants enrolling on day t that will exceed ω days TFAV, and (iv) $\overline{TFAV}_k^{max}$, the population maximum TFAV for each trial k. The optimization functionality of CAPTAIN was implemented using Visual Studio 2010 with an integrated IBM CPLEX solver. The computer used was a 2007 Intel Xeon CPU with 8GB of RAM. For a 60 business day (12 weeks) planning horizon, this program was solved up to a 1% optimality gap in 1 to 6 hours depending on the protocol complexities. As the planning horizon gets larger, the computing time increases exponentially. However, 12 weeks is already considered as a very long planning horizon for our application, since CRUs tend to update their portfolios more frequently.

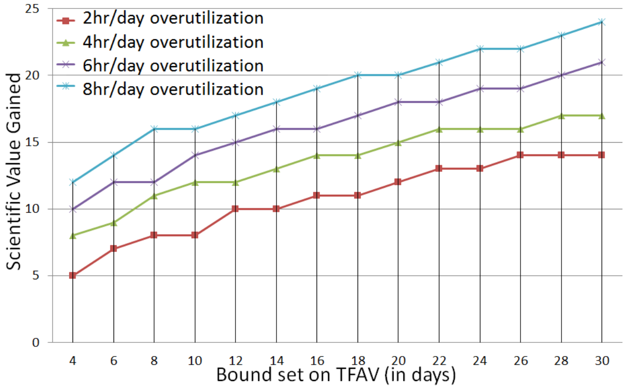
2.5.1 Tradeoff Analysis and Portfolio Selection

We studied the optimal objective values and the associated portfolio selections under various constraints set on daily overutilization and on average TFAV for the candidate trials. We considered 8 candidate trials that could be added to the current portfolio in our planning horizon. The candidate trials also have randomly generated parameters according to Table

2.1. The goal of this section is (i) to provide high level validation that the optimization's objective behaves as expected when the constraints on TFAV and overutilization are altered, and (ii) to show that the selection of a portfolio is not a simple, intuitive process (because of the complex interactions between protocols). This motivates the critical need for a decision support system to manage a CRU.

A rough, high level validation of our method is seen in the left part of Fig. 2.6. A CRU can extract greater scientific value from an optimized portfolio when the bound set on the population average TFAV for all candidate trials is increased by the CRU. By having the ability to delay the access of less delay-sensitive participants, the method is able to integrate more trials in our planning horizon's updated portfolio. Moreover, as seen on the left for a fixed population average TFAV for all candidate trials, the four experiments that increase the overutilization bound show that the CRU can extract more scientific value from the portfolio chosen. For example, by allowing more overutilization hours/expenses, high density trials that were difficult to schedule because of the protocol requirements (e.g., those with very long service times, short time windows for subsequent visits to happen, etc.) may become feasible to offer.

In the right part of Fig. 2.6, we represent the optimal portfolio selections of the candidate set of trials $\{C_1, ..., C_8\}$ each having given bounds on daily overutilization and TFAV. We see that this decision is neither trivial nor intuitive. By increasing the population average TFAV bound for all candidate trials as indicated in the rows (respectively the overutilization bounds per day, identified for each table), we see that the optimal portfolio selection is not always just monotonically adding new trials to the optimal portfolio selection with lower TFAV bounds (resp. overutilization bounds). For example, with an overutilization bound of 4 hours per day, we would only choose candidate trial C2 when the TFAV bound is 5 days. But when the TFAV bound is increased to 15 days, trial C2 is replaced by C1, C3 and C4



- Total overutilization limit of 2 hours per day

Average TFAV bound\Candidate trial	C1	C2	C3	C4	C5	C6	C7	C8
5 days		√						
15 days			√	√				
25 days		√	√	√				

- Total overutilization limit of 4 hours per day

Average TFAV bound\Candidate trial	C1	C2	C3	C4	C5	C6	C7	C8
5 days		√						
15 days	√		√	√				
25 days			√	√			√	

- Total overutilization limit of 6 hours per day

Average TFAV bound\Candidate trial	C1	C2	C3	C4	C5	C6	C7	C8
5 days			√	√				
15 days			√	√			√	
25 days		√	√	√			√	

Figure 2.6: Left: Tradeoff Analyses of the Scientific Value Gained vs. the bound set on population mean TFAV (in days) for all candidate trials and the overutilization bound in hours per day over all nurses; Right: Optimal portfolio selection with respect to population mean TFAV for all trials when the overutilization bound is 2, 4 and 6 hours per day over all nurses

(and we see similar complex behavior when the constraint on overutilization varies). This motivates the utility of the CAPTAIN methodology, since the managerial staff is not able to accurately predict how a given portfolio would impact the site's performance metrics.

2.5.2 Improvements

To assess the benefits of CAPTAIN, we again assumed that no trials would be added at the beginning of our planning horizon, because this allows us to compare the difference between an optimized BRPlan and the current practice of FAS scheduling for a given portfolio generated by the randomized test suite (Table 2.1). To illustrate how using CAPTAIN as a booking/planning tool to schedule participants (ignoring the portfolio selection component) is itself very helpful to control and improve the CRU metrics. This is done by maximizing aggregated resource utilization/workloads over all days of our planning horizon, and letting the solver find an optimized BRPlan that will satisfy the constraints set on mean overutilization, population mean TFAV, TFAV-exception, and population maximum TFAV for each trial. Note that the objective could also be to minimize on of our TFAV metrics (mean, maximum and/or, TFAV-exception) weighted across all active trials.

Fig. 2.7 presents a visual illustration of the capabilities and potential of CAPTAIN for one representative problem instance (portfolio of trials) from our randomized test suite in Table 2.1. The upper left of Fig. 2.7 shows the population average TFAV, the population standard deviation TFAV, and the population maximum TFAV for each trial in a CRU that operates under the current FAS daily level booking policy with a total overutilization limit of 4 hours per day (over all human resources). This was illustrated in the example of Fig. 2.3 - left, but we added the population maximum TFAV for each trial on the graph. In this problem instance, the FAS policy resulted in a total of 1.4 overutilization hours per day on average over all resources (given an overutilization limit of 4 hours per day over all resources). The capability of CAPTAIN allows the user to input any desired target on mean overutilization (per resource and per day), but in order to show in a fair way the improvements made by CAPTAIN, we kept the mean overutilization to the same level as the FAS policy in Fig. 2.7. The upper right of Fig. 2.7 shows the population average TFAV, the population standard deviation TFAV (from simulation after the BRPlan has been optimized), and the population maximum TFAV for each trial in a CRU that operates under CAPTAIN's daily level BRPlan methodology and achieves on average the same overutilization than the FAS policy while also meeting certain targets on population average TFAV for each trial. The horizontal dashed lines show the population average TFAV targets we set on each trial in our optimization. Although we are able to achieve such targets (which gives us the control to prioritize needier participants and/or high-valued trials), we realize that the population standard deviation TFAV for trials with "poor" access (e.g., trials 8-13) tends to be larger than those under FAS policy. In contrast, trials with "good" access (e.g., trials 1-5) will have lower TFAV variability than under the FAS policy. The explanation is that trials with higher population average TFAV targets will have a larger number of successive days with no BRPlan reservations (since they are not prioritized) which significantly increases TFAV variability between participants. Decision makers may want to avoid such scenarios. For example, it is likely that guaranteeing 25 days access on average to their participants will result in a high number of them waiting longer than that due to variability.

One will often want to reduce the TFAV-exception rate for each trial to guarantee that no more than a certain percentage of type k participants will exceed a given target. We added those constraints to the optimization (while keeping the previous overutilization and population average TFAV constraints) by selecting a TFAV-exception target of $\left\lceil \overline{TFAV}_k^{avg} + \sqrt{\overline{TFAV}_k^{avg}} \right\rceil$ and a TFAV-exception rate of 10% for each trial k. The results are illustrated in the lower left of Fig. 2.7. Although we see a slight increase in the population standard deviation of trials with high access, we were able to decrease the access variability for participants in trials that have a larger average TFAV target. However, the population maximum TFAV of those trials has increased compared to the population maximum TFAV of the upper right of Fig. 2.7, even though we have decreased the population standard deviation of those trials. Notice that by definition, the TFAV-exception controls the percentage (or number) of participants exceeding a TFAV target, but not the number of days that those participants will exceed the target. Hence, in the lower left of Fig. 2.7, we may have a small number of participants from trials 8-13 that exceed the $\left\lceil \overline{TFAV}_k^{avg} + \sqrt{\overline{TFAV}_k^{avg}} \right\rceil$ target. However the few that do are likely to exceed it by a lot, thereby increasing in the population maximum TFAV. Again, decision makers may want to avoid such situations and also be able to control the population maximum TFAV such that there isn't even a single patient that is predicted to wait extremely long for first visit.

In the lower right of Fig. 2.7, we see the portfolio TFAV results of a BRPlan that assures (i) population average TFAV targets per trial, (ii) TFAV-exception targets for each trial, and (iii) population maximum TFAV targets per trial (indicated by solid horizontal lines).

Of course, reducing the population maximum TFAV while keeping the TFAV-exception % the same will result in a decrease in the population standard deviation.

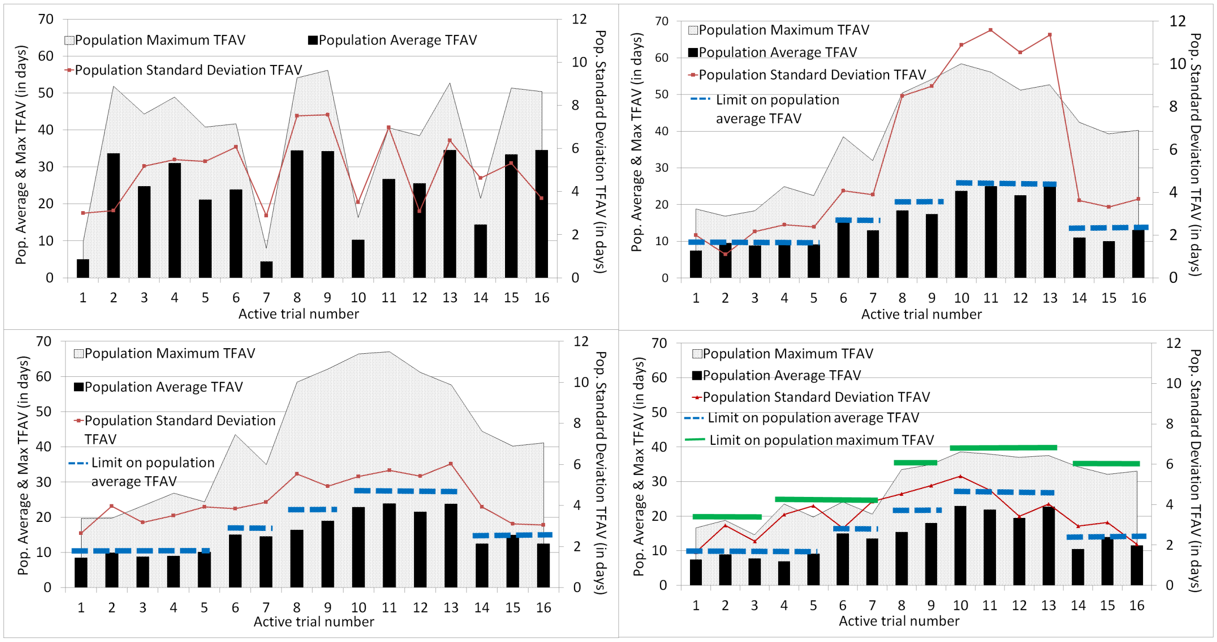
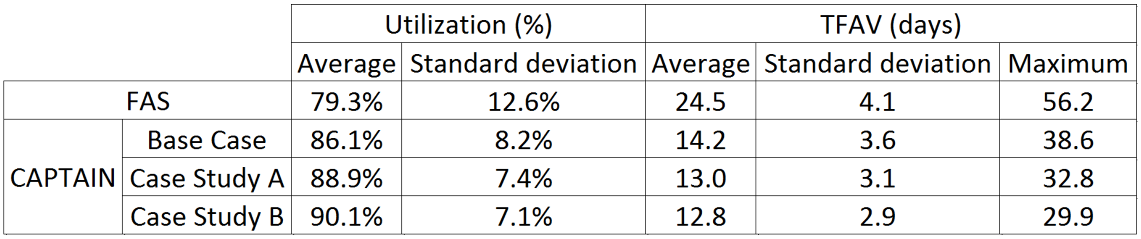


Figure 2.7: Simulation and optimization analysis of CAPTAIN's improvements and controls over TFAV variability compared to current practice under the same overutilization level of 1.4 hours per day over all resources. Upper left - FAS scheduling; Upper right - BRPlan with average TFAV constraints indicated by dashed lines; Lower left - BRPlan with constraints of upper right graph and a TFAV-exception of 10%; Lower right - BRPlan with constraints of lower left graph and a TFAV maximum as indicated by solid lines

We observe significant benefits from using the BRPlan booking methodology rather than the FAS policy. With a mean overutilization limit of 1.4 hours, the BRPlan was able to reduce the population maximum TFAV (the longest time any participant will have to wait) from over 55 days (using the FAS policy, upper left of Fig. 2.7) down to 40 days (lower right of Fig. 2.7). Also, we see that we were able to limit the population average TFAV for each trial to the desired limit, therefore significantly increasing access to care for more urgent participants. By adding constraints on the TFAV-exception rate, we are able not only to reduce TFAV variability between participants of the same trials, but we are also able to give the managerial staff some valuable information: they will then be able to advertise that a given trial guarantees y % (e.g., 90 %) of its participants will have access within x days (e.g., 15 business days) after enrollment. Having this information may increase enrollment (by increasing the perceived value/service), increase participant satisfaction, and decrease overall trial duration. We notice that under both scheduling methodologies (FAS and BRPlan), we have the same amount of daily capacity at our disposition, but we are able to limit average waiting times to significantly shorter amounts while also reducing access variability for each trial. This happens because the BRPlan is able to reduce the workload predictable and unpredictable variability by packing participant visits more efficiently while the FAS rule generates big "holes" (and workload variability) in the schedule. The BRPlan levels the daily workload (and reduces its variability) while also providing overall better access, which results in an increase in resource utilization for a transient planning horizon (see also the FAS and CAPTAIN-Base Case rows in Table 2.4).



		Utilization (%)		TFAV (days)		
		Average	Standard deviation	Average	Standard deviation	Maximum
FAS		79.3%	12.6%	24.5	4.1	56.2
CAPTAIN	Base Case	86.1%	8.2%	14.2	3.6	38.6
	Case Study A	88.9%	7.4%	13.0	3.1	32.8
	Case Study B	90.1%	7.1%	12.8	2.9	29.9

Table 2.4. Calculations of (i) the mean total resource utilization (averaged over all days of our planning horizon), (ii) the standard deviation of the total resource utilization (averaged over all days of our planning horizon), (iii) the mean TFAV in days averaged over all 16 active trials, (iv) the standard deviation of the TFAV averaged over all 16 active trials, and (v) the maximum TFAV averaged over all 16 active trials, for the following cases: (a) the FAS policy, (b) CAPTAIN's Base Case (BRPlan optimization with a fixed NSPlan), (c) CAPTAIN's Case Study A (BRPlan and NSPlan optimization with 0, 4, 8, 12 hours shift options) and (d) CAPTAIN's Case Study B (BRPlan and NSPlan optimization with 0, 4, 6, 8, 10, 12 hours shift options). Note that the constraints specified by the lower right part of Fig. 2.7 were enforced for the CAPTAIN results.

Taking this one step further, we now allow the MIP to also optimize the NSPlan. The number of hours that each nurse will work on each day will now be optimized (under the same parameters and constraints in the lower right part of Fig. 2.7) while obeying a certain set of rules (defined generally as Δ in Section 2.3.5). In the following case studies, we use the same set of rules reported in the description of our test suite (Section 2.4.1). In case study A, we consider three types of shifts (not including the 0-hour shift): 4, 8, or 12 hour shifts. In case study B, we have more shift options/flexibility: 4, 6, 8, 10, 12 hour shifts. Our FAS case study and CAPTAIN Base Case consider a fixed (suboptimal) NSPlan (the same one used in our calculations in the upper left of Fig. 2.7 for FAS, and lower right of Fig. 2.7 for the CAPTAIN Base Case). Note again that in all of our 4 cases (rows in Table 2.4), we are considering the same volume of arrivals and the same nurse total capacity over all days of our transient planning horizon. We see that the largest marginal improvement from the FAS policy (in terms of the average and standard deviation of the utilization and TFAV metrics) is the optimization of a BRPlan while keeping the nurse schedules the same (see the first two rows of Table 2.4).

But we can provide even more improvements by also creating an NSPlan (that obeys the same set of nurse scheduling rules) that is more flexible and appropriate for the optimized BRPlan. Our forecasting methodologies create the relation between fulfilled demand and offered workloads on skills, and the NSPlan links the right amount of nurse capacity based on the skill sets needed on each day. Although it creates only small improvements on our metrics, it is worth noting that having more shift options (therefore flexibility in scheduling nurse capacity with our NSPlan) provides an increase in utilization (and workload smoothing), while increasing access for patients (see Case Study A and Case Study B in Table 2.4). In equilibrium, higher utilization generates longer waiting times for a fixed throughput; however, in our transient case, we are able to increase throughput (throughout our planning horizon) by providing shorter waiting times which translates into higher utilization within our transient planning horizon.

We have explained the use of CAPTAIN for CRUs which are faced with the decision of which portfolio mix of candidate trials should be performed (in addition to the current portfolio), given a fixed infrastructure and set of nurses with specific skills. Because we validated the ability of CAPTAIN to accurately optimize the transition from a prior set of appointments on the books (i.e. under FAS), this also assures us that CAPTAIN can handle transitions caused by future periods (see Appendix C in Section 2.7.3 for further details). This decision support tool could also be used for physical capacity expansions and hiring purposes. It can answer the question: to what level of capacity should we expand our resources in order to meet our objectives for scientific value, service/access levels for participants, and staff overutilization? In other words, given any portfolio of trials and service level constraints, CAPTAIN could determine the minimum staff and room capacities required to conduct this portfolio.

2.6 Conclusion

In contrast to current practice in CRUs, which is often based on common sense and experience, this new planning model, CAPTAIN, increases operational efficiency (and thus cost effectiveness), increases the ability to extract scientific value by thoughtfully selecting the clinical trials performed, and provides a high level of access with limited waiting times that are planned to match the participants' needs. This planning system coordinates care resources and participant visits via novel forecasting and optimization algorithms. This paper provides a proof of concept for a general decision support tool that can enable a CRU to gain control over key performance metrics. Based on a planning horizon, CAPTAIN (i) determines the optimal mix of protocols to perform, (ii) forecasts and controls the workloads that will be placed on resources (staff and rooms), (iii) forecasts and controls overutilization in part by creating efficient nurse schedules and participant/staff coordination, (iv) provides a "system optimal" daily level scheduling plan via the Booking Reservation Plan (BRPlan),

(v) optimizes the allocation of resource capacity, (vi) forecasts and controls the participants' waiting times to first visit, and (vii) answers "what if" questions. As a non-stationary model, CAPTAIN can begin with a base case representing the current conditions and all the existing appointments made to then select the potential clinical research trials that should be conducted without generating excessively long access delays (to first visit) and high overutilization costs. It generates a BRPlan to guide appointment scheduling. The plan is time-varying, maintaining effective operations from one planning period to the next.

Whereas this preliminary work targets strategic planning, future research can focus on appointment scheduling at the time of day level. Therefore, as future research, the methodology will benefit from another layer of modeling and resource allocation to bridge from the planning level to the execution level to schedule time of day appointments.

2.7 Appendix

2.7.1 Appendix A: Notations

Symbol	Description
D	Length (in days) of the CAPTAIN planning horizon
$\mathcal{P}$	Set of active trials in our planning horizon
$\mathcal{P}^{cand}$	Set of candidate trials that can be activated in our planning horizon
$\mathcal{P}^0$	Set of trials that have been active before our planning horizon
T	Time horizon (in days) of current practice booking for the current portfolio $\mathcal{P}^0$
M_k	Research/scientific (or financial) value of conducting trial k
h_k	Binary variable equal to 1 if candidate trial k is activated, and 0 otherwise
$\Theta_{k,d}$	Integer number of slots reserved for type k 1^{st} visits on day d
$\overline{TFAV}_k^{max}$	bound set on the population maximum TFAV (in days) for type k participants
$\alpha_{k,d,t}$	Random number of type k participants enrolled on day t and booked for 1^{st} visit on day d (under the BRPlan)
$A_{k,t}$	Random number of type k participants enrolling on day t
I	Sample space of the daily number of enrollments random variable of a given type
$\bar{I}$	Maximum number of daily enrollments for a given trial
$P_{k,t}(a)$	Probability of having a enrollments of type k on day t
$\overline{P}_{k,t}(a)$	Probability of having a total of a enrollments of type k on the days in $\{0, 1, ..., t-1\}$

Symbol	Description
$\overline{\alpha}_{k,d,t}$	Mean number of type k participants enrolled on day t and booked for 1^{st} visit on day d (under the BRPlan)
$\mu_{k,d}$	Random number of type k participants booked for 1^{st} visit on day d
$\overline{\mu}_{k,d}$	Mean number of type k participants booked for 1^{st} visit on day d
$TFAV_k^{avg}$	Expected TFAV under the BRPlan averaged over all type k participants
$\overline{TFAV}_k^{avg}$	Input bound set on the population average TFAV (in days) for type k participants
$TFAV_{k,t}^{except}(\omega)$	Input bound on the type k ratio of participants enrolling on day t that exceed ω days TFAV
$p_{k,t}^{except}(\omega)$	Input bound set on the mean percentage of type k participants exceeding ω days TFAV
$\mathcal{R}$	Set of all physical resources (rooms, beds, chairs,...)
$\mathcal{S}$	Set of all skills that the nursing staff can perform
$\mathcal{N}$	Set of all nurses
V_k	Number of visits required for trial k
$\Omega_{k,v}$	Time window set of days where a type k visit v can be scheduled after type k visit $v' < v$ without provoking a protocol violation
$\overline{w}_{k,v,r}$	Hours of workload induced on resource r by a type k participant on visit v
L^{max}	Duration (in days) from the first visit to the last visit of the longest trial

Symbol	Description
$PRRM_{k,d_1}(d)$	Vector determining the number of hours a randomly sampled type k participant starting visit 1 on day d_1 will impose on a given resource on day d
$\overline{PRRM}_{k,d_1,r}(d)$	Mean number of hours required on day d and resource r by a type k participant starting visit 1 on day d_1
$W_r^{init}(d)$	Initial workload on resource r and day d induced by participants enrolled prior to the start of our planning horizon l
$\overline{W}_r(d)$	Workload induced (in hours) on resource r (room, nurse or skill) on day d
$K_r(d)$	Capacity (in hours) of resource r (room or nurse) on day d
Δ	Set specifying the clinic's rules on nurse planning and scheduling
$x_{s,n}(d)$	Mean number of hours assigned to nurse n in order to perform skill s on day d
$\mathcal{N}(s)$	Set of nurses that can perform skill s
$\overline{O}_r^l(d)$	Bound set on overutilization (in hours) of resource r (room or nurse) on day d
$TFAV_k^l$	Time to First Available Visit for participants in trial k in planning horizon l
$\overline{\mu}_{k,d}^{sim}$	Simulated mean number of type k participants booked for 1^{st} visit on day d

2.7.2 Appendix B: Proof of selected results

Proof of Lemma II.4

We will show this result by induction over the 2 dimensional space $\{(d,t) : d \in \{1, ..., D+$

$\overline{TFAV}_k^{max}\}, t \in \{0, ..., D\}, t < d\}$. Assume that Lemma 3.1 holds for all (d_1, t_1) with $t_1 < d_1$, $t_1 < t$, and $d_1 < d$.

To show it is true for (d, t) with $t < d$, we start by using the definition of the random variable $\alpha_{k,d,t}$ (see Eq. 2.2). Eq. 2.20 simply uses the definition of the elements in set $\mathcal{A}_{k,d,t}(\infty)$. In Eq. 2.21, we realize that the random variables (i) $\alpha_{k,d,l}$, with $0 \leq l \leq t-1$, (ii) $\alpha_{k,d,l}$, with $t+1 \leq l \leq d-1$, and (iii) $A_{k,t}$, are all deterministic when conditioned on $(A_{k,t}, A_{k,t-1}, ..., A_{k,0})$, which allows us to take the expected value inside the minimum in Eq. 2.22. By noticing that the random variables in item (i) are independent of $(A_{k,t}, A_{k,t-1}, ..., A_{k,l+1})$ and by using the definition of the elements in set $\mathcal{A}_{k,d,t}(\infty)$, we get Eq. 2.23.

$$\alpha_{k,d,t}^{a_t,a_{t-1},...,a_0} = \mathbb{E}\left[\alpha_{k,d,t} | (A_{k,t}, A_{k,t-1}, ..., A_{k,0}) = (a_t, a_{t-1}, ..., a_0)\right] \tag{2.20}$$

$$= \mathbb{E}\left[\min\{\Theta_{k,d} - \sum_{l=0}^{t-1} \alpha_{k,d,l}, A_{k,t} - \sum_{l=t+1}^{d-1} \alpha_{k,l,t} | (A_{k,t}, A_{k,t-1}, ..., A_{k,0}) = (a_t, a_{t-1}, ..., a_0)\right] \tag{2.21}$$

$$= \min\left\{\Theta_{k,d} - \sum_{l=0}^{t-1} \mathbb{E}\left[\alpha_{k,d,l} | (A_{k,t}, A_{k,t-1}, ..., A_{k,0}) = (a_t, a_{t-1}, ..., a_0)\right], \right.$$
$$\left. \mathbb{E}\left[A_{k,t} - \sum_{l=t+1}^{d-1} \alpha_{k,l,t} | (A_{k,t}, A_{k,t-1}, ..., A_{k,0}) = (a_t, a_{t-1}, ..., a_0)\right]\right\} \tag{2.22}$$

$$= \min\left\{\Theta_{k,d} - \sum_{l=0}^{t-1} \alpha_{k,d,l}^{a_l,a_{l-1},...,a_0}, a_t - \sum_{l=t+1}^{d-1} \alpha_{k,l,t}^{a_t,a_{t-1},...,a_0}\right\} \tag{2.23}$$

Eq. 2.23 implies that constraints 2.24 and 2.25 hold. However they are not equivalent since we require at least one of those two constraints to be binding. The BRPlan has to book every enrolling participant to the first available reserved slot, which would not be true if both constraints on $\alpha_{k,d,t}^{a_t,a_{t-1},...,a_0}$ were unbinding. To assess this issue, we create a binary helper variable, $b_{k,d,t}^{a_t,a_{t-1},...,a_0}$, with $d > t$, such that it is equal to 1 if there is enough reserved capacity of type k remaining on day d to schedule the participants that enrolled on day t and still haven't been scheduled up to d; $b_{k,d,t}^{a_t,a_{t-1},...,a_0}$ is equal to 0 otherwise (i.e.,

$\mathbb{1}\{\Theta_{k,d} - \sum_{l=0}^{t-1} \alpha_{k,d,l}^{a_l,a_{t-1},\ldots,a_0} \geq a_t - \sum_{l=t+1}^{d-1} \alpha_{k,l,t}^{a_t,a_{t-1},\ldots,a_0}\})$.

$$\alpha_{k,d,t}^{a_t,a_{t-1},\ldots,a_0} \leq \Theta_{k,d} - \sum_{l=0}^{t-1} \alpha_{k,d,l}^{a_l,a_{l-1},\ldots,a_0} \tag{2.24}$$

$$\alpha_{k,d,t}^{a_t,a_{t-1},\ldots,a_0} \leq a_t - \sum_{l=t+1}^{d-1} \alpha_{k,l,t}^{a_t,a_{t-1},\ldots,a_0} \tag{2.25}$$

$$\Theta_{k,d} - \sum_{l=0}^{t-1} \alpha_{k,d,l}^{a_l,a_{l-1},\ldots,a_0} - \left(a_t - \sum_{l=t+1}^{d-1} \alpha_{k,l,t}^{a_t,a_{t-1},\ldots,a_0}\right) \leq B \cdot b_{k,d,t}^{a_t,a_{t-1},\ldots,a_0}, \tag{2.26}$$

$$a_t - \sum_{l=t+1}^{d-1} \alpha_{k,l,t}^{a_t,a_{t-1},\ldots,a_0} - \left(\Theta_{k,d} - \sum_{l=0}^{t-1} \alpha_{k,d,l}^{a_l,a_{l-1},\ldots,a_0}\right) \leq B \cdot \left(1 - b_{k,d,t}^{a_t,a_{t-1},\ldots,a_0}\right) \tag{2.27}$$

$$\alpha_{k,d,t}^{a_t,a_{t-1},\ldots,a_0} \geq \Theta_{k,d} - \sum_{l=0}^{t-1} \alpha_{k,d,l}^{a_l,a_{l-1},\ldots,a_0} - B \cdot b_{k,d,t}^{a_t,a_{t-1},\ldots,a_0} \tag{2.28}$$

$$\alpha_{k,d,t}^{a_t,a_{t-1},\ldots,a_0} \geq a_t - \sum_{l=t+1}^{d-1} \alpha_{k,l,t}^{a_t,a_{t-1},\ldots,a_0} - B \cdot \left(1 - b_{k,d,t}^{a_t,a_{t-1},\ldots,a_0}\right) \tag{2.29}$$

Given a large enough B (e.g., 20), constraints 2.26 and 2.27 assure that $b_{k,d,t}^{a_t,a_{t-1},\ldots,a_0}$ takes the values specified by the definition we just provided. Based on this definition, constraint 2.28 (resp. constraint 2.29) will now make sure that constraint 2.24 (resp. 2.25) will be binding when $b_{k,d,t}^{a_t,a_{t-1},\ldots,a_0} = 0$ (resp. $b_{k,d,t}^{a_t,a_{t-1},\ldots,a_0} = 1$), hence when the available capacity remaining is smaller than the demand still needed to be scheduled (resp. more available capacity than remaining demand) defined in constraint 2.26 (resp. 2.27). Because of our induction hypothesis on $\alpha_{k,d_1,t_1}^{a_t,a_{t-1},\ldots,a_0}$ for $t_1 < d_1$, $t_1 < t$, and $d_1 < d$ being linear in our decision variables Θ, the system of constraints 2.24-2.29 linearizes Eq. 2.23 which can be integrated in a deterministic MIP. The basis of induction requires us to show that $\alpha_{k,1,0}^{a_t,a_{t-1},\ldots,a_0} = \min\{\Theta_{k,1}, a_0\}$ can be expressed by a set of linear constraints in $\Theta_{k,1}$. By using the same methods used for (d,t) above, the basis of induction follows immediately. By induction, we have showed that this is true for all (d,t) with $t < d$, and proved Lemma II.4.

Proof of Theorem II.5

Given any BRPlan schedule, we must estimate $\overline{\alpha}_{k,d,t}$, the mean number of type k participants that enrolled on day t and are booked for their first visit on day d. In Eq. 2.30, we condition $\alpha_{k,d,t}$ subject to a_l, $l = 0, ..., t$, the number of type k enrollments that occurred on day l from the beginning of our planning horizon up to day t. In Eq. 2.31, we use the assumption that the number of enrollments on a given day is independent of anything else in our system model (specifically, the number of enrollments on other days), and we also partition $(\mathbb{Z}^+)^{t+1}$ into the sets I^{t+1} and $(\mathbb{Z}^+ \setminus I)^{t+1}$ in our sum over all $(a_t, ..., a_0) \in (\mathbb{Z}^+)^{t+1}$.

$$\begin{aligned}\overline{\alpha}_{k,d,t} = & \sum_{(a_t,a_{t-1},...,a_0)\in(\mathbb{Z}^+)^{t+1}} \mathbb{E}\left[\alpha_{k,d,t} | (A_{k,t}, A_{k,t-1}, ..., A_{k,0}) = (a_t, a_{t-1}, ..., a_0)\right] \\ & \cdot \mathbb{P}\left((A_{k,t}, A_{k,t-1}, ..., A_{k,0}) = (a_t, a_{t-1}, ..., a_0)\right) \qquad (2.30) \\ = & \sum_{(a_t,a_{t-1},...,a_0)\in I^{t+1}} \mathbb{E}\left[\alpha_{k,d,t} | (A_{k,t}, A_{k,t-1}, ..., A_{k,0}) = (a_t, a_{t-1}, ..., a_0)\right] \cdot \prod_{l=0}^{t} P_{k,l}(a_l) \\ + & \sum_{(a_t,a_{t-1},...,a_0)\in(\mathbb{Z}^+\setminus I)^{t+1}} \mathbb{E}\left[\alpha_{k,d,t} | (A_{k,t}, A_{k,t-1}, ..., A_{k,0}) = (a_t, a_{t-1}, ..., a_0)\right] \cdot \prod_{l=0}^{t} P_{k,l}(a_l) \\ & \qquad (2.31)\end{aligned}$$

The second sum in Eq. 2.31 will converge to 0 as $\overline{I}$ goes to infinity. Moreover, by using the definition of the set $\mathcal{A}_{k,d,t}(\infty)$ (see Definition 3.2), we can express the mean number of type k participants that enrolled on day t and are booked for first visit on day d, as $\overline{I} \to \infty$:

$$\overline{\alpha}_{k,d,t} = \sum_{\alpha_{k,d,t}^{a_t,a_{t-1},...,a_0} \in \mathcal{A}_{k,d,t}(\infty)} \alpha_{k,d,t}^{a_t,a_{t-1},...,a_0} \cdot \prod_{l=0}^{t} P_{k,l}(a_l) \qquad (2.32)$$

Finally, by using the result from Lemma II.4, and noticing that $P_{k,l}(a_l)$ is independent of Θ, for all $k \in \mathcal{P}, a_l \in I$, and $l \in \{0, ..., t\}$, we have shown that $\overline{\alpha}_{k,d,t}$ can be expressed linearly in our decision variables Θ.

Proof of Proposition II.6

To compute $\overline{\alpha}_{k,d,t}$, we condition $\alpha_{k,d,t}$ in terms of $(A_{k,t}, A_{k,t-1}, ..., A_{k,j^*(t)}, \sum_{i=0}^{j^*(t)-1} A_{k,i})$ in

Eq. 2.33. As $\bar{I} \to \infty$, and by the definition of $\mathcal{A}_{k,d,t}(m)$, Eq. 2.34 follows. That is, $\overline{\alpha}_{k,d,t} =$

$$\sum_{(a_t,\dots,a_{j^*(t)},a)\in\mathbb{Z}^{t-j^*(t)+2}} \mathbb{E}\left[\alpha_{k,d,t} | (A_{k,t}, A_{k,t-1}, \dots, A_{k,j^*(t)}, \sum_{i=0}^{j^*(t)-1} A_{k,i}) = (a_t, a_{t-1}, \dots, a_{j^*(t)}, a)\right]$$
$$\cdot \left(\prod_{l=j^*(t)}^{t} P_{k,l}(a_l)\right) \cdot \overline{P}_{k,j^*(t)}(a) \tag{2.33}$$

$$= \sum_{\alpha_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a}\in\mathcal{A}_{k,d,t}(m)} \alpha_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a} \cdot \left(\prod_{l=j^*(t)}^{t} P_{k,l}(a_l)\right) \cdot \overline{P}_{k,j^*(t)}(a) \tag{2.34}$$

$$\approx \sum_{\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a}\in\hat{\mathcal{A}}_{k,d,t}(m)} \hat{\alpha}_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a} \cdot \left(\prod_{l=j^*(t)}^{t} P_{k,l}(a_l)\right) \cdot \overline{P}_{k,j^*(t)}(a). \tag{2.35}$$

Finally, we can use our approximation $\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a} \in \hat{\mathcal{A}}_{k,d,t}(m)$ for the elements $\alpha_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a} \in \mathcal{A}_{k,d,t}(m)$ in Eq. 2.35. What remains to be shown is that $\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a}$ can be expressed in a MIP linearly in the decision variables Θ. Rewriting Eq. 2.4, we get:

$$\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a} = \min\Bigg\{ \overbrace{a_t - \sum_{l=t+1}^{d-1} \hat{\alpha}_{k,l,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a}}^{U_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a}},$$
$$\underbrace{\Theta_{k,d} - \sum_{l=0}^{t-1} \sum_{(a_{j^*(t)-1},\dots,a_{j^*(l)})\in I^{j^*(t)-j^*(l)}} \hat{\alpha}_{k,d,l}^{a_l,a_{l-1},\dots,a_{j^*(l)},a-\sum_{i=j^*(l)}^{j^*(t)-1} a_i} \left(\prod_{n=j^*(l)}^{j^*(t)-1} P_{k,n}(a_n)\right) \overline{P}_{k,j^*(l)}(a)}_{V_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a}} \Bigg\}, \tag{2.36}$$

and $\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a}$ can now be expressed linearly in Θ in constraints 2.37-2.42.

$$\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a} \le V_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a} \tag{2.37}$$

$$\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a} \le U_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a} \tag{2.38}$$

$$V_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a} - U_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a} \le B \cdot b_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a}, \tag{2.39}$$

$$U_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a} - V_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a} \le B \cdot \left(1 - b_{k,d,t}^{a_t,a_{t-1},\dots,a_{j^*(t)},a}\right) \tag{2.40}$$

$$\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a} \geq V_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a} - B \cdot b_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a} \tag{2.41}$$

$$\hat{\alpha}_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a} \geq U_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a} - B \cdot \left(1 - b_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a}\right) \tag{2.42}$$

Eq. 2.36 implies that constraints 2.37 and 2.38 hold, but they are not equivalent. We require at least one of those two constraints to be binding, because the BRPlan has to book every enrolling participant to the first available reserved slot. We created the parameter B and the binary helper variable, $b_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a} = \mathbb{1}\{U_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a} \leq V_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a}\}$, with $d > t$, such that it is equal to 1 if there is enough reserved type k capacity remaining on day d to schedule the participants that enrolled on day t and still haven't been scheduled up to d; $b_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a}$ is equal to 0 otherwise. This is defined in constraint 2.39 and 2.40. Given a large enough B (e.g., 20), constraints 2.39 and 2.40 assure that $b_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a}$ takes the required values. Based on this definition, constraint 2.41 (resp. constraint 2.42) ensures that constraint 2.37 (resp. 2.38) will be binding when $b_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a} = 0$ (resp. $b_{k,d,t}^{a_t,a_{t-1},\ldots,a_{j^*(t)},a} = 1$). That is, constraint 2.37 (resp. 2.38) will be binding when the available capacity remaining is smaller than the demand to be scheduled (resp., more available capacity than remaining demand).

Proof of Theorem II.7

In this result, we want to show that the mean type k TFAV (averaged over all participants of a given trial) can be expressed linearly in our BRPlan decision variables Θ. First, in Eq. 2.43, we take the expectation of the ratio of the total number of days that type k participants wait for their first visit during our planning horizon (numerator), and the total number of enrollments that occurred from days 0 to D (denominator). Eq. 2.44 follows after conditioning this expression subject to the daily enrollments from days 0 to D.

$$TFAV_k^{avg} = \mathbb{E}\left[\frac{\sum_{t=0}^{D}\sum_{d=t+1}^{D+\overline{TFAV}_k^{max}}(d-t)\cdot\alpha_{k,d,t}}{\sum_{i=0}^{D}A_{k,i}}\right] \tag{2.43}$$

$$=\sum_{t=0}^{D}\sum_{d=t+1}^{D+\overline{TFAV}_k^{max}}\sum_{(a_D,a_{D-1},...,a_0)\in I^{D+1}}\prod_{l=0}^{D}P_{k,l}(a_l)$$

$$\cdot\,(d-t)\cdot\mathbb{E}\left[\frac{\alpha_{k,d,t}}{\sum_{i=0}^{D}A_{k,i}}|(A_{k,D},A_{k,D-1},...,A_{k,0})=(a_D,a_{D-1},...,a_0)\right] \tag{2.44}$$

$$=\sum_{t=0}^{D}\sum_{d=t+1}^{D+\overline{TFAV}_k^{max}}\sum_{(a_D,a_{D-1},...,a_0)\in I^{D+1}}\frac{1}{\sum_{i=0}^{D}a_i}\cdot\prod_{l=0}^{D}P_{k,l}(a_l)$$

$$\cdot\,(d-t)\cdot\mathbb{E}\left[\alpha_{k,d,t}|(A_{k,t},A_{k,t-1},...,A_{k,0})=(a_t,a_{t-1},...,a_0)\right] \tag{2.45}$$

$$=\sum_{t=0}^{D}\sum_{d=t+1}^{D+\overline{TFAV}_k^{max}}\sum_{(a_t,a_{t-1},...,a_0)\in I^{t+1}}(d-t)\cdot\mathbb{E}\left[\alpha_{k,d,t}|(A_{k,t},A_{k,t-1},...,A_{k,0})=(a_t,a_{t-1},...,a_0)\right]$$

$$\cdot\sum_{(a_D,a_{D-1},...,a_{t+1})\in I^{D-t}}\frac{\prod_{l=0}^{D}P_{k,l}(a_l)}{\sum_{i=0}^{D}a_i} \tag{2.46}$$

$$=\sum_{t=0}^{D}\sum_{d=t+1}^{D+\overline{TFAV}_k^{max}}\sum_{\alpha_{k,d,t}^{a_t,a_{t-1},...,a_0}\in\mathcal{A}_{k,d,t}(\infty)}(d-t)\cdot\alpha_{k,d,t}^{a_t,a_{t-1},...,a_0}\cdot\sum_{(a_D,a_{D-1},...,a_{t+1})\in I^{D-t}}\frac{\prod_{l=0}^{D}P_{k,l}(a_l)}{\sum_{i=0}^{D}a_i}. \tag{2.47}$$

In Eq. 2.44, the denominator inside our conditional expectation becomes deterministic since it is conditioned on all the daily enrollments in our planning horizon, hence we can express it outside the expectation in Eq. 2.45. Note now that the conditional expectation of the random variable $\alpha_{k,d,t}$ is independent on future enrollment realizations $(A_{k,D},A_{k,D-1},...,A_{k,t+1})=(a_D,a_{D-1},...,a_{t+1})$: we can now ignore those realizations in the conditional expectation, which gives us Eq. 2.45. In Eq. 2.46, the sum over the set I^{D+1} is separated into two sums: one over the set I^{t+1}, and the other over the set I^{D-t}. Finally by using the definition of the set $\mathcal{A}_{k,d,t}(\infty)$ (see Definition II.2) and our result derived from Lemma II.4, we have proved Theorem II.7.

Proof of Theorem II.10

We consider the workloads induced on resource r and day d of our planning horizon by (i) participants enrolled prior to our horizon (the parameter $W_r^{init}(d)$), and (ii) participants that will eventually enroll in our planning horizon. To capture the workloads induced by participants who haven't enrolled yet, consider that scheduling under the BRPlan generates a random integer number of type k participants booked for a first visit on each day d_1 denoted μ_{k,d_1}. We must check if each participant i that will be booked for a first visit on each day d_1 prior to day d will induce a stochastic workload on resource r, $d - d_1$ days after their first visit, based on PRRM realizations. For example, if $\mu_{k,d_1} = 3$, we will sum three independent realizations of the type k PRRM to determine the number of hours of resource r needed for future days (e.g., $d - d_1$ days after the first visit). Hence, we can compute the offered workload for each skill or room resource r, d days after the beginning of period l, where $d \in \{1, 2, ..., D + L^{max} + \max_k \overline{TFAV}_k^{max}\}$ as:

$$W_r(d) = W_r^{init}(d) + \sum_{k \in \mathcal{P}} \sum_{d_1=1}^{\min\{d, D+\overline{TFAV}_k^{max}\}} \sum_{i=1}^{\mu_{k,d_1}} PRRM_{k,d_1}^i(d) \cdot e_r, \tag{2.48}$$

where $PRRM_{k,d_1}^i(d) \cdot e_r$ is the i^{th} realization of the random variable $PRRM_{k,d_1}(d) \cdot e_r$. By linearity of the expectation we get:

$$\bar{W}_r(d) = W_r^{init}(d) + \sum_{k \in \mathcal{P}} \sum_{d_1=1}^{\min\{d, D+\overline{TFAV}_k^{max}\}} \mathbb{E}\left[\sum_{i=1}^{\mu_{k,d_1}} PRRM_{k,d_1}^i(d) \cdot e_r\right] \tag{2.49}$$

since $\mathbb{E}\left[W_r^{init}(d)\right] = W_r^{init}(d)$. Those quantities are deterministic since all visits of a trial are scheduled at the enrollment date (prior to our planning horizon). We note that (i) $\bar{\mu}_{k,d_1} = \mathbb{E}\left[\mu_{k,d_1}\right] < \infty, \forall k \in \mathcal{P}, \forall d_1 = 1, ..., D + \overline{TFAV}_k^{max}$ since the number of enrolling participants (and therefore, scheduled participants) on any day is finite; (ii)$\{PRRM_{k,d_1}^i(d) \cdot e_r\}_{i=1,2,...}$ all have the same expectation for a fixed patient type $k \in \mathcal{P}$, $d_1 \in \{1, ..., D + \overline{TFAV}_k^{max}\}$, $d \in \{1, ..., D + L^{max} + \max_k \overline{TFAV}_k^{max}\}$ and $r \in \mathcal{R} \cup \mathcal{S}$; and (iii) the PRRM random variables that we defined are drawn from protocol data and do not depend on $\mu_{k,d}, \forall k \in$

$\mathcal{P}, \forall d = 1, ..., D + \overline{TFAV}_k^{max}$. Hence, by these three facts and Wald's equation, we get Eq. 2.48 and prove Theorem II.10.

2.7.3 Appendix C: Practical implementation and transience

In this paper, we studied the effects of our planning tool over a finite planning horizon by considering the initial transition from the current FAS scheduling policy to the BRPlan. Thus, the portfolio selection and capacity planning decisions in our planning horizon are dependent on an initial state of the system determined by the existing appointments (booked via the FAS policy) of participants that enrolled prior to CAPTAIN's period of planning. Possible questions are: how is the transition between planning periods handled? How would CAPTAIN deal with unplanned portfolio updates within a planning horizon?

We define period l as the interval of time between the l^{th} and $(l+1)^{th}$ trial portfolio update (or equivalently, CAPTAIN run). A portfolio update should be performed when the CRU wants to add a portfolio of candidate trials to their workload, or when at least one current trial ends. CAPTAIN is able to construct an optimal transient/non-homogeneous BRPlan and NSPlan that takes into account those portfolio updates. The maximum length of any period will be D days. We define $\mathcal{P}^{l-1}$ as the current portfolio (or set) of active trials in the CRU during period $l-1$, $\mathcal{P}^{l}_{cand}$ as the portfolio of candidate trials that are to be considered for addition at the beginning of period l, and $\mathcal{P}^{l-1}_{out}$ as the portfolio of trials that ended during period $l-1$.

The inputs for the l^{th} run (period 1) of CAPTAIN will be defined by the set Γ^l: (i) $\mathcal{P}^{l-1}$, the portfolio of active trials in period $l-1$ (possibly using an aggregate model of low enrollment trials for computational simplicity); (ii) $\mathcal{P}^{0}_{out}$, the set of trials that ended before the l^{th} run but might still have participants receiving service in period l; (iii) $\mathcal{P}^{l}_{cand}$, the candidate portfolio of trials that the CRU wishes to consider at the beginning of period l; (iv) $W_r^{l,init}(d)$, the committed resource r (nurses and rooms) workload on each day d of

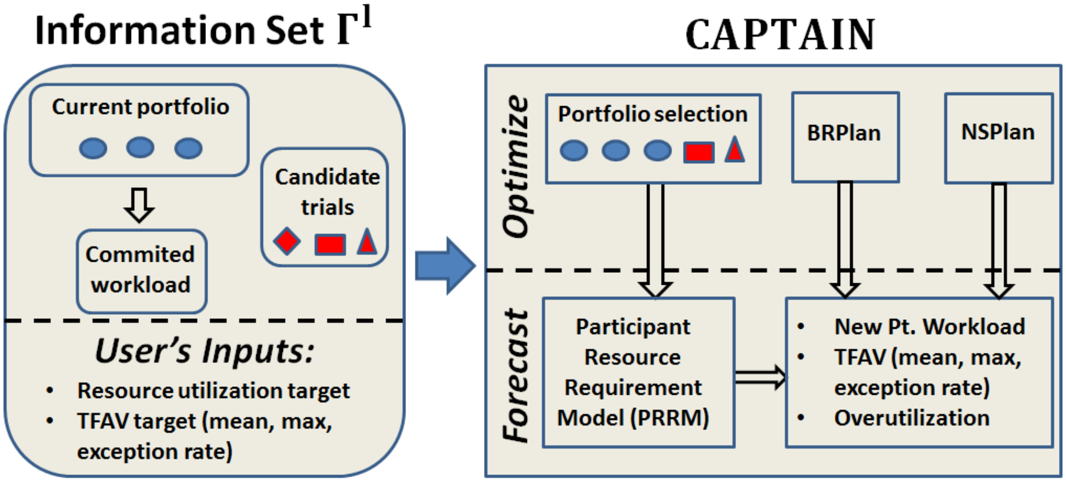

Information Set Γ^1
Current portfolio
Commited workload
Candidate trials
User's Inputs:
Resource utilization target
TFAV target (mean, max, exception rate)
CAPTAIN
Optimize
Portfolio selection
BRPlan
NSPlan
Forecast
Participant Resource Requirement Model (PRRM)
New Pt. Workload
TFAV (mean, max, exception rate)
Overutilization

Figure 2.8: CAPTAIN's functionality with information inputs on the left and CAPTAIN forecasting and optimization on the right

the l^{th} period/planning horizon (i.e., day $T + d$) induced by the participant bookings that were made in the past (i.e., from previous periods); and (v) the CRU's desired constraints on resource overutilization and/or TFAV target (mean, maximum, and/or TFAV exception) per trial. Conceptually, CAPTAIN combines optimization with an online metrics forecasting capability to achieve the targets requested by the managerial staff. D days (or sooner) after the start of period l, the CRU will update the inputs of CAPTAIN, restart planning for the next D days, and construct the information set Γ_{l+1}.

As illustrated in Fig. 2.8, at the beginning of period l, we use the information set Γ^l as an input to CAPTAIN. The modeled metrics are (i) the offered workload on each (physical or human) resource for every day of period l, (ii) the resource utilization (and overutilization) for every day of period l, and (iii) the population average, the mean percentage of patients exceeding a TFAV limit and/or, the population maximum TFAV for each trial in period l. The forecasting methods combine (i) the participant first visit bookings (see Section 2.3.2) generated by reserving resource capacity with our Booking Reservation Plan methodology (the BRPlan), (ii) the Participant Resource Requirements Model (the PRRM), which captures the participants' trajectory for each trial in portfolio $\mathcal{P}^l$ that will be conducted in period l (see Section 2.3.4), and (iii) the nursing staff's daily staffing levels and planned allocations (the NSPlan, see Section 2.3.5).

2.7.4 Appendix D: Mixed Integer Program

$$\max \sum_{k \in \mathcal{P}} M_k h_k, \tag{2.50}$$

$$\overline{\alpha}_{k,d,t} = \sum_{\hat{\alpha}^{a_t,a}_{k,d,t} \in \hat{\mathcal{A}}_{k,d,t}(1)} \hat{\alpha}^{a_t,a}_{k,d,t} \cdot P_{k,t}(a_t) \cdot \overline{P}_{k,t}(a), \tag{2.51}$$

$$\forall t \in \{0,...,D\}, \forall d \in \{t+1,...,D+\overline{TFAV}^{max}_k\}, \forall k \in \mathcal{P},$$

$$\hat{\alpha}^{a_t,a}_{k,d,t} \leq a_t - \sum_{l=t+1}^{d-1} \hat{\alpha}^{a_t,a}_{k,l,t}, \tag{2.52}$$

$$\hat{\alpha}^{a_t,a}_{k,d,t} \leq \Theta_{k,d} - \sum_{l=0}^{t-1} \sum_{(a_{t-1},a_{t-2},...,a_l) \in I^{(t-l)}} \hat{\alpha}^{a_l,a-\sum_{i=l}^{t-1} a_i}_{k,d,l} \cdot \left(\prod_{n=l}^{t-1} P_{k,n}(a_n) \right) \cdot \overline{P}_{k,l}(a), \tag{2.53}$$

$$\Theta_{k,d} - \sum_{l=0}^{t-1} \sum_{(a_{t-1},a_{t-2},...,a_l) \in I^{(t-l)}} \hat{\alpha}^{a_l,a-\sum_{i=l}^{t-1} a_i}_{k,d,l} \cdot \left(\prod_{n=l}^{t-1} P_{k,n}(a_n) \right) \cdot \overline{P}_{k,l}(a) - \left(a_t - \sum_{l=t+1}^{d-1} \hat{\alpha}^{a_t,a}_{k,l,t} \right)$$
$$\leq B \cdot b^{a_t,a}_{k,d,t}, \tag{2.54}$$

$$a_t - \sum_{l=t+1}^{d-1} \hat{\alpha}^{a_t,a}_{k,l,t} - \left(\Theta_{k,d} - \sum_{l=0}^{t-1} \sum_{(a_{t-1},a_{t-2},...,a_l) \in I^{(t-l)}} \hat{\alpha}^{a_l,a-\sum_{i=l}^{t-1} a_i}_{k,d,l} \cdot \left(\prod_{n=l}^{t-1} P_{k,n}(a_n) \right) \cdot \overline{P}_{k,l}(a) \right)$$
$$\leq B \cdot \left(1 - b^{a_t,a}_{k,d,t}\right), \tag{2.55}$$

$$\hat{\alpha}^{a_t,a}_{k,d,t} \geq \Theta_{k,d} - \sum_{l=0}^{t-1} \sum_{(a_{t-1},a_{t-2},...,a_l) \in I^{(t-l)}} \hat{\alpha}^{a_l,a-\sum_{i=l}^{t-1} a_i}_{k,d,l} \cdot \left(\prod_{n=l}^{t-1} P_{k,n}(a_n) \right) \cdot \overline{P}_{k,l}(a) - B \cdot b^{a_t,a}_{k,d,t} \tag{2.56}$$

$$\hat{\alpha}^{a_t,a}_{k,d,t} \geq a_t - \sum_{l=t+1}^{d-1} \hat{\alpha}^{a_t,a}_{k,l,t} - B \cdot \left(1 - b^{a_t,a}_{k,d,t}\right) \tag{2.57}$$

$$\forall t \in \{0,...,D\}, \forall d \in \{t+1,...,D+\overline{TFAV}^{max}_k\}, a_t \in I, a \in \{0,...,t\bar{I}\}, \forall k \in \mathcal{P},$$

$$\overline{\mu}_{k,d} = \sum_{t=0}^{d-1} \overline{\alpha}_{k,d,t}, \forall k \in \mathcal{P}, \forall d \in \{1,...,D+\overline{TFAV}^{max}_k\}. \tag{2.58}$$

$$\bar{W}_r(d) = W_r^{init}(d) + \sum_{k\in\mathcal{P}} \sum_{d_1=1}^{\min\{d, D+\overline{TFAV}_k^{max}\}} \bar{\mu}_{k,d_1} \cdot \sum_{v=1}^{V_k} \frac{\mathbb{1}_{\{(d-d_1)\in\Omega_{k,v}\}}}{|\Omega_{k,v}|} \bar{w}_{k,v,r}, \tag{2.59}$$

$$\forall r \in \mathcal{S}\cup\mathcal{R}, \forall d \in \{1,2,...,D+L^{max}+\max_k \overline{TFAV}_k^{max}\},$$

$$\bar{W}_r(d) \leq K_r(d) + \bar{O}_r(d), \forall r\in\mathcal{R}, \forall d \in \{1,2,...,D+L^{max}+\max_k \overline{TFAV}_k^{max}\}, \tag{2.60}$$

$$\bar{W}_s(d) = \sum_{n\in\mathcal{N}(s)} x_{s,n}(d), \forall s\in\mathcal{S}, \forall d\in\{1,2,...,D+L^{max}+\max_k TFAV_k^{max}\}, \tag{2.61}$$

$$\bar{W}(n,d) = \sum_{s\in\mathcal{S}} x_{s,n}(d), \bar{W}(n,d) \leq K_n(d) + \bar{O}_n(d), \tag{2.62}$$

$$\forall n\in\mathcal{N}, \forall d\in\{1,2,...,D+L^{max}+\max_k \overline{TFAV}_k^{max}\},$$

$$\sum_{t=0}^{D} \sum_{d=t+1}^{D+\overline{TFAV}_k^{max}} \sum_{\hat{\alpha}_{k,d,t}^{a_t,a}\in\hat{\mathcal{A}}_{k,d,t}(1)} (d-t)\cdot\hat{\alpha}_{k,d,t}^{a_t,a} \cdot \sum_{(a_D,a_{D-1},...,a_{t+1})\in I^{D-t}} \frac{\overline{P}_{k,t}(a)\cdot\prod_{l=t}^{D} P_{k,l}(a_l)}{a+\sum_{i=t}^{D} a_i}$$

$$\leq \overline{TFAV}_k^{avg}, \forall k\in\mathcal{P} \tag{2.63}$$

$$\sum_{d=t+\omega+1}^{D+\overline{TFAV}_k^{max}} \sum_{\hat{\alpha}_{k,d,t}^{a_t,a}\in\hat{\mathcal{A}}_{k,d,t}(1)} \frac{\hat{\alpha}_{k,d,t}^{a_t,a}}{a_t}\cdot P_{k,t}(a_t)\cdot\overline{P}_{k,t}(a) \leq p_{k,t}(\omega), \forall k\in\mathcal{P}, \forall t\in\{0,...,D\}, \forall\omega\in\mathcal{T} \tag{2.64}$$

$$h_k \in \{0,1\}, \forall k\in\mathcal{P}, \tag{2.65}$$

$$\Theta_{k,d}\in\mathbb{Z}^+, \forall d\in\{1,...,D+\overline{TFAV}_k^{max}\}, \forall k\in\mathcal{P}, \tag{2.66}$$

$$\{(K_n(1),...,K_n(D+L^{max}+\max_k TFAV_k^{max}) : n\in\mathcal{N}\}\in\Delta \tag{2.67}$$

CHAPTER III

An Outpatient Planning Optimization Model for Integrated Care and Access Management

3.1 Introduction

For scheduled healthcare delivery environments, the prevailing first come first served approach to appointment scheduling does not support that relative urgency varies across types or classes of patient visits. This research contributes a method toward a more rational and planned approach to healthcare delivery. Access delay, or the length of time required to get an initial appointment, is a key performance metric that characterizes how a care delivery system manages its resources and its workload. While emergency care has developed elaborate methods to differentiate the access delay based on patient acuity (i.e., degree of urgency with which they need care, roughly speaking), only simple and relatively crude methods have been developed for appointment-based service operations. For example, primary care providers will reserve some appointment slots during the flu and cold season for urgent sick visits, whereas annual physical exams are scheduled months in advance and are sometimes steered into summertime visits. To account for the complexity of many outpatient care networks, this intuition must be extended, deepened, and ultimately integrated into a sophisticated decision support method to reduce waste and enhance patient access. In particular, our work is applied to outpatient services with different patient classes, each of which may be offered different levels of access. We will refer to this methodology and a research software instantiation as OPT for "Outpatient Planning Technology."

To appreciate the need for a system such as OPT, consider this illustrative example. Chris happens to live more than 100 miles away from a major medical center. After repeated visits to his primary care physician, referral to several specialists, and many diagnostic tests, there is still no clear diagnosis available to Chris. The doctors suspect that the condition is serious. Chris and the primary care doctor agree that it is wise to travel to a medical center with the expertise and resources to obtain a diagnosis and treatment plan. Chris might be asking himself: how long will it take to get an appointment? Medically, Chris will worry what irreversible damage he may sustain during this wait. The first visit is important in our methodology, and the wait from request until the time of that appointment is referred to as *access delay*. Chris also realizes that it will take some random number of visits/consultations and tests to get a diagnosis and treatment, which we refer to this as an *itinerary of care*. A complicating factor is that Chris's full treatment needs (quantity, timing, and services required) are not known at the time of the initial appointment request, which may result in downstream appointment demand exceeding the capacity available at the institution. This leads to a stochastic network patient flow problem complicated by the need to exert admission control to optimize the patient flows (aggregated by patient type). A key motivation to control access delay becomes clear by realizing that while some patient types do not require short waits, urgent cases (e.g., an episode of care to follow up on an indication of advanced breast cancer) will benefit medically from quick diagnosis and treatment. For example, the Mayo Clinic, Rochester, serves a diverse patient population both in terms of conditions treated but also geography and urgency. New patients that travel from across the nation or around the world might require more resources at a higher level of urgency as compared to local established patients. As is often the case in healthcare, the same set of resources serves patients with varying access delay sensitivities/requirements. Patients with severe conditions will come to Rochester for fast diagnosis and access to surgery. The same resources, however, will be used to serve patients with the same condition that can easily travel back and forth to the clinic and do not require fast access to all services. On the other hand, patients who present a suspicion of cancer will need to gain rapid access to diagnostic testing, surgeon consultation and, if necessary, surgery.

Access delay is a key performance metric that characterizes how a care delivery system manages its resources and its workload. The *itinerary of care* is from the time a patient requests service to the last care visit that ends an itinerary (care pathway). Optimizing this access delay metric with different targets for different patient types presents important research challenges that have received little attention for outpatient care networks (e.g., see for single-unit works [45], [44], [51], [15]). From the perspective of the healthcare organization, additional critical metrics must be considered such as (1) medical service (or clinician) utilization, (2) what case mix by patient type can/should be serviced, (3) the daily workload hours that exceed resource capacity (i.e. either delayed to the next day, or performed in overtime), and (4) the probability that the workload induced on medical services will exceed the capacity of key resources. To the best of our knowledge, no methodology yet exists to address these critical issues of access to patients such as Chris while considering the complex and type-dependent tradeoffs among the metrics that the provider must manage.

3.2 Motivation

As the complexity and interconnectivity of systems increases, proper planning and flow management to coordinate a network of resources is becoming increasingly important. This endeavor is complicated by the fact that requests for appointments and patient itineraries are stochastic in many such real-world networks. Additionally, entities that enter the system may have different completion deadlines and different priorities or service level targets for meeting those deadlines. This type of system is particularly well-represented in healthcare, where the paradigm of integrated care delivery is gaining importance.

Due to increasing cost pressure, there are considerable financial incentives to transition to larger, more coordinated network care models such as those found at the Mayo Clinic, the Cleveland Clinic, and Kaiser Permanente. With Accountable Care Organizations expected to play an increasingly important role in health care, effective coordination across health services will only grow in importance as providers try to improve their health outcomes (see [52]). As health care moves toward consolidation, the resulting healthcare networks will serve a diverse patient population both in terms of conditions treated but also geography and urgency. This leads to the same set of resources serving patients with varying time sensitivity to completing their treatment. An example of this is Mayo Clinic's Spine Care Midwest initiative. Patients with severe pain will come to Rochester for fast diagnosis and access to surgery. The same resources, however, will be used to serve patients with chronic back pain that can easily travel back and forth to the clinic and do not require fast access to all services. Another example is surgical patients with cancer. These patients often have higher priority over surgical patients that do not have cancer. In urology, patients may present with a suspicion of cancer and they need to gain rapid access to diagnostic testing, surgeon consultation and, if necessary, surgery.

3.2.1 Current approach and implications

A simple First-Come-First-Served (FCFS) approach to care dominates current practice, causing patients with more urgent care needs to be treated the same as those who can wait longer for an appointment without a negative health impact. However, in some medical specialties at a partner healthcare provider, physicians will assign slots in their schedule to reserve capacity for specific types of patients. This is done for a few of reasons: (1) some physicians specialize and need to reserve capacity for patients who might materialize and require a given subspecialty; (2) our partnering hospital is a top research clinic, and doctors want to see patients who fit the requirements of ongoing medical studies; (3) physicians generally prefer seeing patients they already had contact with (established) rather than going through the diagnosis of a new patient. When those patients don't materialize, however, the slots go unused, and the impact can be large in terms of utilization and capacity going to waste. Patients may have their appointments delayed because they don't meet the criteria specified for a given slot. Having designated slots is a reflection of the desire of the practice to manage patient mix, but this approach to the problem may not be efficient.

Based on roughly one year of data from our partnering hospital for a subset of the medical departments, it was possible to observe a posteriori the performance of the current scheduling approach. In Fig. 3.1, we observe the tail distribution of the access delay to a root appointment of an itinerary of care in three medical departments based on patient urgency. More specifically the graph in Fig. 3.1 shows the probability of exceeding n weeks of access delay ($n = 1, .., 6$) based on (1) the patient type (urgency level) and (2) the medical department required for the first visit appointment in the itinerary. In Fig. 3.2, for each department $d \in \{GIM, GI, Neurology\}$, we observe (1) the daily resource capacity of the department (in hours), (2) the daily average total workload (in hours) induced on the department based on current practice scheduling, and (3) the daily mean and standard deviation of the internal referral workload (i.e., workload induced by patients that started their itinerary in another department $d_1 \neq d$ but were referred to department d for a downstream appointment in their itinerary).

Here, we can see that urgent vs. non-urgent access delays vary significantly by medical department. This is not surprising since demand for specific services will vary across the different patient classes considered here. Heterogeneous demand distributions will generate

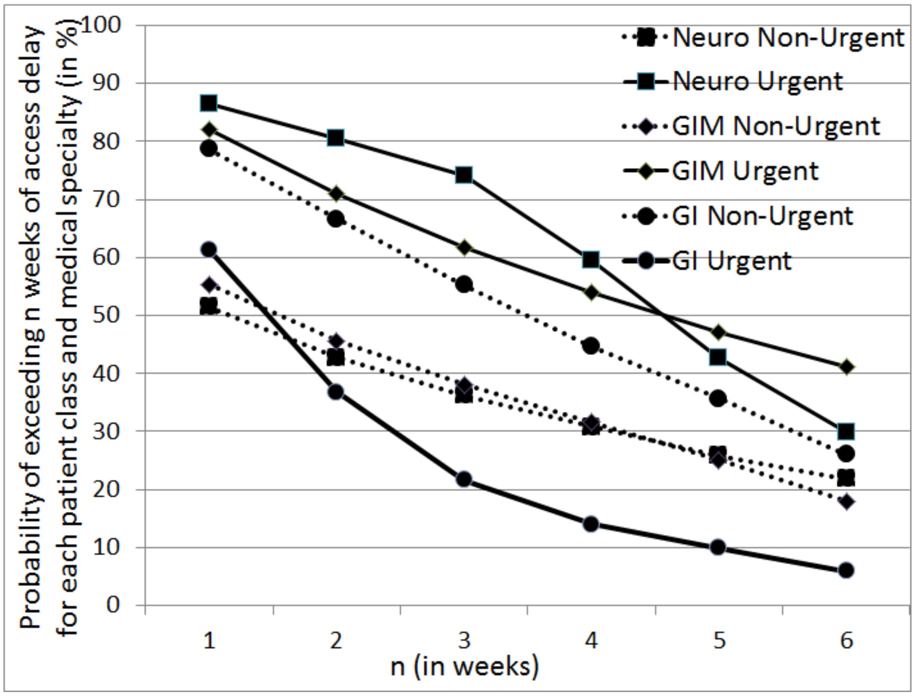


Figure 3.1: Access delay (in weeks) tail distribution to the first visit of an itinerary for different patient types in GI and GIM based on historical data

uncontrolled and/or undesirable differences in access delay across the patient classes if capacity is not properly allocated and access delays are not properly forecasted. Based on historical data, new patients tend to have many more resource requirements (e.g., number of appointments, referrals to other departments, etc.) than non-urgent cases because physicians will need to obtain more information for diagnosis and/or treatment. In the cases of Neurology and GIM, we see that this heavier load on resources tends to have a negative impact on access delay. This is undesirable because a delayed diagnosis for urgent patients can be very harmful. Neurology as a medical service tends to admit patients with much more serious conditions (compared to GIM) and we can see that the tail distribution on access delay tends to be lighter than GIM, which shows that this service do take urgency into account when scheduling (as opposed to just using a FCFS rule). For example, the GI department has recently followed an initiative to prioritize urgent patients, which explains their significantly shorter access delays. However, the consequences weren't properly forecasted, and resulted in much higher access delays for non-urgent patients (compared to GIM and Neurology). This motivates the need for an integrated approach in capacity planning to understand the tradeoffs between the competing access delay metrics by patient class and medical specialty.

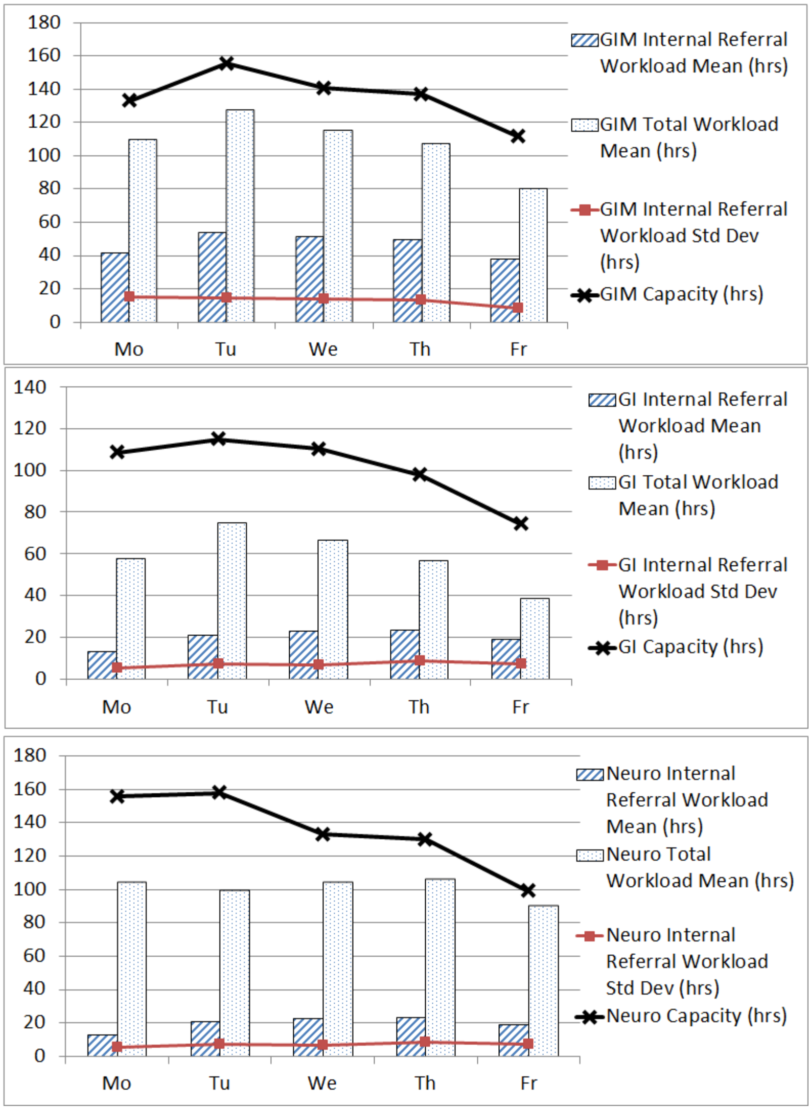


Figure 3.2: Current practice internal referral workload and total workload in 3 medical departments

In GIM, we notice that roughly 40% of the total workload comes from internal referrals from other departments (i.e., downstream appointments in a patient's itinerary) and that the standard deviation of that referred workload is much larger than other departments. This variability of internal referral demand/workload makes it difficult to predict how many initial appointments (urgent/non-urgent visits) to schedule, and how many slots to reserve for internally referred demand. Most of this referred demand is usually generated a day before or the day of the appointment since referrals usually occur based on results from an earlier visit or test, while first visit demand (new or established) occurs at least a week before the appointment. Hence, it is important not to over commit to initial appointments early on since a significant (and variable) amount of referred workload will be required in GIM shortly thereafter: in current practice, that would result in excessive downstream appointments being performed in overtime (since appointment commitments are non revocable). Moreover, it highlights the fact that admission decisions in other departments can have drastic effects on a department like GIM because, with a decentralized approach to scheduling, admission decisions in GI or Neurology are made independently of the load it could impose on GIM. In Fig. 3.2, we see that GI has very low utilization which is probably due to an improper use of resources (e.g. physicians having inefficient slot reservation templates, or poor downstream resource requirement forecasting). For the neurology department, we observe a roughly constant workload throughout the week but with variable resource capacity: this creates variable utilization, large amounts of overtime on Fridays, and poor patient access.

This high level historical analysis aligns with our intuition about the dynamics of this large stochastic queueing network control problem: (1) each combination of patient class and specialty/department has their own issues to improve upon because of different demand patterns and scheduling policies, and (2) a single scheduling/capacity planning decision for each specialty and patient class affects (a) other patient classes, (b) other departments/specialties, and (c) multiple competing critical metrics/objectives in such a system.

3.2.2 Controlling Access via the Template Process

The above motivates the need for a capacity plan that can provide improved access for a heterogeneous group of patients, each with different levels of urgency and different resources needs along their itinerary through the care network. One can describe the system as a multi-class queueing network with blocking. Each group has a different sensitivity to the time to initial appointment. Our approach is to create a plan for capacity allocation so that each patient type must wait for an available appointment permitted for their type (maintaining FCFS within type), enabling the organization to reliably achieve patient class-dependent delay targets.

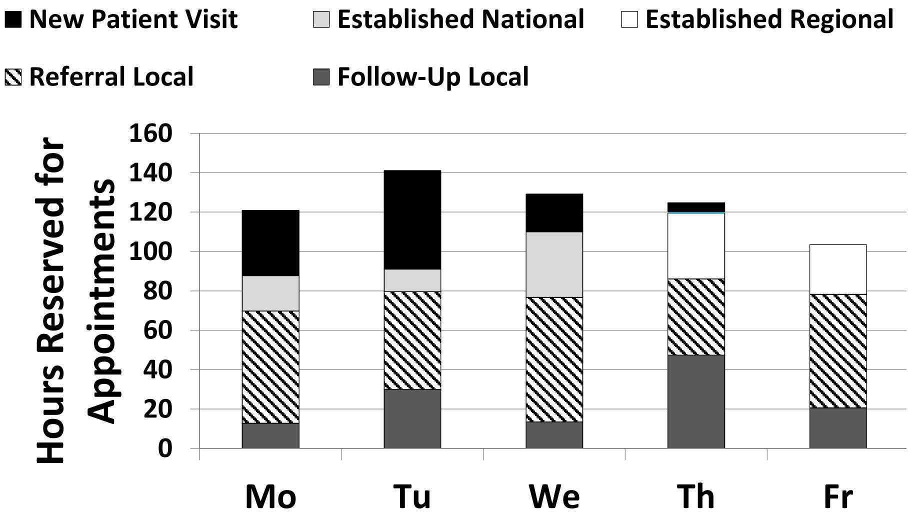


Figure 3.3: One week of a scheduling template for a given medical department/specialty.

OPT's main output to an outpatient practice is a *template* for planned capacity (see example in Fig. 3.3) at the daily level that meets all of the requirements of the practices/services modeled (e.g. GI, GIM, and Neurology). The template can be thought of as a booking plan for how many patients of each type should be scheduled at maximum based on each day of a planning horizon. Fig. 3.3 illustrates one week of a template for the General Internal Medicine (GIM) department at our partnering hospital. Our partnering outpatient institution uses them to capture care provider schedules. These templates can be used at the individual provider level, however, they always exist at the department/medical service level and that is our focus here. OPT will optimize schedule *templates* of outpatient practices by service to control the tradeoff between achieving (1) a high utilization of clinicians' time, (2) short access delays to an initial appointment, (3) a low probability of downstream workload exceeding capacity, and (4) a low number of downstream appointments being performed in overtime.

The templates are populated by the main decision variables of our optimization formulation, a multi-dimensional array $\Theta = \left(\Theta_t^{k,\tau}\right)$, where $\tau \in C$ represents the patient class (e.g. urgency level, geozone, etc.), $k \in K$ represents the department/specialty (or even the subspecialty) required during the root appointment of an itinerary, and $t \in T$ is the day in our planning horizon. The template Θ contains the main decision variables of our optimization formulation, which sets the capacity for initial/root appointments of a care itinerary. The capacity not consumed by the root visits will be used to satisfy subsequent visits ($2^{nd}, 3^{rd}$, etc.) of each patient's itinerary This output template corresponds to existing templates often used in practice by schedulers. For example, the scheduler may be given an upper bound in the number of initial appointments of a given type (by department, subspecialty, urgency, and geozone) to allow on any given day. This upper bound may not always be achieved due to variability in demand, and our models account for this. It allows us to bridge the gap between high level capacity planning, and actual day to day scheduling, while providing accurate forecasts of the most important metrics for decision makers in outpatient practices. EMR systems may implement more sophisticated decision support using the template.

3.2.3 Competing Key Performance Metrics

As mentioned in previous sections, scheduling according to a given template will affect numerous critical performance metrics. Our model incorporates the following competing metrics for each medical department $k \in K$:

1. (M1) Access delay by patient class $\tau \in C$; specifically, we will optimize/control the percentage of type τ patients that exceed n weeks of waiting for a root appointment.
2. (M2-M3) Mean and variance, respectively, of the resource utilization.
3. (M4) Number of planned working hours that exceed resource capacity (or equivalently, mean number of overtime hours).
4. (M5) Probability that the offered workload will exceed resource capacity.

Rather than solving a multi-objective optimization program (which has been shown to be difficult to implement in practice), we will use one metric in our objective with constraints on the others. Scheduling according to the template generated by a given objective (e.g. minimize urgent patient access delay) and constraints (target value for utilization, non-urgent access delay, overtime hours, etc...) will assure that the optimal objective will be achieved while satisfying the targets set on the other metrics. For example, we may want to minimize the access delay metric M1 with $\tau = Urgent$ and $n = 2$ weeks, in department $k = Neurology$. At the same time, the access delay of non-urgent patients in Neurology and any GI or GIM patient should not be negatively affected. Consider that the more root appointments there are for any type of patient, the more workload that it will generate due to downstream resource requirements (especially by increasing access for urgent patients). This will affect M2 and M3 and could lead to the deterioration of metrics M4 and M5. In other words, we are dealing with competing objectives and, in this example, we will make sure to (1) provide an upper bound constraint M1 for all other patient types and medical services, and (2) an upper bound constraint on M2-M5 to make sure we do not over commit our resources to root appointments without proper forecasts of downstream requirements in patients' itineraries.

3.2.4 Differentiation from past research and practice

To the best of our knowledge, the objectives of this research are beyond the capabilities of published research or available commercial products. Our approach differs significantly from the appointment-based scheduling literature (and other areas such as capacity planning, lead time quoting, and revenue management) because it contains multiple technically difficult characteristics such as (1) scope (network vs. single clinic), (2) planning horizon (many day and permitting non-stationarity vs. single day), (3) wait in days for an appointment (explicitly considering working hours by provider by date) rather than minutes of waiting in continuous time after arriving for a visit, (4) incorporation of multiple complex metrics, and (5) service itineraries vs. a single appointment.

This work departs from all commercial software with which we are familiar. Moreover, it differs significantly from the classical outpatient scheduling literature in terms of scope (network vs single clinic) and timing (multiple day planning horizon vs single day). Much of the outpatient scheduling literature focuses on the scheduling of a single clinic, which is often modeled as a queueing system (see the survey paper [7]), and scheduling patients to time slots within a day considering no-shows, doctor availability etc. See these key survey papers [7, 19, 32]). The concept of capacity reservation/allocation is present in some revenue management research (e.g., [2]), but those models lack the features and complexity proposed (see [53, 30, 42]). There has been work in this general area (e.g., [20, 31, 32]) including integrated care in the context of a network of outpatient care services (e.g., [36]).

Recent works for single-unit (non-network) systems have considered priority scheduling and dynamic capacity allocation problems solved via approximate dynamic programming, [29]. Some papers consider multi-priority jobs arriving dynamically that must be scheduled on some future date (or rejected) with holding costs for delays or overtime (see [45, 10, 14]). Other papers consider the fact that each patient (job) may initiate a time series of appointments over multiple days with deadlines/time windows for subsequent appointments (see [15, 31, 51]).

In this research, we rather propose to examine the tactical perspective by allocating capacity for different priority patients on clinicians' schedules in advance of observing requests from individual patients. This is important in outpatient environments, because clinicians typically work according to static calendars as that help them plan their patient care and other activities such as research and administration. Finally, we consider multiple objectives and the percent of patients that meet predetermined, class-specific deadlines for time to initial and subsequent diagnosis and treatment visits. This involves capturing the propagation of workload overflowing from one day to the next as well as downstream visits. Linearizing these processes is another contribution.

Integrated outpatient care also has commonalities with hospital inpatient scheduling problems. Early stochastic models of hospital census includes simulation (see [24, 38, 17, 23]), and probabilistic approaches (see [9]), but capacity optimization was heuristic or intractable. Recent elective patient admission scheduling research has treated the optimization of elective admission schedules for stochastic flows through a network of inpatient hospital resources (e.g. wards) (see [8, 1, 5, 27]). These works, however, consider elective scheduling rather than capacity reservation as is more appropriate for outpatient networks. In this research, we consider how stochastic arrivals fill appointment slots reserved in advance. Further, this research captures delays throughout care episodes, a feature not found in the elective literature.

3.3 “Demand in Progress”

In practice, the provider organization first determines a capacity reservation plan (the *template* in our terminology) and reserves capacity in each medical specialty for each patient class (which can include urgency level/priority) over a finite planning horizon. Then, patients call in to schedule an initial/root appointment according to a random and time-dependent arrival process. The distribution on the daily number of arrivals (i.e., demand for a root appointment) can be general and captured by an empirical distribution based on historical data. Patients will be booked into the earliest reserved slot according to the current availability of the template. Patient demand exceeding the available capacity on a certain day becomes carryover demand that the model attempts to book into the following day and so forth until available capacity is found. To capture these dynamics, *let $\tau \in C$ represent the patient type, where patient type determines both the resources needed to complete a patient's stochastic itinerary as well as the patient's urgency/priority.* For example, patient type includes the status of the patient (urgent vs. non-urgent - important for urgency/priority and stochastic itinerary), the geozone of the patient (national/international vs. local/regional - important for urgency/priority and stochastic itinerary), and the specialty the patient requires in a given service/department $k \in K$ (important for stochastic itinerary). The decision variables $\Theta_t^{k,\tau}$ for planning horizon $t \in \{1, ..., T\}$ are the clinic's capacity allocation plan for patient type τ in service k. Schedulers commonly tend to schedule the root appointment without regard to downstream requirements that may be needed at the time. This is because (1) the root appointment is usually a diagnosis, and access to this visit may be urgent, and (2) very little (if any) information is known about the downstream requirements (referrals, follow-ups) at the time the patient schedules her root/diagnostic appointment. Hence, in this section, we do not consider the modeling of downstream appointments, but this will be analyzed in Section 5.

In our system, demand is either met immediately (scheduled into the current day under consideration if capacity is available) or carryovers to the following day. Each time that a given patient type demand cannot be met on a given day (because previous bookings already filled up the reserved capacity), this demand will be “pushed back” to the next day. We will call this *carryover demand.* To link this to an outpatient practice’s actual process, a scheduler will receive a patient request to start a new itinerary and will consult the scheduling template on each future day to check availability for that patient type based on the capacity plan designed by our methods. We will model this process as the demand will carryover from day to day until there is an availability in the template Θ. The patient will finally be booked by the scheduler on the first day in the future where there is enough available capacity according to the template, so we model this as having her demand carried over from day to day until she is booked. This of course does not make the assumption that the patient will have to wait to know about her appointment until the first day with availability. Our model employing carryover demand simply recreates the booking process of the scheduler in practice. A scheduler will know in real time by consulting the current state of the scheduling template the first day that the patient can be scheduled (according to the template).

On a given day t of our planning horizon, the type τ demand in service k can be split into: (i) the exogenous demand represented by the patient requests for a root appointment that were received on the current day t, and (ii) the carryover demand that represents all previously made requests that were not scheduled up to day t due to capacity/template restrictions. We refer to the combination of (i) and (ii) as the *Demand In Progress (or DIP)* random variable, similar to the concept of the commonly known Work In Progress (WIP) in queueing networks. To maintain realism, we also assume that a patient that requested an appointment in service k earlier than another patient of the same type will have their root appointment happen earlier, since the scheduler will book patients in a first come first served manner based on the limits on available time set by the template. We can think of this booking process as a FIFO queue for each patient class and medical service within the reserved capacity of our template.

Let $DIP^{k,\tau}_{t,w}, t \in \{1, ..., T\}$ be the random variable which represents the amount of type τ DIP demand seeking admission in service k on a given day t in planning horizon (e.g. week) number w (see Eq. 3.1), $\beta^{k,\tau}_{t+1,w}, t \in \{1, ..., T-1\}$ as the type τ carryover demand random variable in service k from day t to day $t+1$ in planning horizon w (Eq. 3.3), and $\beta^{k,\tau}_{1,w}$ as the type τ carryover demand random variable in service k from day T of the prior planning horizon, $w-1$, to day 1 in the current planning horizon w (Eq. 3.2). In this paper, the dynamics of our system will be analyzed over an equilibrium cyclo-stationarity planning horizon of T days, so the template does not depend on w. These random variables will satisfy the following set of equations for any patient type τ and planning horizon w:

$$DIP^{k,\tau}_{t,w} = D^{k,\tau}_{t} + \beta^{k,\tau}_{t,w}, \forall t \in \{1, ..., T\}, \tag{3.1}$$

$$\beta^{k,\tau}_{1,w} = \left(DIP^{k,\tau}_{T,w-1} - \Theta^{k,\tau}_{T}\right)^{+}, \tag{3.2}$$

$$\beta^{k,\tau}_{t+1,w} = \left(DIP^{k,\tau}_{t,w} - \Theta^{k,\tau}_{t}\right)^{+}, \forall t \in \{1, ..., T-1\}, \tag{3.3}$$

where $D^{k,\tau}_{t}$ represents the type τ exogenous demand random variable in service k on day t of our planning horizon (that can be captured by historical data). The DIP on a given week/day, $DIP^{k,\tau}_{t,w}$, is the sum of the exogenous demand, $D^{k,\tau}_{t,w}$, and the carryover demand from the previous day, $\beta^{k,\tau}_{t,w}$. To capture the carryover demand from day t to $t+1$, we take the positive difference of the day t DIP and the amount of carryover demand on that same day.

Given we have a stable system, as $w \to +\infty$, our random variables $DIP^{k,\tau}_{t,w}$, $\beta^{k,\tau}_{t,w}$, and will converge in distribution to the random variables $DIP^{k,\tau}_{t}$ and $\beta^{k,\tau}_{t}$ respectively. Hence, in equilibrium, Eqs. 3.1-3.3 can be rewritten as:

$$DIP^{k,\tau}_{t} = D^{k,\tau}_{t} + \beta^{k,\tau}_{t}, \forall t \in \{1, ..., T\} \tag{3.4}$$

$$\beta^{k,\tau}_{t\oplus 1} = \left(DIP^{k,\tau}_{t} - \Theta^{k,\tau}_{t}\right)^{+}, \forall t \in \{1, ..., T\} \tag{3.5}$$

where the operator $\oplus$ is the modulo T operator defined for convenience to assure that all our variables will have their subscripts in $\{1, ..., T\}$ in our rolling planning horizon (e.g. if $t = T$ then $t \oplus 1 = 1$).

For ease of exposition, let us also define $\alpha_t^{k,\tau}$ to be the amount of $DIP_t^{k,\tau}$ that is scheduled on day t:

$$\alpha_t^{k,\tau} = \min\{DIP_t^{k,\tau}, \Theta_t^{k,\tau}\} = DIP_t^{k,\tau} - \beta_{t\oplus1}^{k,\tau}. \tag{3.6}$$

Then,

$$DIP_t^{k,\tau} = \alpha_t^{k,\tau} + \beta_{t\oplus1}^{k,\tau}. \tag{3.7}$$

Our goal is to translate this set of stochastic non-linear equations (3.4-3.7) into a set of deterministic linear constraints subject to Θ that can capture the probability distributions of the $DIP_t^{k,\tau}$ and $\beta_t^{k,\tau}$ random variables. This information is needed to accurately estimate (based on a given template, Θ) (1) the DIP levels in the planning horizon, (2) the delay to obtain a first visit (which can be calculated by how many times demand carries over), (3) how demand will be fulfilled and scheduled, (4) how capacity decisions will impact demand and workload for downstream visits, and (5) how the institution's resources will be utilized. The dynamics of our system are conceptually similar to the idea of blocking in queueing networks, which is difficult to compute except in special cases, and is non-linear in our main decision variable Θ (i.e., allocated capacity). To overcome this, we propose to develop approximation techniques (see Sections 6.1 and 6.2) that can capture the key system dynamics linearly in order to forecast and control our desired metrics (see Sections 4 and 5) in a deterministic Mixed Integer Programming optimization.

3.4 Access delay

Access delay is a crucial metric for healthcare organizations, as most recently and vividly demonstrated by the public firestorm surrounding long wait lists at Veteran's Administration

(VA) hospitals. To quantify and reduce access delay, this section describes the mathematical approach and challenges at the root of this problem.

Each specialty/service may have different targets/bounds on the waiting time for a first visit appointment to respond appropriately to the urgency of each patient type. This problem induces multi-dimensional tradeoffs, as increasing access for one type of patient necessarily means reserving more capacity for that patient and thereby reduces access for other patient types. Note that our methodology clarifies what is possible and what is not.

We constrain user-selected quantiles of the access delay distribution. This sets a service level bound on the expected fraction of patients of type τ in service k to exceed a desired target (set by the healthcare organization) on access delay. To do so, let $p_n^{k,\tau}$ be the bound we set on the percentage of type τ patients in service k that will exceed $TFAV_n^{k,\tau}$ days of waiting for an initial/root appointment, with $TFAV$ standing for "time to first available visit." The subscript n allows us to set multiple bounds for each patient type. For example, we may want (for $n = 1$) 20% ($p_1^{k,\tau} = 0.2$) of type τ patients to get an appointment within 4 days in service k ($TFAV_1^{k,\tau} = 4$), (for $n = 2$) 50% ($p_2^{k,\tau} = 0.5$) of type τ patients to get an appointment within 7 days in service k ($TFAV_2^{k,\tau} = 7$), and similarly for other desired performance constraints. Note that the number of those constraints, which corresponds to the control we set on the access delay, can vary based on the case and the decision maker's needs. Note that if $TFAV_n^{k,\tau}$ is increasing in n, we will want $p_n^{s,d,z}$ to be decreasing in n. Otherwise, some constraints will have no effect.

To capture these service level metrics, we begin by defining $\delta_{t,n}^{k,\tau}$ as the number of open slots left in our template from day t up to day $t + TFAV_n^{k,\tau}$ after all demand prior to day t has been scheduled:

$$\delta_{t,n}^{k,\tau} = \left(\left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t \oplus l}^{k,\tau} \right) - \beta_t^{k,\tau} \right)^+ .$$

This is the positive difference of (a) the total number of type k, τ slots in an empty template from day t to day $t+TFAV_n^{k,\tau}$ and (b) the number of type k,τ carryovers to day t. We will consider the expected fraction of type τ patients requesting an appointment in service k on day t that exceed $TFAV_n^{k,\tau}$ days of waiting for a root appointment, denoted as $G_{t,n}^{k,\tau}$:

$$G_{t,n}^{k,\tau} = \mathbb{E}\left[\frac{\left(D_t^{k,\tau} - \delta_{t,n}^{k,\tau}\right)^+}{D_t^{k,\tau}}\right],$$

which we will linearize in Section 3.7.2. This quantity could be used in multiple ways. For example, we may want to minimize this quantity for urgent patient types. Alternatively, for other patient types we may constrain $G_{t,n}^{k,\tau}$ to be smaller than a percent (service level $p_n^{k,\tau}$) of the total demand on day t:

$$G_{t,n}^{k,\tau} \leq p_n^{k,\tau}. \tag{3.8}$$

3.5 Workload

We develop a computationally efficient offered load approach to forecast downstream demand during a patient's itinerary based on a stochastic location process parameterized with historical patient flows that are a good representation of the actual visit precedence relationships for a patient's care path (e.g. a general surgery consult must occur at least one day after the CT scan visit to allow ample time to read the CT image). Based on our experience working with data with our partnering hospital, this is a good assumption for some specialties and patient types. On the other hand, it is less realistic for some other services and patient types because the historical time series data is contaminated with endogenous downstream delays inherent in the system that obscure the true itinerary. We propose to use the stochastic location process in an offered load model with infinite capacity, yet historical patient flow data are clearly capacitated. To eliminate this endogeneity, we take advantage of the fact that demand for healthcare services typically follows a seasonal pattern, which is confirmed by the data from the partnering hospital. Thus to approximate the true stochastic location process probabilities, we use data from low utilization months during which time patients rarely have difficulty getting an appointment. This approximates uncapacitated flow because we are trying to capture the itinerary, not the volume of flow.

We construct a model for workload (utilization) at each specialty within the care network based on the capacity reservation plan, Θ. This quantity is needed to assure that we respect daily capacity constraints while estimating/controlling two critical metrics: (1) the probability that the offered workload will exceed capacity, and (2) the expected daily workload overflow, which can be seen as the daily number of overtime hours that have to be spent on downstream appointments (see Section 6.4).

3.5.1 Downstream Demand distributions

Based on historical data, we construct a downstream resource requirement (or "location") probability function capturing the timing (precedence) and type of all subsequent appointments across the network of clinical specialties after a patient's first visit; tailored to each patient type.

We consider K to be the set of medical specialty services at the care provider institution. Since a patient could have appointments at multiple medical specialties on a single day, we need to consider a vector state space for the stochastic location process. Let this vector state space be $\mathcal{S}^0 = \{[a_1, a_2, \ldots, a_1, \ldots, a_{|K|}] : a_k \in \mathbb{Z}^+, \forall k \in K\}$. We let the full state space be $\mathcal{S} = \mathcal{S}^0 \bigcup \{\Delta\}$, where Δ represents that the patient has no appointments (e.g. has returned home, has not yet become a downstream patient, or has no visits on a given day within his/her itinerary). We allow each a_k to take values in $\mathbb{Z}^+$ since in the outpatient setting, a patient can have more than one appointment in a medical specialty on a given day.

Let $L_{t_1}^{k_1,\tau}(t)$ be the $\mathcal{S}$-valued function that represents the unhindered (no delays) appointment(s) needed at time t during a care episode for a patient of type τ that started her itinerary in service k_1, where t_1 is the day that the patient started her itinerary. For notational convenience we let

$$\mathbb{P}(L_{t_1}^{k_1,\tau}(t) \cdot \mathbf{e}_k = m) = r_{t_1}^{k_1,\tau,k}(m, t - t_1), \tag{3.9}$$

where $r_{t_1}^{k_1,\tau,k}(\cdot,\cdot)$ is calculated based on historical data, and $\mathbf{e}_k$ is a column vector with all 0's and a 1 in the k^{th} row.

3.5.2 Offered workload formulation

Based on the formulations in this past section, we can now express the steady state mean offered workload on day t of our planning horizon in each medical department. First, let M_k be the maximum number of time slots a patient can require within a day in specialty k.

Theorem III.1. *The steady state mean offered workload in service k on day t (under the capacity reservation plan Θ) can be computed as:*

$$\overline{W}_t^k = \sum_{\tau \in C} \sum_{k_1 \in K} \sum_{t_1=1}^{T} \mathbb{E}[\alpha_{t_1}^{k_1,\tau}] \cdot \sum_{j=0}^{\infty} \sum_{m=1}^{M_k} m \cdot r_{t_1}^{k_1,\tau,k}(m, t - t_1 + jT) \cdot \mathbf{e}_k \cdot S_k \tag{3.10}$$

Proof. By combining $L_{t_1+jT}^{k_1,\tau}(t)$ with $\alpha_{t_1,j}^{k_1,\tau}$, the number of type τ initial appointments in specialty k_1 on day $t_1 \in \{1,...,T\}$ of planning horizon j, we can determine the demand stream (and also workload by multiplying by the appointment length required) for downstream appointments at each specialty t days later.

Therefore, for medical specialty k, the offered workload random variable (which includes initial and subsequent appointments for each patient type) on day t of planning horizon (e.g. week) w under the booking reservation scheme Θ can be expressed as:

$$W_{t,w}^k = \sum_{\tau \in C} \sum_{k_1 \in K} \sum_{j=0}^{w} \sum_{t_1=1}^{T} \sum_{i=1}^{\alpha_{t_1,j}^{k_1,\tau}} L_{t_1+jT,i}^{k_1,\tau}(t + wT - (t_1 + jT)) \cdot \mathbf{e}_k \cdot S_k, \tag{3.11}$$

where S_k is the length of a slot in specialty k, and $L_{t_1+jT,i}^{k_1,\tau}(t + wT - (t_1 + jT))$ is the $i^t h$ i.i.d instance of the $L_{t_1+jT}^{k_1,\tau}(t + wT - (t_1 + jT))$ random variable. Notice that the first inner sum in Eq. 3.11 considers $\alpha_{t_1,j}^{k_1,\tau}$, all type (k_1, τ) patients that were scheduled for their root appointment on day t_1 of week $j \leq w$. For each one of those patients, we will multiply a given realization of $L_{t_1+jT}^{k_1,\tau}(t+wT-(t_1+jT))$ with $e_k \cdot S_k$, which will determine the workload each one of those patients will impose in specialty k, $t+wT-(t_1+jT)$ days later (hence, on day t of week w). For notational convenience, if $t+wT-(t_1+jT)$ is negative then $L_{t_1+jT}^{k_1,\tau}(t+wT-(t_1+jT))$ will be 0. The second and third inner sums are over all days weeks w from 0 to w, and all days t_1 in weeks $j = 0, ..., w$. Therefore, this captures all type (k_1, τ) patients that were scheduled for a root appointment earlier than day t of planning horizon w, and the sum of their resource requirements in specialty k on day t of week w. The last two outer sums consider all patient types $\tau \in C, k_1 \in K$.

As $w \to \infty$, Eq. 3.11 the steady state mean workload on day t of our planning horizon can be computed as follows:

$$\overline{W}_t^k = \sum_{\tau \in C} \sum_{k_1 \in K} \sum_{t_1=1}^{T} \sum_{j=0}^{\infty} \mathbb{E} \sum_{i=1}^{\alpha_{t_1,j}^{k_1,\tau}} L_{t_1+jT,i}^{k_1,\tau}(t+wT-(t_1+jT)) \cdot \mathbf{e}_k \cdot S_k \tag{3.12}$$

$$= \sum_{\tau \in C} \sum_{k_1 \in K} \sum_{t_1=1}^{T} \mathbb{E}[\alpha_{t_1}^{k_1,\tau}] \cdot \sum_{j=0}^{\infty} \sum_{m=1}^{M_k} m \cdot r_{t_1}^{k_1,\tau,k}(m, t-t_1+jT) \cdot \mathbf{e}_k \cdot S_k \tag{3.13}$$

Eq. 3.12 calculates the limit of Eq. 3.11 as $w \to \infty$ and takes the expectation of this expression. Then, using Wald's equation, the expected value of a random sum, will be equal to the expected value of $\alpha_{t_1,j}^{k_1,\tau}$ multiplied by the expected value of the $L_{t_1+jT,i}^{k_1,\tau}(t+wT-(t_1+jT)) \cdot \mathbf{e}_k$ random variable. Using Eq. 3.9, and using the fact that our calculations are made in equilibrium, Eq. 3.13 follows, and we proved Theorem III.1. □

To calculate the variance in the number of appointment slots requested for service k on day t of our planning horizon, we will need to compute the variance of the number of type τ patients scheduled in service k on day t according to our template Θ.

Theorem III.2. *The steady state variance of the offered workload in specialty k on day t of our steady state planning horizon is given by*

$$\begin{aligned}\tilde{W}_t^k = \sum_{t_1=1}^{T} \sum_{\tau} \sum_{k_1 \in K} \sum_{j=0}^{\infty} \Bigg[& Var\left[\alpha_{t_1}^{k_1,\tau}\right] \left(\sum_{m=0}^{M_k} m \cdot r_{t_1}^{k_1,\tau,k}\left(m, t-t_1+jT\right) \cdot \mathbf{e}_k \cdot S_k \right)^2 + \mathbb{E}\left[\alpha_{t_1}^{k_1,\tau}\right] \\ & \cdot \sum_{m=0}^{M_k} \Bigg(m^2 \cdot S_k^2 \cdot r_{t_1}^{k_1,\tau,k}\left(m, t-t_1+jT\right) \cdot \mathbf{e}_k \left(1 - r_{t_1}^{k_1,\tau,k}\left(m, t-t_1+jT\right) \cdot \mathbf{e}_k\right) \\ & - \sum_{m<q\le M_k} 2mq \cdot S_k^2 \cdot r_{t_1}^{k_1,\tau,k}\left(m, t-t_1+jT\right) \cdot \mathbf{e}_k \cdot r_{t_1}^{k_1,\tau,k}\left(q, t-t_1+jT\right) \cdot \mathbf{e}_k \Bigg) \Bigg]\end{aligned} \tag{3.14}$$

Proof. Following the general idea of the proof in Theorem III.1, we can formulate the steady state variance of the specialty k workload random variable on day t of our planning horizon as:

$$\tilde{W}_t^k = \sum_{\tau \in C} \sum_{k_1 \in K} \sum_{t_1=1}^{T} \sum_{j=0}^{\infty} Var\left[\sum_{i=1}^{\alpha_{t_1,j}^{k_1,\tau}} L_{t_1+jT,i}^{k_1,\tau}(t + wT - (t_1 + jT)) \cdot \mathbf{e}_k \cdot S_k \right] \tag{3.15}$$

Notice that $(L_{t_1+jT,i}^{k_1,\tau}(t + wT - (t_1 + jT)) \cdot \mathbf{e}_k)_i$ is a sequence of independent random variables which are also independent of $\alpha_{t_1,j}^{k_1,\tau}$, the number of patients scheduled. Moreover, the variance of $L_{t_1+jT,i}^{k_1,\tau}(t + wT - (t_1 + jT)) \cdot \mathbf{e}_k$ is the same for every i. Then, the variance of this random sum can be expressed as follows:

$$\begin{aligned} & Var\left[\sum_{i=1}^{\alpha_{t_1,j}^{k_1,\tau}} L_{t_1+jT,i}^{k_1,\tau}(t + wT - (t_1 + jT)) \cdot \mathbf{e}_k \cdot S_k \right] \\ & = \mathbb{E}[\alpha_{t_1,j}^{k_1,\tau}] \cdot Var[L_{t_1+jT,i}^{k_1,\tau}(t + wT - (t_1 + jT)) \cdot \mathbf{e}_k \cdot S_k] \\ & + \mathbb{E}[L_{t_1+jT,i}^{k_1,\tau}(t + wT - (t_1 + jT)) \cdot \mathbf{e}_k \cdot S_k]^2 \cdot Var[\alpha_{t_1,j}^{k_1,\tau}] \end{aligned} \tag{3.16}$$

By using the same arguments as the proof of the previous theorem, Eq. 3.14 follows. □

3.6 DIP distribution adjustment combining optimization and simulation

3.6.1 Discrete Event Simulation

In order to validate the approximations used in our template optimization approach, we constructed a high-fidelity discrete event simulation in Visual C++. The simulation inputs include: (1) an optimized weekly template in equilibrium, (2) the daily resource capacity, (3) the stochastic location functions for downstream appointments based on historical data, (4) the empirical demand distribution for each patient type τ based on historical data, and (5) the historical internal referral workload mean and variance from patients starting their itineraries outside of the three departments we consider (GI, GIM and Neurology). We will consider a 500 day warm-up period to reach equilibrium. The metrics considered in this simulation will be: (1) the expected percentage of type τ patients exceeding ω days of waiting, (2) the mean and variance of the number of type τ patients scheduled by day of week (as well as the carryover patients), (3) the mean and variance of the offered workload by day of week, (4) the expected number of overtime hours by day of week, and (5) the probability of exceeding the daily resource capacity. The metrics we wish to validate will be computed over an evaluation horizon/period that reached equilibrium by using the batch mean method. In addition, the simulation considers a large number of replications (of the warm-up and the evaluation period) for the evaluation of the mean and variance of our metrics, which we shown to be enough to have our sample means and variances converge to the true means and variances. The number of replications, the number of batches, and their sizes in the evaluation period, have been selected to guarantee a 95% confidence interval on our sample metrics (for their means and variances/standard deviations).

At the end of the day, we consider all arrivals of all types that occurred within the day, and the template scheduling function determines the first available **reserved** slots in our schedule by keeping track of the number of remaining open reserved slots for each patient type. Once all patients that requested their initial appointment on that day have been scheduled (on a future day), we update our metrics. We will use realizations of the stochastic location distributions to determine the workload they will impose on days following the first visit, due to their downstream appointments. Note that if any of the appointments scheduled (initial or downstream) generate a workload that makes the aggregate offered load go over the daily capacity, then the appointment will be treated as "overtime" workload. Once the simulation reaches the end of its horizon, and has gone through all its replications, we can then compare our simulated metrics to the results we got from the optimization, and compute the percentage errors.

3.6.2 Simulated DIP distributions

On any day t, we are adding a large number of random variables (each determining the amount of carry over demand from previous days l to t with $l < t$) to the exogenous demand on day t. But we should note that those random variables that we are summing are not i.i.d, therefore using the Central Limit Theorem to assume the DIP random variables follow a Normal distribution has its limitations. In some situations e.g., in heavy traffic, the skew of the distribution tends to increase: this leads to an inexact characterization of the DIP distributions if we assumed normality, therefore errors in our metrics' forecasting (especially for the second moment terms) and a suboptimal template. We conducted some statistical analysis on DIP distributions given a wide range of templates Θ. We considered historical demand patterns, and first simulated the booking process in a low traffic setting (for a template that would generate low utilization/waiting times). Then, we simulated the booking process assuming a historical template, which is a representation of a heavy traffic scenario. We estimated the DIP distributions under those capacity planning scenarios and compared it to Normal distributions (or log-normal) with same means and variances. In Fig.

3.4 we represent the low traffic scenario while the heavy traffic case can be found in Fig. 3.5.

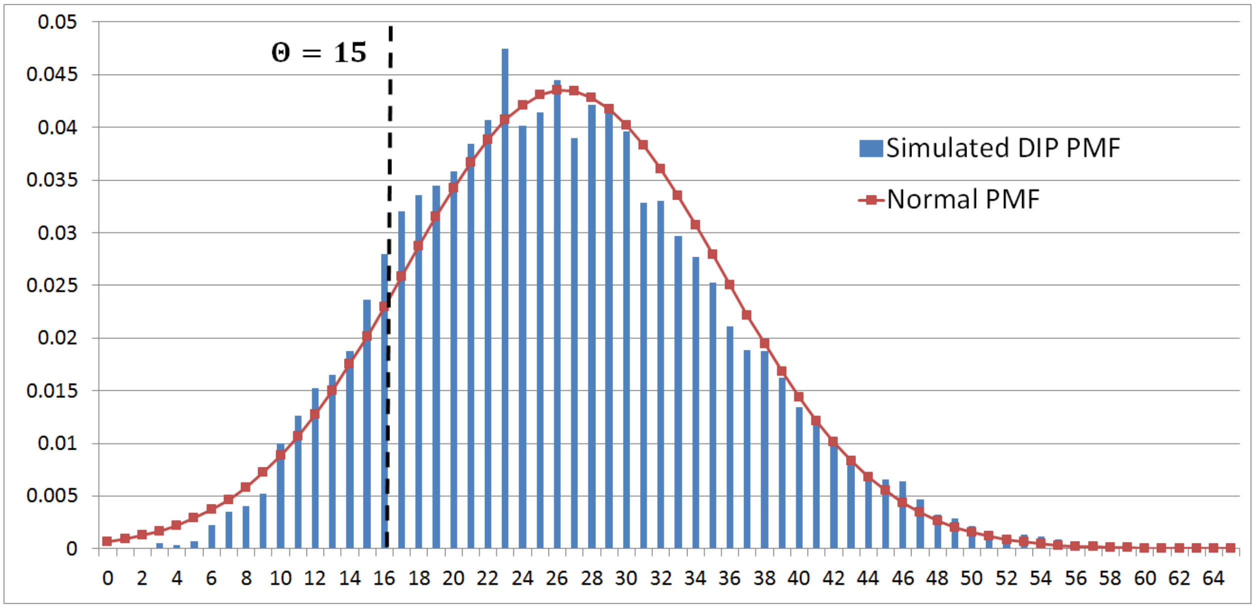


Figure 3.4: Simulated Monday DIP distribution with mean of 26.36 and standard deviation of 9.16 approximated by a Normal distribution (26.36, 9.16) and its effects on the Tuesday carryover demand relative to the truncation factor Θ

$\Theta_1^{t,\tau} = 15$	DIP Simulated Dist.	Normal Dist.	APE	AE
$\mathbb{P}(DIP_1^{t,\tau} > \Theta_1^{t,\tau})$	0.86	0.86	0.06	0.00
$\mathbb{E}(\beta_2^{t,\tau})$	11.68	11.83	1.27	0.15
$\sqrt{Var(\beta_2^{t,\tau})}$	8.65	8.33	3.72	0.32

Table 3.1: Absolute Percentage Error (APE) and Absolute Error (AE) of the probability of exceeding Θ = 15, the mean Tuesday carryover demand, and Tuesday carryover demand standard deviation when we approximate the Monday DIP distribution by a Normal

In Table 3.1 we compared three metrics that would be impacted by a Normal approximation to the DIP: the probability of exceeding the number of reservations $\Theta_1^{t,\tau}$, the mean carryover demand and the carryover demand standard deviation. The accuracy of those metrics will vary based on Θ's relative value to the distribution of the DIP. In the first two cases ($\Theta = 15, 25$), when our template is below the mean or equal to the mean of our

DIP distribution, we observe high accuracy in our metrics. When looking at the Absolute Percentage Error (APE) and Absolute Error (AE) of those three metrics, we can see that the Normal approximation to the DIP distribution behaves quite well. But if we now look at the heavy traffic case (Fig. 3.5 and Tables 3.2 and 3.3), which is a much more realistic representation of reality, we can see that the Normal approximation generates large APEs for the three metrics we care about. This is mainly due to the increase in skewness of the true DIP distribution, which now behaves closer to a log-normal distribution.

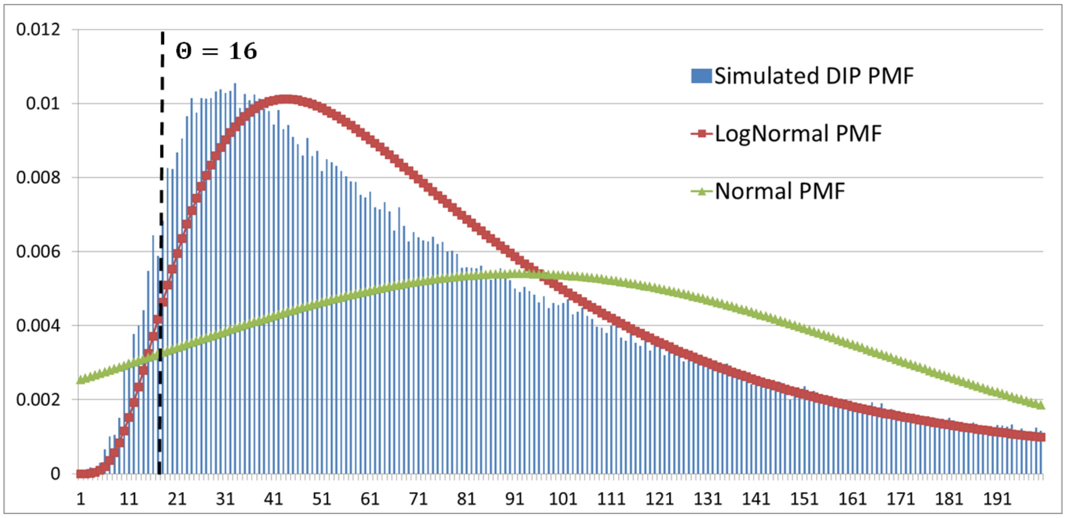


Figure 3.5: Simulated Monday DIP distribution with mean of 91.91 and standard deviation of 73.91 approximated by a Normal distribution (91.91, 73.91) and a log-normal distribution (91.91, 73.91) and its effects on the Tuesday carryover demand relative to the truncation factor Θ

$\Theta_1^{t,\tau} = 16$	DIP Simulated Dist.	Normal Dist.	log-normal Dist.
$\mathbb{P}(DIP_1^{t,\tau} > \Theta_1^{t,\tau})$	0.964	0.846	0.978
$\mathbb{E}(\beta_2^{t,\tau})$	76.037	81.756	74.143
$\sqrt{Var(\beta_2^{t,\tau})}$	73.771	67.454	64.443

Table 3.2: Probability of exceeding Θ = 16, the mean Tuesday carryover demand, and Tuesday carryover demand standard deviation for the Monday DIP distribution when (1) it has been simulated (i.e., the true distribution), (2) approximated by a Normal, and (3) approximated by a log-normal distribution. (1) and (2) both generated with mean and standard deviation equal to the simulated DIP

$\Theta_1^{t,\tau} = 16$	Normal APE	log-normal APE
$\mathbb{P}(DIP_1^{t,\tau} > \Theta_1^{t,\tau})$	12.261	1.427
$\mathbb{E}(\beta_2^{t,\tau})$	7.522	2.490
$\sqrt{Var(\beta_2^{t,\tau})}$	12.642	8.725

Table 3.3: Absolute Percentage Error (APE) of the probability of exceeding $\Theta = 16$, the mean Tuesday carryover demand, and Tuesday carryover demand standard deviation when we approximate the Monday DIP distribution by a Normal and a log-normal distribution, both with mean and standard deviation equal to the simulated DIP

Our model will adjust the DIP distributions in a recursive manner via simulation to determine a truly optimal template with low metrics' forecasting errors. The first step of our methodology and algorithm will be to solve the optimization program assuming a historical distribution on the DIP random variables (e.g. see the simulated bars in Fig. 3.5), and the following steps of the algorithm will use simulation to determine the correct distribution associated with the optimized template (of a given algorithm iteration). In most of our case studies occurring in clinical practice, an adjustment in the probability mass function of the DIP random variables needs to occur for accurate forecasts of our metrics, especially for the variance/standard deviation estimates (since it relates to the tail behavior of the distribution).

3.6.3 Discretization of the DIP distributions

The type τ DIP random variable in service k on a given day t of the planning horizon, $DIP_t^{k,\tau}$ has mean $\overline{DIP}_t^{k,\tau}$ and variance $\tilde{DIP}_t^{k,\tau}$ satisfying the following: $\overline{DIP}_t^{k,\tau} = \overline{D}_t^{k,\tau} + \overline{B}_t^{k,\tau}$, $\tilde{D}_t^{k,\tau} = \tilde{D}_t^{k,\tau} + \tilde{B}_t^{k,\tau}$, where $\overline{D}_t^{k,\tau} = \mathbb{E}\left[D_t^{k,\tau}\right]$, $\overline{B}_t^{k,\tau} = \mathbb{E}\left[\beta_t^{k,\tau}\right]$, $\tilde{D}_t^{k,\tau} = Var\left[D_t^{k,\tau}\right]$, and $\tilde{B}_t^{k,\tau} = Var\left[\beta_t^{k,\tau}\right]$. First, by the linearity of the expected value function, the mean of the DIP random variable will clearly be the sum of the mean exogenous and carryover demands (See Eq. 3.4). Moreover, note that the exogenous demand $D_t^{k,\tau}$ is independent to the carryover demand $\beta_t^{k,\tau}$ on a given day t, since the number of patient requests for an appointment on day t is independent of anything else in our system (including the carryover demand). This allows us to sum up the variance of those two random variables to get the variance of the DIP. Hence, we only need to express the expected value ($\overline{B}_t^{k,\tau}$) and variance ($\tilde{B}_t^{k,\tau}$) of the $\beta_t^{k,\tau}$ random variables in order to get an estimate on the DIP distributions. However, the expectation and variance of the carryover demand (as well as the probability of carrying over) for a generally distributed DIP random variable is non-linear in the admission schedule Θ for most distributions.

We propose an approximation based on Riemann integration. Let $I = \{1, 2, \ldots, \overline{I}\}$ be an index that creates a discrete grid with $\overline{I}$ sections. The grid need not (and in application is not chosen to be) linear. Thus we have a one-to-one mapping function $m(i) : I \to \mathcal{M}$ that maps the integer values of I to the grid values $\mathcal{M}$ (see a grid example in Table 3.4).

$i \in I$	1	2	3	4	5	6	7	8	9	10	11	12	13
$m(i) \in \mathcal{M}$	-3.1	-1.8	-1.2	-0.6	-0.2	-0.1	0.0	0.1	0.2	0.6	1.2	1.8	3.1

Table 3.4: Sample grid mapping from the set I to the grid $\mathcal{M}$

Definition III.3. For a given grid $m(i), i \in I$, and any feasible template Θ, let $\Psi_t^{k,\tau}(i)$ be a data input to the optimization determining the probability of having the random variable $DIP_t^{k,\tau}$ be in the interval $\left[\overline{DIP}_t^{k,\tau} + m(i) \cdot \sqrt{D\tilde{I}P_t^{k,\tau}}, \overline{DIP}_t^{k,\tau} + m(i+1) \cdot \sqrt{D\tilde{I}P_t^{k,\tau}}\right]$.

$\Psi_t^{k,\tau}(i)$ will be calculated before an optimization run either by using historical data or by using discrete event simulation (See Section 3.6.4 for more explanations). We justify the assumption that $\Psi_t^{k,\tau}(i)$ does not vary regardless of which feasible template the Mixed Integer Program converges to with:

1. the changes in DIP means and variances (when the template is optimized) are calculated linearly online (see Section 3.7.1) and are incorporated in the grid design (See Definition 6.1). Hence, assuming invariance in $\Psi_t^{k,\tau}(i)$ does not mean we assume that

our optimized template will have the same DIP distributions as historically, but rather, it will belong to the same distribution family with different means and variances.

2. as we go through iterations of the DIP distribution adjustment algorithm (Section 3.6.4), our input $\Psi_t^{k,\tau}(i)$ to the MIP will get closer to the true DIP distribution associated with the optimized template. This is due mainly to the fact that as we go through the iterations of our algorithm, the variations in the "optimal" template will decrease.

One grid design that has nice intuition and behaves well in practice is to make the grid such that each interval contains an equal amount of probability density on a partition of the interval $[0, +\infty)$ of our DIP distribution. An example of such a discretization is shown in Fig. 3.6 with the grid values of Table 3.4, where our decision variables are represented in bold. Given a linear approximation of the square root function (since mean and variance can be calculated linearly, see Section 3.7.1), those approximations allow us to estimate linearly the discretized distributions of $DIP_t^{k,\tau}(i)$. In the limit, as $\overline{I} \to \infty$ and $m(i+1)-m(i) \to 0, \forall i \in I$, this approximation will converge to the exact DIP distribution because it is equivalent to Riemann integration.

We propose a linearizing approximation for the square root function. Let $D\tilde{I}P_t^{k,\tau}$ be the quantity that we want to take the square root of (e.g., type τ variance of the DIP on day t). Let $D\hat{I}P_t^{k,\tau}$ be an initial guess of $\sqrt{D\tilde{I}P_t^{k,\tau}}$ (e.g., standard deviation). We approximate the true $\sqrt{D\tilde{I}P_t^{k,\tau}}$ by applying one step of the Newton's method: $\sqrt{D\tilde{I}P_t^{k,\tau}} \approx 0.5 \cdot (D\tilde{I}P_t^{\tau}/D\hat{I}P_t^{\tau} + D\hat{I}P_t^{\tau}))$. In previous work, [27] showed that this approximation is highly effective in modeling healthcare workloads when $D\hat{I}P_t^{k,\tau}$ is chosen to be the historical workload standard deviation.

Now, we can define a separate variable, $DIP_t^{k,\tau}(i) \geq 0$, for the realization of $DIP_t^{k,\tau}$ at

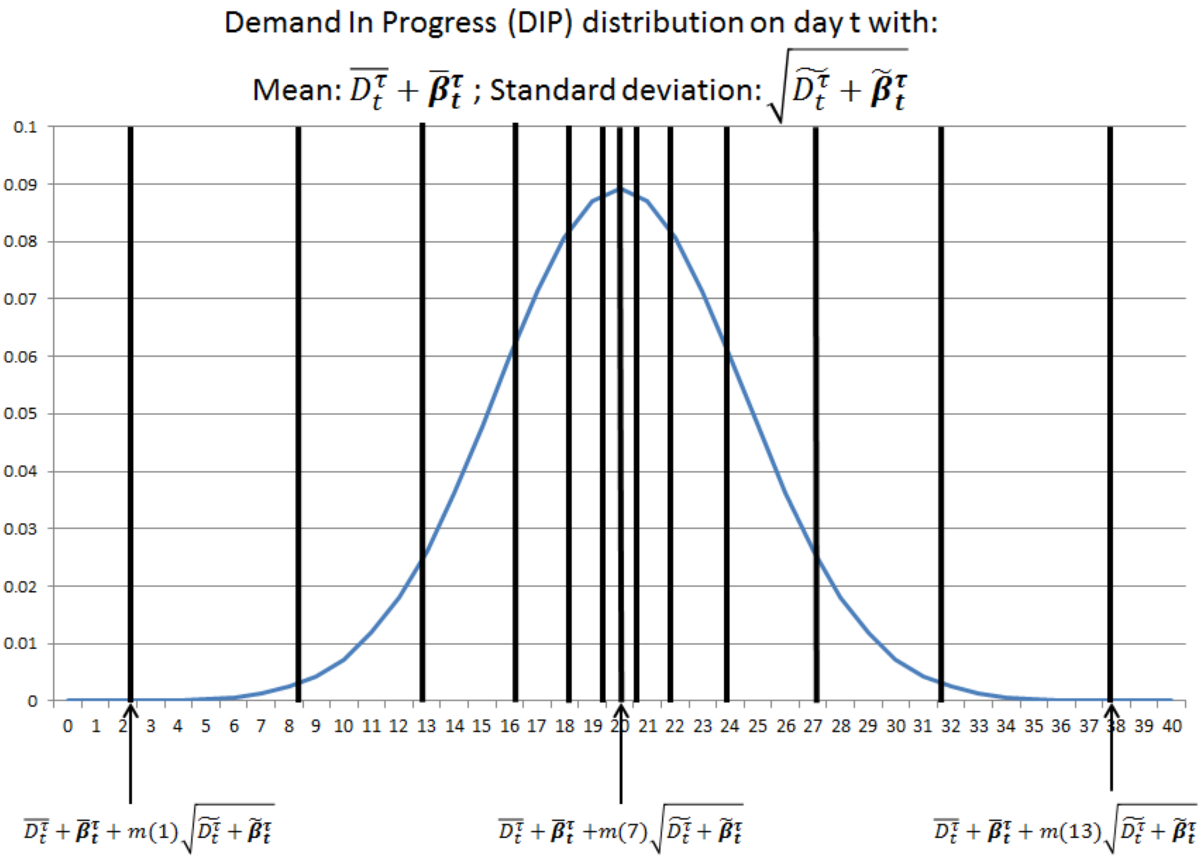


Figure 3.6: Applying the grid approximation to a DIP Normal distribution

each grid level i based on the grid discretization and Newton approximations made above:

$$DIP_t^{k,\tau}(i) := \overline{DIP}_t^{k,\tau} + \frac{1}{2}m(i) \cdot \left(\frac{D\tilde{I}P_t^{k,\tau}}{D\hat{I}P_t^{k,\tau}} + D\hat{I}P_t^{k,\tau} \right) + \epsilon_t^{k,\tau}(i), \tag{3.17}$$

where

$$\epsilon_t^{k,\tau}(i) \leq -\left(\overline{DIP}_t^{k,\tau} + \frac{1}{2}m(i) \cdot \left(\frac{D\tilde{I}P_t^{k,\tau}}{D\hat{I}P_t^{k,\tau}} + D\hat{I}P_t^{k,\tau} \right) \right) + M \cdot b_t^{k,\tau}(i), \tag{3.18}$$

$$\epsilon_t^{k,\tau}(i) \leq M \cdot (1 - b_t^{k,\tau}(i)), \tag{3.19}$$

and

$$-M \cdot \left(1 - b_t^{k,\tau}(i)\right) \leq \overline{DIP}_t^{k,\tau} + \frac{1}{2}m(i) \cdot \left(\frac{D\tilde{I}P_t^{k,\tau}}{D\hat{I}P_t^{k,\tau}} + D\hat{I}P_t^{k,\tau} \right) \tag{3.20}$$

$$\overline{DIP}_t^{k,\tau} + \frac{1}{2}m(i) \cdot \left(\frac{D\tilde{I}P_t^{k,\tau}}{D\hat{I}P_t^{k,\tau}} + D\hat{I}P_t^{k,\tau} \right) \leq M \cdot b_t^{k,\tau}(i). \tag{3.21}$$

Notice that in Eq. 3.17, we also add a term $\epsilon_t^{k,\tau}(i) \geq 0$ which will assure that $DIP_t^{k,\tau}(i) \geq 0, \forall i \in I$ since the DIP random variable cannot take negative values. We choose M to be a large enough integer, and create a binary variable $b_t^{k,\tau}(i)$ where constraints 3.20 and 3.21 assure that $b_t^{k,\tau}(i) = 1$ if $\overline{DIP}_t^{k,\tau} + \frac{1}{2}m(i) \cdot \left(\frac{D\tilde{I}P_t^{k,\tau}}{D\hat{I}P_t^{k,\tau}} + D\hat{I}P_t^{k,\tau}\right) \geq 0$ and 0 otherwise. Hence $\epsilon_t^{k,\tau}(i)$ will either take the value 0 or $-\overline{DIP}_t^{k,\tau} - \frac{1}{2}m(i) \cdot \left(\frac{D\tilde{I}P_t^{k,\tau}}{D\hat{I}P_t^{k,\tau}} + D\hat{I}P_t^{k,\tau}\right) \geq 0$ thanks to constraints 3.18 and 3.19 based on the binary variable $b_t^{k,\tau}(i)$.

3.6.4 The DIP distribution adjustment algorithm

We will employ the following algorithm to capture the true discretized DIP distributions of a given optimized BRPlan.

1. Input an objective and constraints on the other metrics into the Mixed Integer Program (see Section 3.7).

2. Solve the optimization program described in Section 3.7 associating the historical densities $\Psi^{k,\tau}(i)[0]$ to grid points $DIP_t^{k,\tau}(i)[0]$ (see Section 3.6.3) and get the optimal template $\Theta^*[0]$, and the resulting optimized/controlled metrics.

3. Set $n = 1$.

4. Input the iteration $n-1$ template $\Theta^*[n-1]$ into the discrete event simulation described in Section 6.1 and compute the simulated density $\Psi_t^{k,\tau}(i)[n]$ at grid point $DIP_t^{k,\tau}(i)[n-1]$.

5. If $\sum_{i \in I} |\Psi_t^{k,\tau}(i)[n] - \Psi_t^{k,\tau}(i)[n-1]| \leq \kappa, \forall k, \forall \tau$ then go to step 9.

6. Input $\Psi_t^{k,\tau}(i)[n]$ into the Mixed Integer Program with the same objective and constraints.

7. Solve the optimization program now associating the iteration n simulated density $\Psi_t^{k,\tau}(i)[n]$ to grid points $DIP_t^{k,\tau}[n]$.
8. Set $n = n + 1$ and go to Step 4.
9. Compute the statistics of our simulated performance metrics and validate our analytical model by calculating (1) the Absolute Percentage Error (APE), and (2) the Absolute Error (AE) on the desired performance metrics.
10. End Algorithm

At iteration $n = 0$, CPLEX determines the optimal template (given an objective and constraints on other metrics) assuming a historical discretized distribution $\Psi^{k,\tau}(i)[0], i \in I$ on the DIP random variables. Note that this optimized template is not truly optimal since the densities in between grid points are likely to change if our new template varies significantly from the historical one. At iteration $n = 1$, we first use discrete event simulation to observe how patient demand will fill the template slots and record the simulated true density in between the $n = 0$ optimized grid points (i.e. in between $m(i + 1)$ and $m(i)$ DIP standard deviations above/below the $n = 0$ DIP mean). If our historical distribution assumption is "good enough", we stop the algorithm and validate the statistics on our modeled performance metrics. Otherwise, we input into the optimization program the adjusted densities (simulated at $n = 1$) in between two grid points from iteration $n = 0$. When optimizing again at $n = 1$, the DIP grid points will not be the same as the $n = 0$ DIP grid points since the mean and standard deviation of our DIP will change with an updated optimized template. So our adjusted densities will not be assigned to the exact same DIP values but will still reflect the amount of density within a certain standard deviation factor above/below the mean. This information is valuable to capture the correct shape of the distribution while still being able to update its mean and variance when the optimized template changes. We then repeat the same procedure until our optimized template stabilizes while now using the true distributions of the DIP random variables.

3.7 Deterministic Linear Optimization

3.7.1 Linear parameterization the DIP distribution

Our goal is to compute the mean and variance of the DIP random variables for any day $t \geq 0$ (and, equivalently the mean and variance of the carryover random variables) linearly in our decision variables Θ. We will show this by induction. If we can show that this is true for all $t \geq 0$, then this result will also apply to our equilibrium planning horizon with $t \in T$.

Initialization On day 1, we start with an empty system, meaning there is no carryover demand. The DIP random variable will be equal to the exogenous demand on day 1 (see Eq. 3.4) which can be expressed linearly in our decision variables Θ (since the exogenous demand is independent of our template).

Induction hypothesis We assume that we can express the DIP mean and variance on day t linearly in our decision variables Θ.

Induction step We would like to compute the distribution of the DIP random variable on day $t+1$. To estimate the mean $\overline{DIP}_t^{k,\tau}$ and variance $D\tilde{I}P_t^{k,\tau}$ of the DIP random variables linearly in Θ, it is sufficient to express linearly the mean $\overline{B}_{t+1}^{k,\tau}$ and variance $\tilde{B}_{t+1}^{k,\tau}$ of the carryover random variables on day $t+1$ (see Eq. 3.4).

$\beta_{t+1}^{k,\tau}$ will be a truncation of the DIP demand on day t, a distribution with mean $\overline{DIP}_t^{k,\tau}$ and variance $D\tilde{I}P_t^{k,\tau}$, with a lower-bound truncation at $\Theta_t^{k,\tau}$, which is then translated by $\Theta_t^{k,\tau}$: this means that $\beta_{t+1}^{k,\tau}$ will be a mixed random variable and have a mass point at 0 equal to the probability of the DIP demand on day t being below $\Theta_t^{k,\tau}$, and will have the same distribution as the tail distribution of the DIP when the DIP random variable is larger than $\Theta_t^{k,\tau}$ (i.e., for $\beta_{t+1}^{k,\tau} > 0$). An example of such a distribution can be seen in Fig. 3.7 with a truncation at 15 (Fig. 3.7 left) and a truncation at 25 (Fig. 3.7 right) of an assumed normally distributed DIP.

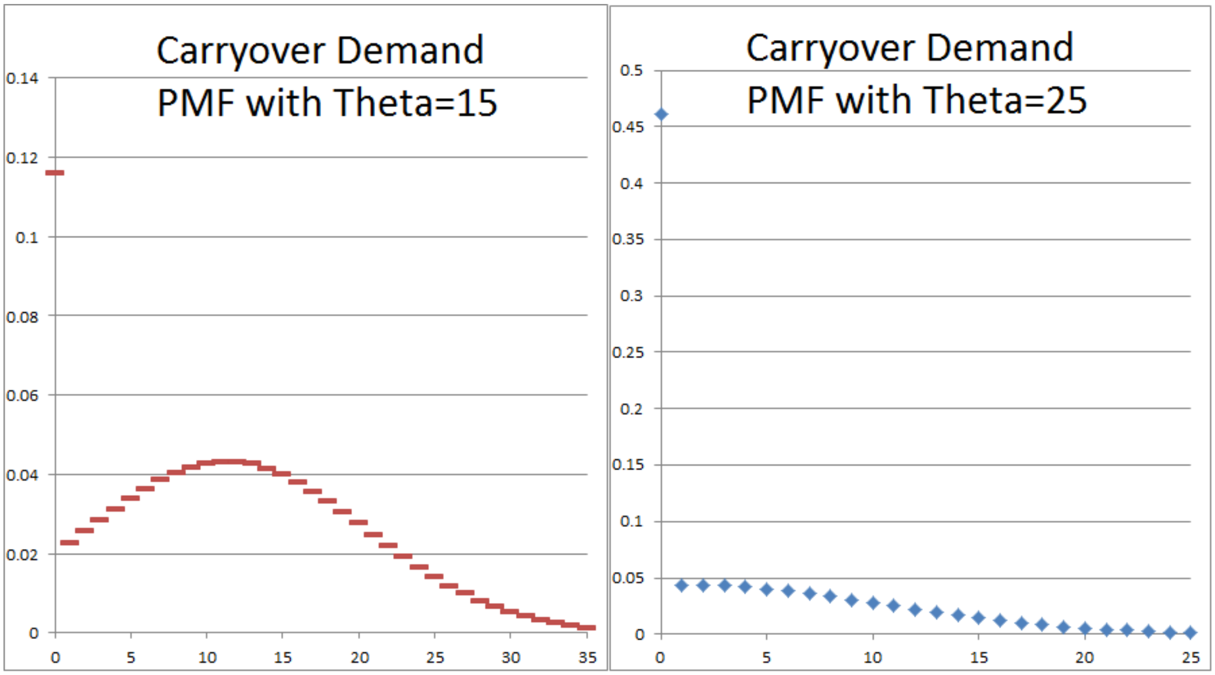


Figure 3.7: Probability mass function of the carryover demand to the next day when the current day DIP has a Normal distribution with mean 26.36 and standard deviation 9.16 with (left) a template $\Theta = 15$ and (right) a template $\Theta = 25$

$DIP_t^{k,\tau}(i)$ (see Eq. 3.17) is linear in our decision variables Θ by the induction hypothesis: $\overline{DIP}_t^{k,\tau}$ and $D\tilde{I}P_t^{k,\tau}$ can be estimated linearly in Θ. We can now express the carryover demand level at grid level/realization i of the DIP based on Eq. 3.5 by:

$$\beta_{t+1}^{k,\tau}(i) = \left(DIP_t^{k,\tau}(i) - \Theta_t^{k,\tau}\right)^+ . \tag{3.22}$$

To linearize $\beta_{t+1}^{k,\tau}(i)$, we start by defining a binary decision variable $y_t^{k,\tau}(i,l)$ that equals 1 when $DIP_t^{k,\tau}(i) - \Theta_t^{k,\tau} > l$, and 0 otherwise. The following constraints on $y_t^{k,\tau}(i,l)$ will assure that this definition is satisfied:

$$-M \cdot \Big(1 - y_t^{k,\tau}(i,l)\Big) \leq DIP_t^{k,\tau}(i) - \Theta_t^{k,\tau} - l \tag{3.23}$$

$$DIP_t^{k,\tau}(i) - \Theta_t^{k,\tau} - l \leq M \cdot y_t^{k,\tau}(i,l). \tag{3.24}$$

By noticing that $y_t^{k,\tau}(i,0)$ is expressed linearly in Θ in constraints 3.23 and 3.24, we can now transform Eq. 3.22 into linear constraints with respect to our decision variables:

$$\beta_{t+1}^{k,\tau}(i) \geq DIP_t^{k,\tau}(i) - \Theta_t^{k,\tau} \tag{3.25}$$

$$\beta_{t+1}^{k,\tau}(i) \leq DIP_t^{k,\tau}(i) - \Theta_t^{k,\tau} + M \cdot \Big(1 - y_t^{k,\tau}(i,0)\Big) \tag{3.26}$$

As $\overline{I} \to +\infty$ and $m(i+1) - m(i) \to 0, \forall i \in I$, the following linear expression with respect to our decision variables Θ,

$$\sum_{i \in I} \beta_{t+1}^{k,\tau}(i) \Psi(i), \tag{3.27}$$

will converge to the daily mean carryover demand $\bar{B}_{t+1}^{k,\tau}$. This uses the concept of Riemann integration.

Theorem III.4. *As $\overline{I} \to +\infty$ and $m(i+1) - m(i) \to 0, \forall i \in I$, the following linear expression with respect to our decision variables Θ,*

$$\sum_{l=0}^{\infty} \left[2l \cdot \sum_{i \in I} y_t^{k,\tau}(i,l) \cdot \Psi(i) \right] - \sum_{(l_1,l_2) \in (\mathbb{Z}^+)^2} \sum_{(i_1,i_2) \in I^2} z_t^{k,\tau}(i_1,i_2,l_1,l_2) \cdot \Psi(i_1) \cdot \Psi(i_2),$$

will converge to the daily carryover demand variance $\tilde{B}_t^{k,\tau}$, where $z_t^{k,\tau}(i_1,i_2,l_1,l_2)$ satisfies the following constraints:

$$z_t^{k,\tau}(i_1,i_2,l_1,l_2) \leq y_t^{k,\tau}(i_1,l_1), \tag{3.28}$$

$$z_t^{k,\tau}(i_1,i_2,l_1,l_2) \leq y_t^{k,\tau}(i_2,l_2), \tag{3.29}$$

$$z_t^{k,\tau}(i_1,i_2,l_1,l_2) \geq y_t^{k,\tau}(i_1,l_1) + y_t^{k,\tau}(i_2,l_2) - 1. \tag{3.30}$$

Proof. We want to express the variance, $\tilde{B}_t^{k,\tau}$, of our carryover random variables linearly in our decision variables. First, note that we can show by recurrence that:

$$\mathbb{E}[(\beta_t^{k,\tau})^2] = \sum_{l=0}^{\infty} 2l \cdot \mathbb{P}\left(\beta_t^{k,\tau} > l\right). \tag{3.31}$$

Moreover, by using our previously defined binary variables $y_t^{k,\tau}(i,l)$ (see constraints 3.20 and 3.21), Eq. 3.31 can be rewritten as:

$$\mathbb{E}[(\beta_t^{k,\tau})^2] = \sum_{l=0}^{\infty} \left[2l \cdot \sum_{i \in I} y_t^{k,\tau}(i,l) \cdot \Psi(i)\right] \tag{3.32}$$

Next, we can also compute the mean of our carryover demand random variable as a linear combination of the CDF (see Eq. 3.27 for the standard way of computing the mean):

$$\begin{aligned}\mathbb{E}[\beta_t^{k,\tau}] &= \sum_{l=0}^{\infty} \mathbb{P}\left(\beta_t^{k,\tau} > l\right) \\ &= \sum_{l=0}^{\infty} \sum_{i \in I} y_t^{k,\tau}(i,l) \cdot \Psi(i) \end{aligned}\tag{3.33}$$

The variance of $\tilde{B}_t^{k,\tau}$, of our overflow random variables can now be expressed as:

$$\tilde{B}_t^{k,\tau} = \sum_{l=0}^{\infty} \left[2l \cdot \sum_{i \in I} y_t^{k,\tau}(i,l) \cdot \Psi(i)\right] - \left[\sum_{l=0}^{\infty} \sum_{i \in I} y_t^{k,\tau}(i,l) \cdot \Psi(i)\right]^2 \tag{3.34}$$

$$= \sum_{l=0}^{\infty} \left[2l \cdot \sum_{i \in I} y_t^{k,\tau}(i,l) \cdot \Psi(i)\right] - \sum_{(l_1,l_2) \in (\mathbb{Z}^+)^2} \sum_{(i_1,i_2) \in I^2} y_t^{k,\tau}(i_1,l_1) \cdot y_t^{k,\tau}(i_2,l_2) \cdot \Psi(i_1) \cdot \Psi(i_2) \tag{3.35}$$

Finally, replacing the binary products $y_t^{k,\tau}(i_1,l_1) \cdot y_t^{k,\tau}(i_2,l_2)$ by a binary variable $z_t^{k,\tau}(i_1,i_2,l_1,l_2)$ satisfying Eqs. 3.28-3.30, we have expressed the variance of the carryover random variable linearly in our decision variables (see Eqs. 3.23-3.26). □

We have shown the ability to derive the mean $\overline{B}_{t+1}^{k,\tau}$ and variance $\tilde{B}_{t+1}^{k,\tau}$ of the carryover random variables on day $t+1$ linearly in Θ. By induction, the DIP mean and variance will be linear in the template Θ for any $t \geq 0$. This result holds for any day $t \in T$ in our equilibrium planning horizon.

Discussion: Although, we were able to calculate the variance $\overline{B}_{t+1}^{k,\tau}$ of our carryover random variables linearly in our decision variables Θ, tractability issues arise when solving this model with commercial solvers such as CPLEX. This is mostly due to the number of binary variables $z_t^{k,\tau}(i_1, i_2, l_1, l_2)$ that need to be optimized. For example, considering only 10 grid points (which is already a small sample size of possible DIP realizations), and the possibility to have at maximum 500 patients carryover (which in practice, could go up to the 1,000) would lead to $10^2 \times 500^2 = 30,000,000$ binary variables each each patient specialty k, urgency level τ, and day of the week t.

Therefore, we rely on an approximation of the variance in order to make our problem tractable. To approximate the variance of carryover and admitted/fulfilled demand, we condition on each realization $DIP_t^{k,\tau}(i)$ of our DIP random variable at grid level i and divide the total DIP variance proportionally between the carryover and admissions. By using our binary variable $y_t^{k,\tau}(i,0)$ that equals 1 if-f $DIP_t^{k,\tau}(i) \geq \Theta_t^{k,\tau}$, and defining $\tilde{A}_t^{k,\tau}$ as the variance of the number of type (k,τ) patients scheduled on day t, we can linearly approximate our variances in a tractable way as follows:

$$\tilde{B}_t^{k,\tau} \approxeq \sum_{i \in I} y_t^{k,\tau}(i,0) \cdot \Psi(i) \tag{3.36}$$

$$\tilde{A}_t^{k,\tau} \approxeq \sum_{i \in I} (1 - y_t^{k,\tau}(i,0)) \cdot \Psi(i) \tag{3.37}$$

We now leverage the dynamics developed in this section to measure the critical metrics discussed in the introduction and previous sections, particularly (i) the access delay from initial appointment request to the earliest time that appointment can be feasibly scheduled, (ii) the offered resource workload/utilization, and (iii) the downstream demand blocking/overtime.

3.7.2 Access delay formulation

In this subsection, we linearly approximate $G_{t,n}^{k,\tau}$ with respect to our decision variables, Θ. Recall that:

$$G_{t,n}^{k,\tau} = \mathbb{E}\left[\frac{\left(D_t^{k,\tau} - \delta_{t,n}^{k,\tau}\right)^+}{D_t^{k,\tau}}\right]$$

First observe that the numerator of inside this expression can be expressed as the positive difference of two random variables: (i) the exogenous demand on day t, $D_t^{k,\tau}$ and (ii) the number of remaining slots left in our template up to day $t + TFAV_n^{k,\tau}$ after all demand prior to day t has been scheduled, $\delta_{t,n}^{k,\tau}$. We can then divide this positive difference by random variable (i) to get a ratio. Because the exogenous demand is independent of past demand/decisions, these two random variables are independent. Hence, we are dealing with two streams of random variables, that are independent. We can use our grid approximation just as before, while making sure we capture the convolution of the two random variables.

We condition $\delta_{t,n}^{k,\tau}$ on the event that the total demand is $DIP_{t\ominus 1}^{k,\tau}(i)$ on day $t-1$ (see Eq. 3.17), then we get:

$$\delta_{t,n}^{k,\tau}(i) = \left(\left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t\oplus l}^{k,\tau}\right) - \beta_t^{k,\tau}(i)\right)^+, \forall i \in I, \forall n \in N, \tag{3.38}$$

and define $\gamma_{t,n}^{k,\tau}(i,j)$ as the percentage of type τ patients requesting an appointment in service k on day t that exceed $TFAV_n^{k,\tau}$ days of waiting for their appointment given that (a) there are j type τ requests in service k on day t, and (b) there are $\delta_t^{k,\tau}(i)$ remaining type τ slots in the template for service k after all demand prior to day t has been scheduled. $\gamma_{t,n}^{k,\tau}(i,j)$ can be expressed as follows:

$$\gamma_{t,n}^{k,\tau}(i,j) = \frac{\left(j - \delta_{t,n}^{k,\tau}(i)\right)^+}{j} \cdot 100\%, \forall i \in I, \forall j \in J, \forall n \in N \tag{3.39}$$

$$= \left(1 - \frac{\delta_{t,n}^{k,\tau}(i)}{j}\right)^+ \cdot 100\%, \forall i \in I, \forall j \in J, \forall n \in N \tag{3.40}$$

where the set $J \subset \mathbb{N}$ represents the sample space of all the $D_t^{k,\tau}$ random variables (excluding the outcome equal to 0) but with $|J| < \infty$: $|J|$ will be chosen such that $\mathbb{P}(D_t^{k,\tau} = |J|) < \kappa$ where κ is very close to 0.

In order to have linear constraints, we need to alter equations 3.38 and 3.40. In the optimization, there is an incentive to keep $\gamma_{t,n}^{k,\tau}(i,j)$ small to meet the access delay constraints (or because we are minimizing it in our objective). This allows us to replace Eq. 3.40 by constraints 3.41.

$$\gamma_{t,n}^{k,\tau}(i,j) \geq \left(1 - \frac{\delta_{t,n}^{k,\tau}(i)}{j}\right) \cdot 100\%, \gamma_{t,n}^{k,\tau}(i,j) \geq 0, \forall i \in I, \forall j \in I, \forall n \in N \tag{3.41}$$

The same cannot be said for Eq. 3.38, since the optimization has the incentive to increase $\delta_{t,n}^{k,\tau}(i)$ in order to get a smaller $\gamma_{t,n}^{k,\tau}(i,j)$ that will meet the access delay constraints. Therefore, we introduce a binary variable $x_{t,n}^{k,\tau}(i)$ that equals 1 when $\left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t\oplus l}^{k,\tau}\right) - \beta_t^{k,\tau}(i) \geq 0$ and equals 0 otherwise. The following constraints will assure $x_{t,n}^{k,\tau}(i)$ will take the correct values:

$$-M \cdot (1 - x_{t,n}^{k,\tau}(i)) \leq \left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t\oplus l}^{k,\tau}\right) - \beta_t^{k,\tau}(i), \forall i \in I, \forall n \in N \tag{3.42}$$

$$M \cdot x_{t,n}^{k,\tau}(i) \geq \left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t\oplus l}^{k,\tau}\right) - \beta_t^{k,\tau}(i), \forall i \in I, \forall n \in N \tag{3.43}$$

With the help of those binary variables, we can now have a set of linear inequalities equivalent to Eq. 3.38:

$$\delta_{t,n}^{k,\tau}(i) \geq \left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t\oplus l}^{k,\tau}\right) - \beta_t^{k,\tau}(i), \forall i \in I, \forall n \in N \tag{3.44}$$

$$\delta_{t,n}^{k,\tau}(i) \leq \left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t\oplus l}^{k,\tau}\right) - \beta_t^{k,\tau}(i) + M \cdot (1 - x_{t,n}^{k,\tau}(i)), \forall i \in I, \forall n \in N \tag{3.45}$$

$$\delta_{t,n}^{k,\tau}(i) \leq M \cdot x_{t,n}^{k,\tau}(i), \forall i \in I, \forall n \in N \tag{3.46}$$

We can then employ our grid approximation, but this time capturing the joint distribution between $D_t^{k,\tau}$ and $\delta_{t,n}^{k,\tau}$ (which is simplified by their independence). Because the approximation still takes the form of Riemann integration, as above, the result will converge to $G_{t,n}^{k,\tau}$ in the limit:

$$G_{t,n}^{k,\tau} = \sum_{i \in I} \sum_{j \in J} \gamma_{t,n}^{k,\tau}(i,j) \cdot \Psi_t^{k,\tau}(i) \cdot F_t^{k,\tau}(j), \forall n \in N \tag{3.47}$$

where $F_t^{k,\tau}(j)$ is the probability that there are j exogenous type τ requests for service k on day t, based on historical data.

3.7.3 Workload excess and overtime

From Eqs. 3.4, 3.7, and 3.27, the linear expression (subject to Θ) of the mean number of type τ patients scheduled for a first visit in service k, $\bar{A}_t^{k,\tau}$, follows immediately in Eq. 3.48.

$$\bar{A}_t^{k,\tau} = \overline{DIP}_t^{k,\tau} - \bar{B}_{t\oplus 1}^{k,\tau} = \overline{D}_t^{k,\tau} + \bar{B}_t^{k,\tau} - \bar{B}_{t\oplus 1}^{k,\tau} \tag{3.48}$$

This linear formulation of the daily mean number of patients scheduled can now be replaced in Theorems III.1 and III.2, and our expression of the mean workload (Theorem III.1) is now shown to be linear in our decision variables. Similarly, the linear expression of $\tilde{A}_t^{k,\tau}$ (see Eq. 3.37) now incorporated into Theorem III.2 shows that the variance on the daily workload can be expressed linearly in our decision variables Θ as well.

Next we use the offered load, W_t^k, to approximate the number of patient hours that will have to be performed in overtime due to insufficient capacity at specialty k on day t. To calculate the offered workload random variable, we propose to adapt the discrete grid approximation from Section 3.6.3 to estimate the workload distribution at each specialty in the network under template Θ. We approximate the distribution of the offered workload W_t^k by a Normal distribution with means and variances expressed in Theorems III.1 and III.2. This approximation has been commonly used in healthcare settings to describe workloads and census distributions in a wards of a hospital. Based on the Normal distribution properties, the probability that our random variables will have realizations within one standard deviation above the workload mean (i.e., in the interval $[\overline{W}_t^k, \overline{W}_t^k + \sqrt{\tilde{W}_t^k})$ is 34.1 %. For the previously defined grid $m(i)$, the probability of being in the interval $[\overline{W}_t^k, \overline{W}_t^k + m(i) \cdot \sqrt{\tilde{W}_t^k}]$ is known from the standard normal CDF as $\Phi\left(m(i+1)\right) - \Phi\left(m(i)\right)$.

Using this approximation, we are able to capture $o_t^k(i)$, the number of downstream workload hours that are overflowed (e.g. processed in overtime) at a given grid point i on day t at service k:

$$o_t^k(i) = \left(\overline{W}_t^k + m(i) \cdot \sqrt{\tilde{W}_t^k} - C_t^k\right)^+ \tag{3.49}$$

$$= \left(\overline{W}_t^k + \frac{1}{2}m(i) \cdot \left(\frac{\tilde{W}_t^k}{\hat{W}_t^k} + \hat{W}_t^k\right) - C_t^k\right)^+, \tag{3.50}$$

where C_t^k is the total capacity of specialty k on day t, and $\hat{W}_t^k$ is the standard deviation historical estimate of the total workload on day t in specialty k. Eq. 3.50 follows from the one step Newton approximation detailed in Section 3.6.3.

Moreover, as $\overline{I} \to \infty$ and $m(i+1) - m(i) \to 0, \forall i \in I$, $\overline{O}_t^k$ defined as:

$$\overline{O}_t^k = \sum_{i \in I} o_t^k(i)\Big(\Phi\left(m(i+1)\right) - \Phi\left(m(i)\right)\Big) \tag{3.51}$$

will converge the expected overflow/overtime hours of the offered workload. Note that we can also limit the violation probability of exceeding service k's capacity on a given day by amount q_t^k. First, we select the smallest $i^* \in I$ such that $1 - \Phi(m(i^*+1)) \leq q_t^k$, then we can constrain the workload level at this grid point i^* to be less or equal than capacity:

$$\overline{W}_t^k + \frac{1}{2}m(i^*) \cdot \left(\frac{\tilde{W}_t^k}{\hat{W}_t^k} + \hat{W}_t^k\right) \leq C_t^k. \tag{3.52}$$

3.8 Numerical case study

The following case study illustrates the decision support that our methodology can provide. By optimizing the template, our model provides a managerially focused perspective of the operations and can support "what if" analysis. It can also be used to help implement tactical level changes in scheduling, but that is not the intended focus of this paper. The template process must be validated and show improvements over current scheduling policies.

In this section, we go through a possible exchange between our team and the care provider in the implementation process. This tool and our template recommendations need to be tuned to the managerial staff's specific needs while also providing a wide range of information that will help them in their own decision process of which template to use. In Section 3.8.1, we use the information provided by the managerial staff to recommend one specific template.

3.8.1 Generating a template to fit specific needs

First, we consider the case where the managerial staff knows exactly what performance level they wish to achieve, but are lacking the engineering tools to build such a template that would assure that all their constraints are met. To provide them with the template that would satisfy those constraints, our first step will be to ask them which objectives they want to optimize and which constraints they want to satisfy. Consider the following scenario:

- We wish to minimize the mean access delay for our urgent cases, but it is important that:
- The mean access delay for less urgent cases does not exceed 5 weeks, and
- Patient downstream visits in their itinerary are not performed in overtime too often:
  - The offered workload generated from the template does not exceed capacity more than with 10% probability, and
  - The expected number of downstream visits performed in overtime are less than 5 per day across the medical specialty

In addition to the above inputs, a second input to this model is the historical data that will calibrate the parameters of the Mixed Integer Program. This includes: (1) the cumulative distribution function of the various demand streams (for each patient class), (2) the downstream demand distribution that will characterize patient flow (see Section 3.5.1), (3) the total daily capacity for each medical specialty.

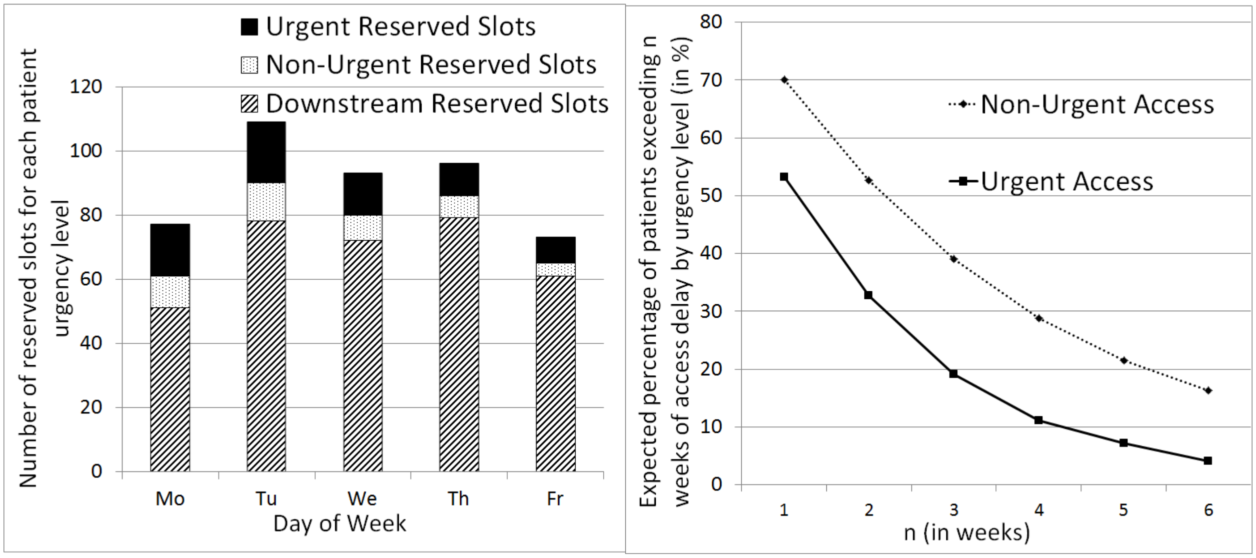


Figure 3.8: (Left) Optimal template for the two urgency classes considered and the downstream appointments of those patients' itinerary, and (Left) violation probability on access delay for the two urgency levels

With all the above inputs, we can run our model to determine a template that will satisfy the targets of the medical specialty. Note that in this case study, we only consider one department for a proof of concept. The results are provided in Fig. 3.8. After providing this template, one natural question that will arise from the managerial staff's side is: how accurate is your template?

3.8.2 How accurate is your template?

Before actual implementation of a template, every health care institution will require some analysis on how well this methodology predicts the critical metrics they are trying to optimize and/or control. Does the template really generate such performance if it was implemented in practice? Rather than the option of implementing the template first and monitoring the performance metrics to see if our forecasts match reality as it unfolds, it is more common to first use simulation in order to reassure care providers that the methodology is accurate enough that it deserves a pilot.

In Section 3.6.1, we detailed how we modeled the simulation used here for validation. The goal was to capture most of the elements in the scheduling process to ensure that the simulation would be a good representation of reality. After inputing the optimized template from Section 3.8.1, we use the simulation to compute the performance of the metrics that are critical to the institution.

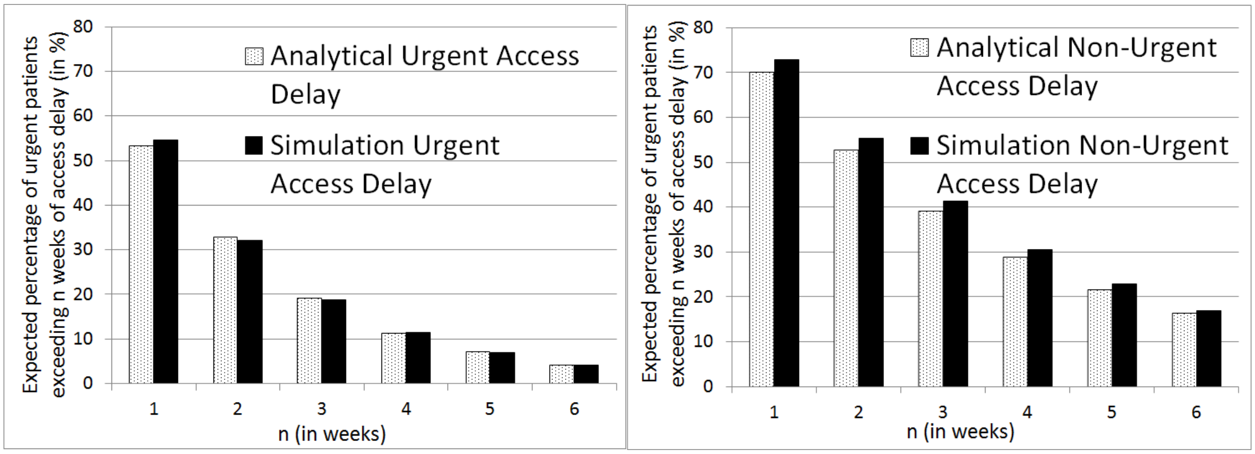


Figure 3.9: Analytical model and simulation comparison for the access delay of (Left) Urgent cases, and (Right) Non-urgent cases

n (in weeks)	1	2	3	4	5	6
Urgent Access Delay APE (%)	2.58	1.91	1.01	2.18	3.98	1.95
Non-Urgent Access Delay APE (%)	3.90	4.84	5.47	5.39	5.67	4.16

Table 3.5: Absolute Percentage error of the expected percentage of urgent and non-urgent patients exceeding n weeks of access delay

In Fig. 3.9 we show the differences in the access delay metric between our analytical estimations and the simulation. The absolute percentage errors of our estimations are also summarized in Table 3.5. We can see a slight raise in our APEs around 4 and 5 weeks access delay (around the mode of the access delay distribution) since most of the density lies around those points, and this results in increased errors if our density approximations are slightly off.

Similarly, Fig. 3.10 and Table 3.6 show the validation results for the mean and standard deviation of our workloads. Our estimations for mean and standard deviation of the workloads is very accurate mostly due to the predictability of those metrics. Based on historical data, patients tend to wait more than a day for an appointment, therefore resulting in a high probability of our variable $\alpha_t^{k,\tau}$, the daily number of patients scheduled for each type, being equal to our template $\Theta_t^{k,\tau}$. This also results in a low and predictable variance around $\alpha_t^{k,\tau}$. Our mean estimation of the workload is almost exact due to the predictability of patients filling up the slots we reserved for them. Our estimations of the workload standard deviation have more errors around them largely due to the approximation we had to make on variance to make our problem tractable.

Day of Week	Mo	Tu	We	Th	Fr
Mean Workload APE (%)	0.29	0.28	0.04	0.15	0.36
Workload St.Dev. APE (%)	3.13	2.4	4.01	3.57	1.49

Table 3.6: Absolute Percentage error of the expected percentage of urgent and non-urgent patients exceeding n weeks of access delay

If the health institution is satisfied with the accuracy of our forecasting approximations, the next question that might come up is: how would our template change if we were to vary the targets we imposed on our metrics? Can we investigate the tradeoffs between our metrics?

3.8.3 Can we investigate various control scenarios and their impact?

Effective sensitivity analysis is a critical part of meeting the advisory and managerial decision support goals. Based on our interactions with physicians and the managerial staff,

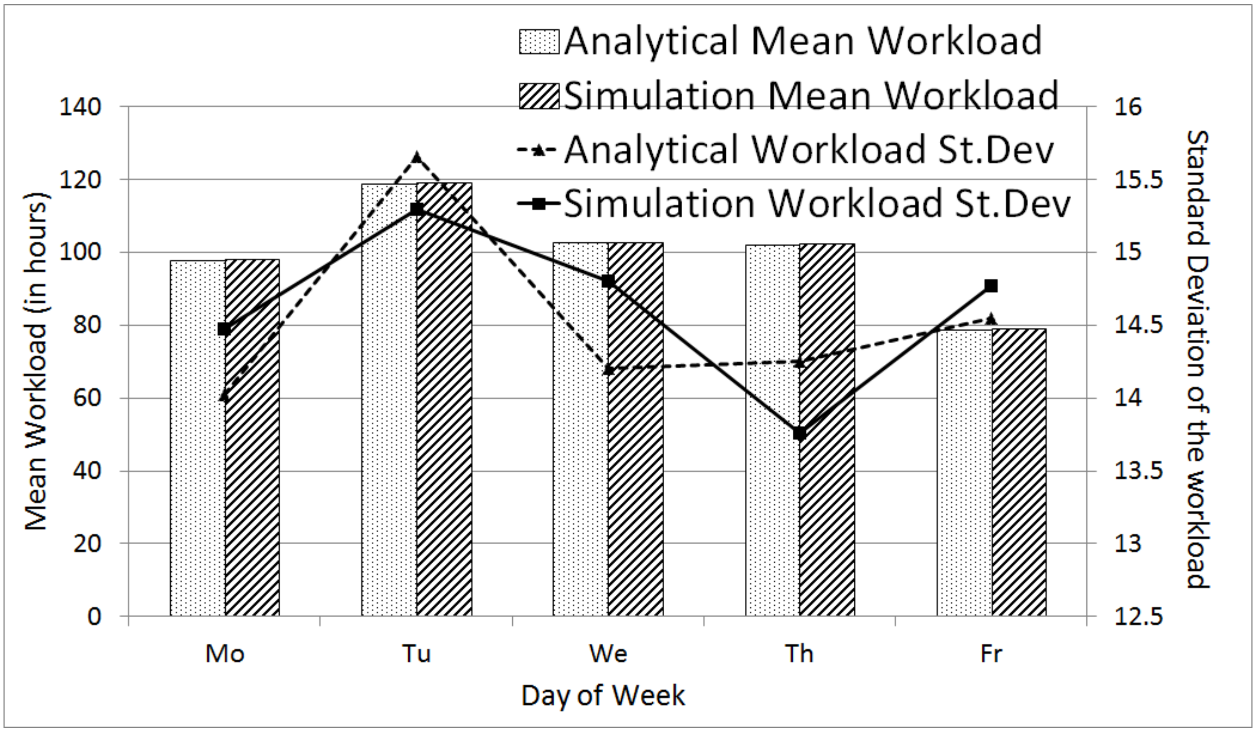


Figure 3.10: Analytical model and simulation comparison for the access delay of (Left) Urgent cases, and (Right) Non-urgent cases

it seems that this sensitivity analysis feature is the most important component of this new methodology. Rather than being blindly referred to one template, a health institution would rather analyze many different template scenarios and see their impact on the metrics. This allows their experience, and general knowledge of their department to be a valuable input to our tradeoff analysis component, and give them the control and information needed to make the "best" decision.

To give a very detailed example, we can identify how increasing the volume of new patients in internal medicine impacts physicians in neurology due to the cross-flow between them. It can deal with the sensitivity of how many more clinicians are needed to increase our throughput for breast cancer patients by some percentage. In particular, what-if questions for capacity management will, for the first time, be able to be addressed. For example: How many urgent cases can be accommodated in a week? What will the downstream impacts be?

First, we will show how we can achieve a higher level of control on patient access if needed. Then, we will give an example on the impact of varying constraints on one metric, and how it affects other metrics and the template.

Imposing constraints on the access distribution

After seeing the results of the first template we provided to the medical specialty (see Fig. 3.8), they realize that even though their initial constraints were met (for the mean access of the two patient urgency levels), it is unacceptable that more than 10% of their urgent patients have to wait longer than three weeks to get a root appointment. But the managerial staff realizes that achieving this new goal will negatively impact the non-urgent access levels. Hence, they would like to add another constraint assuring that no more than 20% of their non-urgent cases will have to wait longer than 5 weeks.

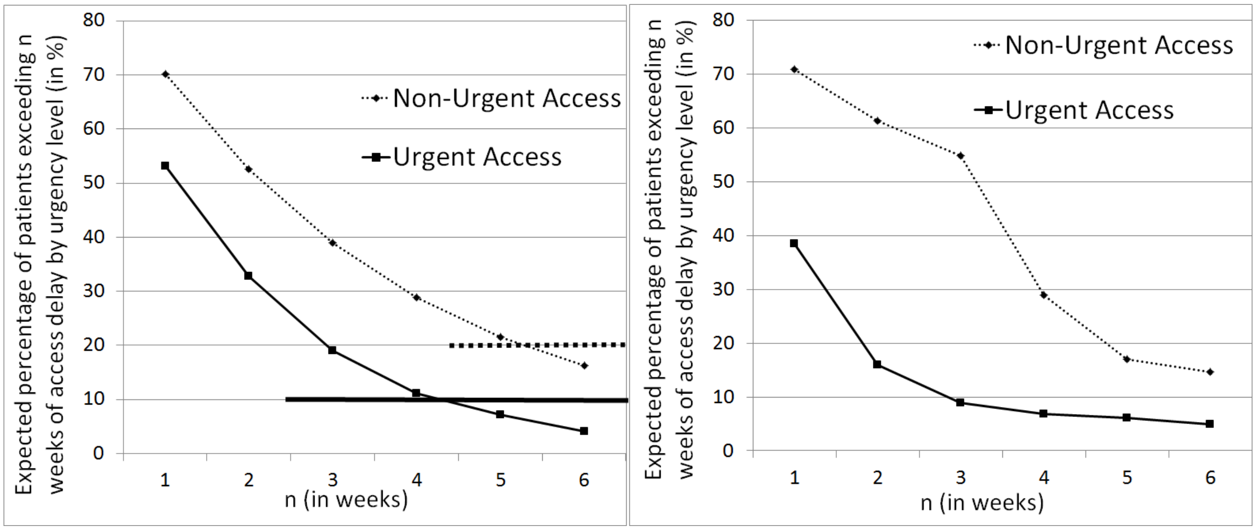


Figure 3.11: Impact on the access delay violation probability curves when two constraints (horizon lines in the left part of the figure) are imposed on the system. (Left) Access delay curves under the constraints defined in Section 3.8.1, and (Right) Access delay curves when we wish that (i) no more than 10% of the urgent patients wait longer than three weeks to get a root appointment, (ii) no more than 20% of the non-urgent patients wait longer than five weeks to get a root appointment

In Fig. 3.11, we investigate the impact on the access delay distribution of those two urgency levels when those additional constraints are imposed on the system: in the left part of the graph, we can see (1) the constraint that assures that no more than 10% of the urgent cases will exceed three weeks access delay for a root appointment, represented by the horizontal solid line, and (2) the constraint that assures that no more than 20% the non-urgent cases will exceed three weeks access delay for a root appointment, represented by the horizon dashed line. After running our optimization, the results are represented in the right part of Fig. 3.11. We first see how the access delay curve drops for the urgent cases that wait less than three weeks in order to meet those new targets. This significantly increases the number of non urgent cases that will wait longer than three weeks for a root appointment (from 40% to 54%). But we see a noticeable drop in the non-urgent access delay curve after three weeks: this is because our constraint on urgent access has been met, and the second constraint on non-urgent access delay starts to kick in.

This allows the health institution to have more control over their access delay metrics, and influence the shape of those violation probability access curves.

Varying targets for our constrained metrics

Another scenario is to investigate the impact on mean access for urgent cases when the constraint on the mean access for less urgent cases varies. The health institution may want to see the tradeoffs between those two metrics, and how it affects the template, in order to make the best control decision for their patients. As we discussed, the access delay metrics across various urgency patient types are competing metrics, and increasing the access for urgent patients will have negative impact on non-urgent access delays. In Fig. 3.12, we show how the optimized template changes when the constraint on mean access delay for non-urgent patients is increased from 5 weeks to 6 weeks.

The first thing to notice is that we see a slight overall decrease in slot reservations for

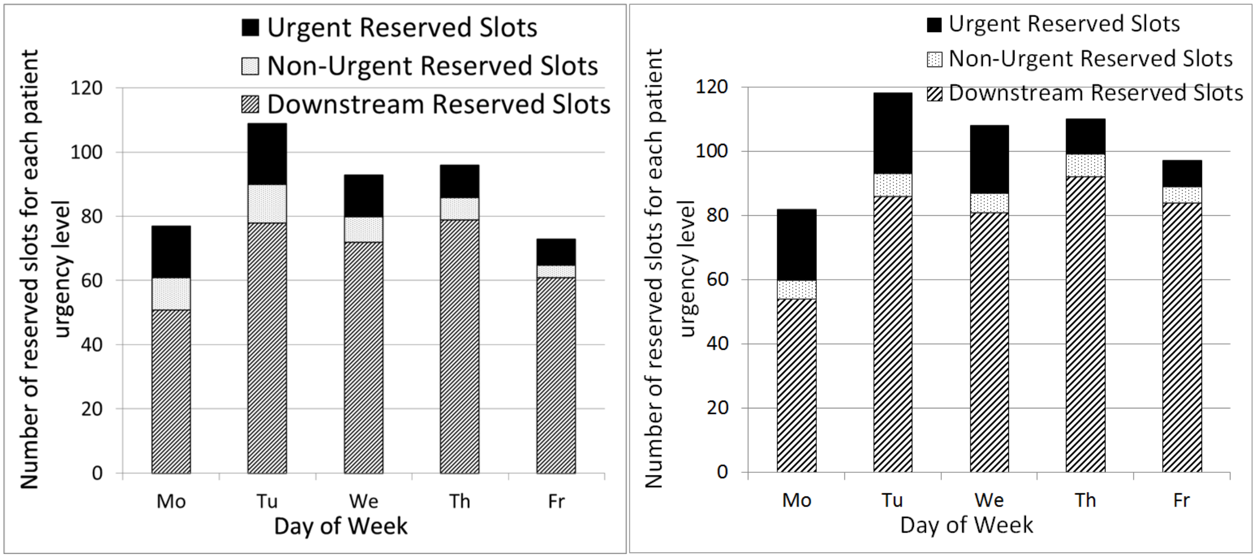


Figure 3.12: (Left) Optimized template under the constraints defined in Section 3.8.1, and (Right) Optimized template when the mean access delay constraint on non-urgent patients is increased to 6 weeks

less-urgent patients since we are have more flexibility of when to schedule them. Reducing the number of slots reserved for them will of course affect their access delay metric, and this is why we are still constraining the mean to be less than 6 weeks. Since the objective of our MIP is to maximize the mean access for urgent patients, we also see an increase in slot reservations for urgent cases. By reserving more slots for them, we will be able to achieve higher access. However, notice that we are also reserving more capacity for downstream appointments. This is largely due to the fact that urgent patients tend to use up more downstream resources than non-urgent cases (due to their condition and/or case complexity). We can see that the increase in capacity reservation for urgent cases (and the reduction of capacity reservation for non-urgent) towards the beginning of the week results in the need to reserve more capacity for downstream appointments towards the middle and end of the week. Note that we are using the same total capacity in this medical specialty under those two cases. Although the graph shows an increase in slot reservations (on the right side), this is largely due to the fact that downstream appointment slots tend to be much shorter than root appointments (for urgent and non-urgent patients).

This analysis is very useful for a health institution since intuition and experience are unable to understand the correlation between the number of urgent, non-urgent, and downstream reservations needed to satisfy a set of constraints on our metrics.

We can also investigate the impact on the access delay for urgent cases when the constraint on the probability of exceeding capacity varies or the constraint on overtime is changed. We omit those results here, but the same procedure could be applied.

Therefore, by being able to test different scenarios, and vary the constraints they impose on the system, the sensitivity analysis feature of this methodology could provide valuable information that could be used to make the right template decision in order to achieve the institution's needs.

3.9 Conclusions

Today's appointment based services such as healthcare have very limited research supporting effective control of waits for an appointment. In contrast to common-sense approaches to capacity plans for how departments or individual clinicians create a template plan for allocating their mix of services/procedure types by day of week, our research takes a fundamental operations-based approach. We model service itineraries and control access delays even with multi-visit itineraries of service. These problems naturally appear to be complex queueing network admission control and scheduling problems, but we model and optimize them using mixed integer programing, which solves relatively larger problems, utilize commercial solvers, and offer modeling capabilities that are useful for these problems. It will greatly increase the ability to manage complex tradeoffs involving (1) operational efficiency and cost metrics, and (2) offer integrated outpatient care with increased value. We emphasize the (2a) control of access delay times (the key being reduced access delays for urgent patients while accepting that some patient types may wait longer) and (2b) more rapidly executed care itineraries/pathways that complete the episode of care sooner. These objectives are not new to leading organizations; however, effective methods to accomplish them are in their infancy. This research and those methodologies will be a useful step toward gaining better managerial control over outpatient healthcare delivery and making it more effective at limiting the amount of access delay for various types of services or patients.

CHAPTER IV

Patient flow modeling and control for differentiated itinerary of care lengths

4.1 Introduction

Fast diagnosis and an early start to treatment is a critical aspect of care delivery that can impact the fate of many patients. Due to the rapid progression of life threating diseases (e.g., many types of cancer), the longer it takes to establish the diagnosis and determine the appropriate treatment plan, the higher the probability this could have irreversible repercussions on the patient's health state (see [35]). Poor patient access significantly affects the main stakeholders' experience. "Patients are harmed in the process of delay, not only through wasted time, but through unnecessary suffering, and through adverse medical outcomes. Health care providers are harmed through the added cost and reduced efficiency resulting from the complications of handling delayed patients" states [22]. Moreover, long waits for care has been shown to have adverse effects on patient satisfaction (see [13]). Most health care institutions are well aware of the need to improve patient access to their resources. As [43] points out, "Delays and restricted access are properties of poorly designed, costly systems. The same changes that reduce delays and increase access can also reduce costs." Health operations management has been shown to have the potential to significantly improve patients and providers' experience in health care systems while also reducing costs (e.g., see

[55]). But this usually requires a system redesign that mitigates the issues/limitations of common practices [48].

Within the scope of our research, the length of time it takes to get a diagnosis can be broken up into two periods of time: (1) the time from a primary care physician referral to a medical specialty root appointment, and (2) the length of time for a medical specialty to provide a clear diagnosis for the patient. After the root appointment (item (1)) for cases that are identified as "more urgent", in Chapter III, we now focus on the length of time it might take to establish a diagnosis (item 2).

In this chapter, we will forecast and control a metric that we call the "**itinerary flow time**." An itinerary refers to a set of diagnostic tests/visits and a follow-up visit the patient has to go through before a diagnosis is established and a treatment plan is determined (if needed). The time it takes from the conclusion of the patient's initial specialty visit to a diagnosis at the end of the follow-up appointment is the itinerary flow time. The itinerary can vary significantly based on the medical specialty, the patient condition (e.g., urgency level), and the information (or lack of information) we have on the patient. For example, a patient seeking a diagnosis in the breast diagnostic clinic will likely require more tests/visits than a patient in General Internal Medicine (GIM). Even within a medical specialty, a patient with a medical record that hints that it could be urgent (e.g., family history) will likely need more tests/visits than another. Moreover, patients with very little medical record information (e.g., new patients to the institution) will tend to require more testing and information building to get to a diagnosis.

The dependence of the itinerary flow time on the number of visits required by specific patient characteristics (e.g., urgency, medical specialty, etc...) is something we cannot control because this number of visits required is determined by medical need. Although we do address/consider this dependence in our models, we are much more concerned about another factor that affects itinerary flow time: the resource capacity management of the health institution's services. We will use the term "**access delay**" to a specific service/visit to capture the time it takes the patient to get an appointment in a downstream service. The above operational outcomes are the result of the institution's capacity management decisions.

Finally, we will also consider a metric that we call the "**diagnostic flow time**." This captures the time it takes for a patient to go through all the diagnostic tests/services required in her itinerary, leaving out the amount of time it will take to get a follow-up visit with the specialty physician to reveal a diagnosis and the next steps. A good estimate of this metric can add significant value for the specialty physician and the patient. Currently, at the time of the initial appointment, it is difficult for the specialty physician to determine when a follow-up visit should occur, given that diagnostic test results need to be in hand prior to it. They currently have to rely on experience and limited information. If the follow-up appointment is scheduled to a time before the patient is able to go through all the diagnostic services required then a reschedule needs to happen. This is undesirable for the specialty clinic because its an inefficient use of their resources (which might result in under-utilization). If the estimate for the follow-up visit is too early, this will result in further delay to reschedule the follow-up (since in the meantime, other patients will have been scheduled to use those same resources). But, being conservative and scheduling the follow-up visit too far into the future will result in unnecessary delays for the patient to determine their diagnosis. Our models will provide this valuable information which should improve service levels for the patient, and help the institution make better/more efficient use of their resources.

By controlling these *three* critical access metrics (itinerary flow time, diagnostic flow time, and access delay), our methodology is able to provide better service levels for the patients by reducing their wait time to a diagnosis (and possibly, start of treatment). However, we should note that we do not consider the following research question that might also have an impact on the level of service provided to patients: how can we limit the number of patient trips to the hospital during an itinerary of care (e.g., coordinate resources such that multiple diagnostic tests can be done on the same day)? Sometimes, it might be more desirable to have a condensed itinerary of care (e.g., all downstream visits happen 11 days from now) vs. visits that are spread out but lead to a shorter itinerary flow time (e.g., diagnostic visits happen on days 1, 4 and 7, and the follow-up visit on day 10). Considering this additional layer of service level adds significant complexity to this problem, especially since the preference between shorter itinerary flow times vs. condensed (few hospital visits) itineraries will vary case by case. Another practical concern can be the desire to avoid itineraries of care that extend over a weekend. Because many services shut down or reduce their capacity on weekends, it can be desirable to have itineraries complete in the same week in which they begin, an issue treated in [26].

In this research, we assume that getting to a follow-up visit (diagnosis) as soon as possible is the most important metric. We do not provide methods to limit the probability of requiring a weekend stay to complete the itinerary. The number of visits to the hospital, and the interval between them is only a secondary issue to a patient: rapid diagnosis, which enables an earlier start of treatment and better health outcomes is taken as paramount. To optimize the other issues mentioned above, it is likely important to incorporate the explicit coordination of visits between diagnostic services. Because our model is a planning model that provides a decision input that could be used by real time schedulers, our methods do not prevent the scheduler from taking such patient preferences in to account. Certainly, the scheduler may defer an appointment until such as time as two or more visits can be provided on the same day or even back to back during the day. The scheduler can delay the root appointment to the next available Monday or Tuesday to increase the chances of itinerary completion within a week. Note that our methods can control the access delay at each downstream service, which can serve as a proxy to limit how spread out the itinerary will actually be. However, we acknowledge that in this research, we will limit our scope to the three metrics defined in the previous paragraphs, and leave it to future research to analyze the complexities involved in patients' preferences of when each visit should happen in relation to others and the undesirability of itineraries that cross one or more weekends.

In Section 4.2, we identify the scope of our research and describe the high level characteristics of the system considered. Section 4.4 formulates the itinerary flow time, diagnostic flow time, and access delay random variables when downstream resource capacity is not differentiated by patient urgency levels. In other words, we consider the case (which is a close representation of current practice) where the total capacity at each service is available to any patient (as opposed to have reserved resource capacity for urgent cases). We then simulate patient flows through their resource needs and provide motivation to provide "fast track" access to urgent cases via optimization. In Section 4.5, we differentiate patient demand and capacity by urgency level and optimize this capacity reservation scheme in a Mixed Integer Program (MIP). We translate the nonlinear stochastic equations developed in Section 4.4 into a set of deterministic linear constraints subject to our decision variables. Finally, a numerical case study is presented in Section 4.6 to show the benefits of our Mixed Integer Program (Section 4.5) by comparing the results to those of the simulation in Section 4.4.

4.2 Problem description

The scope of this research is on a system optimization model for patients' episodes of care through the diagnostic process without regard to any particular disease. Our models operate at the daily level and focus on a time period that starts when a patient is first referred to a specialty service $k \in \mathcal{K}$ (e.g. Neurology, Medical Oncology, Internal Medicine, etc...) and ends when a diagnosis is determined (during the a follow-up visit with the specialty physician in medical specialty k). An example of the patient flow structure considered here can be seen in Fig. 4.1. Patients can either be referred internally to the specialty service needed, or could be referred from another hospital that doesn't have the necessary resources and expertise to diagnose the patient. This referral visit to a specialty service k is what we call a root appointment in a patient itinerary. The access delay from the referral request to this root appointment has been studied in Chapter III. It is rare that diagnostic tests are performed during this initial visit: the specialty physician usually meets the patient, reviews the medical record, asks questions, and informs the patient about the next steps to be taken.

The physician will then refer the patient to a set of diagnostic services that will provide more information on the patient's condition. The set of all diagnostic services s in the health care institution will be represented as $\mathcal{S}$ (e.g. blood draw, MRI, X-Ray, CT Scan, etc...). Every visit/service required after the root appointment (i.e. diagnostic service or follow-up specialty visit) will be referred to as *downstream services/appointments* and will belong to the set $\mathcal{S} \cup \mathcal{K}$. By using the model developed in Chapter III, we can estimate the distribution of the number of patients requiring those downstream services (see the $\alpha_t^{k,\tau}$ random variable for specialty k patients with urgency level τ on day t). The urgency level τ belongs to a discrete set $\mathcal{T}$ (e.g. $\tau = 1$ for urgent, $\tau = 2$ for non-urgent). In principle, a scalar τ is sufficient to include other dimensions of patient type such as new patients versus established patients, or national vs. local patients. Since demand for diagnostic services is generated after the patient's specialty k root visit, it is defined by the random variable $\alpha_t^{k,\tau}$ that represents the total number of specialty k, urgency level τ patients scheduled for a root appointment on day t. It will be modeled using the empirical distribution defined by our grid approximation of the set of outcomes $\{\alpha_t^{k,\tau}(i) : i \in I\}$ (see Chapter III). That is, $\mathbb{P}(\alpha_t^{k,\tau} = \alpha_t^{k,\tau}(i)) = \Psi_t^{k,\tau}(i), \forall i \in I$. If the Chapter III model is not used, one could use historical data on the number of root appointments scheduled in each specialty k to capture the downstream demand.

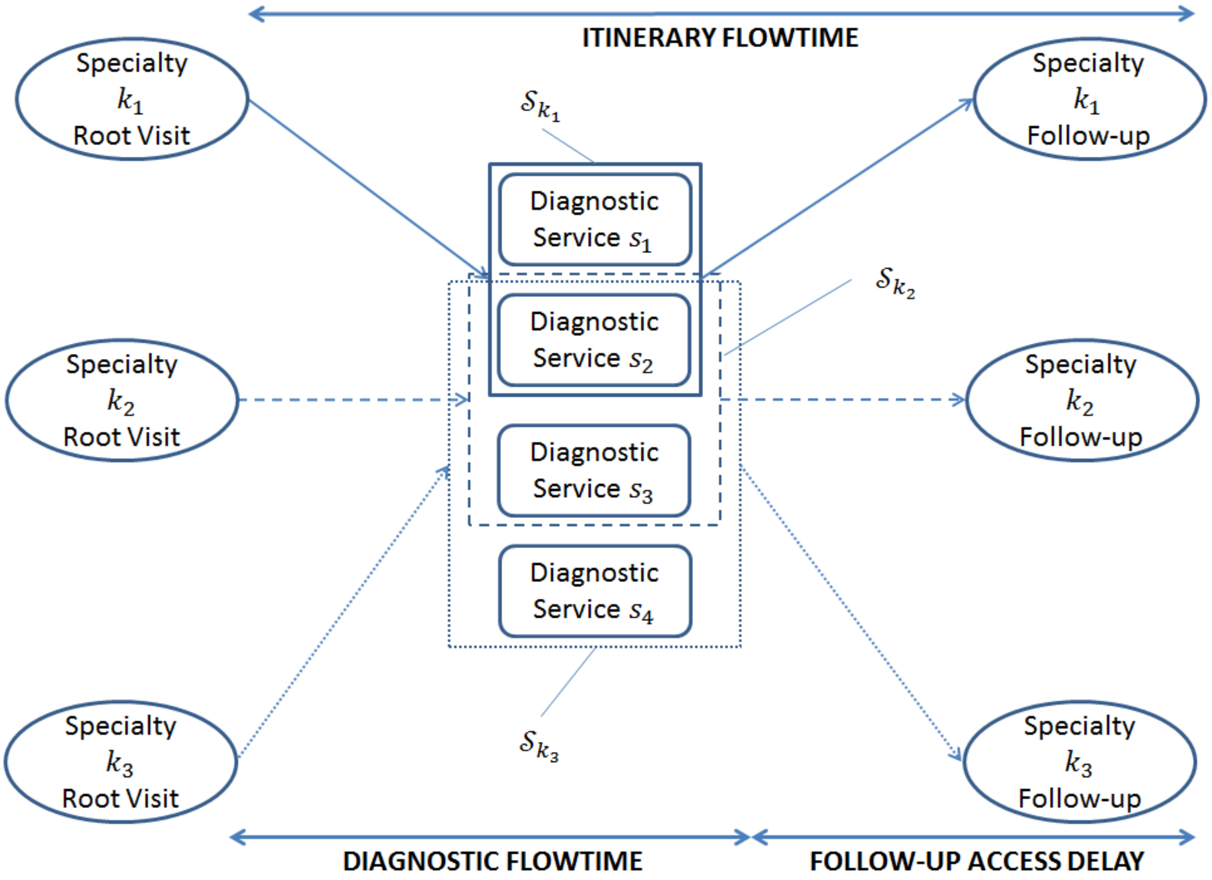


Figure 4.1: Visit precedence structure example in a patient's itinerary

The set of diagnostic services required can be captured probabilistically using historical data. Generally speaking, we can create sets $\mathcal{S}_k \subseteq \mathcal{S}, \forall k \in \mathcal{K}$ that represent the list of diagnostic services that specialty k patients have a nonzero probability of needing at some point in their itinerary. Based on our historical data analysis to identify the patients' resource requirements through the network, it is clear that not all patients of specialty k will require every diagnostic service in the set $\mathcal{S}_k$. We incorporate in our diagnostic flowtime formulation the fact that diagnostic services can be "skipped" with a given probability. This probability is calculated based on historical data as the random variable $\kappa_s^{\tau,k}$ that follows a Bernoulli r.v. with the probability of success determined by $p_s^{\tau,k}$ (see an example in Table 4.1). This random variable is assumed to be independent of everything else in our system, and also does not depend on our decision variables since it is based on historical data.

Specialty k /Diagnostic Service s	Blood Draw	X-Ray	CT Scan	MRI	ECG	EEG
Neurology	0.29	0.62	0.26	0.39	0	0.88
Medical Oncology	0.44	0.71	0.24	0	0	0.19
Cardiology	0.66	0.33	0.72	0.32	0.96	0
General Internal Medicine	0.92	0.16	0.12	0.24	0.26	0.29

Table 4.1: Probability $p_s^{\tau,k}$ that a specialty patient of urgency level τ requires this set of diagnostic services during their itinerary

The diagnostic services required are assumed to be performed in parallel without any precedence constraints (see Fig. 4.1). This assumes that all diagnostic testing appointments are independent of each other, and can be performed in any sequence. The parallel visit precedence structure for diagnostic services is considered a reasonable assumption by our partnering hospital. In future work, it may be possible to extend the methods developed here to more complex visit precedence structures for the diagnostic appointments portion of a patient's itinerary; however, we caution that the data requirements to create general visit precedence structures may be quite challenging.

Discussion: Note that in practice, it is possible that a diagnostic test is canceled dynamically after the results from another diagnostic service become available. Although we only consider static resource requirement distributions (e.g., Table 4.1), we are able to capture the probability that a test might not be needed after all, or "skipped" (based on historical data). Our methodology does already capture some of these dynamics but future research could be more precise by incorporating dynamic updating of the resource requirement distributions based on the outcomes of diagnostic tests. This is beyond our scope however, and this work considers only static resource requirements based on historical data.

Finally, once all the diagnostic services have been performed on the patient, a follow-up appointment will be necessary with the specialty physician to reassess the situation, determine a diagnosis and possibly start treatment. Note that the follow-up location is always consistent with the patient's medical specialty: a specialty k patient will always have a follow-up visit in medical specialty k. The scope of this research does not consider the treatment portion of a patient's itinerary of care, because it is often performed by a different set of resources (e.g., chemotherapy, radiotherapy, infusion, etc...). Here, we assume the patient will be scheduled for her follow-up visit as soon as all required diagnostic appointments have been performed.

It is important to note that those diagnostic services are usually not performed by physicians, but rather by nurses, technicians, and physician assistants. Therefore the capacity $\overline{C}_t^s$ of a diagnostic service $s \in \mathcal{S}$ is modeled as separate from the capacity at a given medical specialty. We define $\overline{C}_t^k, k \in \mathcal{K}$ as the total capacity available in specialty service $k \in \mathcal{K}$ on day t. We assume that, as explained in Chapter III, the capacity has been set aside for the root visits of all specialty patients and each urgency type: this is what we called our template $\Theta_t^{k,\tau}$. Then, for the downstream follow-up visits of their itinerary, we only need to consider the remaining capacity available at each specialty service $k \in \mathcal{K}$: $C_t^k = \overline{C}_t^k - \sum_\tau \Theta_t^{k,\tau}$. As mentioned above, this does not apply to the capacity of diagnostic services $s \in \mathcal{S}$.

4.3 Literature Review

Most of the literature in outpatient scheduling focuses on a single clinic, and does not consider the visit precedence in patient flows through their itinerary of care (see the survey paper [7]). Most research uses queueing theory approaches (e.g., [16]) or simulation (e.g., [6]). Queueing models are simpler to use, and have well known waiting time formulas. However, queueing theory has its limitations when we need to capture complex patient flows through a network of resources, or when we consider a system at a discrete daily level (as opposed to continuous time). In those cases, discrete event simulation is the most appropriate modeling tool.

In recent years, the allocation of outpatient resources has emerged as one of the most critical/challenging research topics in outpatient scheduling (see [56]). They state: "most researchers consider cost and revenue as their goals. However, from the customer perspective, patients' waiting time is the most important issue to consider in an outpatient scheduling system." [16] analyzes the correlations between access delays, resource workload/utilization and the number of servers in classic M/M/s queueing systems (with some extensions). They study how the number of servers affects waiting times, and how much capacity should be available to meet access targets (and consider flexible/time varying capacity by hour of day to meet demand patterns). Similarly to most of the published literature, waiting time is defined as the time from the arrival to the healthcare practice to the time of treatment (measured in minutes/hours). We consider the time from a request to the occurrence of the appointment, at the daily level.

To consider access delays at the daily level, we need to also consider the concept of blocking in queueing networks (e.g., ICU blocking [40]) and resource bottleneck analysis (see [37]). As opposed to the common idea of blocking (where demand is lost), we consider the "overflow" (demand in queue that exceeds the daily capacity) as carryover to the next day. This concept was used in outpatient practices, where patients required a number of service tasks (with various visit precedence structures) that had to be completed by the end of a work week for urgent cases (see [26]). They use phase type distributions to capture the length of an itinerary while approximating the blocking probability at each resource/visit.

Our model considers: (1) access delays measured in days (as opposed to minutes/hours within a day), (2) multi-class patient demand varying by day of week, (3) patients' probabilistically flowing through a network of resources, (4) a specific visit precedence structure including visits in parallel and in series, (5) resource/capacity reservation by patient class varying by day of week, (6) itinerary flow times through the stochastic queueing network, and (7) resource bottleneck analysis. See the more detailed literature review of Chapter III for additional background.

4.4 Itinerary flow time forecasting to approximate common practice

In this section, we will express the itinerary flow time random variable $\Xi_t^{\tau,k}$ for a specialty $k \in \mathcal{K}$ and urgency level $\tau \in \mathcal{T}$ patient starting her itinerary on day $t = 1, 2, ...$ Here, we forecast the itinerary flow times assuming that the total capacity within a downstream service $s \in \mathcal{S} \cup \mathcal{K}$ (i.e., diagnostic and specialty services) will be available to serve all patients, regardless of the different urgency level τ. This means that all urgency levels are treated the same, and requests for a diagnostic or specialty service will be processed in the order they were received. This simple/intuitive policy is very close to common practice, and to analyze its impact on itinerary flow times for various urgency level classes gives us the motivation to provide "fast track" access to higher urgency patients by reserving a portion of the total capacity exclusively for them at a given downstream service.

Based on the visit precedence structure described in the previous section, the itinerary flow time can be expressed as:

$$\Xi_t^{\tau,k} = \overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} + \chi_{t+1+\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k}}^{\tau,k} \tag{4.1}$$

where $\chi_t^{\tau,k}$ is the access delay of a patient in specialty service $k \in \mathcal{K}$ that requested an appointment on day t and $\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k}$ is the diagnostic flow time through the diagnostic services in the set $\mathcal{S}_k$ of a type k patient having her root appointment on day t. Notice that the itinerary flow time random variable of a specialty k, urgency level τ patient finishing her root appointment on day t will be the sum of (1) the diagnostic flow time of a specialty k, urgency level τ patient that starts diagnostic testing on day t, and (2) the access delay to specialty service k given this follow-up visit is requested on day $t + \overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} + 1$ (i.e., after the patient has go through all diagnostic services required). Here, we assume that a follow-up visit to specialty service k has to happen at least one day after the patient finished the last diagnostic service required. In practice, it is not common to have a follow-up visit with the specialty physician on the same day of the last diagnostic service required, since most of these services take some time to have the results processed and analyzed by the physician.

In the following subsections, we express the access delay and diagnostic flow time random variables in order to get the itinerary flow time random variable.

4.4.1 Access delay random variable for a downstream service

In order to calculate the itinerary flow time of patients in this stochastic queueing network, we first start by expressing the access delay that will occur at each diagnostic and specialty service. Given that a patient requires a downstream service on day t, how many days will she have to wait for an appointment? We extend the definition of the access delay random variable $\chi_t^{\tau,s}$ to all services (diagnostic and follow-up specialty) with $s \in \mathcal{S} \cup \mathcal{K}$.

The access delay in a downstream service $s \in \mathcal{S} \cup \mathcal{K}$ will be a function of the future available capacity at the time of the request. For the policy considered in this section, a day t request for a downstream service s will be scheduled on the first day in the future that has enough available capacity. Hence, it is important to keep track of the current availability in our schedule (for days t and later) by calculating the capacity remaining on each of those days after all earlier patient requests have been scheduled. In our model, the capacity at each downstream service can be measured in number of patients per day since appointments in each service are the same predictable length (e.g., a blood draw will always be scheduled in a 15 minute time slot, an MRI appointment will always be scheduled in a 45 min time slot).

Hence, at the time of the request on day t, one can get a snapshot of the capacity available on days $t' \geq t$ by considering the number of patient requests that came on or prior to day t and have been scheduled on or after day t. We call this random variable L_t^s, which can be thought of our downstream demand in progress similar to the DIP in Chapter III. Since all urgency levels τ are using the same capacity at service s and their appointment lengths are the same, there is no need to differentiate L_t^s by urgency level τ. But this notation will be updated to $L_t^{\tau,s}$ in Section 4.5 when we allocate capacity at each downstream service s based on the urgency level $\tau \in \mathcal{T}$. At the time of a patient request on day t, the amount of the available capacity for downstream service $s \in \mathcal{S} \cup \mathcal{K}$ on day $t' \geq t$ will then be:

$$\left(\sum_{m=0}^{t'} C_{t+m}^s - L_t^s + 1 \right)^+ \tag{4.2}$$

Discussion: The above expression of L_t^s tries to capture the number of patient requests that are "ahead" of a request that arrived on day t. But this variable only captures snapshots at the daily level, which makes it unclear how many requests arrived earlier in the day than the current request which we are forecasting the access delay. The random variable L_t^s captures the number of patients "in queue" at service s at the beginning of day t after all the day t demand has arrived. Hence, this model is conservative and places the patient request considered as the last one of the day, which will have the longest access delay out of all the day t arrivals. In practice, this seems like a reasonable way to forecast access delay since the patient could be told a conservative estimate of how long she is likely to wait for a service regardless of when the request is made at the hourly level on day t.

Now that we know the state of our available capacity for the following days, the access delay random variable of the last request of day t follows directly by identifying the first day that has non-zero capacity.

$$\chi_t^{\tau,s} = \arg\min_{l\geq 0}\left\{\left(\sum_{m=0}^{l} C_{t+m}^{s} - L_t^s + 1\right)^{+} > 0\right\} \tag{4.3}$$

$$= \arg\min_{l\geq 0}\left\{\sum_{m=0}^{l} C_{t+m}^{s} \geq L_t^s\right\} \tag{4.4}$$

Eq. 4.4 follows simply because L_t^s is an integer random variable. Note that the access delay forecasting provided in this research can be used on its own to identify the main bottlenecks in the system. With this information, each downstream service (diagnostic and follow-up specialty) has the ability to make the proper capacity expansion adjustments to mitigate the delays they are imposing on patient flow in the system as a whole.

Note that the access delay random variable for a given downstream service $s \in \mathcal{S} \cup \mathcal{K}$ does not depend on the urgency level $\tau \in \mathcal{T}$ in this policy since the capacity at each downstream service is pooled to service all urgency levels in order of request. We still consider the superscript τ here since this random variable will depend on the urgency level in the next section. In subsection 4.4.4, we will express the random variable L_t^s, which we call the downstream demand in progress, to fully characterize this access delay metric.

4.4.2 Diagnostic flow time random variable

We now focus on the diagnostic flowtime formulation $\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k}$, that captures for a specialty k, urgency level τ patient, the time from the root appointment of an itinerary to the end of the set of diagnostic services the patient has to perform. We incorporate the possibility that a specialty k patient will require multiple diagnostic visits after her initial appointment with the set $\mathcal{S}_k \subset \mathcal{S}$. Moreover, based on historical data, the need for a service $s \in \mathcal{S}_k$ is not deterministic and depends on the urgency level τ of the patient (e.g., more urgent patient usually require more diagnostic tests which leads to higher probabilities). We defined $\kappa_s^{\tau,k}$, $s \in \mathcal{S}_k$ as a Bernoulli random variable that equals 1 if and only if a specialty k, urgency level τ patient requires diagnostic service s in her care itinerary. Note that $\kappa_s^{\tau,k} = 0, \forall s \notin \mathcal{S}_k$ by definition of the set $\mathcal{S}_k$. Hence, the diagnostic flow time for an urgency level τ and specialty k patient finishing her root appointment on day t is expressed as:

$$\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} = \max_{s \in \mathcal{S}_k} \kappa_s^{\tau,k} \cdot \chi_t^s \tag{4.5}$$

$$= \max_{s \in \mathcal{S}_k} \left\{ \kappa_s^{\tau,k} \cdot \arg\min_{l \geq 0} \left\{ \sum_{m=0}^{l} C_{t+m}^s \geq L_t^s \right\} \right\}. \tag{4.6}$$

In Equation 4.5, we characterize the diagnostic flow time as the maximum of the access delays at each diagnostic service s that is required in the care itinerary of a type (τ, k) patient on day t. We assume here that all diagnostic requests are made on the same day t as the root appointment, which is common in practice since the specialty physician usually internally refers the patient to diagnostic services at the end of this initial visit. Multiplying the projected access delay at each diagnostic service $s \in \mathcal{S}_k$ by the Bernoulli random variable (which is binary) assures that we do not consider the delays that might occur in diagnostic services that are not required for this specific patient. Finally, 4.6 uses the access delay formulation established in 4.4.

As mentioned earlier, proper forecasting of the diagnostic flow time is critical in practice. By using the modeling tools developed here, the physician will then be able to schedule a follow-up appointment with his/her patient using this diagnostic flow time forecast. In current practice, the physician either (1) schedules the follow-up many weeks later with extra lead time to make sure time is left for the patient to complete all required diagnostic appointments, or (2) has to reschedule the follow-up appointment when the patient is delayed longer than expected at some diagnostic services. Case (1) is not ideal for the patient, especially for urgent ones. Typically, a patient will want to know the test results and get a diagnosis as soon as possible without having to wait longer than needed for a follow-up visit. Case (2) is very inefficient for the healthcare provider and usually wastes resources and time due to rescheduling mechanisms.

4.4.3 Itinerary flow time random variable

Although forecasting (and later, controlling) the access delay and diagnostic flow time metrics can have practical value for a medical practice (as mentioned in the previous subsections), the main metric of concern is arguably the itinerary flow time. This metric of course depends on the other two, but summarizes the true service level provided to a patient during her entire itinerary of care.

Based on Eqs. 4.1, 4.4, and 4.6, we formulate the itinerary flow time random variable of a specialty $k \in \mathcal{S}_k$ and urgency $\tau \in \mathcal{T}$ patient starting her itinerary on day t:

$$\begin{aligned}
\Xi_t^{\tau,k} &= \overline{\overline{\chi}}_t^{\mathcal{S}_k} + \sum_{n=0}^{\infty} \mathbb{1}\{\overline{\overline{\chi}}_t^{\mathcal{S}_k} = n\} \cdot \chi_{t+n+1}^{k} && (4.7)\\
&= \max_{s \in S_k} \left\{ \kappa_s^{\tau,k} \cdot \arg\min_{l \geq 0} \left\{ \sum_{m=0}^{l} C_{t+m}^{s} \geq L_t^s \right\} \right\}\\
&\quad + \sum_{n=0}^{\infty} \mathbb{1}\left\{ \max_{s \in S_k} \left\{ \kappa_s^{\tau,k} \cdot \arg\min_{l \geq 0} \left\{ \sum_{m=0}^{l} C_{t+m}^{s} \geq L_t^s \right\} \right\} = n \right\} \cdot \chi_{t+n+1}^{k} && (4.8)
\end{aligned}$$

In Eq. 4.7, we first use Eq. 4.1. To avoid having a random variable in a subscript, we consider every possible realization of the diagnostic flow time random variable $\overline{\overline{\chi}}_t^{\mathcal{S}_k} = n, n \geq 0$. Based on the actual realization of $\overline{\overline{\chi}}_t^{\mathcal{S}_k}$, we use the indicator variable to determine which day the follow-up visit in specialty k will be, and use our access delay formulation. Eq. 4.8 follows by replacing our expressions by Eqs. 4.4 and 4.6. We now have expressed the three random variables we wish to forecast/control as a function of the demand in progress L_t^s on day t at each downstream service $s \in \mathcal{S} \cup \mathcal{K}$. We will characterize the demand in progress random variables in the next subsection.

4.4.4 Demand in progress random variable at each downstream service

The demand in progress random variable L_t^s defined in subsection 4.4.1 captures the main dynamics of our system. It identifies the number of requests in downstream service

$s \in \mathcal{S} \cup \mathcal{K}$ that have arrived before or on day t but haven't been processed prior to day t. This quantity solely depends on the daily capacity at downstream service s and on the daily demand random variable for that service. If we let $D_{t,k}^{\tau,s}$ be the day t number of requests in downstream service $s \in \mathcal{S} \cup \mathcal{K}$ for type (τ, k) patients, we can formulate L_t^s recursively as:

$$L_{t+1}^s = (L_t^s - C_t^s)^+ + \sum_{k \in \mathcal{K}} \sum_{\tau \in \mathcal{T}} D_{t+1,k}^{\tau,s} \tag{4.9}$$

where we have the initial condition $L_0^s = 0$. In Eq. 4.9, the demand in progress on day $t+1$ is the sum of (1) the demand in progress from the previous day that could not be processed on day t due to capacity restrictions C_t^s, and (2) the new requests (over all specialty patients and all urgency levels) on day $t+1$ for a service k. (2) will depend on the number of patients that either completed their root appointment on day $t+1$ which produces new requests for diagnostic services $s \in \mathcal{S}$, or that completed their set of required diagnostic services on day t and will then request a follow-up appointment the next day in their specialty service $s \in \mathcal{K}$.

First, let's focus solely on this demand $D_{t,k}^{\tau,s}$ in the diagnostic services $s \in \mathcal{S}$. The type (τ, k) demand random variable $D_{t,k}^{\tau,s}$ in diagnostic service $s \in \mathcal{S}$ on day t can be expressed as:

$$D_{t,k}^{\tau,s} = \sum_{j=1}^{\alpha_t^{\tau,k}} \mathbb{1}\{s \in \mathcal{S}_k\} \cdot \kappa_s^{\tau,k} \tag{4.10}$$

where $\alpha_t^{\tau,k}$ was defined in Section 4.2 as the number of urgency level τ patients scheduled for their root appointment in specialty k on day t. Hence, each one of those patients will require diagnostic service s based on our Bernoulli random variable $\kappa_s^{\tau,k}$.

Now, to determine $D_{t,k}^{\tau,s}, s \in \mathcal{K}$, the number of urgency level τ, specialty k new requests on day t for a follow-up appointment in specialty service s (note that in this case s is always equal to k), we will need to use the diagnostic flow time random variable (see Eq. 4.6) as follows:

$$D_{t,k}^{\tau,s} = \sum_{l=0}^{t-1} \sum_{j=1}^{\alpha_{t-l-1}^{\tau,k}} \mathbb{1}\left\{ \overline{\overline{X}}_{t-l-1}^{\tau,\mathcal{S}_k} = l \right\} \tag{4.11}$$

$$= \sum_{l=0}^{t-1} \sum_{j=1}^{\alpha_{t-l-1}^{\tau,k}} \mathbb{1}\left\{ \max_{s \in S_k} \left\{ \kappa_s^{\tau,k} \cdot \arg\min_{l \geq 0} \left\{ \sum_{m=0}^{l} C_{t+m}^s \geq L_t^s \right\} \right\} = l \right\} \tag{4.12}$$

In Eq. 4.11, we consider all the patients $\alpha_{t-l-1}^{\tau,k}$ that started their itinerary on day $t-l-1$ and had a diagnostic flow time of l days. Hence, each of these patients considered will generate a follow-up demand for their specialty service k on day t. If we do this for every possible diagnostic flow time $l = 0, ..., t-1$ and link it to patients that started their itineraries on $t-l-1$, then we will have considered all the demand for a follow-up appointment on day t. In Eq. 4.12, we simply use Eq. 4.6 to replace the diagnostic flow time random variable.

In this section, we formulated (1) the demand in progress random variable at each downstream service $s \in \mathcal{S} \cup \mathcal{K}$. This allowed us to determine (2) the access delay random variable at each downstream service, which in turn helped us to express (3) the diagnostic flow time random variable for each specialty k, urgency level τ patient class. Finally, (2) and (3) determined (4) the itinerary flow time random variables for each specialty k, urgency level τ patient class. All metrics were expressed as a function of the demand in progress random variables L_t^s of (1), which we determined recursively as a function of the exogenous demand in the system.

4.4.5 Forecasting current metrics via simulation

Now that we have defined our three metrics and expressed them as a function of the basic demand random variables, we will focus on calculating the "historical" itinerary flow time. By historical, we mean that no differentiated access based on the urgency level τ is provided at each downstream service, and the capacity is available to be used by all urgency levels for each downstream service (as we assumed in the previous subsections). This will allow us to quantify common practice system performance, identify service bottlenecks, and the impact that this current policy has on itinerary flow times for urgent cases. Specifically, we are interested in observing the differences in itinerary flow times across patient urgency levels.

We constructed a simulation that reproduces patient demand for diagnostic testing, and mimics patients' flows through the diagnostic services and specialty follow-up visits according to the distributions described in this section. This simulation is at the daily level of granularity. The main inputs to our simulation will be (1) the daily demand distribution for downstream appointments, which is equivalent to the distribution of the daily number of patients scheduled for a root appointment in each specialty $k \in \mathcal{K}$, (2) the total daily capacity (in number of patients per day) at each diagnostic service $s \in \mathcal{S}$, (3) the remaining daily capacity (in number of patients per day) available for follow-up appointments at each specialty service $k \in \mathcal{K}$: the total capacity available at specialty service k minus the amount allocated to root appointments (see Chapter III)), and (4) each specialty k, urgency level τ patient diagnostic service requirement distribution (i.e., the sets $\mathcal{S}_k$ and the probabilities $p_s^{\tau,k}$).

With all this information, we can simulate probabilistic patient arrivals (i.e., the start of diagnostic testing) for specialty k and urgency level τ, and determine upon arrival their probabilistic diagnostic service needs according to their type (τ, k). The realization of patients' specific resource needs will route them to the queue of their required diagnostic services where a first come first served approach across all patient specialties and urgency levels will be used. In other words, the simplifying approximation to our daily load model exploits the unmodeled hours in each day by "cloning" the patients that require multiple diagnostic services, and send each clone to one of these services. On every day t, and at each diagnostic service $s \in \mathcal{S}$, the minimum of the demand in progress L_t^s and capacity C_t^s will be processed. For each patient, the simulation waits until every clone has been processed, and then generates a request for a follow-up appointment in the specialty service k determined by their type k (i.e. the specialty of their root appointment). Again, a first come first served approach is used there, and finally, the patient leaves the system.

The outputs of the simulation in equilibrium will be the critical metrics discussed in this section: (1) the itinerary flow time distribution and mean for each specialty k, urgency level τ patient, (2) the diagnostic flow time distribution and mean for each specialty k, urgency level τ patients, and (3) the access delay distribution and mean at each downstream service $s \in \mathcal{S} \cup \mathcal{K}$. We will also be interested in estimating the distributions, means, and standard deviations of the daily demand in progress random variables for each urgency level τ in the downstream services considered. This information will be used in Section 4.5.3 to feed iterations of our optimization model. Note that all our metrics are estimated in a cyclo-stationary model at the weekly level in equilibrium.

We compute our desired metrics over an evaluation horizon that reached equilibrium (e.g., after a warm-up period of 500 days), and use a "batch means output analysis" method to get the desired statistics. We also consider a number of independent replications (warm-up and evaluation horizons) in order to get more accurate estimates of the metrics' distributions. We selected the number of replications, the number of batches, and their sizes in a way that guarantees a 95% confidence interval on our sample metrics (means and variances) indicating sufficient accuracy that we don't need to put confidence intervals in our reports.

Based on preliminary analysis, we show the itinerary flow time distribution in Fig. 4.2 for a given patient specialty, and look at the distribution differences between non-urgent and urgent cases. On the left side of Fig. 4.2 we represent the probability mass function of the itinerary flow time: each point represents the probability (y-axis) that the patient's itinerary flow time will be a certain number of days (x-axis). On the right side of Fig. 4.2 we represent the cumulative distribution function of the itinerary flow time: by taking any point on the x-axis (which represents a given number of days), one could determine the probability that the patient has already completed her itinerary by that time. Note that the diagnostic flow time distributions can be computed similarly, and such graphs could provide the necessary information for the specialty physician on the cumulative probability that the patient has finished all required diagnostic services by a given day. Further, the access delay distribution at each downstream service (diagnostic or specialty) could also be represented with similar graphs to determine the bottlenecks in the current system, and understand what specific services are the major drivers of itinerary flow times.

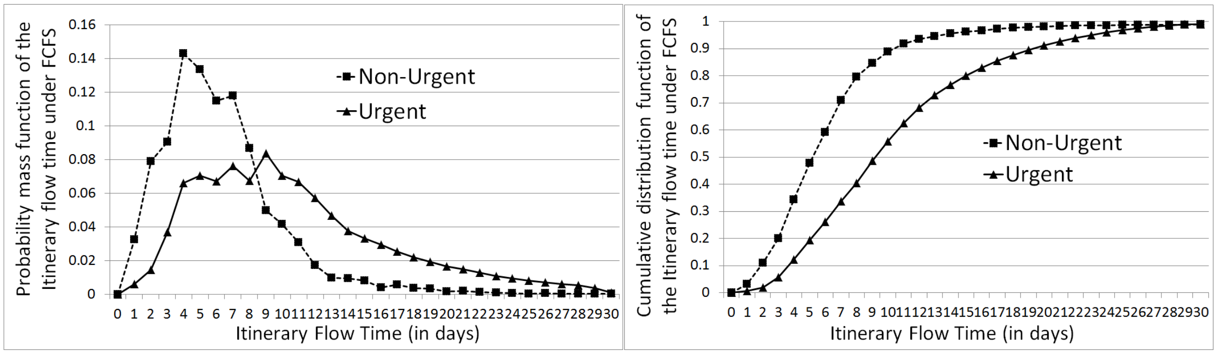


Figure 4.2: Quantification of FCFS scheduling via simulation to represent (Left) the probability mass function, and (Right) cumulative distribution function of the itinerary flow time random variable differentiated by patient urgency level

Based on these initial results, we see that urgent cases are usually more complex and require more services which affects their diagnostic and itinerary flow times significantly. By being treated the same as less urgent cases, we see slots being used by patients who could wait a few days longer without affecting their health, while urgent cases' flow times increase nonlinearly as a result.

This motivates the critical need to allocate capacity at each of the downstream services (diagnostic or specialty for follow-up) based on the urgency level of the patient. Exercising the control resulting from an efficient capacity management methodology will allow health care institutions to provide the shortest itinerary flow time possible for their urgent cases given a certain service level for other less urgent cases (e.g., the probability that itineraries of less urgent cases exceed two weeks has to be less than 10%).

4.5 Linearization for the control of itinerary flow time in a Mixed Integer Program

Similar to the template optimization in Chapter III, we now want to divide/allocate each service's capacity between the different patient urgency levels to provide differentiated access delays for diagnostic and specialty services. As we discussed in Section 4.4, the access delay at each service is going to be a major driver of the diagnostic flow time and itinerary flow time metrics. In this section, we will identify ways to capture the distributions of these random variables (see Section 4.4), their means, linearize them subject to our decision variables, and optimize them in a Mixed Integer Program.

4.5.1 Decision variables

Let the set $\{1, ..., T\}$ be a finite planning horizon (e.g. one week). The health care institution will reserve capacity for each patient urgency level at each of the downstream services (diagnostic and specialty for follow-up) over this planning horizon. The dynamics of our system will be analyzed over an equilibrium cyclo-stationary planning horizon of T days. This is a very convenient approach for purposes of this paper. On the other hand, in practice one could update the cyclo-stationary model periodically or one could even employ a time-varying planning model and a rolling planning horizon. Since demand patterns and capacity are usually similar from week to week, this model assumes $T = 5$ to represent one week.

We define the **downstream admission plan** decision variables $\Omega := (\Omega_t^{s,\tau}, s \in \mathcal{S} \cup \mathcal{K}), \tau \in \mathcal{T}, t \in \{1, 2, ..., T\}$ as the number of patient slots in downstream s allocated specifically to the urgency level τ during diagnostic testing and follow-up appointments. We do not reserve capacity in each diagnostic service $s \in \mathcal{S}$ based on the patient's specialty $k \in \mathcal{K}$ to keep a fair access to all diagnostic resources. Rather, the focus here is to provide "fast track" access to downstream resources for the more urgent patients (rather than basing this on specialty/condition). But each specialty will have different downstream resource requirements so it is important to model their care paths/resource needs individually for an accurate representation of patient flows.

The total capacity at each diagnostic service $s \in \mathcal{S}$ will be optimally divided across urgency levels to meet the constraints and optimize the objective set by the health institution. Therefore, $C_t^s = \sum_{\tau \in \mathcal{T}} \Omega_t^{s,\tau}, \forall s \in \mathcal{S}$. As explained earlier, we assume that capacity for specialty services was already set aside for root appointments (in Chapter III). The remaining capacity available for follow-up visits for specialty $k \in \mathcal{K}$ can also be divided across urgency levels: $C_t^k = \sum_{\tau \in \mathcal{T}} \Omega_t^{k,\tau}$.

4.5.2 High level description of the Mixed Integer Program

The goal of this methodology is to analyze tradeoffs between type τ patients based on our scheduling template $\Omega_t^{s,\tau}$ for each one of the following metrics:

- (M1) The "violation" probability that the itinerary flow time exceeds l days for the typical specialty k, urgency level τ patient that started her itinerary on day t.
- (M2) The mean itinerary flow time of a specialty k, urgency level τ patient given the itinerary started on day t.
- (M3) The "violation" probability that the diagnostic flow time exceeds l days for the typical specialty k urgency level τ patient that started her itinerary on day t.
- (M4) The mean diagnostic flow time of a specialty k, urgency level τ patient given the

itinerary started on day t.

- (M5) The "violation" probability that the access delay exceeds l days in downstream service s for an urgency level τ patient that started her itinerary on day t.
- (M6) The mean access delay in downstream service s for an urgency level τ patient given the itinerary was started on day t.

Multi-objective optimization has been shown to be difficult to implement in practice. For example, the weighted method could be used here, but determining the desired weights while understanding the relationships between all metrics is difficult for a health institution. Therefore, we will use one metric as an objective with constraints on the others. This methodology has value for a health institution since it is flexible and can provide answers to many "what if" scenarios based on various constraint and objective choices. For example, we could minimize M1 for urgent patients with a constraint on M2, M4, and M6 for less urgent cases. The main output from the optimization will be a weekly schedule $\Omega_t^{s,\tau}$. If used for a long period of time, this schedule will guarantee that the optimal objective is achieved while satisfying the constraints on the other metrics.

Based on the previous section, we can formulate the itinerary flowtime distribution differentiated by the urgency level τ as a function $f_t^{\tau,k}$ of our capacity reservation plan $\Omega := (\Omega_t^{s,\tau}, s \in \mathcal{S} \cup \mathcal{K}), \tau \in \mathcal{T}, t \in \{1, 2, ..., T\}$:

The formulations in Section 4.4 of the random variables that determine metrics M1-M6 were non-linear with respect to the capacity C_t^s, hence are non-linear subject to our capacity reservation plan Ω. We will show how to express M1-M6 linearly in our decision variables Ω to incorporate those metrics in a deterministic Mixed Integer Program as an objective or constraint. Moreover, we will need to translate these stochastic system dynamics into a set deterministic constraints that can be incorporated in our MIP.

4.5.3 Deterministic and linear approximation of demand in progress

Our ultimate goal is to translate all the non-linear (subject to Ω) stochastic equations expressed in Section 4.4 into a set of deterministic linear constraints. As was discussed in Section 4.4, the randomness in our metrics is exclusively due to the stochasticity of the daily demand in progress variables $L_t^s, s \in \mathcal{S} \cup \mathcal{K}$ at each downstream service s. Note that in our controlled system where capacity is specifically allocated to urgency levels, the demand in progress now has to be defined as a function of the urgency level $\tau\mathcal{T}$: we update our notation to $L_t^{\tau,s}, s \in \mathcal{S} \cup \mathcal{K}, \tau \in \mathcal{T}$. Hence, the first step is to transform the stochastic demand in progress dynamics into deterministic and linear constraints while still capturing this randomness in the system dynamics.

First, using the same concepts as Eq. 4.9, we formulate the stochastic demand in progress recursive equations for each downstream service $s \in \mathcal{S} \cup \mathcal{K}$ under a controlled system where capacity is differentiated by urgency level τ:

$$L_{t+1}^{\tau,s} = (L_t^{\tau,s} - \Omega_t^{\tau,s})^+ + \sum_{k \in \mathcal{K}} D_{t\oplus 1,s}^{\tau,k} \tag{4.13}$$

where the operator $\oplus$ is the modulo T operator to assure that all our variables will have their subscripts in our cyclo-stationary planning horizon $\{1, ..., T\}$ (e.g., if $t = T$ then $t \oplus 1 = 1$). Notice now that the capacity available to serve the urgency level τ demand in progress $L_t^{\tau,s}$ is limited to our decision variables $\Omega_t^{\tau,s}$. To translate these recursive equations into a deterministic set of constraints, an estimate of the demand in progress distribution is necessary.

Capturing the distribution of the demand in progress random variable analytically can be quite challenging, and is non-linear in our decision variables $\Omega_t^{\tau,s}$. Based on the simulation analysis in Section 4.4.5, we can get a good estimation of the equilibrium distribution of the demand in progress for each urgency level in an uncontrolled system where capacity is not allocated based on urgency levels τ. However, by optimizing a capacity reservation plan Ω based on urgency levels, the type τ demand in progress distribution will not be the same as in an uncontrolled system. Therefore, we will rely on a heuristic developed in Chapter III that allows us to use information provided via simulation on the demand in progress distribution and adjust/scale it to the new schedule controlled by Ω.

First, we construct a discrete grid $m(i), i \in I$ taking values in $(-\infty, \infty)$ similarly to Table 3.4. In the simulation, we extract the probability $\phi_t^{\tau,s}(i)$ that the demand in progress (having mean $\mu_t^{k,\tau}$ and standard deviation $\sigma_t^{k,\tau}$) for patient specialty k, urgency level τ on day t of our planning horizon takes values between $\mu_t^{k,\tau} + m(i) \cdot \sigma_t^{k,\tau}$ and $\mu_t^{k,\tau} + m(i+1) \cdot \sigma_t^{k,\tau}$. We will assume that this uncontrolled demand in progress distribution can now be adjusted/scaled to characterize the demand in progress controlled by our decision variable Ω. Specifically, after calculating $\overline{L}_t^{\tau,s}$, the demand in progress mean under schedule Ω, and $\tilde{L}_t^{\tau,s}$ its variance under schedule Ω, we can define $L_t^{\tau,s}(i)$, the demand in progress at grid level i, as follows:

$$L_t^{\tau,s}(i) = \overline{L}_t^{\tau,s} + m(i) \cdot \sqrt{\tilde{L}_t^{\tau,s}}. \tag{4.14}$$

Our assumption above implies that the random variable $L_t^{\tau,s}$ under schedule Ω takes values between $L_t^{\tau,s}(i)$ and $L_t^{\tau,s}(i+1)$ with probability $\phi_t^{\tau,s}(i)$. We omit the linear formulations of $\overline{L}_t^{\tau,s}$ and $\sqrt{\tilde{L}_t^{\tau,s}}$ subject to our decision variables Ω in this chapter since this was developed and analyzed in Chapter III.

With this methodology, we can now capture the randomness in the demand in progress variables as well as their distribution, by considering $|I|$ realizations $L_t^{\tau,s}(i)$ (based on the discrete grid $m(i)$) and assigning a probability $\phi_t^{\tau,s}(i)$ to each one.

4.5.4 Access delay linearization

In this subsection, we focus on expressing the access delay distribution for a downstream service $s \in \mathcal{S} \cup \mathcal{K}$ linearly in our decision variables Ω. The access delay distribution can be controlled in our MIP to assure that the probability of exceeding l days of delay for urgency level τ patients at a service s will be lower than a set target (see metric M5). Although this control is at the micro level (only one specific service in an itinerary), it could be a good way to identify the service bottlenecks in the health system.

Moreover, this subsection's main goal is to lay down some ground work that can then be used in the following subsections to express the diagnostic and itinerary flow times linearly in Ω. Since the access delays at the diagnostic and specialty services are the main drivers of flow times in the system (see Section 4.4), we start by building our modeling formulations at this micro level (diagnostic and specialty service) to leverage this foundation in the diagnostic and itinerary flow time linear formulations.

The probability of having the access delay in service $s \in \mathcal{S} \cup \mathcal{K}$ exceeding l days for urgency level τ patients requesting a (diagnostic or specialty) service on day t is calculated as follows:

$$\begin{aligned}
\mathbb{P}\left(\chi_t^{\tau,s} > l\right) &= \mathbb{P}\left(\operatorname*{arg\,min}_{l' \geq 0}\left\{\sum_{m=0}^{l'} \Omega_{t \oplus m}^{\tau,s} \geq L_t^{\tau,s}\right\} > l\right) && (4.15)\\
&= \mathbb{P}\left(L_t^{\tau,s} > \sum_{m=0}^{l} \Omega_{t \oplus m}^{\tau,s}\right) && (4.16)\\
&= \sum_{i \in I} \mathbb{P}\left(L_t^{\tau,s} > \sum_{m=0}^{l} \Omega_{t \oplus m}^{\tau,s} | L_t^{\tau,s} = L_t^{\tau,s}(i)\right) \cdot \mathbb{P}\left(L_t^{\tau,s} = L_t^{\tau,s}(i)\right) && (4.17)\\
&= \sum_{i \in I} \mathbb{1}\left\{L_t^{\tau,s}(i) > \sum_{m=0}^{l} \Omega_{t \oplus m}^{\tau,s}\right\} \cdot \phi_t^{\tau,s}(i) && (4.18)\\
&= \sum_{i \in I} x_{t,l}^{\tau,s}(i) \cdot \phi_t^{\tau,s}(i) && (4.19)
\end{aligned}$$

where $x_{t,l}^{\tau,s}(i)$ is a binary helper variable that satisfies the following linear equations in our decision variables Ω, for a large integer M:

$$M \cdot x_{t,l}^{\tau,s}(i) \geq L_t^{\tau,s}(i) - \sum_{m=0}^{l} \Omega_{t \oplus m}^{\tau,s} \tag{4.20}$$

$$-M \cdot (1 - x_{t,l}^{\tau,s}(i)) \leq L_t^{\tau,s}(i) - \sum_{m=0}^{l} \Omega_{t \oplus m}^{\tau,s}. \tag{4.21}$$

In Eq. 4.15, we first use the definition of the access delay random variable provided in Eq. 4.4. The probability that the access delay random variable is greater than l days is equivalent to the probability that the type τ demand in progress at service s is greater than the total capacity allocated for urgency level τ patients from days t to $t+l$, as expressed in Eq. 4.16. In Eq. 4.17, we condition this probability on the realizations of the type τ demand in progress random variable which makes $L_t^{\tau,s}$ deterministic. Therefore this probability is now an indicator function equal to 1 if and only if the urgency level τ demand in progress realization $L_t^{\tau,s}(i)$ is greater than the total capacity allocated to urgency level τ patients from day t to $t+l$ (Eq. 4.18). Finally $x_{t,l}^{\tau,s}(i)$ can be interpreted as the conditional probability of exceeding l days of access delay for an urgency level τ patient who requests service s on day t given that the type τ demand in progress level is $L_t^{\tau,s}(i)$ on day t. This conditional probability is actually either equal to 0 or 1 since the access delay becomes deterministic. Eq. 4.20 and Eq. 4.21 assure that $x_{t,l}^{\tau,s}(i)$ takes the appropriate binary value based on $L_t^{\tau,s}(i)$. The above equations allow us to express and control the access delay distribution in our MIP since all constraints will be linear in our decision variables Ω.

Finally, we can also calculate the expected value for the access delay of an urgency level τ patient requesting an appointment in service $s \in \mathcal{S} \cup \mathcal{K}$ on day t (see metric M6) as follows:

$$\mathbb{E}\left[\chi_t^{\tau,s}\right] = \sum_{l=0}^{\infty} \mathbb{P}\left(\chi_t^{\tau,s} > l\right) \tag{4.22}$$

$$= \sum_{l=0}^{\infty} \sum_{i \in I} x_{t,l}^{\tau,s}(i) \cdot \phi_t^{\tau,s}(i) \tag{4.23}$$

where Eq. 4.22 follows from the definition of the expected value, and Eq. 4.23 replaces this probability by Eq. 4.19.

4.5.5 Diagnostic flow time linearization

The diagnostic flow time distribution will be a function of the access delays at the diagnostic services $s \in \mathcal{S}$. As can be seen in Eq. 4.5, we will have to take the maximum of multiple access delay distributions at the required diagnostic services. This subsection will focus on how to linearly compute this metric subject to our decision variables Ω while keeping our problem tractable (such that common optimization solvers like CPLEX will be able to solve our problem).

A specialty k, urgency level τ patient finishing her root appointment on day t of our planning horizon will have to go through each element of the set $\mathcal{S}_k$ with a probability $p_s^{\tau,k}$ of requiring diagnostic service $s \in \mathcal{S}_k$. Our methodology wants to capture the probability that the diagnostic flow time will exceed l days, which will give us the diagnostic flow time distribution. To do this, we would need binary variables that inform us under any given scenario (realizations of our random variables) if the diagnostic flow time exceeds the level l or not. This "excess probability" will depend on (1) the demand in progress realizations $L_t^{\tau,s}(i_s), i_s \in I$ at each diagnostic service $s \in \mathcal{S}_k$ that might be required for a specialty k patient, (2) the combination of diagnostic services $s \in \mathcal{S}_k$ that are required for the specific patient (i.e., the realizations of the $\kappa_s^{\tau,k} \in \{0,1\}, \forall s \in \mathcal{S}_k$), and (3) the integer violation level $l = 1, 2, ..., \bar{l}$, where $\bar{l}$ is chosen such that the probability of the diagnostic flow time exceeding $\bar{l}$ is close to 0. Hence, our model would need to consider $\mathcal{T} \cdot \mathcal{K} \cdot \bar{l} \cdot I^{|\mathcal{S}_k|^2}$ scenarios/binary variables in order to get an estimate on the diagnostic flow time distribution of a specialty k, urgency τ patient. In a mid-size health care institution, patients could have diagnostic flow times that last up to a month ($\bar{l} = 30$), since for example, MRIs have long waiting times. Being very conservative, let's say we only model 5 diagnostic services, two urgency levels and two specialty patient types. We need at least 10 grid levels $i \in I$ for a reasonable representation of the downstream demand in progress random variables $L_t^{\tau,s}(i)$. Then, even with a very small problem, we will require binary variables of size: $2 \times 2 \times 30 \times 15^{10^2} = 1.2 \times 10^{27}$. This is obviously a too large number of binary variables for CPLEX to handle. Hence, we will approximate the diagnostic flow time distribution in a tractable way.

We start by making the following assumption:

Assumption IV.1. *The demand in progress random variables across all diagnostic services s, $\{L_t^{\tau,s} : s \in \mathcal{S}\}$ are assumed to be mutually independent.*

Discussion: This assumption is good under heavy traffic. In such settings, there will be multi-day access delays to a root appointment in a medical specialty. Based on our industry partners' data and the results presented in Chapter III, the access delays for a root appointment in a specialty $k \in \mathcal{K}$ are rarely one day or less, and tend to be on the order of multiple weeks. This means that the demand in progress for a root appointment (see $DIP_t^{\tau,k}$ in Chapter III) will be large, which leads to a near 0 probability of having the template $\Theta_t^{\tau,k}$ (for root appointments) exceed the demand in progress level $DIP_t^{\tau,k}$. In other words, $\alpha_t^{\tau,k}$, the number of specialty k, urgency level τ patients scheduled for a root appointment on day t in our planning horizon will closely match the planned number of root appointment admissions $\Theta_t^{\tau,k}$ (with probability close to 1), therefore its variance will be close to 0. $\alpha_t^{\tau,k}$ also is the demand for downstream appointments (since they are generated at the time of the root appointment with the patient), and based on various practical/realistic scenarios analyzed through simulation, the demand $\alpha_t^{\tau,k}$ does indeed have very low variance. Hence, most of the variability in $L_t^{\tau,s}$ comes from the number and location of the clones generated for subsequence diagnostic visits (i.e., the variability around our Bernoulli random variable $\kappa_s^{\tau,k}$). The random variables $\{\kappa_s^{\tau,k} : s \in \mathcal{S}\}$ being mutually independent, and having $\alpha_t^{\tau,k}$

being close to deterministic allows us to make a good case for Assumption IV.1: $\{L_t^{\tau,s} : s \in \mathcal{S}\}$ are mutually independent across all diagnostic services.

Based on this assumption, we now introduce Lemma IV.2, which allows us to capture the diagnostic delay distribution linearly in our decision variables, Ω, for the simple case where the number of diagnostic services in the health institution is limited to two. The proofs of this lemma, and the following theorems and corollaries, can be found in 4.8.

Lemma IV.2. *Under Assumption IV.1 and assuming $\mathcal{S}_k = \{s_1, s_2\}$, the diagnostic flow time distribution for an urgency level τ, specialty k patient in a system with two diagnostic services is computed as:*

$$\mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right) = \sum_{i \in I} \sum_{j \in \{0,1\}} y_{t,l}^{(2),\tau,k}(i,j) \cdot \phi_t^{\tau,s_2}(i) \cdot \mathbb{P}\left(\kappa_{s_2}^{\tau,k} = j\right) \tag{4.24}$$

where $y_{t,l}^{(2),\tau,k}(i,j)$ can be expressed as a set of linear constraints subject to our decision variables Ω and is computed as follows:

$$y_{t,l}^{(2),\tau,k}(i,j) = \max\left\{p_{s_1}^{\tau,k} \cdot \sum_{i' \in I} x_{t,l}^{\tau,s_1}(i') \cdot \phi_t^{\tau,s_1}(i'), j \cdot x_{t,l}^{\tau,s_2}(i)\right\}, \tag{4.25}$$

and this distribution can be expressed as a set of linear constraints subject to our decision variables Ω.

Now that we have shown how to calculate the diagnostic flow time distribution in the simpler case where a health institution has only two possible diagnostic services, we generalize our methodology to a set of diagnostic services $\mathcal{S}_k, k \in \mathcal{S}$ that can be of any dimension for each specialty k patient. We can leverage the ideas, methods, and equations of Lemma IV.2 to the general case without increasing the dimensionality of our problem.

Theorem IV.3. *Under Assumption IV.1, the diagnostic flow time distribution for an urgency level τ, specialty k patient starting the itinerary on day t (Metric M3) is computed* *as:*

$$\mathbb{P}\left(\overline{\overline{X}}_t^{\tau,\mathcal{S}_k} > l\right) = \sum_{i\in I}\sum_{j\in\{0,1\}} y_{t,l}^{(|\mathcal{S}_k|),\tau,k}(i,j)\cdot\phi_t^{\tau,|\mathcal{S}_k|}(i)\cdot\mathbb{P}\left(\kappa_{|\mathcal{S}_k|}^{\tau,k}=j\right) \tag{4.26}$$

where $y_{t,l}^{(n),\tau,k}(i,j)$ *can be expressed as a set of linear constraints subject to our decision variables* Ω *and is computed recursively as follows* $\forall n = 3,...,|\mathcal{S}_k|$*:*

$$y_{t,l}^{(n),\tau,k}(i,j) = \max\left\{\sum_{i'\in I}\sum_{j'\in\{0,1\}} y_{t,l}^{(n-1),\tau,k}(i',j')\cdot\phi_t^{\tau,s_{n-1}}(i')\cdot\mathbb{P}\left(\kappa_{s_{n-1}}^{\tau,k}=j'\right), j\cdot x_{t,l}^{\tau,s_n}(i)\right\}, \tag{4.27}$$

with $y_{t,l}^{(2),\tau,k}(i,j)$ *satisfying Eq. 4.25 and where* $\mathcal{S}_k = \{s_1, s_2, ..., s_{|\mathcal{S}_k|}\}$.

From Theorem IV.3, we can formulate metric M4, the expected value for the diagnostic flow time of an urgency level τ and specialty k patient given the itinerary started on day t as follows:

$$\begin{aligned}
\mathbb{E}\left[\overline{\overline{X}}_t^{\tau,\mathcal{S}_k}\right] &= \sum_{l=0}^{\infty}\mathbb{P}\left(\overline{\overline{X}}_t^{\tau,\mathcal{S}_k} > l\right) && (4.28)\\
&= \sum_{l=0}^{\infty}\sum_{i\in I}\sum_{j\in\{0,1\}} y_{t,l}^{(|\mathcal{S}_k|),\tau,k}(i,j)\cdot\phi_t^{\tau,|\mathcal{S}_k|}(i)\cdot\mathbb{P}\left(\kappa_{|\mathcal{S}_k|}^{\tau,k}=j\right) && (4.29)
\end{aligned}$$

where Eq. 4.28 follows from the definition of the expected value, and Eq. 4.29 uses the result from Theorem IV.3 (see Eq. 4.26).

4.5.6 Itinerary flow time linearization

Finally, we will now formulate the itinerary flow time distribution linearly in our decision variables Ω. The methods developed in IV.5 will use results from Theorem IV.3, and the definition provided in Eq. 4.19. The key difficulty here is to formulate linearly the access delay for the follow-up visit in specialty service k given the realizations of the diagnostic flow time of the patient, which affects when the patient starts waiting for a follow-up appointment.

We will start with the following assumption.

Assumption IV.4. *Given diagnostic testing ends on day $t \oplus l_1$, the access delay random variable $\chi^{\tau,k}_{t\oplus l_1 \oplus 1}$ for a follow-up visit in specialty $k \in \mathcal{K}$ is independent of the diagnostic flow time random variable $\overline{\overline{\chi}}^{\tau,\mathcal{S}_k}_t$.*

Discussion The access delay for a follow-up visit will depend on the day of week the request for a follow-up will occur (see Eq. 4.1) since the request for a follow-up is made after all diagnostic tests have been made. We will consider this dependence. However, Assumption IV.4 claims that given we have this information, there will be very little correlation between the diagnostic flow time of a patient that ends diagnostic testing on $t \oplus l_1$ and the access delay from her request for a follow-up visit on day $t \oplus l_1 \oplus 1$ to the occurrence of that follow-up. The use of diagnostic resources is spread around a very large number of patient types in a clinical setting. Hence, the wait for say, an MRI diagnostic test, is unlikely to affect the wait for a follow-up visit in the GI department once all diagnostic tests are done. In practice, your wait in diagnostic resources is mostly due to other patient specialties who need the MRI, as opposed to patients of your same specialty.

Corollary IV.5. *Under Assumptions IV.1 and IV.4, the itinerary flow time distribution for an urgency level τ, specialty k patient starting the itinerary on day t (Metric M1) is computed as:*

$$\mathbb{P}\left(\Xi^{\tau,k}_t > l\right) = \mathbb{P}\left(\overline{\overline{\chi}}^{\tau,\mathcal{S}_k}_t > l\right) + \sum_{l_1=0}^{l} \sum_{i \in I} z^{\tau,k}_{t,l_1,l}(i) \cdot \phi^{\tau,k}_{t\oplus l_1 \oplus 1}(i) \tag{4.30}$$

where $z^{\tau,k}_{t,l_1,l}(i,j)$ can be expressed as a set of linear constraints subject to our decision variables Ω and is computed as follows:

$$\begin{aligned} z^{\tau,k}_{t,l_1,l}(i,j) = \min \Big\{ & \sum_{i' \in I} \sum_{j \in \{0,1\}} \left(y^{(|\mathcal{S}_k|),\tau,k}_{t,l-1}(i',j) - y^{(|\mathcal{S}_k|),\tau,k}_{t,l}(i',j) \right) \cdot \phi^{\tau,|\mathcal{S}_k|}_t(i') \cdot \mathbb{P}\left(\kappa^{\tau,k}_{|\mathcal{S}_k|} = j\right), \\ & x^{\tau,k}_{t\oplus l_1 \oplus 1, l-l_1}(i) \Big\}. \end{aligned} \tag{4.31}$$

From Corollary IV.5, we can formulate metric M2, the expected value for the itinerary flow time of an urgency level τ and specialty k patient given the itinerary started on day t as follows:

$$\mathbb{E}\left[\Xi_t^{\tau,\mathcal{S}_k}\right] = \sum_{l=0}^{\infty} \mathbb{P}\left(\Xi_t^{\tau,k} > l\right) \tag{4.32}$$

$$= \sum_{l=0}^{\infty} \mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right) + \sum_{l_1=0}^{l} \sum_{i\in I} z_{t,l_1,l}^{\tau,k}(i) \cdot \phi_{t\oplus l_1 \oplus 1}^{\tau,k}(i) \tag{4.33}$$

$$= \mathbb{E}\left[\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k}\right] + \sum_{l=0}^{\infty} \sum_{l_1=0}^{l} \sum_{i\in I} z_{t,l_1,l}^{\tau,k}(i) \cdot \phi_{t\oplus l_1 \oplus 1}^{\tau,k}(i) \tag{4.34}$$

where Eq. 4.32 follows from the definition of the expected value, and Eq. 4.33 uses the result from Theorem IV.5 (see Eq. 4.30). Finally, from Eq. 4.28, Eq. 4.34 follows.

4.6 Numerical case study

For validation and tradeoff analysis, we will calibrate/parameterize our model to a system that is a good representation of common practices by using historical data. We focus on only one medical specialty in our optimization for simplicity of exposition, but we still consider the demand of other specialty patients (in aggregate) for diagnostic resources. We will not optimize/control this aggregate demand and assume it will follow historical patterns. We consider the downstream demand distribution that is generated from the scheduling of root appointments (see the template methodology developed in Chapter III). The parameters of this downstream demand are summarized later in this section in Table 4.2.

4.6.1 Tradeoff Analysis

Mean Itinerary flow time tradeoff analysis

First we will consider an optimization formulation that minimizes the mean itinerary flow time of urgent patients with respect to some constraint on the mean itinerary flow time of non-urgent cases. In Fig. 4.3, we first see that with a very tight constraint on the mean itinerary flow time for non-urgent cases (e.g., consider 3 days on the x axis), the access delay for the follow-up visit, 7 days, is a big contribution to the large mean itinerary flow time for urgent cases which is around 15 days. Clearly, the specialty follow-up service will be the bottleneck resource. In this extreme scenario, most of the resource capacity at each downstream service has to be allocated to the non-urgent cases, since the probability of receiving service within a day at each downstream service has to be close to 1 (i.e., the non-urgent demand in progress exceeds the capacity allocated to them with very low probability). Urgent patients' diagnostic flow time will of course be affected by this, but not as much as their waiting time for the follow-up appointment. The reason is that urgent and non-urgent patients always use the same specialty service for a follow-up appointment (with probability 1) but, during the diagnostic portion of the itinerary, the routing probabilities do not send every patient to all diagnostic resources and, can send urgent cases to services not often used by non-urgent cases, which mitigates the congestion seen at the specialty services. As we move along the x axis of the graph, we see that the gap between diagnostic flow time and itinerary flow time diminishes. When we allow poor access to non-urgent cases (e.g., mean itinerary flow times of 10 days), we see the gap between the diagnostic flow time and itinerary flow time curves is very small, and most of patients' delay will be during the diagnostic testing portion of their itinerary since they usually require multiple visits in high demand diagnostic services (since the condition might be serious).

Moreover, in Fig. 4.3, we compare the itinerary flow times of the FCFS policy to our efficient frontier of solutions (when the constraint on mean itinerary flow time of non-urgent patients varies), we realize that we can improve in both itinerary flow time metrics (for urgent and non-urgent cases). The reason we are able to improve in both metrics is because resource requirements between the two classes of patients are not the same. Note that more value seems to be gained by allowing the mean itinerary flow time for non-urgent patients to

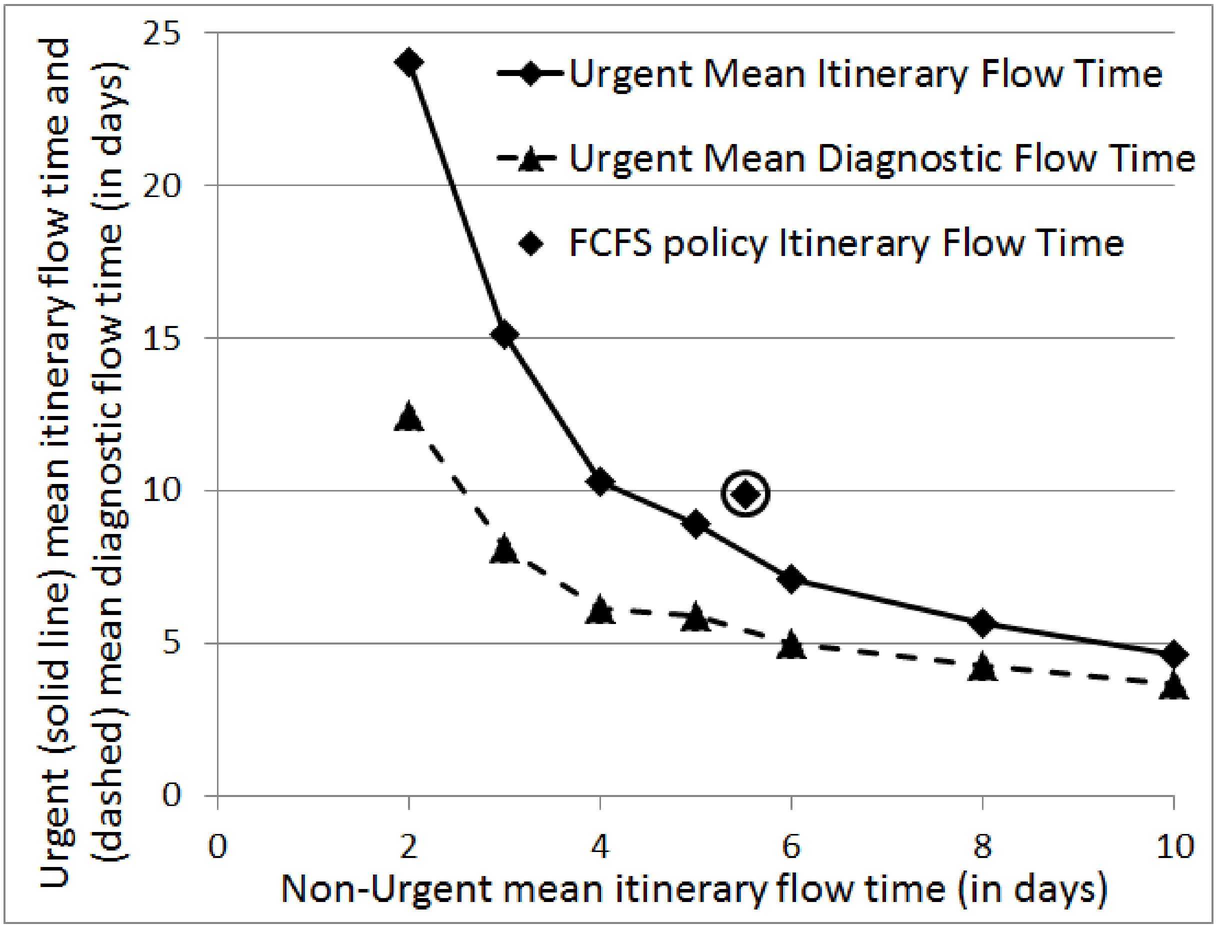


Figure 4.3: Tradeoff Analysis: optimal mean itinerary flow time and mean diagnostic flow time for urgent patients when varying the constraint on the non-urgent mean itinerary flow time, and comparison with the itinerary flow times of the First Come First Served (FCFS) policy (see Fig. 4.2)

be 6 days: this is only increasing it by 0.5 days on average from the FCFS policy, while the mean itinerary flow time of urgent patients is being reduced by 28% (from 9.9 to 7.1 days).

Itinerary flow time distribution tradeoff analysis

Although we were able to provide the desired control on the mean itinerary flow times, this may lead to some undesired behaviors in our system. For example, if we take a closer look at the itinerary flow time distributions associated with the optimized solution $(6, 7.1)$ in Fig. 4.3, we may not get the patient prioritization behavior that we were hoping for. In Fig.

4.4, we illustrate the itinerary flow time distributions (probability that patients complete their itinerary within a certain number of days) of our "controlled-mean" policy that result from optimizing a formulation that constrains only the means and not the tail probabilities of the itinerary flow times (this lead to the $(6, 7.1)$ solution in Fig. 4.3). What happens here is that the optimization will choose to schedule "enough" non-urgent patients quickly for short itinerary flow times (e.g., 2, 3, 4 days), and the rest of the non-urgent patient population (speaking of the right tail of the itinerary flow time distribution) has relatively large itinerary flow times (e.g., 15 or more days): specifically, close to 60% of non-urgent cases have an itinerary flow time of 5 days or less, but only 9% will have an itinerary flow time between 5 and 10 days. This increases the itinerary flow time variability, while still meeting the constraints we imposed on the mean (here, 6 days). At the same time, we see that most urgent cases obtain an itinerary flow time around the mean of 7.1 days: 66% of urgent cases will have an itinerary flow time between 6 and 8 days. This behavior might be undesirable if we wish to have more urgent patients with shorter itinerary flow times (e.g., a higher probability to complete an itinerary within 5 days), even if it means sacrificing a portion of the urgent (or non-urgent) population. Note that the behavior observed here is not guaranteed to happen in all cases, in fact, the exact opposite scenario (between urgent and non-urgent patients) could occur as well. Ultimately, we have very little control on our system if we consider only the mean of our metrics, i.e., the schedule generated by our optimization is quite unpredictable. In Fig. 4.4, we also compare these itinerary flow time distributions generated by the "controlled-mean" policy to the distributions rendered by the FCFS policy.

Based on the results/insights provided in Fig. 4.4, we may want to improve on our solutions by exercising a more specific control on our system. Most non-urgent patients don't necessarily need short itinerary flow times. Controlling the mean itinerary flow time

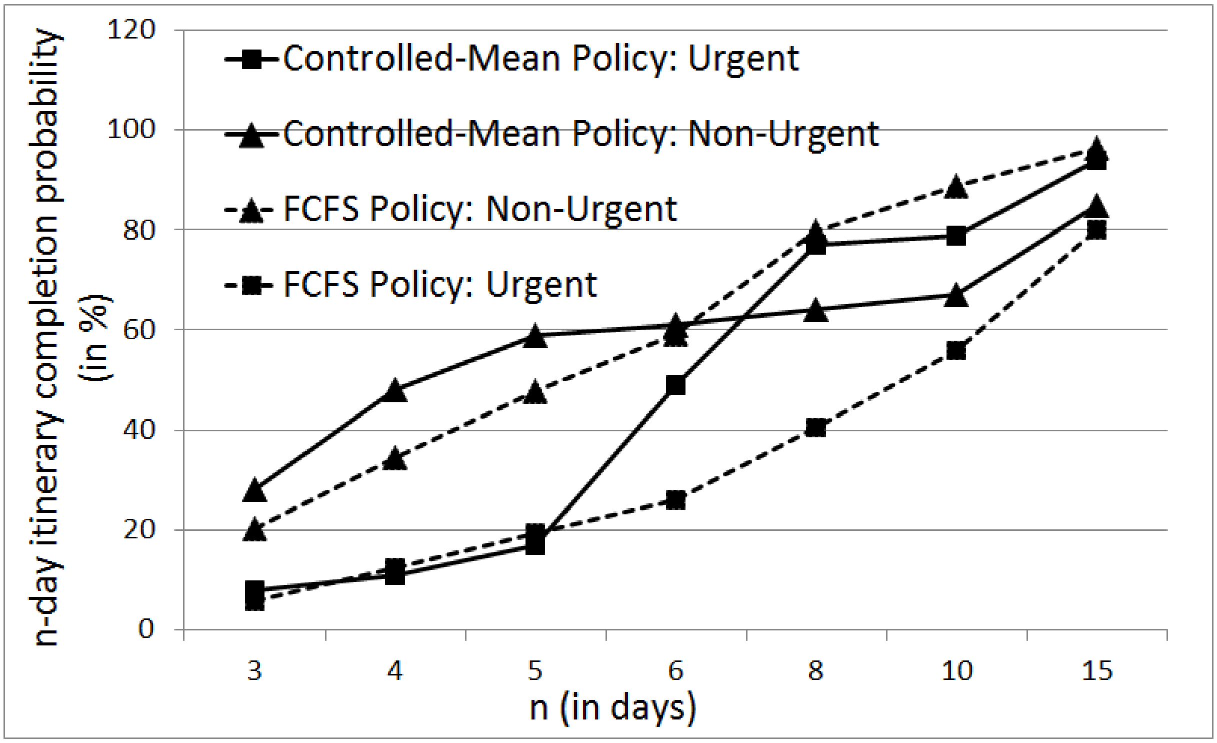


Figure 4.4: "Controlled-mean" policy (which resulted to the solution $(6, 7.1)$ in Fig. 4.3) itinerary completion probability within n days, and comparison to the FCFS policy benchmark (see Fig. 4.2)

of non-urgent cases, and minimizing the mean itinerary flow time for urgent cases might not be the most efficient way to extract the desired performance of our system. To allow more flexibility in our system so as to provide expedited service to a higher number of urgent cases, and improve the overall performance, we can constrain our model in a way that non-urgent itinerary flow times would be less than t days with $x\%$ probability while maximizing the probability that urgent cases will complete their itinerary in t days or less. We will call such optimization formulations "controlled-distribution" policies.

We now maximize the 5-day itinerary completion probability of urgent patients subject to a given constraint on the 5-day itinerary completion probability of non-urgent cases. In Fig. 4.5 we study the tradeoffs between the probability of having an itinerary flow time of one week (5 days) or less for our two patient classes. We provide an efficient frontier of solutions

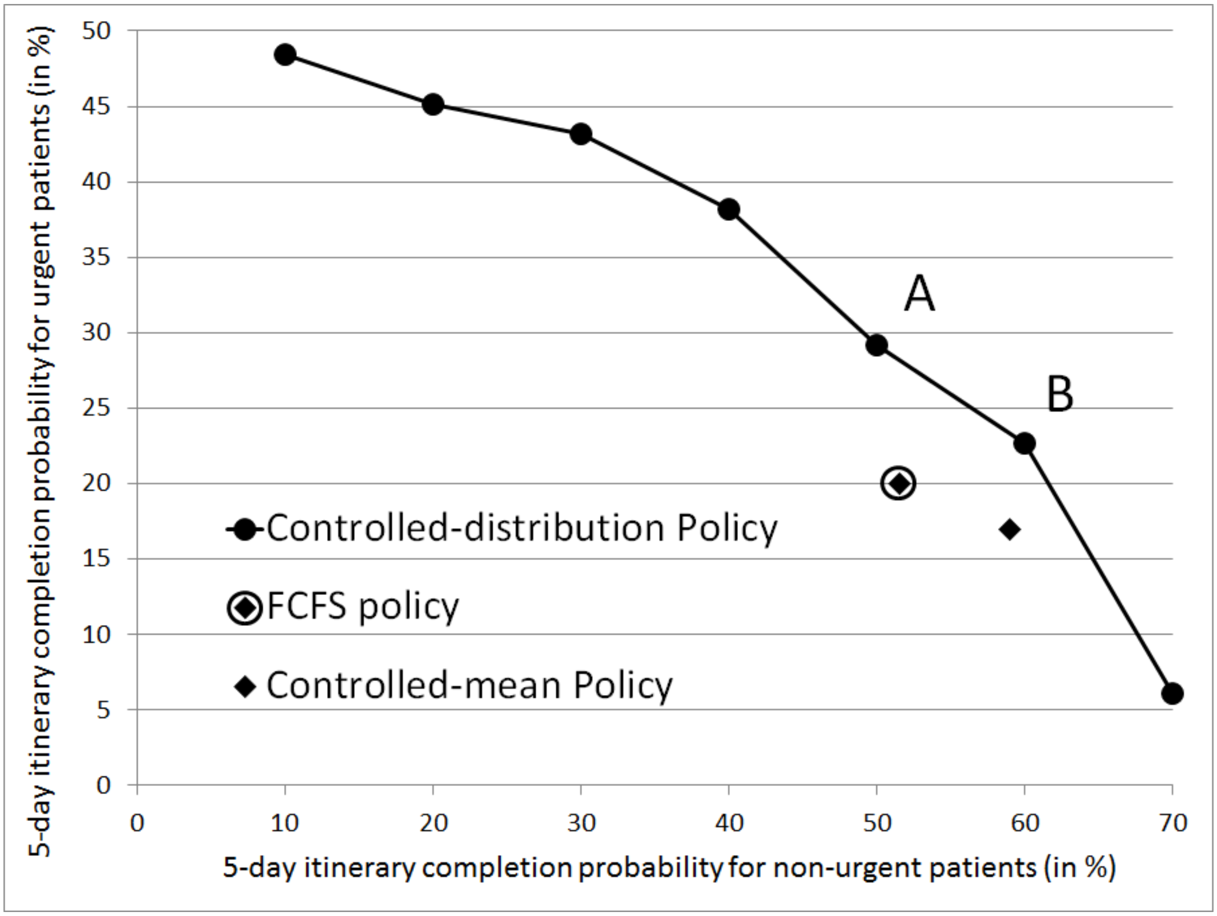


Figure 4.5: Tradeoff Analysis: optimal 5-day itinerary completion probability for urgent patients when varying the constraint on the non-urgent 5-day itinerary completion probability under the controlled-distribution policy, and comparison to the FCFS (Fig. 4.2) and controlled-mean (Fig. 4.4) policies

by varying the constraint set on the probability that non-urgent patients complete their itinerary within 5 days (see the solid curve). The two additional data points compare that efficient frontier with 5-day itinerary completion probability under (1) the FCFS policy, and (2) the controlled-mean policy optimized above. Recall from Fig. 4.3 that the controlled-mean policy renders a 28% decrease in the urgent mean itinerary flow time compared to the FCFS policy while keeping the mean of the non-urgent itinerary flow time nearly the same. However, we can see that in terms of the metrics analyzed in Fig. 4.5 (i.e., probability of completing an itinerary within 5 days), the controlled-mean policy seems to be more desirable than the previously optimized policy (2).

We can see that improvements can be made to our two policies in relation to this chosen metric by using a controlled-distribution policy. If we were to implement solution A, $(50, 34.2)$ on the efficient frontier, this would lead to a 45% increase in the urgent 5-day itinerary completion probability compared to the FCFS policy while seeing only a 3% decrease in non-urgent itinerary completion within 5 days. If we were to consider solution B, $(60, 22.6)$ on the efficient frontier, we would be able to achieve: (1) 13% improvement in our metric for urgent patients, and 16.5% improvement for non-urgent patients compared the FCFS policy, and (2) 33.7% improvement in our metric for urgent patients, and 2.5% improvement for non-urgent patients in comparison the controlled mean policy. However, all these improvements on the 5-day itinerary completion probability comes at a cost on the mean itinerary flow times: solution A renders a mean itinerary flow time of $(8.4, 7.8)$, Solution B has a mean itinerary flow time of $(10.6, 6.9)$, while the FCFS and controlled mean policies achieved $(7.1, 6)$.

As we mentioned in Section 4.5.2, this tradeoff and sensitivity analysis illustrates a key contribution of our method to practice. Rather than providing one capacity reservation plan, it will be useful to let the managerial staff experiment with different objectives and constraints. This approach can provide a rich portfolio of various policies that inform decision-making and tailor the solutions to the application. Our new methodologies will share an efficient frontier of optimal solutions with the health institution, and can provide insights about the dynamics/interactions between all metrics considered.

4.6.2 Validation

In this section, we will validate our methodology. Having used some approximations on the demand in progress distribution (see Section 4.5.3), and having made two independence assumptions (Assumptions IV.1 and IV.4), it is important to analyze the absolute percentage errors of our analytical metrics' estimates. After optimizing our system, which resulted in a capacity reservation plan Ω, we will simulate this new controlled system. This simulation will be similar to the one described in Section 4.4.5, but we will now consider different formulations to yield various control policy solutions for capacity reservations at each downstream service. We compare our analytical estimations of the optimized performance metrics to the simulated ones, to get a sense of how accurate our approximations and how valid our assumptions were.

Having validated the access delay metric for root appointments in Chapter III (see Section in 3.8.2), and therefore the approximations made on the demand in progress distributions at a given service (see Sections 3.6.4 and 4.5.4), we do not to validate these metrics again (i.e. M5 and M6).

Diagnostic flow time

The diagnostic flow time metric's errors will be caused by Assumption IV.1 (independence of the demand in progress across diagnostic services). As we discussed in the paragraph following this assumption, we consider a healthcare scenario where the capacity reserved for root appointments has a high probability of being filled up completely. As long as we are in a setting where the waiting time for a root appointment is in the order of multiple days/weeks, then the approximation we make with this assumption should perform well. In Table 4.2, we show the mean and variance of the number of patients scheduled for a root visit according to a template optimized in Chapter III's case study. The mean number of patients scheduled for a root appointment on day t in our planning horizon, is also the number of patients requiring diagnostic appointments requested on day t for subsequent fulfillment.

As can be seen in Table 4.2, the variance of the number of patients requiring downstream services will be close to 0, due to the multi-week access delay for a root appointment, and the fact that the total reserved capacity is used with high probability. Therefore,

Day of Week	Mo	Tu	We	Th	Fr
Mean Urgent Downstream Demand	15.89	18.52	12.80	9.95	7.99
Mean Non-Urgent Downstream Demand	9.90	11.39	8.57	6.78	3.99
Urgent Downstream Demand Variance	0.64	3.67	0.99	0.16	0.04
Non-Urgent Downstream Demand Variance	0.32	2.97	1.36	0.61	0.01

Table 4.2: Mean and Variance of the number of patients requiring downstream (diagnostic) appointments based on the template developed for root visits in Chapter III

the variability of the demand for a specific diagnostic resource will mostly depend on the Bernoulli/multinomial random variables $\kappa_s^{\tau,k}$ (i.e., relating to the routing probability for downstream resources) and not on the flow of root visits. Since these random variables are independent of each other, and everything else in the system, Assumption IV.1 does not yield large errors in our metric estimations, as we can see in Fig. 4.6.

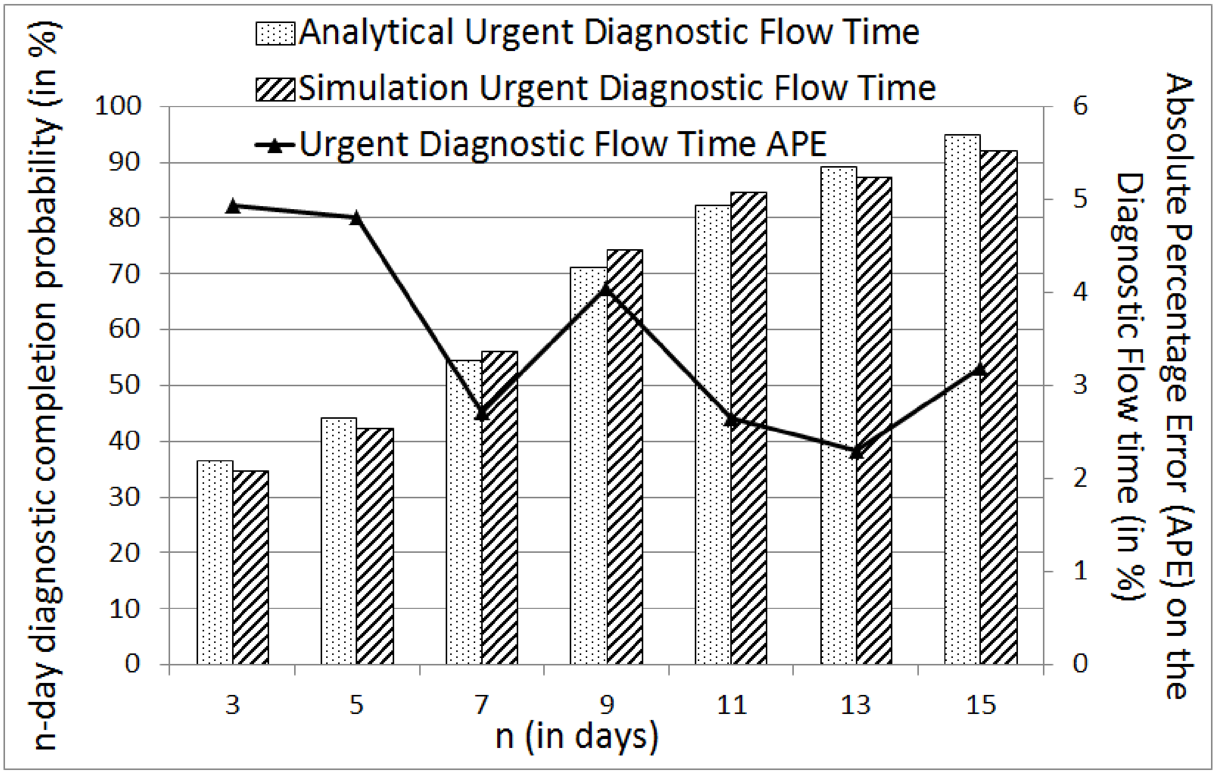


Figure 4.6: Validation of the accuracy of the n-day diagnostics completion probability estimates for the controlled-probability policy solution B in Fig. 4.5

Itinerary flow time

Finally, the validation of the itinerary flow time metric emphasizes the validity of Assumption IV.4 (independence of the demand in progress between diagnostic and follow-up services). In Fig. 4.7, we validate our metric estimates computed in the tradeoff analysis of Fig. 4.5 which led to an efficient frontier of solutions. We do not validate the entire itinerary flow time distribution (similarly to what was done in Fig. 4.5 for the diagnostic flow time distribution) but rather, validate across multiple formulations (i.e., different constraints on non-urgent itinerary flow time) of our optimization program.

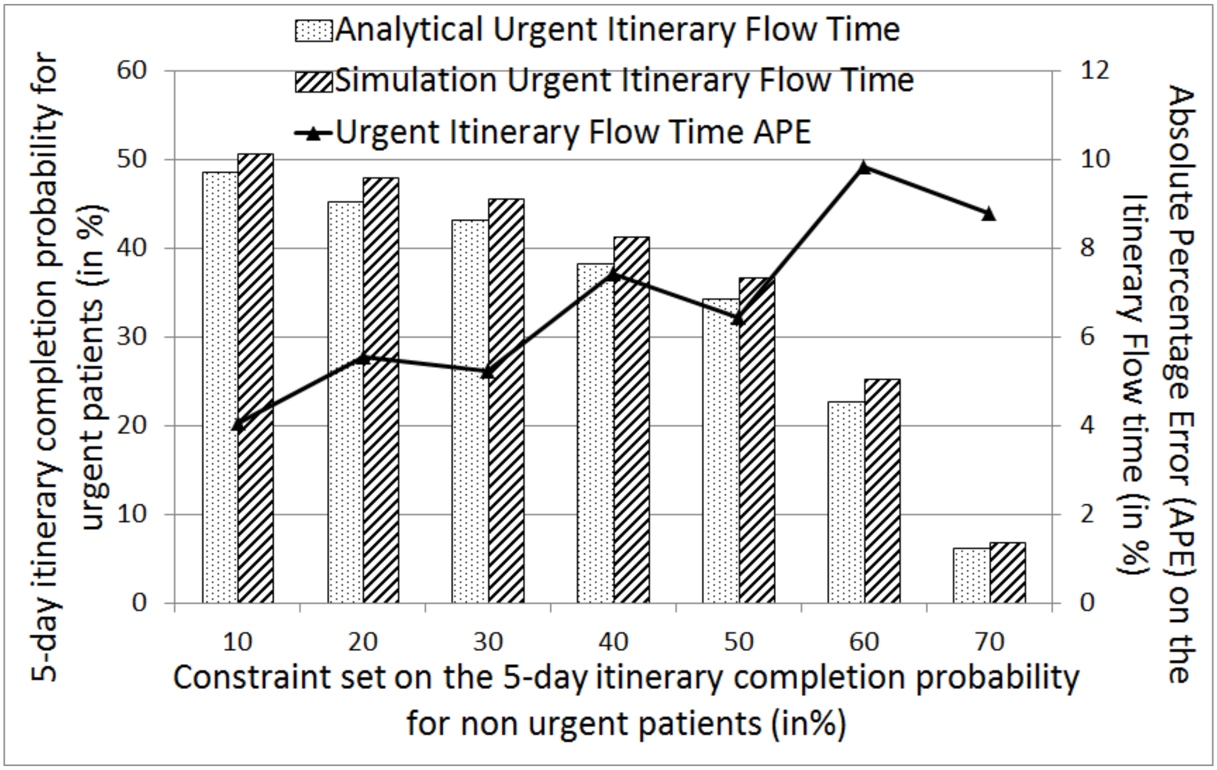


Figure 4.7: Validation of the accuracy of the 5-day itinerary completion probability estimates under multiple controlled-probability policies generated in Fig. 4.5 when varying the constraint on the 5-day itinerary completion probability for non-urgent cases

The larger errors (that go up to 9-10%) for the tightly constrained system (i.e., 5-day itinerary completion probability of 60-70% for non-urgent patients) are partly due to the smaller values obtained for the 5-day itinerary completion probability of urgent patients in those scenarios. But, having a tight constraint on our system also yields higher variability in urgent itinerary flow times (since they will have longer and less predictable itinerary flow times), which makes it more difficult for our model to accurately forecast tail probabilities.

4.7 Conclusion

Today's appointment based services such as healthcare have very limited research supporting effective control of waits for an appointment. This research generates fundamental operations-based methodology to facilitate complex tradeoffs involving (1) cost-efficiency and (2) differentiated levels of wait to serve urgent cases.

This research focused on patient flows through diagnostic and specialty resources/services of a health institution. The motivation for this study was to provide differentiated access to the resources based on the urgency level of the patient. Experiencing delays to a diagnosis, and in many cases, to the start of treatment, can have drastic effects for some patients. Allocating/reserving resource capacity between various urgency levels provides a control that health care institutions can leverage to provide more efficient and effective service to their patients.

We focused our analysis on three critical metrics to forecast and control: (i) the access delay patients experience at each downstream service along their itinerary of care (diagnostic testing and specialty follow-up services), (ii) the diagnostic flow time of patients which captures the time from the root appointment to the end of diagnostic testing, and (iii) the itinerary flow time of patients, which captures the total length of an itinerary of care. Being able to estimate and control those metrics across various patient types can help heath institutions manage their supply and demand in a more efficient way. An optimized capacity reservation plan developed in this research will assure that the performance targets desired by the institution will be met. This capacity plan has a nice simple structure that can be implemented in practice, but optimized methods to do that are beyond our scope. The tradeoff analyses, and what if scenarios provided with this new methodology will also be a valuable tool for decision making in health institutions.

Future research could involve more complex itinerary visit precedence structures. The tools developed here could be extended to these more complex scenarios if the data requirements can be provided.

4.8 Apppendix

4.8.1 Proof of Lemma IV.2

Our goal here is to show that $\mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right)$ can be expressed in terms of linear constraints subject to our decision variables Ω. We first start by using our definition of the diagnostic flow time random variable (see Eq. 4.5) in Eq. 4.35. Based on our assumption that $|\mathcal{S}_k| = 2$, we will refer to the two elements of the set $\mathcal{S}_k$ as s_1 and s_2, and therefore Eq. 4.36 follows.

$$\begin{aligned}
\mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right) &= \mathbb{P}\left(\max_{s\in S_k} \kappa_s^{\tau,k} \cdot \chi_t^s > l\right) && (4.35)\\
&= \mathbb{P}\left(\max\{\kappa_{s_1}^{\tau,k} \cdot \chi_t^{s_1}, \kappa_{s_2}^{\tau,k} \cdot \chi_t^{s_2}\} > l\right) && (4.36)\\
&= \sum_{i\in I} \sum_{j\in\{0,1\}} \mathbb{P}\left(\max\{\kappa_{s_1}^{\tau,k} \cdot \chi_t^{s_1}, \kappa_{s_2}^{\tau,k} \cdot \chi_t^{s_2}\} > l | \kappa_{s_2}^{\tau,k} = j, L_t^{\tau,s_2} = L_t^{\tau,s_2}(i)\right)\\
&\cdot \mathbb{P}\left(\kappa_{s_2}^{\tau,k} = j, L_t^{\tau,s_2} = L_t^{\tau,s_2}(i)\right) && (4.37)\\
&= \sum_{i\in I} \sum_{j\in\{0,1\}} \max\Bigg\{\mathbb{P}\left(\kappa_{s_1}^{\tau,k} \cdot \chi_t^{s_1} > l | L_t^{\tau,s_2} = L_t^{\tau,s_2}(i), \kappa_{s_2}^{\tau,k} = j\right),\\
&j \cdot \mathbb{1}\Big\{L_t^{\tau,s_2}(i) > \sum_{m=0}^{l} \Omega_{t\oplus m}^{\tau,s_2}\Big\}\Bigg\} \cdot \phi_t^{\tau,s_2}(i) \cdot \mathbb{P}\left(\kappa_{s_2}^{\tau,k} = j\right) && (4.38)
\end{aligned}$$

In Eq. 4.37, without loss of generality, we condition our probability on the realizations of the demand in progress L_t^{τ,s_2} and the service/resource requirement $\kappa_{s_2}^{\tau,k}$ random variables in diagnostic service s_2. Based on this conditioning, the requirement of service s_2 along the patient's care path as well as the access delay in service s_2 become deterministic. Specifically, the access delay in service s_2 will be greater than l with probability 1 if (1) s_2 is a diagnostic service required in the patient's itinerary (i.e., $j = 1$), and (2) the demand in progress at service s_2 is greater than the total amount of capacity reserved for type τ patients from day t to $t+l$ (i.e., $L_t^{\tau,s_2}(i) > \sum_{m=0}^{l} \Omega_{t\oplus m}^{\tau,s_2}$). Otherwise, the probability of the access delay in service s_2 being greater than l will be 0. Therefore, the conditional probability of the diagnostic delay exceeding l days (given the information on diagnostic service s_2) will be the probability that the access delay in diagnostic service s_1 exceeds l days if the access delay in s_2 does not exceed l, and will be 1 otherwise. Moreover, since the resource need for service s_2 random variable $\kappa_{s_2}^{\tau,k}$ is independent of the current state of the system (e.g., the current demand in progress at s_2), Eq. 4.38 follows. Using Eq. 4.19, and the independence assumption of the demand in progress across diagnostic services, we can write $\mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right) =$

$$= \sum_{i \in I} \sum_{j \in \{0,1\}} \max\left\{ \mathbb{P}\left(\kappa_{s_1}^{\tau,k} \cdot \chi_t^{s_1} > l\right), j \cdot x_{t,l}^{\tau,s_2}(i) \right\} \cdot \phi_t^{\tau,s_2}(i) \cdot \mathbb{P}\left(\kappa_{s_2}^{\tau,k} = j\right) \tag{4.39}$$

$$\begin{aligned} &= \sum_{i \in I} \sum_{j \in \{0,1\}} \max\left\{ \sum_{i' \in I} \sum_{j' \in \{0,1\}} \mathbb{P}\left(\kappa_{s_1}^{\tau,k} \cdot \chi_t^{s_1} > l | L_t^{\tau,s_1} = L_t^{\tau,s_1}(i'), \kappa_{s_1}^{\tau,k} = j'\right) \right. \\ &\left. \cdot\, \phi_t^{\tau,s_1}(i') \cdot \mathbb{P}\left(\kappa_{s_1}^{\tau,k} = j'\right), j \cdot x_{t,l}^{\tau,s_2}(i) \right\} \cdot \phi_t^{\tau,s_2}(i) \cdot \mathbb{P}\left(\kappa_{s_2}^{\tau,k} = j\right) \end{aligned} \tag{4.40}$$

$$\begin{aligned} &= \sum_{i \in I} \sum_{j \in \{0,1\}} \max\left\{ \sum_{i' \in I} \mathbb{1}\left\{ L_t^{\tau,s_1}(i') > \sum_{m=0}^{l} \Omega_{t \oplus m}^{\tau,s_1} \right\} \cdot \phi_t^{\tau,s_1}(i') \cdot \mathbb{P}\left(\kappa_{s_1}^{\tau,k} = 1\right), j \cdot x_{t,l}^{\tau,s_2}(i) \right\} \\ &\cdot\, \phi_t^{\tau,s_2}(i) \cdot \mathbb{P}\left(\kappa_{s_2}^{\tau,k} = j\right) \end{aligned} \tag{4.41}$$

$$= \sum_{i \in I} \sum_{j \in \{0,1\}} \max\left\{ p_{s_1}^{\tau,k} \cdot \sum_{i' \in I} x_{t,l}^{\tau,s_1}(i') \cdot \phi_t^{\tau,s_1}(i'), j \cdot x_{t,l}^{\tau,s_2}(i) \right\} \cdot \phi_t^{\tau,s_2}(i) \cdot \mathbb{P}\left(\kappa_{s_2}^{\tau,k} = j\right) \tag{4.42}$$

$$= \sum_{i \in I} \sum_{j \in \{0,1\}} y_{t,l}^{(2),\tau,k}(i,j) \cdot \phi_t^{\tau,s_2}(i) \cdot \mathbb{P}\left(\kappa_{s_2}^{\tau,k} = j\right) \tag{4.43}$$

In Eq. 4.40, we now condition on the demand in progress in diagnostic service s_1 and the resource need for service s_1 random variables (while again using the independence property of the demand in progress and $\kappa_{s_1}^{\tau,k}$). This conditional probability (given information on diagnostic service s_1) will be equal to 1 if and only if $j' = 1$ and $L_t^{\tau,s_1}(i') > \sum_{m=0}^{l} \Omega_{t \oplus m}^{\tau,s_1}$ (similarly to the concept that gave us Eq. 4.38 above) which justifies Eq. 4.41. Again, using Eq. 4.19, Eq. 4.42 follows directly. We now have proven 4.24 thanks to 4.25.

Since our goal is to minimize those probabilities of exceeding access/flow time targets, it is sufficient for $y_{t,l}^{(2),\tau,k}(i,j)$ to satisfy the following linear equations in $x_{t,l}^{\tau,s}(i)$:

$$y_{t,l}^{(2),\tau,k}(i,j) \geq j \cdot x_{t,l}^{\tau,s_2}(i) \tag{4.44}$$

$$y_{t,l}^{(2),\tau,k}(i,j) \geq p_{s_1}^{\tau,k} \cdot \sum_{i' \in I} x_{t,l}^{\tau,s_1}(i') \cdot \phi_t^{\tau,s_1}(i'). \tag{4.45}$$

Eqs. 4.19, 4.20, and 4.21 have expressed $x_{t,l}^{\tau,s}(i)$ as a a linear set of constraints (subject to our decision variable Ω), therefore we have shown that $\mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right)$ can be expressed linearly in our Mixed Integer Program.

4.8.2 Proof of Theorem IV.3

We start this proof by assuming that $|\mathcal{S}_k| = 3$, with $s_1, s_2, s_3 \in \mathcal{S}_k$ (see Eq. 4.46). We then condition this probability in Eq. 4.47 on the realizations of the demand in progress L_t^{τ,s_3} and the resource requirement $\kappa_{s_3}^{\tau,k}$ random variables in diagnostic service s_3.

$$\mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right) = \mathbb{P}\left(\max\{\kappa_{s_1}^{\tau,k} \cdot \chi_t^{s_1}, \kappa_{s_2}^{\tau,k} \cdot \chi_t^{s_2}, \kappa_{s_3}^{\tau,k} \cdot \chi_t^{s_3}\} > l\right) \tag{4.46}$$

$$= \sum_{i \in I} \sum_{j \in \{0,1\}} \mathbb{P}\left(\max\{\kappa_{s_1}^{\tau,k} \cdot \chi_t^{s_1}, \kappa_{s_2}^{\tau,k} \cdot \chi_t^{s_2}, \kappa_{s_3}^{\tau,k} \cdot \chi_t^{s_3}\} > l | \kappa_{s_3}^{\tau,k} = j, L_t^{\tau,s_3} = L_t^{\tau,s_3}(i)\right)$$

$$\cdot \mathbb{P}\left(\kappa_{s_3}^{\tau,k} = j, L_t^{\tau,s_3} = L_t^{\tau,s_3}(i)\right) \tag{4.47}$$

$$= \sum_{i \in I} \sum_{j \in \{0,1\}} \max\left\{\mathbb{P}\left(\max\{\kappa_{s_1}^{\tau,k} \cdot \chi_t^{s_1}, \kappa_{s_2}^{\tau,k} \cdot \chi_t^{s_2}\} > l | L_t^{\tau,s_3} = L_t^{\tau,s_3}(i), \kappa_{s_3}^{\tau,k} = j\right),\right.$$

$$\left. j \cdot \mathbb{1}\left\{L_t^{\tau,s_3}(i) > \sum_{m=0}^{l} \Omega_{t \oplus m}^{\tau,s_3}\right\}\right\} \cdot \phi_t^{\tau,s_3}(i) \cdot \mathbb{P}\left(\kappa_{s_3}^{\tau,k} = j\right) \tag{4.48}$$

$$= \sum_{i \in I} \sum_{j \in \{0,1\}} \max\left\{\mathbb{P}\left(\max\{\kappa_{s_1}^{\tau,k} \cdot \chi_t^{s_1}, \kappa_{s_2}^{\tau,k} \cdot \chi_t^{s_2}\} > l\right), j \cdot x_{t,l}^{\tau,s_3}(i)\right\}$$

$$\cdot \phi_t^{\tau,s_3}(i) \cdot \mathbb{P}\left(\kappa_{s_3}^{\tau,k} = j\right) \tag{4.49}$$

$$= \sum_{i \in I} \sum_{j \in \{0,1\}} \max\left\{\sum_{i' \in I} \sum_{j' \in \{0,1\}} y_{t,l}^{(2),\tau,k}(i',j') \cdot \phi_t^{\tau,s_2}(i') \cdot \mathbb{P}\left(\kappa_{s_2}^{\tau,k} = j'\right), j \cdot x_{t,l}^{\tau,s_3}(i)\right\}$$

$$\cdot \phi_t^{\tau,s_3}(i) \cdot \mathbb{P}\left(\kappa_{s_3}^{\tau,k} = j\right) \tag{4.50}$$

$$= \sum_{i \in I} \sum_{j \in \{0,1\}} y_{t,l}^{(3),\tau,k}(i,j) \cdot \phi_t^{\tau,s_3}(i) \cdot \mathbb{P}\left(\kappa_{s_3}^{\tau,k} = j\right) \tag{4.51}$$

The conditional probability in Eq. 4.47 (given the information on diagnostic service s_3) will be equal to 1 if and only if $j = 1$ and $L_t^{\tau,s_3}(i) > \sum_{m=0}^{l} \Omega_{t \oplus m}^{\tau,s_3}$ (similarly to the concept that gave us Eq. 4.38 above). This allows us to transition to Eq. 4.48. In Eq. 4.49, we use the independence assumption of the demand in progress across diagnostic services, and leverage the result presented in Eq. 4.19. Finally, the result from Lemma IV.2 allows us to obtain Eq. 4.50, and from the definition of $y_{t,l}^{(3),\tau,k}(i,j)$ in Eq. 4.27, Eq. 4.51 follows. This proves Theorem IV.3 for the case $|\mathcal{S}_k| = 3$.

We can repeat this procedure up to the true dimension of the set $\mathcal{S}_k$, while still being able to express the diagnostic flow distribution similarly. Notice that $\mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right)$ can be expressed linearly in $y_{t,l}^{(|\mathcal{S}_k|),\tau,k}(i,j)$ and $x_{t,l}^{\tau,|\mathcal{S}_k|}(i)$ similarly to constraints 4.44 and 4.45. Hence, $\mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right)$ can be modeled as linear constraints subject to $y_{t,l}^{(2),\tau,k}(i,j)$ (see Eq. 4.27), therefore linearly subject to our decision variables Ω as well (from the result of Lemma IV.2). This proves Theorem IV.3.

4.8.3 Proof of Corollary IV.5

We use the definition of the itinerary flow time random variable (see Eq. 4.1) to express the itinerary flow time distribution as follows:

$$
\begin{aligned}
\mathbb{P}\left(\Xi_t^{\tau,k} > l\right) &= \mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} + \chi^{\tau,k}_{1\oplus t\oplus \overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k}} > l\right) && (4.52)\\
&= \sum_{l_1=0}^{\infty} \mathbb{P}\left(\chi^{\tau,k}_{t\oplus l_1\oplus 1} > l - l_1 | \overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} = l_1\right)\cdot \mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} = l_1\right) && (4.53)\\
&= \mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right) + \sum_{l_1=0}^{l} \mathbb{P}\left(\chi^{\tau,k}_{t\oplus l_1\oplus 1} > l - l_1 | \overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} = l_1\right)\cdot \mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} = l_1\right) && (4.54)\\
&= \mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right) + \sum_{l_1=0}^{l}\sum_{i\in I} \mathbb{P}\left(\chi^{\tau,k}_{t\oplus l_1\oplus 1} > l - l_1 | L^{\tau,k}_{t\oplus l_1\oplus 1} = L^{\tau,k}_{t\oplus l_1\oplus 1}(i)\right)\cdot \phi_t^{\tau,k}(i)\\
&\cdot\left(\mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l_1 - 1\right) - \mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l_1\right)\right) && (4.55)\\
&= \mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right) + \sum_{l_1=0}^{l}\sum_{i\in I} \mathbb{1}\Big\{L_t^{\tau,k}(i) > \sum_{m=0}^{l-l_1}\Omega^{\tau,k}_{t\oplus m}\Big\}\cdot \phi_t^{\tau,k}(i)\\
&\cdot\sum_{i'\in I}\sum_{j\in\{0,1\}}\left(y_{t,l-1}^{(|\mathcal{S}_k|),\tau,k}(i',j) - y_{t,l}^{(|\mathcal{S}_k|),\tau,k}(i',j)\right)\cdot \phi_t^{\tau,|\mathcal{S}_k|}(i')\cdot \mathbb{P}\left(\kappa^{\tau,k}_{|\mathcal{S}_k|} = j\right) && (4.56)\\
&= \mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right) + \sum_{l_1=0}^{l}\sum_{i\in I} x^{\tau,k}_{t\oplus l_1\oplus 1, l-l_1}(i)\cdot \phi^{\tau,k}_{t\oplus l_1\oplus 1}(i)\\
&\cdot\sum_{i'\in I}\sum_{j\in\{0,1\}}\left(y_{t,l-1}^{(|\mathcal{S}_k|),\tau,k}(i',j) - y_{t,l}^{(|\mathcal{S}_k|),\tau,k}(i',j)\right)\cdot \phi_t^{\tau,|\mathcal{S}_k|}(i')\cdot \mathbb{P}\left(\kappa^{\tau,k}_{|\mathcal{S}_k|} = j\right) && (4.57)
\end{aligned}
$$

In Eq. 4.53, we condition the itinerary flow time probability on realizations of the diagnostic flow time. Given the diagnostic flow time is l_1 days, we are now interested to see the probability of the access delay in specialty k being greater than $l - l_1$ (given the specialty service request for a follow-up happened on day $t \oplus l_1 \oplus 1$). For $l_1 > l$, the first term in Eq. 4.53 will always be 1. Therefore, in Eq. 4.54 we can make the sum go to l and add the probability that the diagnostic flow time itself will be greater than l days. In Eq. 4.55, we first use the independence assumption of the demand in progress random variables across all services (see Assumption IV.1), and then condition our probability on the demand in progress in specialty service k on day $t \oplus l_1 \oplus 1$. This conditional probability will result in a binary variable equal 1 if and only if $L_t^{\tau,k}(i) > \sum_{m=0}^{l-l_1} \Omega_{t\oplus m}^{\tau,k}$, and we use the result from IV.3 to express the diagnostic flow time probability in Eq. 4.56. Finally, using Eq. 4.19, Eq. 4.57 follows.

Since $x_{t\oplus l_1 \oplus 1, l-l_1}^{\tau,k}(i)$ is a binary variable, and it is multiplied by an expression between 0 and 1, we can convert Eq. 4.57 as follows, by using the definition of $z_{t,l_1,l}^{\tau,k}(i,j)$ provided in Eq. 4.31:

$$\mathbb{P}\left(\Xi_t^{\tau,k} > l\right) = \mathbb{P}\left(\overline{\overline{\chi}}_t^{\tau,\mathcal{S}_k} > l\right) + \sum_{l_1=0}^{l} \sum_{i\in I} z_{t,l_1,l}^{\tau,k}(i) \cdot \phi_{t\oplus l_1 \oplus 1}^{\tau,k}(i) \tag{4.58}$$

where $z_{t,l_1,l}^{\tau,k}(i)$ can be expressed as follows, for a large enough $M > 0$:

$$z_{t,l_1,l}^{\tau,k}(i) \le x_{t\oplus l_1 \oplus 1, l-l_1}^{\tau,k}(i) \tag{4.59}$$

$$z_{t,l_1,l}^{\tau,k}(i) \ge \sum_{i'\in I} \sum_{j\in\{0,1\}} \left(y_{t,l-1}^{(|\mathcal{S}_k|),\tau,k}(i',j) - y_{t,l}^{(|\mathcal{S}_k|),\tau,k}(i',j) \right) - M \cdot (1 - x_{t\oplus l_1 \oplus 1, l-l_1}^{\tau,k}(i)) \tag{4.60}$$

Eq. 4.59 assures that $z_{t,l_1,l}^{\tau,k}(i)$ will be 0 when $x_{t\oplus l_1 \oplus 1, l-l_1}^{\tau,k}(i) = 0$, and Eq. 4.60 constrains $z_{t,l_1,l}^{\tau,k}(i)$ to be at least $\sum_{i'\in I} \sum_{j\in\{0,1\}} \left(y_{t,l-1}^{(|\mathcal{S}_k|),\tau,k}(i',j) - y_{t,l}^{(|\mathcal{S}_k|),\tau,k}(i',j) \right)$ when $x_{t\oplus l_1 \oplus 1, l-l_1}^{\tau,k}(i) = 1$. In this case, since our objective is to minimize itinerary flow time, this constraint will be binding. Hence, from Eqs. 4.59 and 4.60 we are able to express Eq. 4.31 linearly in the $x_{t\oplus l_1 \oplus 1, l-l_1}^{\tau,k}(i)$ and $y_{t,l}^{(|\mathcal{S}_k|),\tau,k}(i,j)$ variables. From Theorem IV.3, Eq. 4.20 and Eq. 4.21, we have shown how to compute the the itinerary flow time distribution linearly in our decision variables Ω.

CHAPTER V

Conclusions

To conclude this dissertation, we discuss the integration of clinical care and clinical research under shared resources. This dissertation has focused on providing differentiated access to a set of resources based on patient characteristics and urgency levels. This was done by using operations engineering tools to optimize a patient schedule so that it allocates system capacity in a way that meets the health institution's goals. We considered this general idea in three different settings: (1) Access management in Clinical Research Units (CRUs) where resources are entirely focused on clinical research visits and participants, (2) Access delay control for a root appointment of an itinerary of care in a health institution with resources devoted entirely to clinical care visits and patients, and (3) Itinerary/Diagnostic flow time and access delay control for downstream visits of a care itinerary in a clinical care setting. The access and patient flow management tools we provided in the above three settings are useful independently, and address the entire scope provided in the introduction.

One question that remains is how to manage a system's capacity so that it can provide differentiated access to care patients and research participants, in a shared resource setting where both for clinical care and clinical research visits occur. Specifically, the patient can become a research participant to a clinical trial after waiting for access to an initial appointment in a medical specialty and either before or after going through all the necessary referrals and downstream visits (for disease diagnosis). Patients that have gone through this process will, according to an i.i.d. Bernoulli process, enroll in the clinical trial with probability that based on (1) the patient being eligible to participate in this trial, and (2) the patient accepting to be part of it. The patient can be recruited for a clinical trial either at the initial visit (if there is enough data on the patient to assure that she will meet the trial inclusion criteria) or at the time of the follow up visit (when the diagnostic tests have provided enough information about inclusion/exclusion criteria). If the patient becomes a research participant of a given clinical trial, then she will follow a trial-specific protocol of research visits as described in Chapter II. Note that portfolio selection and admission control (BRPlan) decisions can be optimized to guarantee differentiated access by trial (see Chapter II).

Given a commitment to the ongoing care needs of the patient panel/population, the question regularly arises as to which clinical research studies/trials can be conducted within the limited resources of the organization. With the operational dexterity provided by the different methodologies of this dissertation and their integration with smart health IT, this research can tackle the thorny problem of how to properly plan and manage the integration of clinical care with clinical research when it occurs in the same clinic (such as the Michigan Clinical Research Unit). Many patients benefit greatly from participating in a clinical research program that is integrated into the patient's regular clinical care visits. The capacity planning and multi-visit scheduling methodology can serve a role as an integrating mechanism for accomplishing both the care and research dimensions of many diseases in a manner that maximizes the value to patient and research communities while managing costs by optimizing efficiency.

The goal is to determine which clinical trials should be performed in the site given (i) clinical care patients' access delays to a root appointment are limited to a given target (specified by the patient's urgency level), (ii) clinical care patients' diagnostic and itinerary flow times (waiting times to a diagnosis) satisfy specific targets (e.g., urgent patients will require much tighter constraints), and (iii) clinical research participants will be guaranteed a certain access level (the Time to First Available Visit) based on their type (urgent vs. non urgent) and trial they enroll in.

To develop methods to manage the joint delivery of clinical care and clinical research, methods will need to be put into place to answer the following questions. (1) What is the minimum capacity required for clinical care visits while assuring access delay to a root appointment, diagnostic flow time, and itinerary flow time targets for each of our patient types? (2) Given the remaining capacity available for clinical research visits, which portfolio of clinical trials could/should be conducted in our infrastructure while assuring certain access levels to the trials? Answering these two questions in an integrated way can be quite computationally challenging, and results in non-linearity issues. This is why we built our methodologies tackling each problem independently.

Given that a institution's main priority is to guarantee proper access to resources for clinical care patients, one could answer the above two questions in a multi-stage mixed-integer optimization approach. In a first stage, one could use the methods developed in Chapter III to minimize the capacity that needs to be allocated for clinical care root appointments of an itinerary while still meeting the differentiated access levels desired for various patient types. Then, in a second stage, one could use the methods from Chapter IV to once again minimize the capacity required to perform downstream care visits such that itinerary and diagnostic flow time constraints are met for each patient class. Both steps increase the efficiency as much as possible while addressing the desired access and service/delay levels. This would then result in the minimization of the clinical care capacity required to meet our desired access constraints which is equivalent to maximizing the remaining capacity that can then be allocated for clinical research visits and participants (while still meeting the clinical care standards we set on our system). Once we know the remaining available capacity that can be used for clinical research and the resulting demand streams for clinical research trials (that depend on when/how many patients of each type are scheduled for an initial or follow up appointment, optimized in Stages 1 and 2), we can then determine in Stage 3 how much clinical research demand to accept (by answering the trial portfolio decision) and how to create the BRPlan to provide differentiated access by type/trial. The methods from Chapter II could be used for that last optimization stage.

Combining the ideas and methods developed in this dissertation will allow health institutions to conduct clinical care and clinical research efficiently under shared resources. The new methods help provider organizations to understand (1) how to allocate system resources between (1a) care appointments and research visits, (1b) urgent cases and less urgent ones; (2) how much clinical research workload can be allowed, and (3) how each capacity decision affects patient flow and access to care/research/diagnosis. These methods and insights are fundamentally relevant to any health care institution that decides to tackle these complex problems of access, efficiency, and cost control.

Bibliography

[1] I. Adan, J. Bekkers, N. Dellaert, J. Vissers, and X. Yu. Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health care management science*, 12(2):129–141, 2009.

[2] C. Akkan. Finite-capacity scheduling-based planning for revenue-based capacity management. *European Journal of Operational Research*, 100(1):170–179, 1997.

[3] J.R. Anderson and M. Matessa. An overview of the epic architecture for cognition and performance with application to human-computer interaction. *Human–Computer Interaction*, 12(4):391–438, 1997.

[4] J. Asmundsson, R.L. Rardin, C.H. Turkseven, and R. Uzsoy. Production planning with resources subject to congestion. *Naval Research Logistics*, 56(2):142–157, 2009.

[5] R. Bekker and P.M. Koeleman. Scheduling admissions and reducing variability in bed demand. *Health Care Management Science*, pages 1–13, 2010.

[6] John Bowers and Gilliam Mould. Ambulatory care and orthopaedic capacity planning. *Health Care Management Science*, 8(1):41–47, 2005.

[7] T. Cayirli and E. Veral. Outpatient scheduling in health care: A review of literature. *Production and Operations Management*, 12(4):519–549, 2003.

[8] Vincent S Chow, Martin L Puterman, Neda Salehirad, Wenhai Huang, and Derek Atkins. Reducing surgical ward congestion through improved surgical scheduling and uncapacitated simulation. *Production and Operations Management*, 20(3):418–430, 2011.

[9] M.M. Connors. A stochastic elective admissions scheduling algorithm. *Health Services Research*, 5(4):308–319, 1970.

[10] Alexander Erdelyi and Huseyin Topaloglu. Approximate dynamic programming for dynamic capacity allocation with multiple priority levels. *IIE Transactions*, 43(2):129–142, 2010.

[11] B. Farrell, S. Kenyon, and H. Shakur. Managing clinical trials. *Trials*, 11(1):78, 2010.

[12] J.C. Felli, W.H. Anderson, J.P. Kremidas, and S.J. Ruberg. A semi-markov model for patient progression through clinical trials. *European journal of operational research*, 176(1):542–549, 2007.

[13] Caroline Fogarty and Patricia Cronin. Waiting for healthcare: a concept analysis. *Journal of advanced nursing*, 61(4):463–471, 2008.

[14] Yasin Gocgun and Archis Ghate. Lagrangian relaxation and constraint generation for allocation and advanced scheduling. *Computers & Operations Research*, 39(10):2323–2336, 2012.

[15] Yasin Gocgun and Martin L Puterman. Dynamic scheduling with due dates and time windows: an application to chemotherapy patient appointment booking. *Health care management science*, pages 1–17, 2013.

[16] Linda V Green, Joao Soares, James F Giglio, and Robert A Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, 2006.

[17] J.R. Griffith, W.M. Hancock, and F.C. Munson. *Cost Control in Hospitals*. Lippincott Williams & Wilkins, 1978.

[18] M. Grunow, H.O. Gunther, and G. Yang. Development of a decision support model for scheduling clinical studies and assigning medical personnel. *Health care management science*, 7(4):305–317, 2004.

[19] D. Gupta and B. Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40:800–819, 2008.

[20] D. Gupta, M.K. Natarajan, A. Gafni, L. Wang, D. Shilton, D. Holder, and S. Yusuf. Capacity planning for cardiac catheterization: a case study. *Health policy*, 82(1):1–11, 2007.

[21] Diwakar Gupta and Brian Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, 40(9):800–819, 2008.

[22] Randolph Hall. *Patient flow: reducing delay in healthcare delivery*, volume 206. Springer Science & Business Media, 2013.

[23] W.M. Hancock and P.F. Walter. The use of computer simulation to develop hospital systems. *SIGSIM Simul. Dig.*, 10(4):28–32, 1979.

[24] W.M. Hancock and P.F. Walter. *The "ASCS": Inpatient Admission Scheduling and Control System*. Health Administration Press, Ann Arbor, MI, 1983.

[25] C. Heine, R. Herrler, M. Petsch, and C. Anhalt. Adapt: adaptive multi-agent process planning & coordination of clinical trials. In *Proceedings of the 2003 Americas Conference on Information Systems*, pages 1823–1834. Citeseer, 2003.

[26] J.E. Helm, T. Rohleder, and M.P. Van Oyen. Priority scheduling in a queueing network with an application to itinerary completion at destination medical centers. 2015. working paper, Dept. of Industrial and Operations Engineering, University of Michigan.

[27] Jonathan E. Helm and Mark P. Van Oyen. Design and optimization methods for elective hospital admissions. *Operations Research*, 1(1):Forthcoming, 2014.

[28] Jonathan E Helm and Mark P Van Oyen. Design and optimization methods for hospital admissions. *Operations Research*, 62(6):1265–1282, 2014.

[29] J Herbots, W Herroelen, and R Leus. Single-pass and approximate dynamic-programming algorithms for order acceptance and capacity planning. *Journal of Heuristics*, 16(2):189–209, 2010.

[30] H-M Hsu and W-P Wang. Possibilistic programming in production planning of assemble-to-order environments. *Fuzzy sets and Systems*, 119(1):59–70, 2001.

[31] PJH Hulshof, RJ Boucherie, EW Hans, and JL Hurink. Tactical resource allocation and elective patient admission planning in care processes. *Health Care Management Science*, 16(2):152–166, 2013.

[32] P.J.H. Hulshof, N. Kortbeek, R.J. Boucherie, E.W. Hans, and P.J.M. Bakker. Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health systems*, 1(2):129–175, 2012.

[33] M.A. Johansen, D.K. Mayer, and HC Hoover Jr. Obstacles to implementing cancer clinical trials. In *Seminars in oncology nursing*, volume 7, page 260, 1991.

[34] Pooyan Kazemian, Yue Dong, Thomas R Rohleder, Jonathan E Helm, and Mark P Van Oyen. An ip-based healthcare provider shift design approach to minimize patient handoffs. *Health care management science*, 17(1):1–14, 2014.

[35] John W Kenagy, Donald M Berwick, and Miles F Shore. Service quality in health care. *JAMA*, 281(7):661–665, 1999.

[36] Robert Kocher and Nikhil R Sahni. Physicians versus hospitals as leaders of accountable care organizations. *New England Journal of Medicine*, 363(27):2579–2582, 2010.

[37] Naoru Koizumi, Eri Kuno, and Tony E Smith. Modeling patient flows using a queuing network with blocking. *Health Care Management Science*, 8(1):49–60, 2005.

[38] T. Lim, D. Uyeno, and I. Vertinsky. Hospital Admissions Systems: A Simulation Approach. *Simulation Gaming*, 6(2):188–201, 1975.

[39] B.M. McGarvey, N.J. Dynes, B.C. Lin, W.H. Anderson, J.P. Kremidas, and J.C. Felli. A discrete event model of clinical trial enrollment at eli lilly and company. In *Simulation Conference, 2007 Winter*, pages 1467–1474. IEEE, 2007.

[40] Michael L McManus, Michael C Long, Abbot Cooper, and Eugene Litvak. Queuing theory accurately models the need for critical care resources. *ANESTHESIOLOGY-PHILADELPHIA THEN HAGERSTOWN-*, 100(5):1271–1276, 2004.

[41] C.L. Meinert and S. Tonascia. *Clinical trials: design, conduct, and analysis*, volume 8. Oxford University Press, USA, 1986.

[42] J. Mula, R. Poler, J.P. Garcia-Sabater, and F.C. Lario. Models for production planning under uncertainty: A review. *International journal of production economics*, 103(1):271–285, 2006.

[43] Thomas W Nolan. *Reducing delays and waiting times throughout the healthcare system.* Institute for Healthcare Improvement, 1996.

[44] J. Patrick, M.L. Puterman, and M. Queyranne. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40:800–819, 2008.

[45] Jonathan Patrick, Martin L Puterman, and Maurice Queyranne. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations Research*, 56(6):1507–1525, 2008.

[46] C. Phillips, C. Stein, and J. Wein. Scheduling jobs that arrive over time. *Algorithms and Data Structures*, pages 86–97, 1995.

[47] M. Pinedo. *Scheduling: theory, algorithms, and systems.* Springer Verlag, 2008.

[48] Hajo A Reijers and S Liman Mansar. Best practices in business process redesign: an overview and qualitative evaluation of successful redesign heuristics. *Omega*, 33(4):283–306, 2005.

[49] C. Roland and P. Litka. Clinical trial operations. *Drug and Biological Development*, pages 178–201, 2007.

[50] ER Rubin, D. Lazar, N. Gaich, and D. Haray. The clinical trials landscape: Limitations, strengths, and promise. 2007.

[51] Antoine Sauré, Jonathan Patrick, Scott Tyldesley, and Martin L Puterman. Dynamic multi-appointment patient scheduling for radiation therapy. *European Journal of Operational Research*, 2012.

[52] S.C. Schoenbaum. Accountable care organizations: roles and opportunities for hospitals. *Hospital Practice (Minneapolis)*, 39:140–148, 2011.

[53] KT Talluri and GJ Van Ryzin. *The theory and practice of revenue management*, volume 68. Springer Science & Business Media, 2006.

[54] V.A. Varma, R. Uzsoy, J. Pekny, and G. Blau. Lagrangian heuristics for scheduling new product development projects in the pharmaceutical industry. *Journal of Heuristics*, 13(5):403–433, 2007.

[55] Jan Vissers and Roger Beech. *Health operations management: patient flow logistics in health care.* Psychology Press, 2005.

[56] Xiao-dan Wu, Mohammad T Khasawneh, Juan Hao, and Zhan-ting Gao. Outpatient scheduling in highly constrained environments: A literature review. In *The 19th International Conference on Industrial Engineering and Engineering Management*, pages 1203–1213. Springer, 2013.