Team-based Staffing Optimization for an Urgent and Primary Care Centre

by

Samantha L. Zimmerman

B.Sc., Simon Fraser University, 2019

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Mathematics
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2022

Declaration of Committee

Name:	**Samantha L. Zimmerman**
Degree:	**Master of Science**
Title:	**Team-based Staffing Optimization for an Urgent and Primary Care Centre**
Committee:	**Chair:** Tom Archibald Professor, Mathematics **Tamon Stephen** Co-Supervisor Professor, Mathematics **Alexander (Sandy) Rutherford** Co-Supervisor Adjunct Professor, Mathematics **Caroline Colijn** Committee Member Professor, Mathematics **JF Williams** Examiner Associate Professor, Mathematics

Abstract

Urgent and primary care centres (UPCCs) provide both walk-in services for urgent healthcare needs and booked appointments for longitudinal care. UPCCs utilize multi-disciplinary teams of healthcare professionals who collaborate to provide client care. This thesis develops a new approach to optimize team-based staffing at a UPCC in Vancouver, British Columbia. The core of the approach is a discrete event simulation that estimates client access indicators based on the UPCC operational profile and client visit data. The analysis compares two algorithms that minimize staffing levels subject to access targets given by the time-dependent expected proportion of simulated clients who leave due to a prolonged wait. One approach combines an extension of an iterative, simulation-based algorithm for small-interval staffing with an integer programming formulation for shift-based staffing. Another approach optimizes shift-based staffing through simulation optimization. Both approaches make staffing recommendations to improve care access.

Keywords: Urgent and primary care; Team-based care; Staffing optimization; Discrete event simulation; Simulation optimization

Acknowledgements

This research was conducted on the ancestral and unceded territory of the Musqueam, Squamish, and Tsleil-Waututh Nations.

The research topic in this thesis was defined in collaboration with Vancouver Coastal Health (VCH). In particular, I thank and acknowledge

- Andrew Day, Director of Primary Care
- Lori Benning, Director Strategy Deployment
- The staff team at the Urgent and Primary Care Centre in focus. In particular, I thank the management team including Ana Fernandes, Elaine Baxter, and Tim Henderson
- The VCH Decision Support team including Emilie Gladstone, Paige Knorr, Dylan Shard, and Eduardo Garza
- Michelle DeMoor at Vancouver General Hospital

Partial funding of this research was provided by Mitacs, the province of British Columbia, and the Natural Sciences and Engineering Research Council of Canada.

I thank my supervisors and thesis committee for their feedback and supervision.

I thank my parents, brothers, grandmothers, and roommates for their personal support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Urgent and Primary Care Centres and Team-based Care

Within healthcare, the term *primary care* refers to the provision of clinical services to address integrated, non-specialized care needs in a community context [166]. Community-based primary care has been highlighted widely as one approach to increase health equity in British Columbia (BC), Canada [144, 153]. However, many individuals in BC either do not have a family physician or are often unable to book appointments in a timely manner for unexpected health needs. Many individuals resort to emergency department (ED) visits for urgent, but non-emergent, level care, which can entail long wait times and diminished continuity of care compared to family doctor settings [151]. Furthermore, the challenges associated with accessing healthcare can be intensified for individuals experiencing marginalization [62, 90, 97, 115, 148].

A provincial initiative in BC aims to increase primary care access through additional urgent and primary care centres (UPCCs) that combine two healthcare modalities: urgent care and longitudinal primary care [134, 78]. Urgent care refers to healthcare for conditions requiring medical attention within 12 to 24 hours that do not need ED services. Table 1.1 lists for examples of urgent versus emergent care needs. UPCCs provide urgent care on a walk-in basis over extended hours to offer an alternative to non-emergent ED visits [78]. Longitudinal care services consist of primary care provision through booked follow-up appointments for registered clients. In the context of primary care, a panel is the group of clients who are registered or attached with a healthcare provider or team [19], and empanelment refers to the process of registering clients to a panel. UPCC empanelment is focused on individuals with complex biopsychosocial needs who have marked difficulties accessing fee-for-service family doctors [154], and the aim of the longitudinal care stream is to increase primary care access and health equity. While the two streams of urgent and longitudinal care are both aimed at boosting healthcare access, they are also distinct service paradigms. Urgent care is episodic, short term, and available for the general population, whereas longitudinal care is long term and restricted to clients with specific needs. The incorporation of both urgent and longitudinal care models into a single clinic setting is not operationally straightforward.

Medical Condition(s)	Urgent Care	Emergency Care
Sprains and strains	✓	
High fever	✓	
Suspected stroke or heart attack		✓
Asthma attack	✓	
Poisoning or overdose		✓
Major trauma		✓
Cuts, wounds, or skin conditions	✓	
Dehydration/constipation	✓	
Infections (chest, ear, or urinary tract)	✓	

Table 1.1: Examples of urgent and emergent care [51].

Furthermore, as a part of BC's strategy for primary healthcare, a strong emphasis is put on the expanded implementation of team-based care delivery [78]. Team-based care is recognized internationally as an important factor in effective primary care [19], and had a successful pilot implementation in the context of an urban community health centre in BC [153]. Under the team-based care model, a group of multi-disciplinary healthcare providers and staff collaborate to provide client-centred care. This increases the capacity and efficiency of care provision through workload sharing [60], as well as improves the ability to meet holistic client needs through the inclusion of allied health professionals, including social workers, dietitians, and physiotherapists [153]. While the potential value of team-based care is high, realization of these benefits depends on a successful and context-specific implementation. In the UPCC context, several operational questions are raised by the integration of team-based care within the combination of urgent and longitudinal care services.

1.2 Thesis Overview

The research topic in this thesis was defined in collaboration with the management team at an urgent and primary care centre (UPCC) in Vancouver, BC. One of the key operational questions that the UPCC management raised was how to determine team-based staffing. Current staffing is based on average demand projections made prior to the establishment of the UPCC. The main goal of this thesis project is to develop a new approach to optimize UPCC staffing based on client access indicators and current UPCC data. The analysis determines the staffing levels that are required for several staff disciplines to ensure that modeled client access is maintained in the urgent care stream. The core of my approach is a queueing model and an associated simulation that estimates time-dependent client-centred key performance indicators (KPIs) and incorporates the interaction between different provider disciplines at the UPCC. Optimization over simulation results informs staffing on both a small-interval and shift-based level, to determine the number of each staff discipline needed in each 15-minute interval and on each shift. My analysis does not incorporate the rostering of individual staff to specific shifts.

The analysis in this thesis makes staffing recommendations based on modeled client-centred access indicators and UPCC data. It provides a new approach to optimize urgent care staffing and further quantify team-based urgent care [153]. To do this, I introduce extensions of current stabilization techniques [44] to incorporate multiple staff types and observation based performance measures. My work contributes more broadly to staffing optimization techniques beyond healthcare applications.

This thesis is structured as follows. Chapter 2 gives background on operations research in primary care, queueing models, staffing models, KPI estimation, and simulation optimization. Chapter 3 describes the operational profile of the UPCC for both urgent and longitudinal care. Chapter 4 introduces the staffing models that I use to optimize team-based staffing in the urgent care stream at the UPCC, including: the underlying queueing model and simulation implementation; KPI definitions; and optimization procedures for both small-interval and shift-based staffing. Chapter 5 describes data analysis and parameter estimation. Chapter 6 presents the simulation validation and staffing optimization results, which are both discussed in Chapter 7 along with recommendations for UPCC staffing and broader conclusions.

Chapter 2

Background

2.1 Operations Research in Primary Care

Operations research to improve access and efficient resource utilization in service delivery settings includes two distinct paradigms: capacity planning and demand regulation. Capacity planning approaches optimize the supply of staff or other resources to meet demand, by changing overall staffing levels and the distribution of capacity over time. An alternative approach is to regulate demand to match a given supply, in general by determining how many clients or tasks can be accepted and when they should be scheduled. The suitability and specific mechanisms used in each approach are highly context dependent. In emergency department settings, where patient arrivals are typically unregulated and uncertain, optimization of medical personnel schedules has been extensively studied [25, 34, 81]. In comparison, operations research studies in primary care and other outpatient settings have predominantly focused on optimizing panel design or appointment scheduling procedures [3, 30, 74, 86, 92].

In primary care, one mechanism to match demand with supply is through panel optimization [70, 129, 136]. Effective panel design can adjust the volume of client demand to match available capacity [129], and can also address modeled stochastic outcomes including client no-shows, physician over-time, and client access targets [71, 70, 175]. Panel optimization can incorporate heterogeneity in client needs through analysis of case-mix [136]. Some studies address the joint optimization of panel design and appointment scheduling [108, 174].

Another important factor of primary care performance is appointment scheduling, which can influence the timing of client arrivals so that they align with capacity and improve access [74, 92]. Appointment systems have multiple aspects that can be optimized to improve efficiency and quality of care, for example: the number of appointments scheduled in a day [108, 174]; the duration of each appointment slot [17, 31, 33, 98]; the number of clients scheduled in each time slot [167]; and the relative ordering of different appointment categories [17, 31, 32]. The complexity of optimizing appointment schedules increases when stochasticity is considered in factors including client no-shows [98, 167], cancelations, punctuality, service times [17, 31] and appointment preferences [52]. Many studies suggest that optimum rules for appointment scheduling can be highly context specific [32, 33, 29].

One paradigm to improve primary care access for urgent clients is to offer a combination of both pre-booked appointments as well as either same-day appointments [130] or walk-in services. In these settings, efficiency and quality of care can be increased by optimizing the allocation of appointment slots between pre-booked and same day and/or walk-in visits [105, 152], as well as which slots should be reserved for urgent care or overbooked [29, 28, 99, 142, 167]. Optimal allocation between booked and urgent care is further complicated by the interrelatedness of these care streams, since clients can strategically choose between services [164].

Capacity planning through staffing optimization represents an alternate, "underexposed" [80] approach to improve outpatient healthcare services, especially given the diverse range in operational profiles. Studies that analyze staffing in primary care settings focus on settings such as urgent care centres [160], community health centres [176], private practices [106, 107], family planning clinics [7, 18, 55, 79], and rural health centres [104]. Other scheduling studies for outpatient services include pharmacies [128], telemedical appointments [101], and mobile examination centres [135]. Tan et al. [160] and Zimmerman et al. [176] consider systems with multiple staffing disciplines but only optimize staffing for a single care provider type. Franz et al. [55] schedule multiple types of providers across a network of family planning clinics; however, their integer linear program requires predetermined staffing ratios and requirements that are not known in advance in the UPCC context. Hudgins et al. [79] derive personnel requirements for multiple staffing disciplines in a family planning clinic based on multiplication of average staff time needed for each task, and Boaz [18] optimizes staff-mix using a modified exponential function to estimate the marginal patient costs for each discipline; however, neither approach optimizes staffing based on stochastic client access. Liu and D'Aunno (2012) [106] and Liu et al. (2014) [107] analyze the long-run cost effectiveness of a small number of team-based care configurations. To my knowledge, there are no existing studies that optimize primary care staffing for multiple care provider types using time-dependent client access metrics. Furthermore, the quantification of team-based care is an acknowledged need within primary care [153].

2.2 Queueing Models

This section provides a brief introduction to queueing models, which are used widely in both healthcare modeling [80] and staffing models [46]. Queueing models (or systems) are stochastic processes used to represent clients requesting, waiting for and receiving some type of service [122]. Model clients can represent customers, patients, or any individual requesting service. In this thesis, the terms *model client(s)* or *modeled client(s)* will be used to refer to queueing model representations of clients, which should be distinguished from

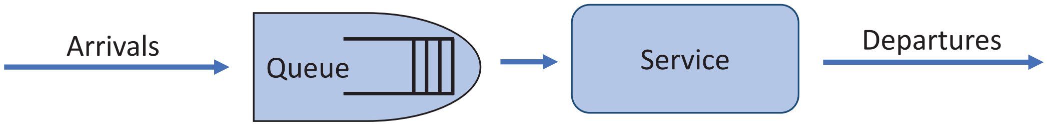

Figure 2.1: Diagram of a single service queueing model. Model clients arrive to request service and wait in a queue if no model servers are available. Once a model server is available they receive service and depart the system after service completion.

real-world clients or customers. Similarly, the phrase *modeled service* refers to queueing model representations of services.

The exact nature of the service represented can vary between model applications, ranging from the provision of healthcare treatment by a physician [73], to talking with a call centre associate [20], to using a telephone line [49]. The main aspect of service modeled in a queueing system is the duration of time until service completion, which is referred to as service time and represented as a non-negative continuous random variable [122]. Service providers are represented as *model servers*, which can be used to represent staffing personnel or other resources and infrastructure [122]. Model servers in the process of providing service are referred to as *busy*. Requests for service are modeled as arrivals, and represented using a stochastic process referred to as the *arrival process* for the number of cumulative arrivals at any point in time [122]. Model client arrivals that find all model servers busy will form or join a queue to wait until a model server becomes available. Once a model server is available, a model client will begin service for a random duration of service time. After a model client's request for service is completed, they will depart from the system. Figure 2.1 depicts a basic queueing system that incorporates arrivals, queueing, service, and departures.

There are three core parameters for a queueing model, namely the arrival process, service time distribution, and number of model servers. If the arrival rate, service rate and number of servers are all constant, the system is considered *homogeneous*. If a queueing model has exponentially distributed service time distributions and arrivals defined by a Poisson process, then it has the Markovian property, which means that future events depend only on the current system state, and not past events [123]. Service times or arrival processes that are not assumed to follow an exponential distribution or Poisson process, respectively, are referred to as general, or *generally distributed* [122]. Phase-type distributions refer to a class of distributions that can be constructed by combining multiple exponential distributions in convolution (series) or mixture (parallel) [77].

Given any queueing model parameterization, there are time-dependent probability distributions for each system performance aspect, including the number of modeled clients in the system, the number of these clients in the queue, and client waiting times. Under certain conditions, queueing systems will approach a state of stochastic equilibrium, where distributions for performance measures become constant in time [122]. Equilibrium distributions are known as *steady-state* or *stationary* distributions, whereas non-equilibrium performance is referred to as *transient* or *time-dependent* system behavior [122]. For some queueing systems there are exact analytical formulae for steady-state distributions, whereas more complex systems require approximation of steady-state values.

Some queueing models can incorporate the event of client *abandonment*, where model clients may leave the queue before receiving service due to an extended wait [119]. In some models, clients may return later to the queue in a *retrial*, whereas in other models they leave the system altogether [119]. Abandonment can be modeled by utilizing a random variable for the amount of time that clients are willing to wait, referred to as *patience* or *willingness to wait*, which has some assumed distribution [119]. Outcomes for model clients, including abandonment and extended waiting, are stochastic performance indicators that can be used to represent or approximate real-world service quality and client access indicators [120].

Another important queueing model parameter is the number of modeled servers. In infinite server queueing models, there are no capacity limits and all model clients receive service immediately upon their arrival [48]. Finite server queueing models are able to address capacity dependent outcomes for model clients, including extended waiting and abandonment. Queueing models can also incorporate a time-dependent number of modeled servers, where capacity changes represent fluctuations in staff scheduling or other changes in infrastructure availability. During a staffing level reduction, model servers that are scheduled to cease operation may remain longer to complete the service time of model clients already being seen, under what is referred to as an *exhaustive* service policy [44, 45, 82]. Alternatively, under a *non-exhaustive* or *preemptive* service policy, interrupted model clients may rejoin the queue and complete their remaining service time with the next available model server [83].

Another queueing model aspect is the *queueing discipline*, which determines the order in which modeled clients are seen. Under a *first come first served* (FCFS) queue discipline, modeled clients are seen in the order of their arrival. Alternatively, some queueing models use a random order of service, or a *last come first served* queue discipline [122]. Some queueing models consider different priority classes of customers and see clients in order of priority, with a FCFS discipline used within each priority class. Under a priority based model, metrics for modeled client outcomes can be different within each priority class [157]. In accumulating priority queues, the priority value of modeled clients increases over time [36, 157].

2.3 Staffing Models

There are several different factors that can be used to inform staffing decisions, including staffing costs, labor regulations, staff scheduling preferences, staff workload [54], and client access indicators. A common staffing model approach is to represent services with a queueing model and optimize the number of model servers, subject to different constraints and objective functions. In particular, the queueing model approach can be used to address stochastic performance measures including client-centred access indicators and system utilization targets. In general, there are four distinct decisions that queue-based staffing models can optimize [22, 46], namely to determine:

1. *Steady-state staffing*: The optimum number of constant model servers required to address steady-state performance measures.
2. *Small-interval staffing*: The optimum number of model servers required in each time interval to address time-dependent performance measures, without consideration for shift options. This problem is sometimes referred to as “stabilization”.
3. *Shift-based staffing*: The optimum number of staff required on each shift in each day in the planning horizon, without consideration for the assignment of individual shifts to staff.
4. *Roster-based staffing*: The optimum workforce size needed based on the rostering or assignment of individual staff to specific shifts.

Further discussion for each of these staffing optimization problems is included in Subsections 2.3.1, 2.3.2, 2.3.3, and 2.3.4, respectively.

2.3.1 Steady-state Staffing

One approach to staffing, or capacity planning, is to identify the number of servers required so that constraints are met on steady-state performance metrics. For systems with exact solutions, required staffing levels can be identified by optimizing over steady-state formulae; however, this approach is less efficient under heavy demand, and complex systems may not have exact formulae. One approximation is the *square-root staffing rule*, as part of the *quality and efficiency driven* (QED) staffing regime developed by Halfin and Whit [75]. This approach is based on meeting targets for the steady-state probability of delay as the arrival rate and number of servers grow arbitrarily large. In this asymptotic regime, the optimum number of staff can be approximately calculated from the arrival rate using a linear and square-root term, with the root term coefficient coming from a root solve involving other system parameters [75]. Garnet et al. [57] and Mandelbaum and Zeltyn [120] extend this regime to consider additional constraints on the probability of abandonment in queues with generally distributed abandonment times. Because of the underlying asymptotic limit, the QED approximations is less accurate in small-scale systems [44] and more accurate in systems with large numbers of servers and high arrival rates, for example call centres [57, 120]. The QED regime has been extended to consider staffing for multiple classes of servers in the specific case where all server types can be used interchangeably by homogeneous model customers [8].

Ahmed and Alkhamis [4] determine ED staffing that maximizes the simulated throughput given constraints on the budget and wait time. They identify constant staffing levels for multiple providers by comparing all possible staffing combinations.

2.3.2 Small-interval Staffing

One approach to small-interval staffing requirements for time-dependent queueing models is the *stationary independent period by period* (SIPP) staffing approach and its extensions [63, 66, 67, 73]. In the SIPP framework, staffing requirements are set to the number of model servers required in stationary queue approximations for each time period [63]. Each proxy steady-state model is based on the arrival rate in its corresponding time interval and is independent of system performance in other periods. The SIPP approach is known to perform poorly in systems with rapidly changing arrival rates, multiple demand peaks, long service times, and limited opening hours [66, 67, 83]. In the extended lag-SIPP approach, arrival rate consideration intervals are lagged by the mean service time to incorporate the carry-over of demand between intervals [66, 67]. Both SIPP and lag-SIPP optimize staffing independently each time interval and do not incorporate the impact staffing decisions in other intervals, which can lead to less accurate results [83]. Furthermore, in systems with finite opening hours and client abandonment, steady-state equations are not able to capture the clients who are not seen because of the end of the day.

Two other approaches to time-dependent small-interval staffing are the *infinite server* (IS) and *modified offered load* (MOL) approximations [52, 91]. In both IS and MOL, staffing decisions are informed by a corresponding time-dependent queueing model with an infinite number of servers. The IS approach chooses staffing levels based on stabilizing the time-dependent probability that a model client needs to wait for service, by directly using the probability that staffing levels are exceeded in the infinite server queue [48, 52, 91]. The MOL approach extends this concept to other performance measures by using proxy steady-state queueing models based on the expected number of model customers in the infinite server system at each point in time [52, 91]. The MOL approach incorporates the interplay between service time for past arrivals under a time-dependent arrival rate, but does not consider the effects of staffing choices in other time intervals. Sinreich and Jabali [155] determine small-interval staffing for multiple medical resources in a ED care network based on the maximum expected number of busy model servers in each interval of an infinite server simulation.

Kim and Ha [95] propose the *consecutive staffing using simulation* (CSS) algorithm, which determines the staffing in each interval separately and sequentially using a linear search in each interval. Corominas and Lusa [39] use a similar technique, which is modified to start with steady-state results as initial conditions.

Feldman et al. [52] propose the *iterative staffing algorithm* (ISA) to stabilize the probability of delay by using simulation estimates for the time-dependent distribution of the number of customers in the system at each point in time. The ISA approach begins with an arbitrarily large number of servers, so that first step of the procedure is equivalent to the IS approach. ISA extends this by re-running the simulation after each staffing update, to incorporate the impact of time-dependent staffing from previous iterations. Defraeye and Van Nieuwenhuyse [44] extend the ISA approach to determine the small-interval staffing levels required to meet a target on the time-dependent probability that a modeled client has a prolonged wait.

2.3.3 Shift-based Staffing

One approach to shift-based staffing is through integer linear programming (ILP)[1], where small-interval staffing requirements are used as model constraints or targets. The first ILP staffing model formulation is Dantzig's set covering formulation [42], which continues to be applied in healthcare contexts [21, 45]. Dantzig's staffing model uses explicit variables for each shift and break option, and identifies the lowest cost shift combination that satisfies all small-interval staffing requirements [42]. To improve performance and flexibility for large numbers of shift options, Dantzig's ILP can be extended to represent shifts implicitly by introducing variables for the start and end time of shifts, as well as break placement [11, 14, 127, 146, 161]. An alternate ILP staffing objective is to minimize the difference between shift-based staffing levels and small-interval staffing targets [22]. The ILP staffing approach addresses performance targets indirectly through the choice of small-interval staffing requirements, and does not incorporate the impact of under or over staffing in each interval on performance measures [45, 84, 83].

An alternate approach to shift-based staffing optimization draws on performance measure estimation for each combination of shifts. Further details about performance measure estimation techniques are discussed in Section 2.4. The stochasticity and non-linearity of performance measure constraints or objectives can be addressed by coupling KPI estimation with either meta-heuristic optimization [84, 173] or ILP-based heuristics [9, 10, 83]. In optimization, a solution is referred to as a *local optimizer* if it has the best objective function value compared to nearby solutions, but is not necessarily a *global optimizer* over the entire

[1] Integer linear programming is an optimization problem with integer decision variables and a linear objective function and constraint set; solutions must satisfy all constraints and either minimize or maximize the objective function value.

set of feasible solutions. Meta-heuristic frameworks coordinate strategic interaction between local optimization searches [2] and mechanisms to explore other neighborhoods in the search space and try to find global optima [58]. Meta-heuristics do not make assumptions about the structure of the problem, and are typically used when the structure of the objective function or constraint set is not well known and could have several local optima. Meta-heuristics balance the trade off between efficiency and accuracy and are not guaranteed find global optima. An example of a meta-heuristic is the *genetic algorithm*, which maintains a population of candidate solutions and combines subsets of these solutions to create new ones based on some measure of fitness [169]. Ingolfsson et al. [84] apply a genetic algorithm to optimize the staffing of police patrol shifts to minimize both cost and extended waiting probabilities. Yeh and Lin [173] use a genetic algorithm to schedule ED nurse shifts that minimize wait times within budget constraints.

Several heuristic algorithms have been developed for shift-based staffing that exploit the specific objective function and constraint structure of the problem to increase the accuracy and efficiency of optimization. One approach is to iteratively combine ILP solutions with performance measure estimates by sequentially adding cutting plane constraints that remove unsatisfactory solutions while maintaining the feasibility of potentially optimum solutions [9, 10, 83]. For example, Atlason et al. [9] iteratively add boundary cutting-plane constraints in staffing intervals with unsatisfactory performance by using estimated sub-gradients[3] of the concave and increasing KPI surface. Atlason et al. [10] extend this approach to KPI functions that are pseudo-concave increasing by using an interior point cutting-plane approach. Ingolfsson et al. [83] base cutting plane constraints for shift-based staffing optimization on an exponential approximation for performance measure improvement.

Defraeye and Van Nieuwenhuyse [45] use an ILP solution for shift-based staffing as an initial conditions in a branch and bound[4] search procedure over simulation results. Castillo et al. [27] present a framework to optimize shift-based staffing for a general combination of competing performance measures, where simulated KPIs for a number of generated plausible schedules are used to estimate an efficient frontier.

[2] Local optimization procedures search close to existing solutions in order to find new solutions that improve the objective function value.

[3] In this context a sub-gradient (or more literally a *super-gradient*) is a hyper-plane which lies above the surface for the KPI value in each interval [9].

[4] The branch and bound approach divides a discrete feasible region using a rooted decision tree with a node for each solution. The optimization search procedure begins with the root and sequentially branches to explore child nodes, while choosing to bound the search by not exploring further sub-nodes whose main node has poor performance [45].

2.3.4 Roster-based Staffing

The translation of shift-based staffing to shift assignment and workforce size is not straightforward because of the need to consider numerous constraints on staff rostering [50]. Labor regulations can include, for example: rules around the maximum and minimum number of shifts per period; the maximum and minimum consecutive number of days working or days of; and the minimum number of hours between two assigned shifts for each worker [34]. Scheduling objectives can include minimizing staffing costs [12, 21, 24, 41, 89], appropriate demand coverage [89, 126, 162] and maximizing the satisfaction of nurse preferences [6, 15, 89, 126, 162] including fairness [12]. Staffing costs can include full and part-time salaries as well as the cost of overtime and call-in staffing. Demand coverage can be incorporated through shift or small-interval based staffing requirements [126, 125, 162] or KPI estimation for each schedule [173, 109]. Nurse preferences and satisfaction can incorporate requests for days on or off. The trade offs between these cost, coverage, and satisfaction can be incorporated using weighted objective functions [89, 126, 138, 162] or multi-objective approaches [15, 47].

A fundamental approach is to formulate rostering as a mathematical program [34] with binary variables representing the assignment of shifts to staff. However, the numerous constraints needed for realistic staff rostering can pose challenges for efficient staffing optimization [25, 50]. One approach to address this is to relax a subset of the constraints and incorporate them into the objective using penalty functions [34] that encourage satisfactory solutions. Meta-heuristic optimization can be used to address non-linear objective functions [173, 109], or improve solution efficiency and tractability as the scale of the rostering application increases. The diverse range of meta-heuristic algorithms applied to rostering including the genetic algorithm [5, 6, 173], tabu search[5] [15, 23, 47], and variable neighborhood search [6] [41, 145, 109] and OptQuest optimization [7] [93, 140, 139]. Hybrid approaches can combine meta-heuristics with mathematical programming to boost the efficiency and accuracy of optimization [41, 145].

The rostering of nurses in hospital settings has been extensively studied, to the extent that the *nurse rostering problem* (NRP) is quintessential operations research problem [25, 34, 50]. However, advances in the literature to the NRP tend to focus on developing new optimization techniques for simplified problems, as opposed to considering the different operational profiles or the complexities needed for implementation [143]. The *nurse re-rostering problem* (NRRP) optimizes schedule adjustments in response to changing staff

[5] Tabu search modifies a local search procedure by uses a short term memory to restrict repeated changes [59].

[6] Variable neighborhood search iteratively identifies the best solution in a particular search region, which is systematically changed to avoid local optima [76].

[7] See Subsection 2.5 for more background on the OptQuest optimization algorithm.

availability in order to minimize disruption [171], navigate fairness [172], and minimize the use of call-in workers [13].

2.4 Performance Measure Estimation in Queueing Models

In queueing models, system performance can be measured by a number of different aspects, including [46]: model client wait times and the time clients spend in the system; whether or not model clients receive service or leave without being seen; the number of clients in the system; system utilization; and model server idle time. All of these aspects are stochastic in nature and have time-dependent distributions. Key performance indicators (KPIs) are used to summarize these distributions, for example using means, outcome probabilities, proportions, and rates [46]. Queueing model KPIs based on client outcomes including abandonment or prolonged waiting can be used to represent or service access and quality. The remainder of this section describes several different methods for the estimation or approximation of time-dependent performance measures in queueing models.

One group of approaches to obtain proximate time-dependent properties is by applying steady-state formulae independently in each considered time interval [52, 63, 64, 65, 69, 87, 91, 121]. The *pointwise stationary approximation* substitutes time-dependent arrival rates directly into steady-state formulae [63, 64, 65, 69] and the modified offered load approximation applies these formulae instead using the expected number of busy servers in in infinite server model[52, 69, 87, 91, 121]. Steady-state based approaches are known to be less accurate for rapidly changing arrival rates [68, 63, 67, 69, 82, 177] and rely on established exact or approximate steady-state formulae [75, 120].

Another group of approaches represent time-dependent queue model properties using a system of differential equations [37, 40, 43, 72, 84, 82, 83, 131, 150, 159]. For Markovian queue models, the *Chapman-Kolmogorov equations* provide an exact system of ordinary differential equations (ODEs) that govern the transitions between each queueing system state [43, 72, 84, 82]. However, numerical solutions to this ODE system of can be computationally inefficient as the system size grows [82, 85]. Approaches that approximate the ODE solution include *randomization*, which truncates the state space [40, 82, 83], and *closure approximation*, which uses a small ODE system to represent only the initial moments for the number of clients in the system [37, 131, 150, 159].

An alternate approach is to apply continuous representations of queueing models. *Fluid approximations* represent queue model properties with a deterministic continuous process that capture the time-dependent difference between queue arrivals (inflow) and departures (outflow) [116], and is suitable when changes in arrival rate dominate the stochastic variation in the queue [112]. *Diffusion approximations* use reflected Brownian motion to approximate the stochastic fluctuation in both the arrival and departure process [116], which can be specified using partial differential equations [132]. Fluid and diffusion models have the flexibility to capture systems with general arrival processes [110, 111, 112, 113, 114], general service distributions [112, 113, 170], general abandonment distributions [110, 111, 112, 113, 114, 170], client retrials [2, 117, 118], networks of services [111, 117], and multiple customer classes [147]. The accuracy of these approaches is reliant on assumptions about whether the queueing system is over, under, or critically loaded [35, 100, 111, 112, 114, 118].

Time-dependent queue model performance can also be estimated by simulation, where the stochastic processes for different events (including arrivals, service initialization, and departures) are sampled using pseudorandom number generation [149]. *Discrete event simulation* (DES) records the simulated system state after each of these generated events, which produces a potential sequence of outcomes [149]. Alternatively, *discrete time simulation* (DTS) updates the simulated system state at regular time intervals [26]. Under both DES and DTS, repeated samples from numerous simulation runs can be combined to produce KPI estimates. DES estimates for performance measures are typically more accurate the DTS estimates [26]; however, the time-step parameter in DTS can be strategically chosen to increase simulation efficiency [26]. Both DES and DTS have the flexibility to accurately capture complex queueing systems.

For example, if the arrival process in a queueing model is a homogeneous Poisson process, then DES will use pseudorandom number generation to repeatedly sample an exponential distribution with a constant rate. These samples are then used as inter-arrival times, which determine the exact times that model customers arrive in a generated event sequence. One approach to simulate a non-homogeneous Poisson process is to generate a homogeneous Poisson process with an inflated rate, then to accept or reject each potential arrival according to the output of a generated Bernoulli experiment with a time-dependent probability of success based on the ratio of the non-homogeneous rate and the inflated homogeneous rate [149]. This approach is referred to as *Poisson thinning*, and relies on the decomposition property of Poisson processes, under which a randomly selected subset of a Poisson process is another Poisson process [123].

2.5 The OptQuest Meta-heuristic

The OptQuest optimization engine is a commercially available tool to optimize stochastic performance measures by integrating simulation and optimization. Published OptQuest applications include financial risk management [14, 16] and scheduling to address wait times [53, 93, 140]. The OptQuest algorithm primarily uses the *scatter search* meta-heuristic framework [102] that builds and updates a small reference set of solutions, chosen for both quality and diversity [61, 103]; subsets of reference set solutions are then combined and improved to produce new solution candidates [61, 103]. Within this framework, the OptQuest engine utilizes a range of methods for combining and generating solutions, including gradient approximations and genetic algorithm procedures [102]. The OptQuest engine draws on machine learning models, for example multiple linear regression[8], to predict simulation outcomes and avoid running simulations on variables that have poor predicted performance [102]. Stochastic constraints are incorporated as penalty functions [102].

[8] Multiple linear regression fits a linear function of input data variables an output data variable by determining the coefficients that minimize the squared error in the predicted output [88].

Chapter 3

UPCC Operational Profile

This chapter describes the operational profile of the UPCC for both urgent and longitudinal care services, in Sections 3.1 and 3.2 respectively. The client flow diagrams represent a simplification of clinic operations, and were developed and affirmed through direct collaboration with UPCC management. To further understand clinic operations, I spent time shadowing different staff at the UPCC. Currently only the urgent care stream is operating at the UPCC, with longitudinal care services scheduled to begin operation in 2022. The description of longitudinal care in Section 3.2 represents the current operational plan for this care stream. Section 3.3 describes the interaction between the two streams of care. Table 3.1 lists types of care providers and staff at the UPCC and their corresponding acronyms, which are used throughout this thesis. At the UPCC, general and nurse practitioners have the same work scope and are collectively referred to as *most responsible practitioners* (MRPs).

Staff Discipline	Acronym
Medical office assistant	MOA
Registered nurse	RN
Nurse practitioner	NP
General practitioner	GP
Most responsible practitioner (Either a GP or NP)	MRP
Social worker	SW
Registered Dietitian	RD
Registered Clinical Counselor	RCC

Table 3.1: Clinical and non-clinical service disciplines at the UPCC and their corresponding acronyms.

3.1 Urgent Care

Urgent care at the UPCC is open seven days of the week, from 8am to 10pm on Mondays through Saturdays, and 9am to 5pm on Sundays and holidays. Registration closes at least an hour before closing time. Figure 3.1 provides a visual representation of client flow through

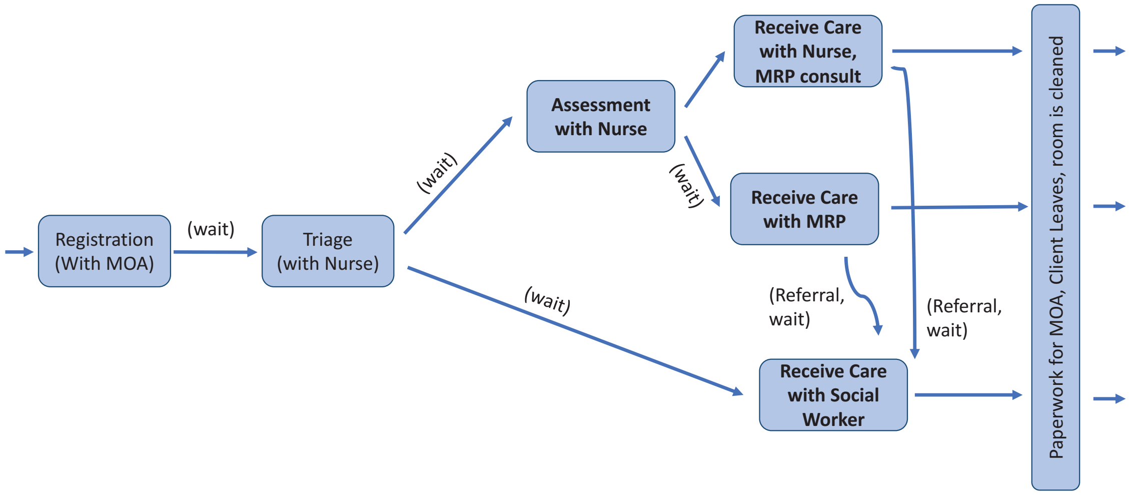

Figure 3.1: Client flow diagram for the urgent care stream. See Table 3.1 for the acronyms of clinical and non-clinical staff.

the urgent care stream. Upon arrival at the UPCC, clients take a number and wait to be registered with an MOA at the front desk. After registration, clients receive triage from an RN, where they are assigned a priority score using the *Canadian triage and acuity scale* (CTAS) [133]. Clients then wait until an exam room becomes available, are assessed by an RN in an exam room, and receive medical care or treatment. Depending on their care needs, clients receive treatment from either an MRP or RN with the consultation of an MRP. There may be a non-negligible wait time between each of these steps due to staff availability. Clients are seen in order of decreasing acuity, with earlier arrivals being seen first within each acuity group. After the client receives treatment, most clients leave the UPCC and their exam room is cleaned by a housekeeper. During treatment or assessment, an RN or MRP may refer clients to the on-site SW. Clients may receive care from a SW during their visit or book an appointment for another day. Clients arriving for a pre-booked SW appointment will bypass registration and triage. Each client visit may generate paperwork that needs to be completed by an MOA, including referrals or prescriptions.

Within urgent care staff team, there are four clinical staff disciplines (RNs, GPs, NPs and SW) and one non-clinical staff discipline (MOAs). The UPCC typically has at least two MOAs on duty at most times, one MOA at the front desk to perform registrations and one MOA to complete client paperwork. The *scope of practice* for RNs is provincially regulated and determines the extent of medical care that they are licensed to provide [1].

Some of the team-based care interactions in the urgent care stream include:

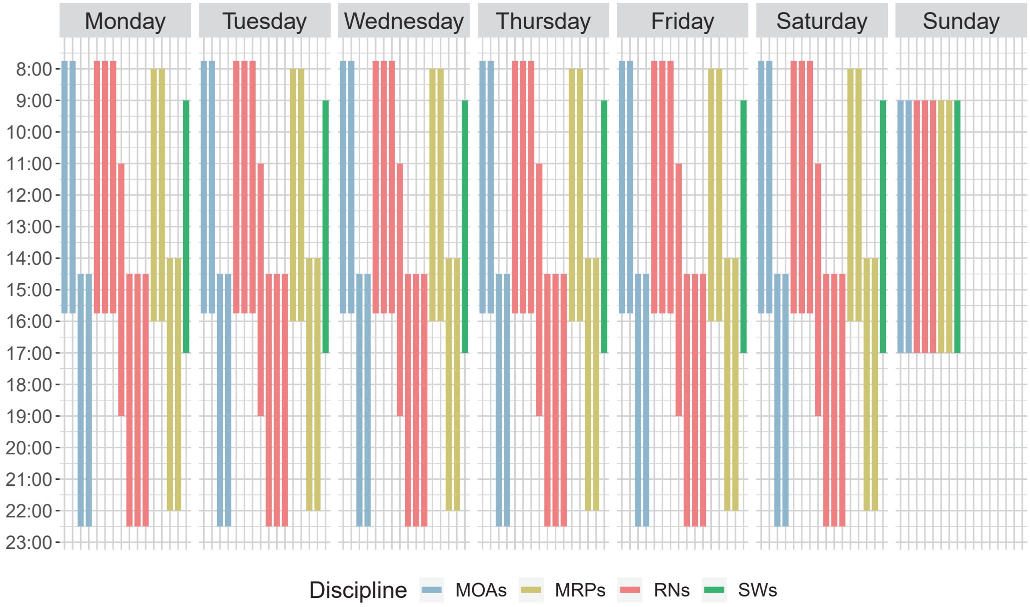

Figure 3.2: Current staff schedule for urgent care at the UPCC.

- The RNs, GPs, NPs and one MOA sit in a single pod of desks and the clinical staff will give and receive feedback on care provision, occasionally working together to provide treatment.
- NPs are able to provide the same scope of practice as GPs.
- RNs work within their full scope of practice, and are able to provide treatment for some client needs and initiate medical orders, working in consultation with an MRP.
- MRPs and RNs can both refer clients to an on-site SW.

Since the UPCC offers urgent care over extended hours, staff are assigned to one of multiple shifts that are combined to provide coverage throughout the day. Table 3.2 lists the currently used shift options for urgent care. Each shift includes one half hour lunch break and two 15 minute coffee breaks. The current UPCC staff schedule is shown on Figure 3.2.

Shift Name	Start Time	End Time
Morning (RNs & MOAs)	7:45 am	3:45 pm
Morning (MRPs)	8 am	4 pm
Day	9:00 am	5:00 pm
Afternoon	11:00 am	7:00 pm
Evening (RNs & MOAs)	2:30 pm	10:30 pm
Evening (MRPs)	2 pm	10 pm

Table 3.2: Shift options for urgent care.

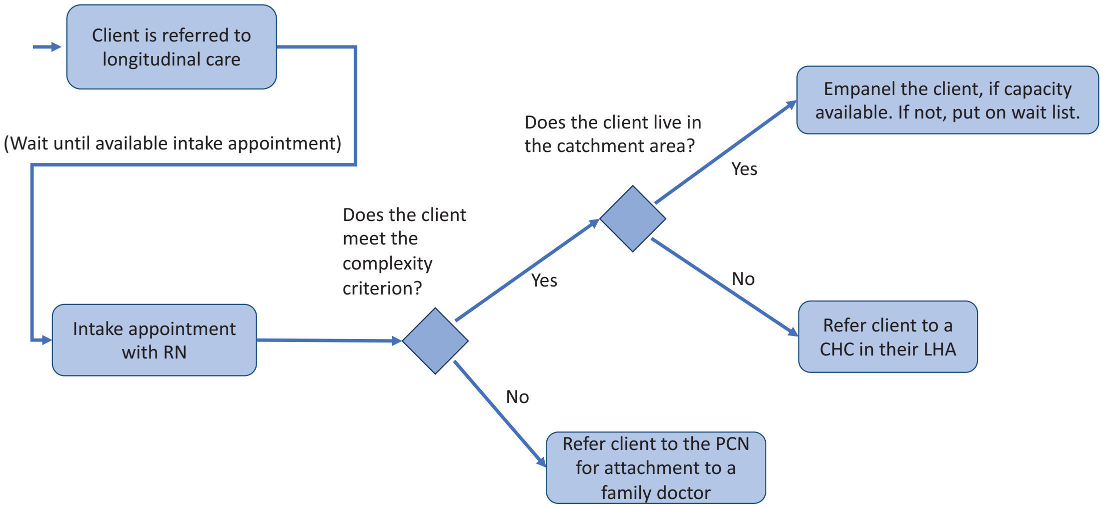
Client is referred to longitudinal care
(Wait until available intake appointment)
Intake appointment with RN
Does the client meet the complexity criterion?
Yes
No
Refer client to the PCN for attachment to a family doctor
Does the client live in the catchment area?
Yes
No
Empanel the client, if capacity available. If not, put on wait list.
Refer client to a CHC in their LHA

Figure 3.3: Diagram of the empanelment process for UPCC longitudinal care. PCN = primary care network, CHC = community health centre, LHA = local health authority.

3.2 Longitudinal Care

Longitudinal care services at the UPCC will be open seven days of the week, from 9:00 am to 5:00 pm. Unlike the urgent care stream in which client flow is episodic and begins upon arrival at the clinic, longitudinal care client interaction involves the processes of empanelment, appointment booking, and then the client visit. These processes are described in Subsections 3.2.1, 3.2.2 and 3.2.3, respectively.

3.2.1 Empanelment and Complexity

Before clients receive care through the UPCC longitudinal care stream, they must be registered as part of one of the UPCC panels. The process of empanelment is illustrated in Figure 3.3. Empanelment is initiated when a potential client is referred to the UPCC through one of four sources: the urgent care stream at the UPCC, the primary care network (PCN), a local emergency department (ED), or another community organization. Referred potential clients have an intake appointment where an RN determines whether or not they meet the criteria for UPCC empanelment. The empanelment criteria are described in Table 3.3, and are based on both the complexity of care needed and the geographic catchment area of the UPCC. If a potential client meets both sets of criteria, they are assigned to the panel of an appropriate UPCC clinician or put on a wait list if capacity is not available. If the potential client does not meet the empanelment criteria, they will be referred for primary care attachment through an appropriate channel, that could be a community health centre

Category	Criteria
Geographic	Home address in the same local health authority as the UPCC or No fixed address, and uses social support in the same local health area
Biopyschosocial Complexity	Meets several of the following criteria: 1. Unattached or poorly attached to a family physician despite a need for primary care. 2. Experiencing a period of functional instability that is challenging to manage within a fee-for-service practice. 3. Multiple social barriers, such as housing instability, poverty, etc. that impact connection to care. 4. Marked difficulties in accessing the fee-for-service health care system due to significant cognitive, behavioral, and/or functional impairment. 5. Inability to maintain lasting personal or professional relationships. 6. Marked difficulties with activities of daily living without access to appropriate supports. 7. Medically complex conditions presenting with chronic disease, concurrent disorders or communicable diseases (for example diabetes, hepatitis, HIV, mental health issues, substance misuse) that are untreated or uncontrolled. 8. High emergency department use for issues that could be addressed in the primary care setting and/or frequent acute care admission/readmission rates. 9. Risk of causing harm to self or others.

Table 3.3: Potential empanelment criteria for longitudinal care at the UPCC. Markers of biopsychosocial complexity are quoted from [154].

(CHC), another UPCC closer to the clients home address, or a local PCN that facilitates attachment to a fee-for-service family physician.

Longitudinal care at the UPCC is targeted for individuals who live in the same area as the clinic, have marked difficulties accessing fee-for-service primary care and have complex biopsychosocial needs given described in Table 3.3 and [154]. Shukor et al. [154] developed a quantitative approach to combine these measures into a single complexity score.

The UPCC intake nurse must evaluate primary care attachment based on the relative complexity and priority of potential clients, and also determine whether a physician has capacity available for a given client. Target panel sizes that do not consider client complexity may underestimate the available capacity, and overwork clinical staff.

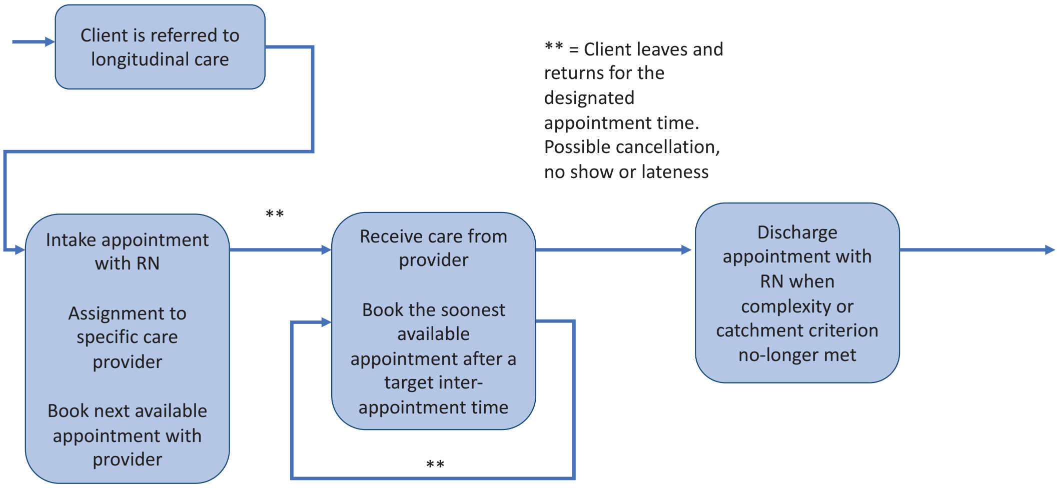
Client is referred to longitudinal care
** = Client leaves and returns for the designated appointment time. Possible cancellation, no show or lateness
**
Intake appointment with RN
Assignment to specific care provider
Book next available appointment with provider
Receive care from provider
Book the soonest available appointment after a target inter-appointment time
**
Discharge appointment with RN when complexity or catchment criterion no-longer met

Figure 3.4: Diagram of the potential appointment booking process for most empanelled clients in longitudinal care.

3.2.2 Appointment Frequency

Figure 3.4 illustrates the big picture appointment booking process for future longitudinal care. The booking process in only for clients who become part of the longitudinal care panel, and starts with client intake. After an initial intake appointment, clients are assigned to an individual care provider in the longitudinal care team based on their overall care needs. This could be an MRP or an allied health worker including a SW, RCC, or RD. Each provider has their own panel of clients. Newly registered clients will book their next appointment with their assigned provider at the end of their intake appointment. Clients will return to the centre for their next appointment, although subject to possibly canceling, being late, or not showing up. At the end of each appointment, they will either book another appointment with some target inter-appointment time (for example, one week, two weeks, one month, three months) or leave without booking an appointment, to call and book one later. This process will repeat for over the long-term, until clients no longer meet empanelment criteria and are referred to another primary care service. The day and time of each scheduled appointment can be impacted by several factors, including: provider staffing and schedules; the set of possible appointment slots offered; the proportion and timing of appointment slots reserved for new clients or same-day appointments; client preferences for appointment selection; and the number of empaneled clients. The time between when an appointment is made and the date that it is booked for is referred to as an indirect wait, and can be used

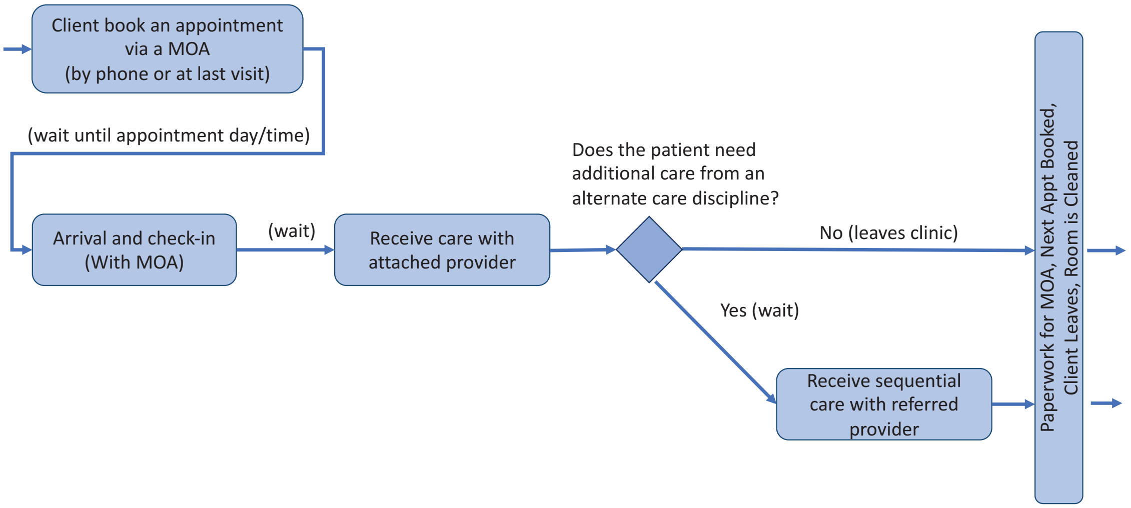
Client book an appointment via a MOA (by phone or at last visit)
(wait until appointment day/time)
Arrival and check-in (With MOA)
(wait)
Receive care with attached provider
Does the patient need additional care from an alternate care discipline?
No (leaves clinic)
Yes (wait)
Receive sequential care with referred provider
Paperwork for MOA, Next Appt Booked, Client Leaves, Room is Cleaned

Figure 3.5: Diagram of the longitudinal care visit process for booked appointments. MOA = medical office assistant.

as a longitudinal care KPI. Indirect waiting can be measured for intake appointments in particular.

3.2.3 Care Visits

Figure 3.5 illustrates the workflow for each longitudinal care visit, which is a sub-component of Figure 3.4. Clients arrive at the clinic in accordance with their scheduled appointment, although they can be early, late, or not show. Clients check-in with the longitudinal care MOA at the front desk, and then wait until the provider that they booked an appointment with, typically their attached provider, is available. After receiving care from this provider, the client leave immediately, or may be referred to see a provider from another discipline for sequential care within the same visit, or to make a follow-up appointment with another provider at the clinic. For example, a client on the panel of an NP may be advised to book an appointment with the RD. Or during a counseling session, a client of an RCC may exhibit the need for medical care and be seen by an RN afterwards as soon as they are available. Both the initial appointment session and possible sequential care will generate paperwork for the MOA, and exam rooms will be cleaned between clients. The duration of time between a scheduled appointment start and when a clients are seen by a provider can be referred to as a *direct wait*, and can be used as a KPI measure for longitudinal care.

3.3 Interaction between Urgent and Longitudinal Care

There is no planned overlap of clinical staff between urgent and longitudinal care teams at the centre. However, some empaneled clients may choose to walk into the urgent care stream. These clients may still be seen by their regular provider if an appointment slot is available, but otherwise will receive treatment from the urgent care team. The volume of registered client crossover may be significant, as an existing CHC that offers longitudinal care to a similar client population reported around 30% walk-in visits [153]. Additionally, client crossover may increase as longitudinal booking lead times increase due to the strategic behavior of clients [164]. The walk-in visits of longitudinal clients could considered into capacity planning by strategically leaving longitudinal appointment slots empty [29] and incorporating the additional urgent care demand into staffing optimization.

Chapter 4

Urgent Care Staffing Optimization

The analysis in this thesis compares multiple staffing models to identify the minimum number of staff in each discipline needed to meet targets on client access indicators in the urgent care stream. To model stochastic client-centred access, I used a queueing network model, described in Section 4.1, which builds on the client flow diagrams in Section 3.1 for urgent care services. The main KPI used in my analysis is the expected proportion of clients who receive care as opposed to leaving without being seen due to an extended wait. This focus is motivated by the care provision goals for the urgent care stream, where services are offered as an alternative to low acuity ED visits and clients who leave without receiving care may either go to the emergency department or have their condition deteriorate. Section 4.3 discuses this choice of access indicator further and defines the notation used.

To evaluate time-dependent client access indicators, I built a DES implementation of the urgent care queueing model. Simulation provides an accurate method to estimate performance measures for the complex network of team-based care interaction [45]. Section 4.2 describes the DES implementation and how it estimates client access indicators for any given staffing level.

To identify team-based staffing requirements, I embedded the DES model into multiple staffing optimization procedures. My analysis focuses on two levels of staffing optimization, namely: (1) small-interval staffing requirements and (2) shift-based staffing requirements. Small-interval staffing optimization for (1) determines the minimum number of staff of each discipline needed on an interval by interval basis throughout the day and week. These requirements are based solely on meeting client access targets and do not consider whether or not staffing can be achieved with standard shift options. Shift-based staffing optimization (2) determines minimum number of staff of each discipline required on each shift to meet simulated client access targets. Sections 4.4 and 4.5 describe different approaches to small-interval and shift-based staffing optimization, respectively. The rostering of individual staff to shifts is not incorporated in this analysis.

My analysis focuses on identifying weekly schedules for urgent care staffing of MRPs, RNs, and MOAs at the front desk. Future work will further incorporate staffing optimization for social workers (SWs), as well as longitudinal care and the cross-over of longitudinal demand into the urgent care stream at the UPCC.

4.1 Urgent Care Queueing Model

Building on the operational profile of the urgent care stream at the UPCC, and making simplifying assumptions, I modeled care access at the UPCC using a queueing network model, which is depicted in Figure 4.1. Each care component (registration, triage, assessment, RN treatment, MRP consult, MRP treatment and social worker visits) is represented as a model service node. At each node, modeled clients join the queue, wait until a staff member of the corresponding discipline is available, and then receive that service. Some staffing disciplines provide for multiple service nodes: MRPs provide both MRP treatment and MRP consult, and RNs provide triage, assessment and RN treatment. Registration is performed by an MOA at the front desk. Modeled client arrivals at the clinic begin when they join the queue for the registration node. This is the only external arrival to the system, as pre-booked SW visits are not incorporated in the current analysis. The model assumes that clients arrive at the registration node according to a non-homogeneous Poisson process, which captures daily and weekly changes in arrival rates. Once registration is completed, model clients join the queue for triage, and after receiving triage they join the queue for assessment from an RN. A constant proportion of clients are routed between receiving care at an MRP treatment node or an RN treatment node, which is preceded by an MRP consult node. After modeled client treatment is completed, a constant proportion of clients either leave the clinic or are referred to visit a SW.

At each node, the service time for the corresponding care component has a general distribution that is independent and identical for each modeled client and staff member. After model clients complete registration, triage or treatment, a cleaning task is initiated for the modeled housekeeper. Room cleaning time is generally distributed and must be completed before another client can be seen in that space. There is a modeled finite number of exam rooms, which can be set arbitrarily large or reflect the current UPCC exam room capacity, for example.

In the model, each visit has an individual stochastic limit for how long the client will wait before leaving without receiving care. This modeled distribution for willingness to wait is assumed to be independently and identically distributed for each visit, from some general distribution. Wait times from each node are accumulated until this limit is reached and the model client may leave the system from any node. Modeled clients will also abandon if they are still in system when the UPCC closes at the end of each modeled day. Staffing levels can fluctuate throughout the day and the service policy is non-exhaustive, which means that when staffing is reduced, a client in service will rejoin the queue to complete their care with the next available staff member of the appropriate discipline. In the current model,

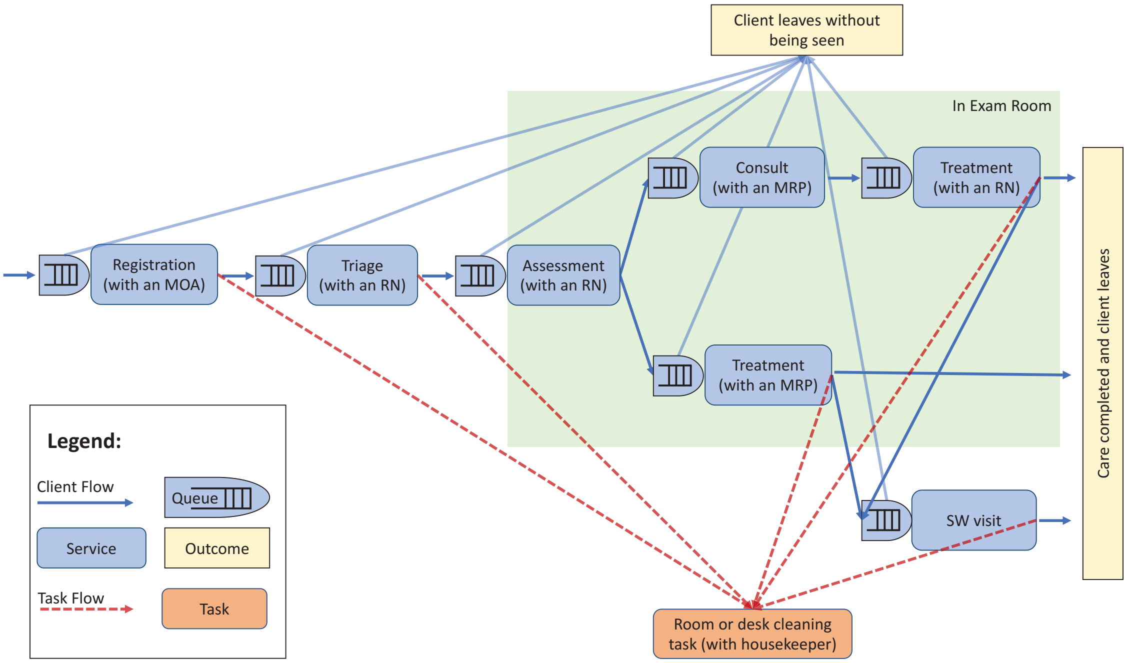

Figure 4.1: Diagram of the queueing model used to represent urgent care at the UPCC. Blue lines indicate modeled client flow through urgent care services and associated queues. Care completion triggers room cleaning tasks indicated here by red dashed arrows.

clients are seen in the order that they join the combined queue for each staffing discipline, and priority differences between clients or components of care are not captured.

4.2 Simulation of the Urgent Care Queueing Model

To estimate performance indicators in the queueing model, I built a DES implementation of the model that can accurately capture the modeled dynamics of team-based urgent care. The DES model is implemented with the AnyLogic[1] modeling software, and a screenshot of the model is shown in Figure 4.2. The simulation has a visual interface with a block or node for each element in modeled client flow. The simulation was presented to management at the UPCC for their feedback in model development.

On the far left of Figure 4.2, the DES model uses a source node to generate simulated agents according to a homogeneous Poisson process. The subsequent choice node is used to implement Poisson thinning, where the client stream is filtered based on a Bernoulli decision with time-dependent probability of acceptance [149]. The source and thinning nodes collectively implement a non-homogeneous Poisson process for model arrivals.

[1] https://www.anylogic.com/

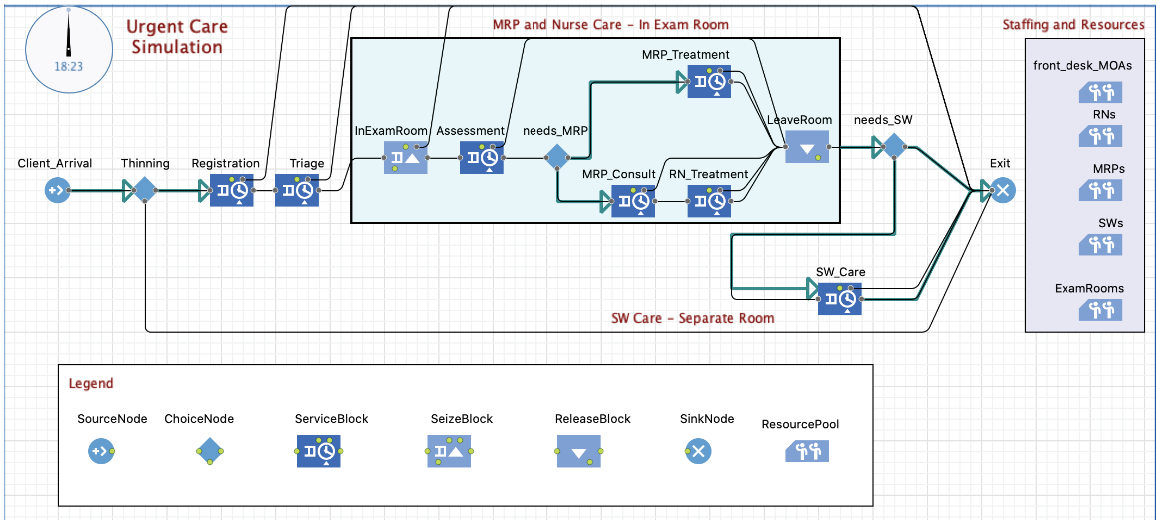

Figure 4.2: Screenshot of the DES implimentation in the Anylogic software.

The DES model simulates client flow at the UPCC by using using seven service blocks that represent the seven care components (registration, triage, assessment, MRP and RN treatment, and SW visits). Each service block captures simulated clients waiting in a queue until a simulated staff member of the corresponding discipline is available; then both are delayed in the block for a pseudo-randomly generated service time. All simulated clients have a pseudo-randomly generated wait limit, after which they will time-out and leave the model through the exit node. Staff are modeled using a separate resource pool for each discipline, with staffing levels that can fluctuate up and down throughout the day. Exam rooms are implemented as a constant size resource pool that simulated clients draw using a seize block before starting assessment. Simulated clients continue to occupy an exam room for the duration of assessment, consult, and treatment, as well as the waiting times in between. Once simulated treatment times are completed, or if a client waiting for treatment/consult reaches their waiting limit, a release block allows the exam room resource to become available for other simulated clients after a room cleaning task.

The DES implementation generates events using the linear congruential method, which produces subsequent pseudorandom numbers with a linear formula adjusted modulo a set interval [96]. The initial number in the sequence, referred to as a seed, is chosen pseudo-randomly so that each simulation run has a pseudo-unique event sequence. Combining outcomes from a large number of simulation runs can provide accurate performance measures estimates.

4.3 Client-centred Access Indicator: Expected Percentage Access

To model urgent care client access, my analysis focuses on quantifying whether or not clients receive care in the queueing model representation of urgent care. This focus is motivated by the care provision goals for the urgent care stream, where services are offered as an alternative to low acuity ED visits and clients who leave without receiving care may either go to the emergency department or have their condition deteriorate.

In the queueing model for urgent care, there are two reasons that a client will leave the UPCC without receiving care. Model clients will leave the clinic if they experience an extended wait time that reaches their individual, stochastic wait limit or if they are still in the system when the clinic closes at the end of the day. For both reasons, the probability of whether or not a model client receives care is time-dependent. This KPI depends on when a model client arrives at the clinic, since the UPCC starts empty each morning and client arrival rates and staffing levels fluctuate throughout the day; furthermore, clients arriving near the end of the day may not be seen before the clinic closes overnight. In queueing models, there are several approaches to quantify time-dependent client outcomes, including: (1) the probability of different outcomes, conditional on model client arrival at instantaneous moments in time [44, 45, 52], and (2) the expected number or proportion of these outcomes, measured over model client arrivals during specific time intervals[9, 27, 73, 168]. The probability based quantification (1) is more common in queueing studies [46], since it can be estimated using steady-state formulae [120] or simulation [44, 45]. However, steady-state results do not capture clients who leave without being seen due to the end of the day. Furthermore, simulation probability estimates require using virtual simulated agents to measure outcomes without affecting staffing resources [44, 45], which is not able to capture outcomes due to non-exhaustive service. To my knowledge, the probability of different client outcomes cannot be measured or estimated from historic or simulation client data. On the other hand, the expected number of clients outcomes can be estimated directly from observation of client or simulated client experiences.

My analysis focuses on the quotient of the expected number of modeled clients receiving care divided by the expected number of modeled arriving clients, which I refer to generally as expected proportion access (EPA) in an extension of the notation used by Wang et al. [168]. Table 4.1 introduces the notation that I use for EPA in this thesis. The notation uses multiple subscripts to distinguishes EPA specific to multiple services and different time-intervals, based on when modeled clients begin queueing. The indexes $k = 0, ..., 5$ are used to denote the six services or care components in the urgent queueing model, namely registration, triage, medical assessment, MRP consult, and RN or MRP provided treatment, respectively. The random variable for the number of modeled clients who begin queueing for care component k_1 in time interval $[t_1, t_2)$ is denoted $N_{k_1}([t_1, t_2))$; similarly, the random

Notation	Definition
K	The set of urgent care service components.
$E[N_{k_1}([t_1, t_2))]$	The expected number of model clients who begin queueing for service k_1 in time interval $[t_1, t_2)$ for all $t_1 < t_2$ and services $k_1 \in K$.
$E[A_{k_1,k_2}([t_1, t_2))]$	The expected number of model clients who begin queueing for service k_1 in time interval $[t_1, t_2)$ and receive service k_2 in any interval, for all $t_1 < t_2$ and services $k_1, k_2 \in K$.
$EPA_{k_1,k_2}([t_1, t_2))$	$\begin{cases} \frac{E\left[A_{k_1,k_2}([t_1,t_2))\right]}{E\left[N_{k_1}([t_1,t_2))\right]}, & \text{if} \quad E[N_{k_1}([t_1, t_2))] \neq 0 \\ 1, & \text{if} \quad E[N_{k_1}([t_1, t_2))] = 0 \end{cases}$
$\widehat{EPA}_{k_1,k_2}([t_1, t_2))$	The quotient of the mean number of clients in the simulation who receive service for care component k_2 (that begin queueing for service k_1 in time interval $[t_1, t_2)$) divided by the mean number of simulated clients who queueing for service k_1 in time interval $[t_1, t_2)$.
$(1 - \alpha_{k_1,k_2})$	A target access proportion for service k_2 that it is desirable for $EPA_{k_1,k_2}([t_1, t_2))$ to be above.

Table 4.1: Definition and notation for expected percentage access (EPA).

variable $A_{k_1,k_2}([t_1, t_2))$ denotes the number of those clients who complete component k_2 in any interval[2]. Note that the time interval that a client is counted in for N_{k_1} and A_{k_1,k_2} depends on the time that they begin to queue for service k_1 and not on the interval that they receive service in. Both $A_{k_1,k_2}([t_1, t_2))$ and $N_{k_1}([t_1, t_2))$ are discrete random variables resulting from a stochastic process representing urgent care at the UPCC, and their expectation is computed over the stochastic processes for model client arrivals, service completions, and abandonment. Furthermore, both expected values are a function of time dependent staffing levels for each care discipline. Expected percentage access $EPA_{k_1,k_2}([t_1, t_2))$ is defined as the quotient of the expectations of $A_{k_1,k_2}([t_1, t_2))$ and $N_{k_1}([t_1, t_2))$, for each $k_2 \geq k_1$ and $t_2 > t_1$ with simulation estimate $\widehat{EPA}_{k_1,k_2}([t_1, t_2))$. Note that the quotient of the expectations A and N is not the same as the expectation of the quotient of these variables, since they are not independent.

The value or estimate for $EPA_{k_1,k_2}([t_1, t_2))$ does not assume that the probability of model client access or abandonment is constant throughout the interval considered. Instead, EPA is a quotient of expected values of an aggregation of observable client outcomes in a

[2] In the model, service or care component completion involves finishing the full service time of registration, triage, assessment, or consult, and completing at least 30 minutes of care for MRP or RN provided treatment

time-dependent stochastic process. The time interval $[t_1, t_2)$ in the introduced notation could be chosen to capture small time intervals, such as a 1-hour or 15-minute interval, which results in what is referred to as a *local* performance measurement; alternatively, the time interval $[t_1, t_2)$ can be chosen capture EPA over an entire planning horizon in what is referred to as a *global* performance measurement [46, 120]. Local EPA measurements can be related to global EPA by using the linearity of expectation for EPA in non-overlapping time intervals, that is

$$\left.\begin{array}{l} EPA_{k_1,k_2}([t_1,t_2)) \geq (1-\alpha_{k_1,k_2}) \\ EPA_{k_1,k_2}([t_3,t_4)) \geq (1-\alpha_{k_1,k_2}) \end{array}\right\} \implies EPA_{k_1,k_2}([t_1,t_2) \cup [t_3,t_4)) \geq (1-\alpha_{k_1,k_2}), \quad (4.1)$$

for all $k_1, k_1 \in K$ and non-overlapping time intervals $[t_1, t_2)$ and $[t_1, t_2)$. However, the converse does not necessarily hold, as poor access in one interval may be outweighed in EPA by a higher access proportion in another interval. Adding multiple constraints to EPA over a series of smaller, local intervals (sometimes referred to as *stabilization*) is a harder performance target to achieve than a global EPA constraint. My staffing optimization approach uses local constraints on EPA in each 15-minute interval to ensure that access is maintained throughout the day.

For an instantaneous moment in time, namely $t_1 = t_2$, the EPA definition above is not defined. Assuming a Poisson arrival process to service k_1, as $t_2 \to t_1^+$ then $EPA_{k_1,k_2}([t_1, t_2))$ will limit towards the probability of a model client receiving care from service k_2, conditional on their arrival to the queue for service k_1 at time t_1. However, given arbitrarily distributed service times for different care components, the arrival processes for each component of the urgent care queueing model are not necessarily Poisson.

To consider client access to multiple urgent care services, the EPA definition has two subscripts, k_1 and k_2, each corresponding to an urgent care service, (including registration, triage, assessment, MRP provided care, MRP consult and RN provided care, referred to with indices 0,1,2,3,4, and 5, respectively). The second subscript k_2 specifies the service that access is being measured for and the first service subscript specifies which clients are being considered for measurement, based on their queue entry time. There are multiple options for measuring service access using this notation. For example, one possibility for measuring medical treatment access is $EPA_{0,5}([t_1, t_2))$, which considers all clients that arrive at the clinic in interval $[t_1, t_2)$ and whether they receive medical treatment. An alternative measure is $EPA_{5,5}([t_1, t_2))$, which will only consider only clients who begin waiting for medical treatment (after assessment) in interval $[t_1, t_2)$. The considered pool of clients will be different for each approach, since $EPA_{5,5}([t_1, t_2))$ will not include clients who leave the clinic prior to assessment. My staffing optimization focuses on $EPA_{k,k}([t_1, t_2))$, $\forall k \in K$, to ensure that efficiency is maintained over each service, and that staffing levels are appropriate for each staffing discipline. Service specific targets $EPA_{k,k}$ can be used to address UPCC arrival based EPA metrics ($EPA_{0,k}$) by incorporating the sequence of care component interdependence. Since clients who access one service all join the queue for the next, it holds that $A_{k,k}(d) = N_{k+1}(d)$, for $k = 0, 1, 2$, where d denotes the time interval corresponding to an entire day of urgent care services. Thus,

$$\frac{EPA_{0,5}(d)}{\gamma} = \prod_{j=0,1,2,5} EPA_{j,j}(d) \tag{4.2}$$

where γ is the proportion of clients requiring MRP care. Similarly,

$$\frac{EPA_{0,4}(d)}{(1-\gamma)} = \prod_{j=0,1,2,3,4} EPA_{j,j}(d) \tag{4.3}$$

where $(1-\gamma)$ is the proportion of clients requiring nurse care.

Equations (4.2) and (4.3) must be considered when forming EPA constraints. Setting a 95% constraint on each $EPA_{k,k}$ value means that the overall access to MRP care could be as low as $0.95^4 \approx 0.81$, and RN care access could be even lower at $0.95^5 \approx 0.77$. One approach to achieve a lower bound of 95% access to nurse and physician care is by placing a $(1-\alpha_{k,k}) = 0.95^{1/5} = 0.9897938$ EPA constraint on each service. Note that equation (4.2) and (4.3) do not necessarily hold for general time intervals, since $A_{k,k}([t_1, t_2))$ is not necessarily the same as $N_k([t_1, t_2))$ whenever there is possibility of a client being in the midst of receiving service k at precise time-point t_2. Nonetheless, combining these equations with equation (4.1) means that overall access targets can be met by interval and service specific ones.

In general, system optimization outputs are affected by the choice of performance indicator constraints or objectives [120]. Mandelbaum and Zeltyn [120] illustrate that focusing solely on client abandonment will often require fewer staff than if constraints are included on wait time metrics as well. For a scheduling based illustration, if there is a scenario where all customers will wait at least an hour for service before they leave without being seen, then there would be no need to schedule staff in the first hour of operation under an abandonment only constraint.

4.4 Small-interval Staffing Requirements

This section describes an approach to approximately identify the minimum number of staff of each discipline required in each 15-minute interval so that meet client access constraints are met in each interval. These small-interval staffing requirements are chosen independently of the set of possible staff shifts, and can change throughout the day to stabilize performance under fluctuating demand. The results can be used as a constraint for shift-based staffing procedures [22, 45], or motivate the introduction of new shift options.

Equation (4.4) gives defines the formulation that I use to optimize small-interval staffing requirements for the urgent care stream. The decision variable is a staffing matrix $S \in$

Weekday(s)	Staffing Intervals	Constraint Intervals
Monday through Saturday	◦ Every 15 minutes from 8:00 to 20:45 ◦ 20:45 to 22:00	◦ Every 15 minutes from 8:00 to 20:45 ◦ 20:45 to 21:30
Sunday	◦ Every 15 minutes from 8:00 to 15:45 ◦ 15:45 to 17:00	◦ Every 15 minutes from 8:00 to 15:45 ◦ 15:45 to 16:30

Table 4.2: Staffing and constraint intervals used for small-interval staffing optimization. Intervals are changed on Sundays to reflect different opening hours. Access constraints are not used for the last half hour that the clinic is open, since it is model clients who begin to queue for treatment during this time may not receive full care regardless of staffing at this time. Since registration closes an hour before the clinic does, access indicators for the last 45 minutes of constrained time are merged. The last 75 minutes of each day is staffed as a single interval.

$Z^{+|M|\times|I|}$, where each component $S_{m,i}$ is the number of modeled staff of discipline m on duty in time interval i, for each staffing disciplines m in the set of considered disciplines M and each interval i in the set of staffing intervals I_s. My analysis focuses on three staffing disciplines: MRPs, RNs, and MOAs at the front desk. The objective function is to minimize the total staffing hours across each staff discipline and time interval, using parameter w_i to incorporate the duration of each time interval. The formulation uses a set of constraints to ensure that simulation client access targets are maintained in each time interval from I_c. In this optimization problem, $\widehat{EPA^S_{k,k}}(i)$ is the simulation estimate for the expected proportion of clients who begin queueing for service k in interval i that receive service k, conditional on staffing matrix S, and $(1-\alpha_{k,k})$ is the access target for modeled service k. Table 4.2 describes the sets of staffing and constraint intervals, which are adjusted to account for the opening hours and registration cut off at the centre.

$$\begin{aligned} \text{minimize} \quad & \sum_{m\in M}\sum_{i\in I_s} w_i s_{m,i} \\ \text{subject to} \quad & \widehat{EPA^S_{k,k}}(i) \geq 1-\alpha_{k,k}, \quad \forall\ k \in K,\ i \in I_c \\ & S \in Z^{+|M|\times|I_s|} \end{aligned} \tag{4.4}$$

4.4.1 Extended Iterative Staffing Algorithm

To approximately solve the small-interval staffing formulation in (4.4), I applied the iterative staffing algorithm (ISA) from Defraeye and Van Nieuwenhuyse [44], which I extended to address team-based care and EPA constraints. The ISA framework makes iterative approximations of small-interval staffing requirements, which draw on simulation output to update staffing levels in each iteration [44, 52]. Feldman et al. [52] introduced the ISA approach to find the number of staff in each interval required to stabilize the instantaneous probability of delay, under which required staffing can be inferred directly from iterative simulation estimates of time-dependent distributions for the number of customers in the system. Defraeye and Van Nieuwenhuyse [44] use an extended ISA(τ) algorithm to stabilize the instantaneous probability of an extended wait beyond duration τ by using a more generalized staffing update procedure that is split into two algorithm phases. The ISA framework offers a valuable approach to optimizing small-interval staffing based on the iterative use of simulation output; however, previous ISA studies only consider instantaneous, waiting based performance measures for a single staff type [44, 52]. I propose an extension of the ISA approach to consider EPA-based staffing, which I refer to as the EPA-ISA approach. I further extend this to address multiple staffing disciplines using a sequential-EPA-ISA approach.

Phase I of the two-phase ISA(τ) algorithm uses a generalized stochastic binary search procedure to iteratively scale staffing levels up or down in each interval based on the difference in KPI between simulation and target values and identify staffing levels close to optimum. Since phase I results may not satisfy all the performance targets, phase II uses a unidirectional stochastic linear search that incrementally increases staffing levels until performance targets are met. In both phases, staffing level updates are conducted simultaneously across each time intervals. For each staffing interval i, staffing is modified based only on simulation KPIs in a corresponding measurement interval $\tilde{i}$, which is chosen based on the time period that performance is directly impacted by staffing in i. This assignment of intervals can depend on the choice of KPI constraint used, and is an important factor in the robust convergence of the algorithm; if there is no arbitrarily large level of staffing in interval i that will satisfy the performance constraint in interval $\tilde{i}$, then the algorithm may not converge.

In the context of a performance target on extended waiting, Defraeye and Van Nieuwenhuyse [44] use a shifted measurement interval of $\tilde{i} = [t_s - \tau, t_f - \tau)$ for each staffing interval $i = [t_s, t_f)$, where the magnitude of the shift corresponds to the KPI waiting threshold τ. If there is an arbitrarily large surge in staff during $[t_s, t_f)$, then client arrivals from $[t_s - \tau, t_f)$ will not wait more than τ time for service in their model, regardless of previous staffing levels. Similarly, if there is an arbitrarily large surge in staff immediately after $\tilde{t}$, then arrivals after $t_f - \tau$ in their model will all start being seen in at least time τ, even if there are no staff at all in $\tilde{t}$. This choice of measurement interval assignment is specific to both the waiting threshold τ and an exhaustive service policy. Under an abandonment based performance measure with generally distributed time until abandonment, some portion of model clients may abandon during time $[t_s - \tau, t_s)$, regardless of staffing in $[t_s, t_f)$. Furthermore, under a non-exhaustive service policy, model clients who begin service in $[t_s, t_f)$ may abandon later if they do not complete sufficient service due to a reduction in staff at t_f.

To extend the ISA framework to use EPA and non-exhaustive service, I introduce a new strategy for measure interval assignment. Unlike ISA(τ), which uses only a single, non-overlapping measure interval for each staffing interval, I assign multiple measure intervals to each staffing interval. Each staffing interval, say i_h, is updated using KPI measures from i_h itself along with the preceding intervals $\{i_{h-\eta_k}, i_{h-\eta_k+1}, .., i_{j-1}\}$, up to a lag factor η_k that depends on the service k. To account for non-exhaustive service, I determine η_k based on the minimum amount of time σ_k that a client needs to spend in service k to be considered seen, divided by the constant interval duration Δ in

$$\eta_k = \lceil \frac{\sigma_k}{\Delta} \rceil > 0 \tag{4.5}$$

The incorporation of a positive lag factor means that each measurement interval will affect staffing decisions in multiple subsequent staffing intervals. For example, in a scenario with base intervals that are 15-minutes long and a service where clients who receive at least 30 minutes of care are counted as being seen, then the EPA-ISA approach will respond to simulation EPA estimates are below target in interval i_h by increasing the staffing in the three intervals i_h, i_{h+1}, and i_{h+2} to ensure that enough staffing is added to complete the service times of the modeled clients in i_j. Furthermore, this approach still meets client demand in its own interval, without a built in delay.

The ISA(τ) algorithm modifies staffing based on the maximum instantaneous probability of extended wait within each measure interval. I extend this approach in the EPA-ISA approach to combine EPA estimates for multiple intervals and multiple services by using the minimum EPA estimate across these measures. For a staff discipline m that provides care for the service components in set K_m, EPA-ISA assesses whether or not targets are not by comparing whether or not

$$\max_{k \in K_m} \left\{ \max_{g \in \big\{\{\max(h-\eta,0),...,\max(h,|I_c|)\}} \{1 - EPA_{k,k}(i_g)\} - \alpha_{k,k} \} \right\} \tag{4.6}$$

is greater than zero or not, for all $h = 1, ..., |I|$. In equation (4.6), the number of measure intervals used is adjusted to account for the beginning and end of the day.

To incorporate optimization for multiple staff disciplines, I present the sequential-EPA-ISA approach, which performs EPA-ISA separately for each staff discipline in a set sequence. This approach is similar to that of Sinreich and Jabali [155], who consider sequential staffing of medical resources in order of an estimated bottleneck factor. In my sequential-EPA-ISA implementation, I optimize staffing for multiple disciplines using the order of interaction with clients as they move through the components of urgent care. My analysis first optimizes staffing for front desk MOAs, then RNs, and finally MRPs. Staffing decisions made for disciplines later in the sequence build on the impact of earlier staffing discipline choices, but not vice versa. Some EPA target values may not be achievable when only considering staffing levels for one staffing discipline at a time or incorporating restricted exam room capacity. To mitigate this, I applied ISA to a scenario with an arbitrarily large number of exam rooms and added a maximum number of iterations to both phase I and phase II of the EPA-ISA approach. Appendix A describes the details of the sequential-EPA-ISA approach to solve problem (4.4).

4.5 Shift-based Staffing Requirements

This section describes two methods to approximately identify the minimum number of staff needed on each shift such that client access targets are met in each 15 minute interval. I formulated the optimization of shift-based staffing using equation (4.7), which has a similar objective function and constraint set to equation (4.4), but a different set of decision variables. This optimization problem uses a shift-based staffing matrix $X \in Z^{|M| \times |J|}$, where each component $X_{m,j}$ is the number of staff of discipline m assigned to shift option j, for all m in the set of considered staffing disciplines M, and for each j in the set of considered shifts J. In this optimization problem, $\widehat{EPA^X_{k,k}}(i)$ is the simulation estimate for the expected proportion of client access to service k in interval i, conditional on staffing matrix X. The formulation minimizes total staff hours by using weight parameters w'_j to account for the active staff hours in each shift j.

$$\begin{aligned} \text{minimize} \quad & \sum_{m \in M} \sum_{j \in J} w'_j X_{m,j} \\ \text{subject to} \quad & \widehat{EPA^X_{k,k}}(i) \geq 1 - \alpha_{k,k} \quad \forall\, k \in K,\, i \in I_c \\ & X \in Z^{+|M| \times |J|} \end{aligned} \tag{4.7}$$

My analysis considers shift-based optimization over two different sets of shift options, one set that does not incorporate staff breaks and one set that does. Optimization without breaks uses the set of base shift options in Table 3.2. The timing of staff breaks at the UPCC is not pre-determined and can change depending on several factors including staff preferences and clinic congestion. To identify optimum shift-based staffing that incorporates the reduction in staff availability due to breaks, I introduced four break timing patterns for each standardized shift at the UPCC, recorded in Table 4.3. However, I still used shift options without breaks for MOAs, since the breaks for MOAs at the front desk are covered by the MOAs who work in the clinical pod. Recommendations on break timing are outside of the scope of my analysis.

My analysis uses two methods to approximately solve the shift-based staffing optimization in equation (4.7). One approach is to use integer linear programming (ILP) to determine the shift combinations that meet the small-interval staffing requirements from EPA-ISA. Alternately, shift-based staffing can be optimized by combining simulation EPA estimates for each staff schedule with a simulation optimization procedure. My analysis compares the

Shift	Break Pattern	Early Coffee Break	Lunch Break	Late Coffee Break
Morning Shift (RN) (7:45–15:45)	1	9:30–9:45	11:00–11:30	13:00–13:15
	2	9:45–10:00	11:30–12:00	13:15–13:30
	3	10:00–10:15	12:00–12:30	13:30–13:45
	4	10:15–10:30	12:30–13:00	13:45–14:00
Morning Shift (MRP) (8:00–14:00)	1	9:45–10:00	11:15–11:45	13:15–13:30
	2	10:00–10:15	11:45–12:15	13:30–13:45
	3	10:15–10:30	12:15–12:45	13:45–14:00
	4	10:30–10:45	12:45–13:15	14:00–14:15
Day Shift (9:00–17:00)	1	10:45–11:00	12:00–12:30	14:00–14:15
	2	11:00–11:15	12:30–13:00	14:15–14:30
	3	11:15–11:30	13:00–13:30	14:30–14:45
	4	11:30-11:45	13:30–14:00	14:45–15:00
Afternoon Shift (11:00–19:00)	1	12:45–13:00	14:00–14:30	16:00–16:15
	2	13:00–13:15	14:30–15:00	16:15–16:30
	3	13:15–13:30	15:00–15:30	16:30–16:45
	4	13:30–13:45	15:30–16:00	16:45–17:00
Evening Shift (RN) (14:45–22:15)	1	16:00–16:15	17:00–17:30	19:30–19:45
	2	16:15–16:30	17:30–18:00	19:45–20:00
	3	16:30–16:45	18:00–18:30	20:00–20:15
	4	16:45–17:00	18:30–19:00	20:15–20:30
Evening Shift (MRP) (14:00–22:00)	1	15:45–16:00	17:15–17:45	19:15–19:30
	2	16:00-16:15	17:45–18:15	19:30–19:45
	3	16:15–16:30	18:15–18:45	19:45–20:00
	4	16:30–16:45	18:45–19:15	20:00–20:15

Table 4.3: Simplified break options for the UPCC used in shift-based staffing optimization.

ILP approach with simulation optimization performed using the OptQuest meta-heuristic. Subsections 4.5.1 and 4.5.2, describe each approach, respectively.

4.5.1 Integer Linear Programming

One approach to address the shift-based staffing optimization in equation (4.7) is to solve a proxy problem where the stochastic EPA constraints in (4.7) are replaced with deterministic constraints given by small-interval staffing requirements. The ILP formulation used in my analysis is shown in equation (4.8) and is based on Dantzig's [42] formulation of shift-based staffing optimization. This formulation constructs shifts explicitly by using the parameter $c_{i,j}$, which is set to 1 if a staff member assigned to shift j would be on-duty during time interval i and zero otherwise.

In the staffing constraints in equation (4.8) sum over $c_{i,j}$ and $X_{m,j}$ values to determine the number of staff who are on-duty in each interval, which must be greater than or equal to the small-interval staffing requirements $S^*_{m,i}$. In my application, I set $S^*_{m,i}$ to the results of the sequential-EPA-ISA approach to ensure that the client access targets in (4.7) are met by any feasible schedule in the ILP. The resulting minimization problem is deterministic and can be optimized using a linear programming solver independently for each staff discipline and day of the week. Since the ILP does not consider the KPI impact of the additional staffing needed for standardized shifts, this approach is known to sometimes overestimate staffing requirements [45, 83].

$$\begin{aligned} \text{minimize} \quad & \sum_{m \in M} \sum_{j \in J} l_j X_{m,j} \\ \text{subject to} \quad & \sum_{j \in J} c_{i,j} X_{m,j} \geq S^*_{m,i} \quad \forall\, m \in M,\, i \in I \\ & X \in Z^{+|M| \times |J|} \end{aligned} \tag{4.8}$$

4.5.2 Simulation Optimization

Another approach to shift-based staffing is to use simulation optimization, that combines meta-heuristic search procedures with simulation EPA estimates for each staffing decision. I applied the OptQuest meta-heuristic to optimize shift-based staffing using the urgent care simulation from Section 4.2. The OptQuest engine uses the scatter search framework [102], which considers new candidate solutions by combining and improving previous solutions from a maintained reference set [61, 103]. In order to balance local and global optimization, the reference set solutions are chosen for both diversity and objective function quality [61, 103]. To address equation (4.7), the OptQuest approach reformulates the set of stochastic access constraints as penalties in the objective function.

To incorporate multiple staffing disciplines in an urgent care context, my analysis considers two different approaches to performing simulation optimization, namely:

1. *Sequential simulation optimization:* Optimizing the schedule separately for each staff discipline, sequentially considering each discipline in order of dependency.
2. *Joint simulation optimization:* Optimizing the schedule for multiple staff disciplines at the same time.

Joint simulation optimization for multiple staff disciplines at the same time can address the interactive nature of team-based care and the inter-discipline impact of staffing decisions. However, the larger number of variables used in joint optimization can pose challenges for efficient and accurate simulation optimization. My analysis compares joint simulation optimization with a simplified sequential simulation optimization.

Under both the joint and sequential approach, I applied the OptQuest engine to solve equation (4.7) independently for each day of the week. My OptQuest implementation uses explicit non-negative integer variables for the number of staff of each discipline on each shift in either Table 3.2 or Table 4.3, depending on whether or not breaks are considered. I set the initial value for these decision variables to the current UPCC staff schedule, and used an upper bound of 5 staff per discipline for shifts without breaks and 2 staff per discipline for shifts with breaks. To improve performance, I added a requirement that at least one staff member of each type is scheduled on each day. Each run of the OptQuest procedure stops after there are no more significant improvements to the objective function[3] or a maximum number of iterations is reached.

[3] The definition of significant change used in this stopping criteria is determined by an OptQuest heuristic.

Chapter 5

Urgent Care Data Analysis and Model Parameter Estimation

To inform parameter choices for my urgent care queueing model and simulation, I analyzed data on UPCC client visits. For model parameters without relevant data available, choices were made using expert opinion from clinic staff and simulation calibration. UPCC client visits are primarily recorded using the Cerner EMR database system [1]. Recorded timestamps for each client visit include the time of registration time, triage time, the time that the client begins being seen by a provider, and an estimated time of client departure. The data extract used in my analysis includes records of client visits made within the 35 weeks after the date that the clinic opened. My analysis does not include data on client visits with social workers.

5.1 Client Arrival Rates

The number of recorded client visits per week in the data extract are shown in Figure 5.1. Weekly client counts start low when the UPCC opens, then increase steadily until plateauing 19 weeks after opening. I analyzed weeks 19 through 35 of the data extract to construct a representative arrival rate profile for the queueing model. For this time period, the mean number of urgent care registrations for each hour of the week is shown in Figure 5.2. I used the mean number of registrations per hour as an hourly arrival rate in the queueing model that defines of a piece-wise constant non-homogeneous Poisson process for model client arrivals. This approach may underestimate the true number and pattern of UPCC client arrival, since clients may arrive then wait and/or leave the UPCC prior to registration.

In Figure 5.2, hourly registration rates are typically highest in the hour immediately after daily opening, with some fluctuation throughout the day before declining shortly before closing. Table 5.1 records daily summaries, and shows that Sunday has the second highest average daily registration rate, despite having the shortest opening hours. Furthermore,

[1] `https://www.cerner.com/`

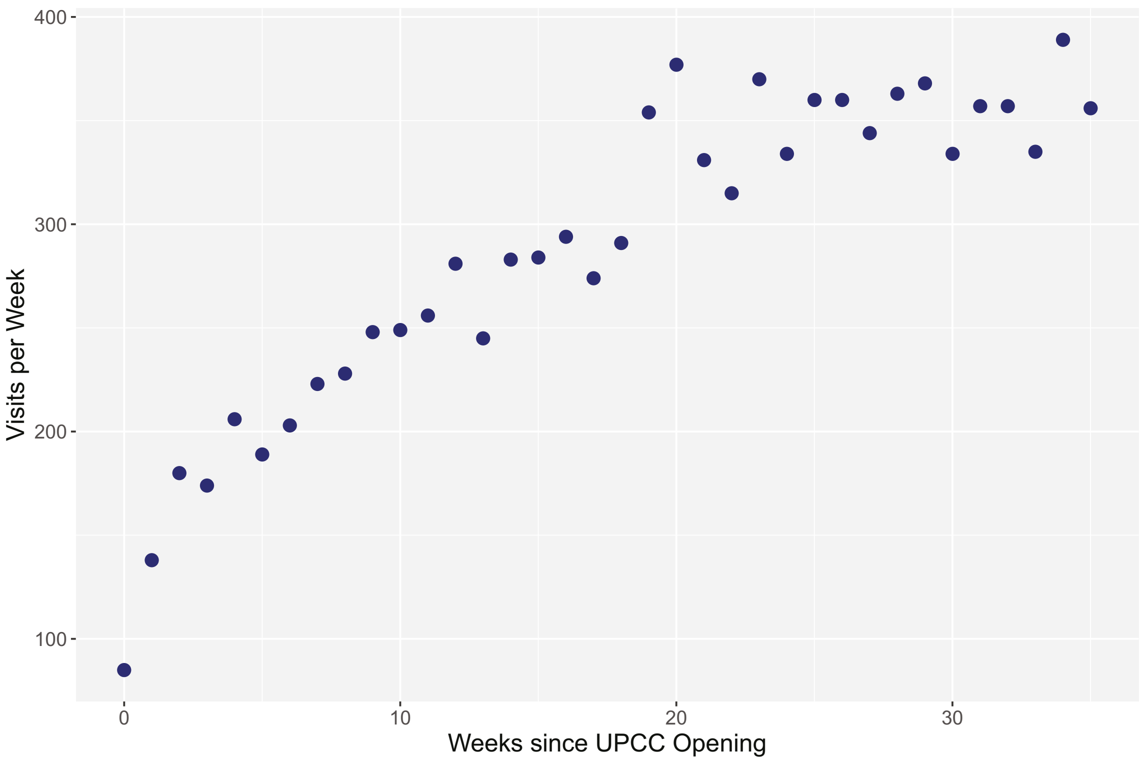

Figure 5.1: Urgent care visits per week in the data extract, with weeks numbered starting from the date the UPCC opened.

the highest average number of registrations per hour occurs during 9 to 10am on Sunday mornings.

5.2 Service Duration

5.2.1 Medical Care or Treatment Time

In the UPCC data, timestamps are included for both the time that any client's visit file is initially opened by an MRP (referred to as *seen time*) and the approximate time that each client leave the clinic (referred to as *departure time*). The interval between seen and departure time can be reflective of the time that taken to provide medical care or treatment, although there are several data quality concerns that need to be considered to estimate treatment times.

One concern raised by the UPCC management team is that recorded client departure times may not be accurate, especially for records were departure is recorded a very long time after the initial file was opened. For example, the longest recorded duration between seen and departure time in the data extract is over 16 hours, which is longer than the clinic is open in a day. Another concern is that there can be a large overlap in the recorded

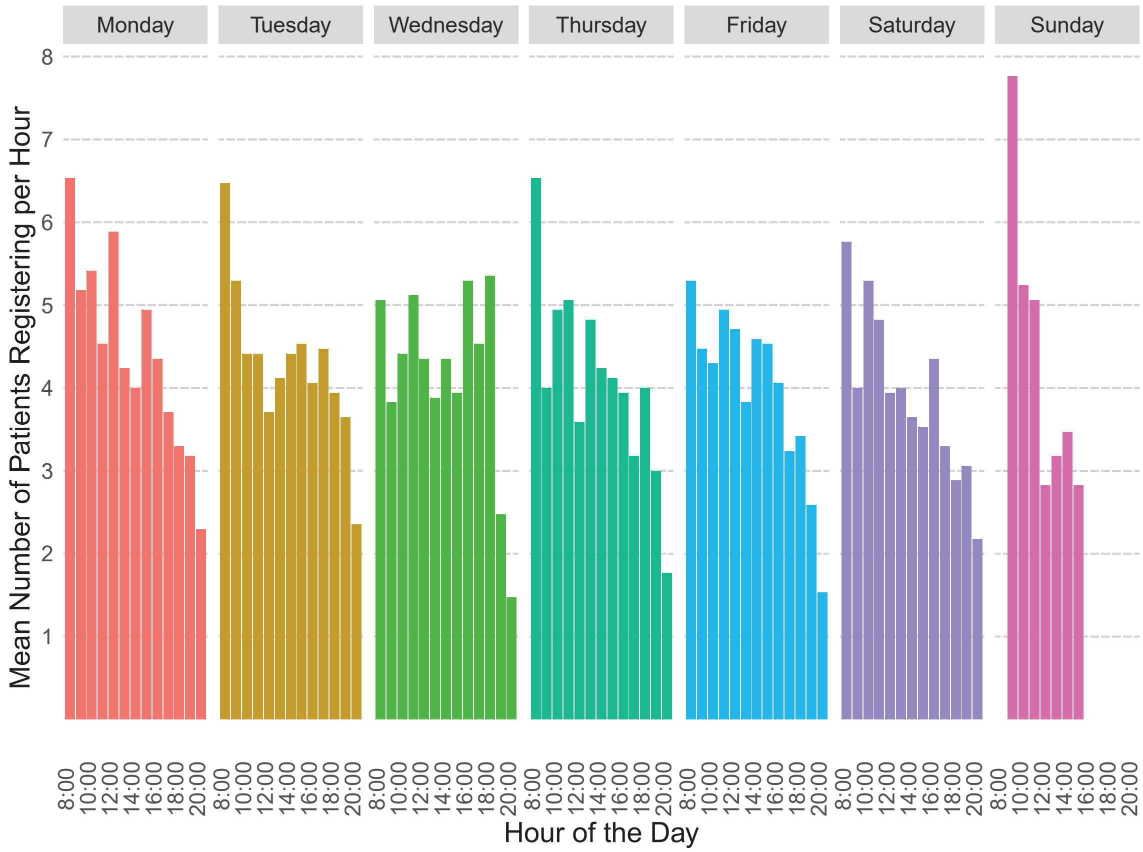

Figure 5.2: Mean number of urgent care registrations by hour of day and day of week for weeks 19 to 35 of UPCC operation. Clients who register before clinic opening or after standard registration closing are included in the mean visit calculations for the hour immediately after opening or immediately before standard registration closing, respectively.

Weekday	Mean Visits per Day	Mean Daily Visits per Hour
Monday	57.5	4.43
Tuesday	55.8	4.29
Wednesday	54.1	4.16
Thursday	53.2	4.09
Friday	51.5	3.96
Saturday	50.8	3.90
Sunday	30.4	4.34

Table 5.1: Mean number of registrations per day and daily mean visits per hour for each day of the week. Each calculation uses data from weeks 19 to 35 of UPCC operation. Daily mean visits per hour is calculated as the mean number of visits divided by the total opening hours in that day.

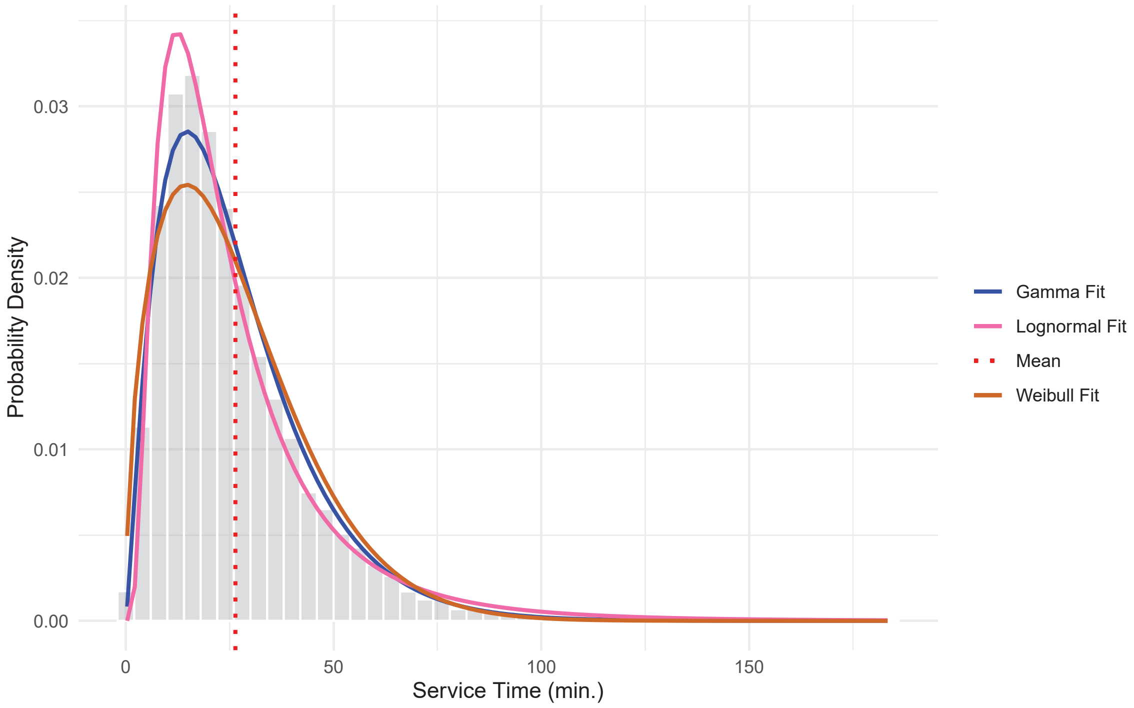

Figure 5.3: Histogram of estimated treatment times with three fitted distributions. Estimated treatment times include 9930 data points from week 0 to 35 of UPCC operation, after filtering out entries with over 200 minutes between seen and departure times and then adjusting for overlapping care provision.

treatment times for clients receiving care from the same provider. In the data extract, 47% of client visits are recorded as having overlapping treatment time with at least one other client receiving care from the same provider.

To account for these data quality concerns and estimate treatment times in the data, I removed outliers with extremely long duration between seen and departure, and then adjusted each entry to account for overlapping care provision. Appendix B gives more details on my adjustment approach, which reduces treatment times evenly during any overlap period. A histogram of the resulting care time estimates is shown in Figure 5.3. To identify a representative service time distribution, I fit three different continuous distributions to the estimated treatment time data using maximum likelihood estimation (MLE) in R. The fitted distributions are included in Figure 5.3, and Table 5.2 records their summary statistics and parameters. Compared to the fitted log-normal and Weibull distributions, the fitted gamma distribution matches the mean and standard deviation of estimated data the best. The queueing model represents the treatment or care provision time for both MRP and RNs using a gamma distribution with a shape of 2.25 and a scale of 11.67 minutes.

	Estimated Treatment Data	Gamma Fit	Log-normal Fit	Weibull Fit
Mean (mins)	26.37	26.37	27.09	26.52
SD (mins)	18.71	17.54	22.61	17.74

Parameters		Shape = 2.25
		Scale = 11.67
			Meanlog = 3.03
			SDlog = 0.73
				Shape = 1.52
				Scale = 29.43

Table 5.2: Mean and standard deviation (SD) for the estimated care provision time data, compared to fitted gamma, log-normal and Weibull statistics. The given parameters for each distribution are fit to the estimated care provision time data using MLE. Estimated treatment times include 9930 data points from week 0 to 35 of UPCC operation, after filtering out entries with over 200 minutes between seen and departure times and then adjusting for overlapping care provision.

5.2.2 Registration, Triage, Assessment, and Consult Time

While the start time of registration is recorded for each client, the corresponding end time is not captured. An expert opinion provided by an MOA at the UPCC estimates that registration it takes on average 2 minutes to complete registration, and that the standard set of questions involved can limit the duration of registration. Based on this input, the queueing model represents the distribution of time needed for registration with a continuous uniform distribution between 1 and 3 minutes.

Similarly to registration, only the start time of triage is recorded for client visits. An expert opinion provided by a UPCC manager estimates that triage typically takes 5 to 7 minutes. Accordingly, the queueing model represents the distribution of time needed for triage with a continuous uniform distribution between 5 and 7 minutes. This time scale is consistent with an estimated average triage time of 4 minutes in an emergency department study [163].

Neither start nor end time are recorded for the medical assessment of clients by nurses. The management team at the UPCC suggested that client assessment takes on average 5 minutes, and mentioned that there could be substantial variability in assessment times. The distribution of assessment time in the model is represented using a gamma distribution with a mean of 5 minutes, and the same shape as the treatment time distribution. I also used expert opinion to represent the distribution of time needed to consult an MRP prior to RN provision of care, for which the queueing model uses a continuous uniform distribution between 4 and 7 minutes.

5.2.3 Room Cleaning Time

Table 5.3 lists the continuous distributions that the queueing model uses to represent the time needed for each room or desk cleaning task, which were chosen based on consultation with a manager and housekeeper at the UPCC.

Cleaning Task	Representative Time Distribution (minutes)
Registration desk	Uniform(0.25,0.75)
Triage room	Uniform(1,3)
Exam room	Uniform(1,3)

Table 5.3: Modeled time distribution for cleaning tasks.

5.3 Client Willingness to Wait

The interval between registration and seen time in the UPCC can be reflective of the client wait times, although this interval does not capture the time spent by clients waiting for registration, and also includes triage and assessment times. Some client visit records do not include a seen time or provider care information and indicate that a client left the UPCC without receiving care. Data records for clients who leave without being seen still include an estimated time of client departure; the interval between registration and departure can be reflective of the willingness to wait of these clients. However, estimated wait times for clients who do receive care provide additional information that these clients were willing to wait at least those amounts of time. To combine these two sets of data and estimate the distribution of clients' willingness to wait, I applied the Kaplan-Meier (K-M) method, which is frequently used in lifetime analysis data [94, 156] including customer patience distributions [20]. The K-M estimator for the survival distribution of willingness to wait is approximated by the fraction of visits still waiting for each length of time.

Figure 5.4 displays the K-M estimate of the cumulative distribution function (CDF) for willingness to wait based on the records in the UPCC data extract. To model willingness to wait, I used a truncated gamma distribution with parameters fit to the K-M CDF using minimization of squared error [2] and the maximum observed wait as a the truncation value. The fitted parameter values are recorded in Table 5.4 and the resulting CDF is plotted alongside the K-M CDF in Figure 5.4. The queueing model represents abandonment by using this truncated gamma distribution to independently generate willingness to wait for each modeled client visit, which is a simplification of the complex human behavior associated with waiting.

[2] This parameter optimization was performed using the it optim function in R, which implements a quasi-Newton method for constrained continuous optimization

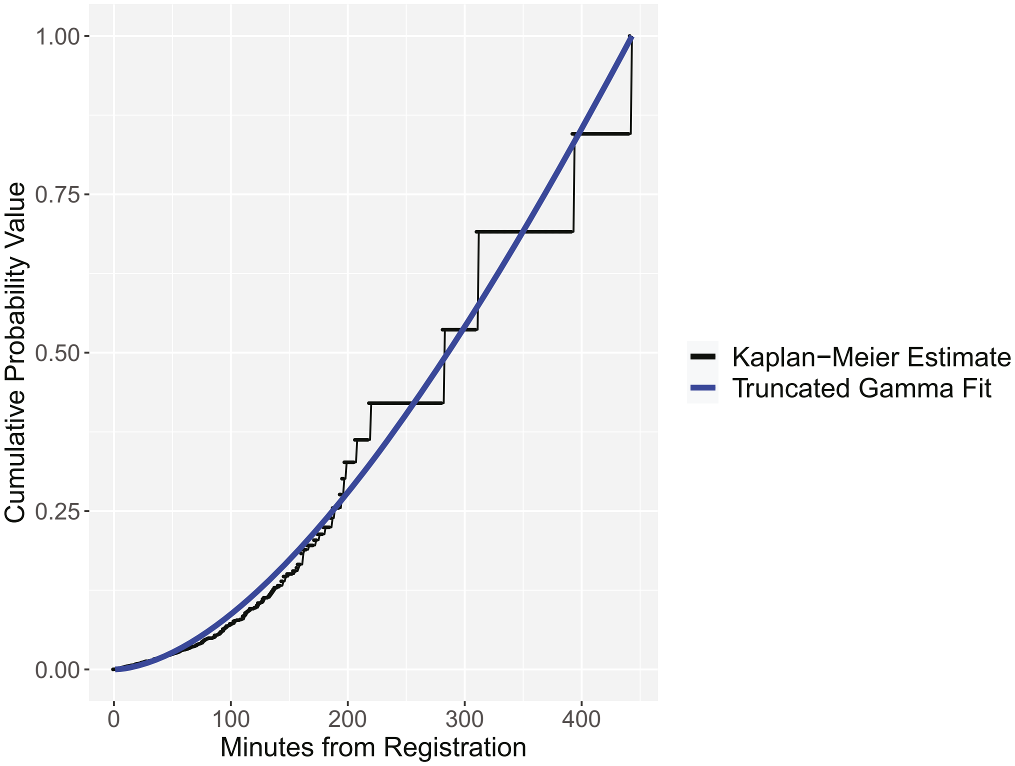

Figure 5.4: Kaplan-Meier estimates for the CDF of the UPCC clients willingness to wait for urgent care. The K-M method was applied to 35 weeks of UPCC data for the time between registration and departure for clients who abandon, and the time between registration and seen time for clients who receive care.

Parameter	Value	(units)
Shape	1.7	
Scale	21.9	(hours)
Maximum	7.4	(hours)
Mean	4.5	(hours)

Table 5.4: Fitted parameters for a truncated gamma distribution used to represent client willingness to wait.

Chapter 6

Results

6.1 DES Calibration and Validation

To calibrate and validate the simulation model, my analysis compares EPA estimates from the current schedule DES to the EPA estimates from the UPCC data. This comparison focuses on hourly EPA values for the expected proportion of clients receiving care from either an RN or MRP, for each opening hour throughout the week based on simulated arrival time or recorded registration time. The current urgent care staff schedule in the DES is from Figure 3.2 combined with the break options defined Table 4.3, specifically: break options 1 and 2 for each pair of MRP shifts; break option 1 for each afternoon RN shift; break options 2,3, and 4 for each triple of RN morning or evening shifts; break options 1,2, and 4 for the triple of RN day shift on Sundays; and shifts without breaks for MOAs at the front desk, since their breaks are covered by the other MOA on duty.

To calibrate the proportion of clients requiring RN care, I ran the simulation for the current schedule using different values for this proportion from the set $\{0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45\}$. Figure 6.1 shows the simulation EPA estimates for each of these parameter choices alongside EPA estimates from UPCC data. Compared to the other values tested, the simulation EPA estimates are closest[1] to the data when the proportion of clients requiring RN care is 0.4. I used the value of 0.4 for this parameter throughout the rest of the analysis in this thesis.

Figure 6.2 and Table 6.1 compare the resulting EPA estimates from DES and the UPCC data using this parameter choice. Over the entire week, the EPA values are similar for both DES and the data, although the simulation EPA estimates tend to overestimate the data. Figure 6.2 shows that 57% of the data points are within the DES interquartile range (IQR); the data EPA estimates show more intra-day and intra-week variability than the simulation, since they are the result of intricate human behavior compounded by data quality factors. For example, UPCC management commented that some clients who arrive around lunchtime

[1] Using the mean squared difference in EPA estimates.

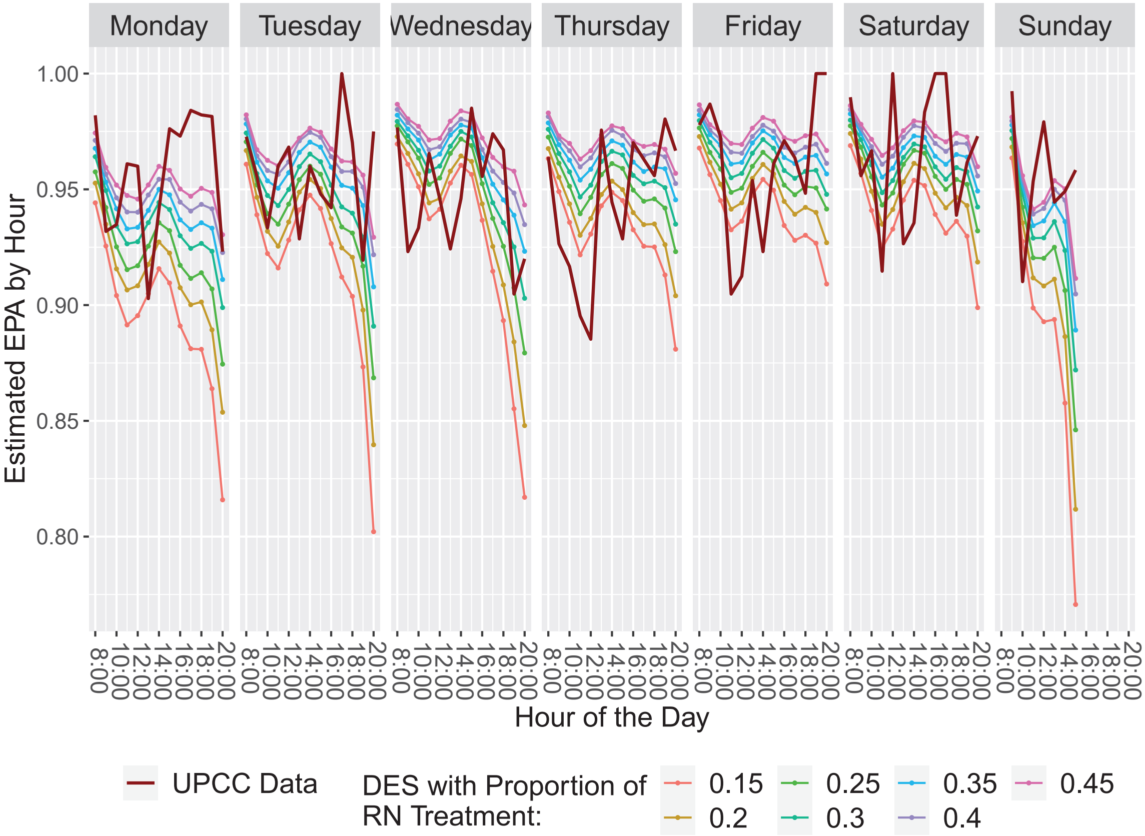

Figure 6.1: Comparison of EPA estimates from UPCC data and DES results, using seven different parameter values for the proportion of visits that are within RN nurse scope. Data estimates are based on the last 17 weeks of UPCC operation, with EPA calculated as the mean number receiving care divided by the mean number registering, in each hour of the day and week. The DES estimates were calculated similarly based on arrivals for 85,000 simulated weeks. The current schedule in the DES uses break options from Table 4.3.

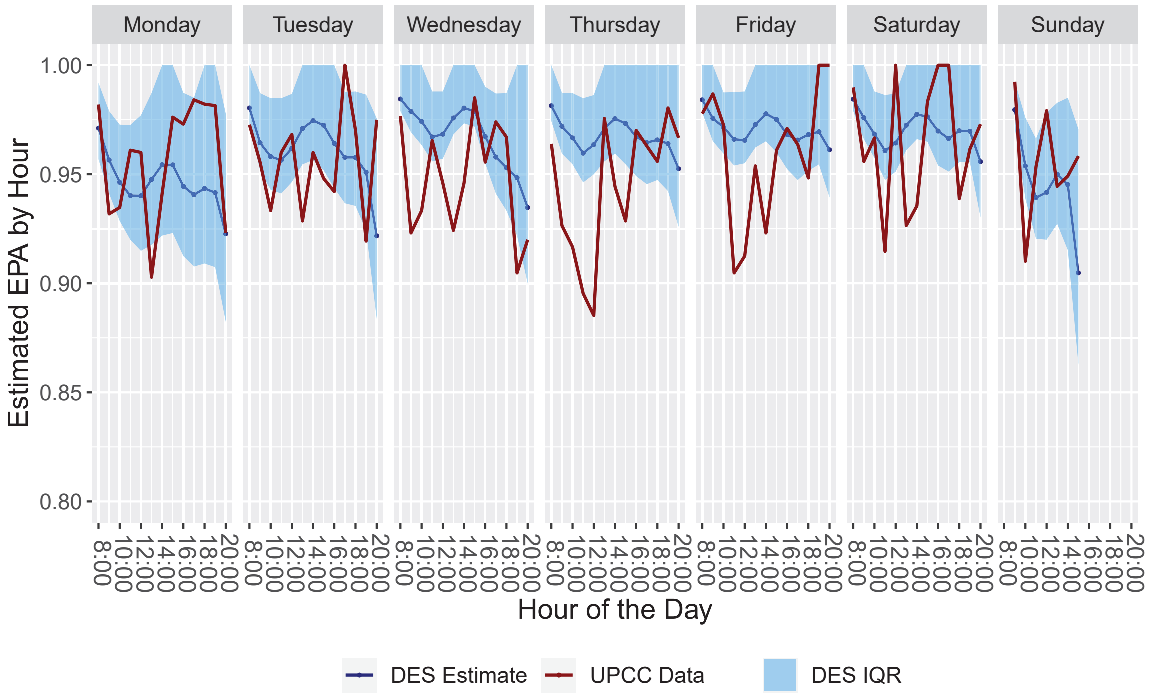

Figure 6.2: Validation plot comparing expected percentage access (EPA) estimates from the UPCC data and DES. Data estimates are based on the last 17 weeks of UPCC operation, with EPA calculated as the mean number receiving care divided by the mean number registering, in each hour of the day and week. The DES estimates were calculated similarly based on arrivals for 85000 simulated weeks. The IQR for the simulation was calculated from individual EPA estimates from each group of 17 simulated weeks.

who will leave promptly if they cannot receive care before they need to return to work. This can been seen in the data as a substantial drop in data EPA estimates around 12 or 1pm. The simulation model does not incorporate time-dependent fluctuations in the distributions of treatment times and patience times, and is able to capture general daily trends in EPA, without over-fitting to these specific weeks of data and detailed human behavior that they are comprised of. The parameters and assumptions of the DES model are sufficient to provide a test case for the analysis of staffing algorithms. Future application of the model can investigate the sensitivity of the simulation and staffing models to different parameter choices.

6.2 Small-interval Staffing Requirements

I applied the sequential-EPA-ISA algorithm (described in Subsection 4.4.1 and Appendix A) to determine staffing requirements for each 15-minute interval that meet EPA targets (prob-

Source	Weekly EPA Estimate
UPCC Data	0.955
DES	0.964

Table 6.1: Weekly EPA estimates from the UPCC data and DES. The UPCC data estimate is from the last 17 weeks of UPCC data, with weekly EPA calculated as the mean number receiving care per week divided by the mean number registering per week. The DES estimates were calculated similarly based on arrivals for 85,000 simulated weeks.

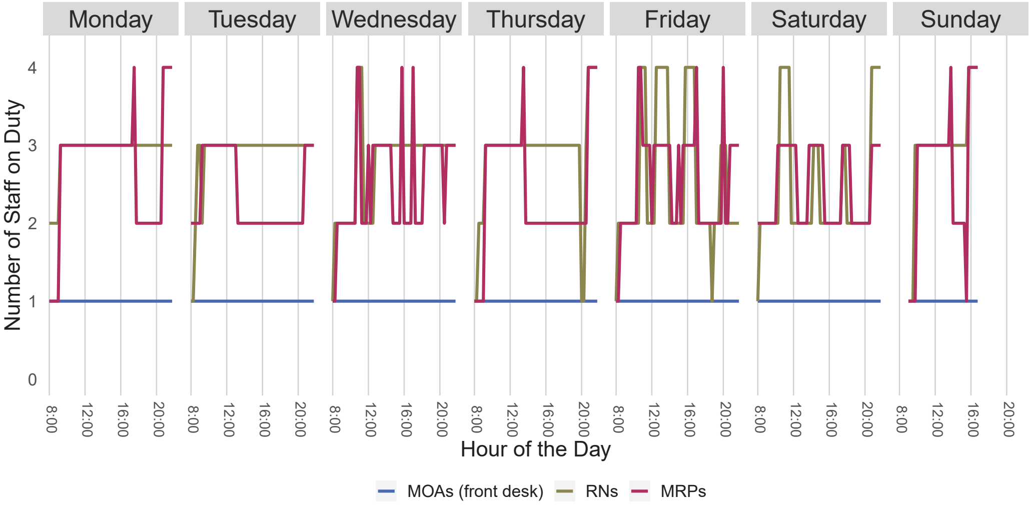

Figure 6.3: Required staffing levels, determined using the sequential-EPA-ISA approach, for 3 staffing disciplines over each 15-minute interval of the week. Each iteration of the sequential-EPA-ISA approach evaluated EPA using 50,000 simulation days. The total staff hours used are 92 MOA hours, 235.25 RN hours and 232 MRP hours.

lem (4.4)). Figure 6.3 shows the identified results, found independently for each day of the week. The simulation EPA estimates under these staffing levels are shown in Figure 6.4.

The small-interval staffing requirements in Figure 6.3 consistently recommend having a single MOA at the front desk. In these results, the number of recommended RNs and MRPs on duty varies from 1 to 4 staff, with a recommendation of 3 RNs and MRPs on duty for most time-intervals of the day and week. The total ISA recommended staff hours are highest on Mondays (Table 6.2), which is the day of the week with the highest number of average arrivals (Table 5.1). Both RN and MRP staffing results show intra-day fluctuation to meet patterns in client demand, with several staffing peaks aligned during or after peaks in client arrivals. For example, on Saturdays, the staffing peaks at 10:30am, 2:15pm, and 5:30pm align after the hourly arrival rate peaks at 10am, 1pm and 4pm. ISA staffing requirements are also strongly influenced by the opening and closing of the clinic; high arrival rates in the first hour after opening each day are mitigated by preceding opening of the UPCC, and staffing requirements are still low in these hours. On several days of the week, the ISA staffing requirements have an additional peak at the end of the day to maintain access

Weekday	Total ISA Staff Hours		
	MOA (font desk)	RN	MRP
Monday	14	40.8	38.0
Tuesday	14	40.5	33.2
Wednesday	14	38.5	35.5
Thursday	14	40.5	34.0
Friday	14	35.2	35.5
Saturday	14	34.0	34.2
Sunday	14	23.8	21.5

Table 6.2: Total ISA recommended staff hours for each day of the week.

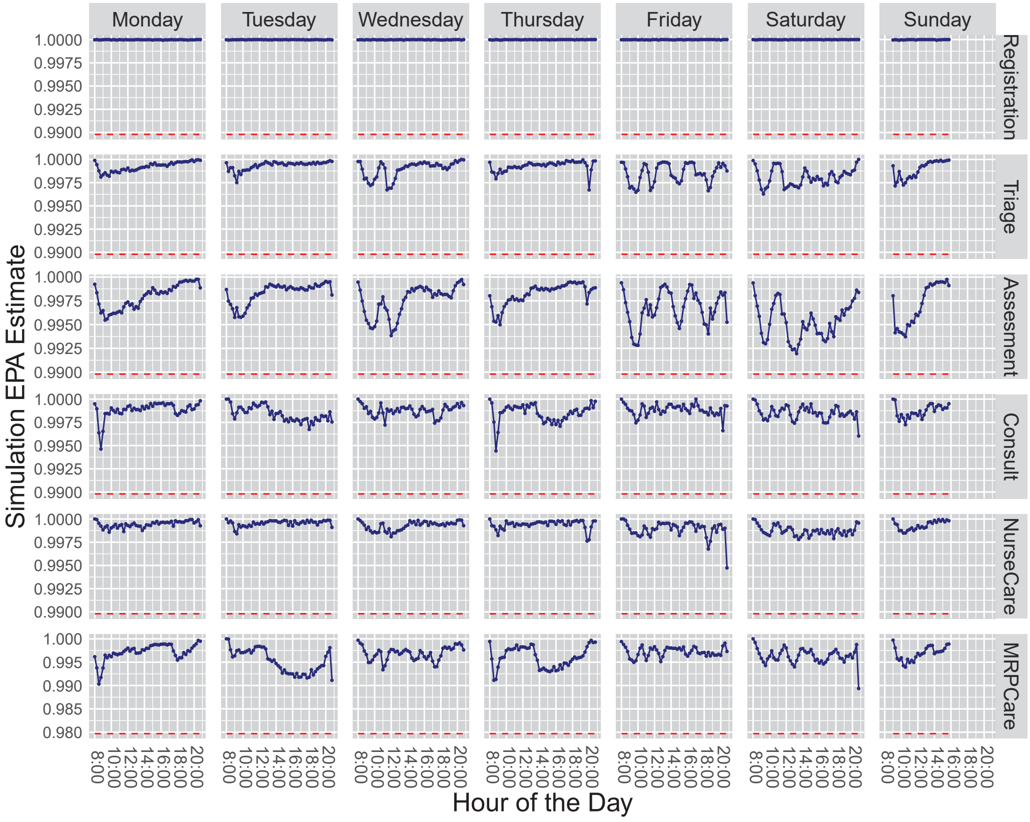

Figure 6.4: Simulation EPA estimates given the ISA staffing levels shown in Figure 6.3, for each component of urgent care and 15-minute interval. Each EPA estimate is evaluated using 50,000 simulated days, with an arbitrary large number of exam rooms (100). The red dashed lines represent EPA targets for each care component, which are chosen to yield an overall 95% access rate—see Section 4.3 for more detail on this choice.

Day	Discipline	Number Phase I Simulation Calls	Number Phase II Simulation Calls
Monday	MOA	1	0
	RN	5	1
	MRP	3	2
Tuesday	MOA	1	0
	RN	3	1
	MRP	3	3
Wednesday	MOA	1	0
	RN	3	1
	MRP	3	1
Thursday	MOA	1	0
	RN	3	2
	MRP	3	2
Friday	MOA	1	0
	RN	10	2
	MRP	3	1
Saturday	MOA	1	0
	RN	3	1
	MRP	5	3
Sunday	MOA	1	0
	RN	3	0
	MRP	4	3

Table 6.3: Number of simulation calls per phase when applying the sequential-EPA-ISA approach across 7 weekdays and 3 staffing disciplines. Each used simulation call estimated EPA by using 50,000 simulated days.

targets before closing time. Under the ISA staffing levels, simulation EPA estimates are displayed in Figure 6.4 and demonstrate that the ISA approach is able to consistently meet EPA targets across each service and time interval.

Table 6.3 records the numbers of simulation calls used in my application of sequential-EPA-ISA, specified by each phase in the two phase algorithm (see Subsection 4.4.1 and Appendix A for a detailed description of these phases). Each simulation call estimated EPA using 50,000 simulated days and all evaluates took around 6.5 minutes to run[2], and these simulation runs were performed without parallelization. For each day, ISA staffing requirements were found in under 20 total simulation calls across three staff types.

6.3 Shift-based Staffing

I applied both integer linear programming and simulation optimization to determine shift-based staffing requirements for two different scenarios, one where staff breaks are not con-

[2] Computation times were evaluated on a 2.3 GHz quad-core Intel core i7 processor.

sidered in the solution, and one where they are. The results for these two scenarios are presented in Subsections 6.3.1 and 6.3.2, respectively.

6.3.1 Shifts without Breaks Considered

I first optimized shift-based staffing for the set of shift options in Table 3.2, without the incorporation of breaks. I applied the integer linear programming (ILP) formulation in equation (4.8) to find the minimum number of staff on each of these shifts needed to cover the small-interval staffing requirements yielded from ISA (shown in Figure 6.3). I also applied simulation optimization to directly solve Problem (4.7) by using the OptQuest meta-heuristic to identify the minimum shift-based staffing required to meet targets on simulated EPA in each 15-minute interval. Figure 6.5 compares the original schedule with both the ILP and OptQuest results. Two results are included for simulation optimization, one in which staffing for RNs and MRPs are optimized in sequence, and one in which staffing for RNs and MRPs are jointly optimized. Each application of simulation optimization used the original staff schedule as a starting point, and had a timeout limit of 500 simulation calls. Each simulation call uses 50,000 simulated days to estimate EPA values in each time interval, assuming an arbitrary large number of exam rooms (100 rooms). I did not consider OptQuest optimization of the number of MOAs at the front desk, since the small-interval staffing requirements in Figure 6.3 are at the lowest possible value for this group. The ILP and simulation optimization approaches were all solved independently for each day of the week, with Sunday shifts restricted to day shifts because of the reduced opening hours. Figure 6.5 displays the three resulting optimized schedules along with the original schedule. For each schedule, the total staffing levels and simulation performance measures are compared in Table 6.4, with time-dependent EPA estimates for each schedule shown in Figure 6.6.

In Figure 6.5, the original schedule typically staffs 3 RNs and 2 MRPs on each morning and evening shift on Mondays through Saturdays, with an additional RN on an afternoon shift to provide break coverage. The optimized schedule from ILP has similar RN staffing levels; however, it substantially increases MRP staffing by adding 1-2 MRPs to each morning and evening shift, with highest new MRP staffing on Wednesday, Thursday, and Friday. The optimized schedule obtained by sequentially applying OptQuest sightly decreases RN staffing levels from the original schedule by removing 1-2 RN shifts from six days of the week, with lowest new RN staffing on Thursday, Friday, and Saturday. The sequential OptQuest results have 6 RN shifts on Monday through Wednesday, and 5 on Thursday through Saturday, as well as 6 MRP shifts on Monday, Tuesday, Thursday, Friday and Saturday and 5 on Wednesday. The optimized schedule obtained from applying OptQuest jointly (to both RNs and MRPs) is similar to the sequential results, but sometimes substitutes MRP for RN staffing on Thursdays through Saturdays, and adds a further additional RN shift on Wednesdays, Thursdays, and Saturdays.

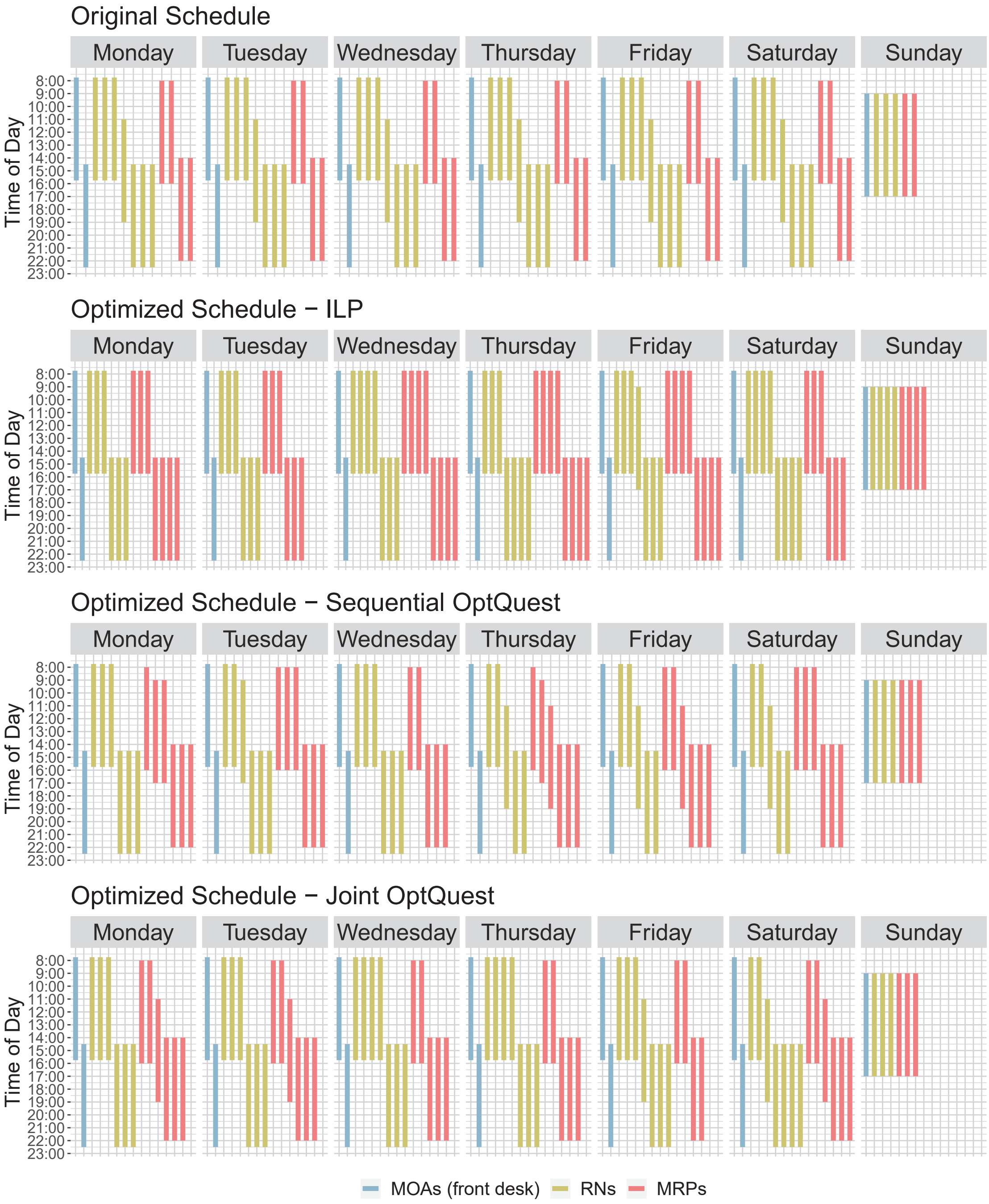

Figure 6.5: Original and optimized shift-based staffing for a scenario that does not consider staff breaks.

Metric		Original Schedule	Optimized Schedules		
			ILP	Sequential OptQuest	Joint OptQuest
Number of Shifts	MOAs	13	13	13	13
	RNs	45	45	34	43
	MRPs	26	47	38	33
Total Active Staff Hours	MOAs	99.5	99.5	99.5	99.5
	RNs	346.0	345.0	262.0	331.75
	MRPs	208.0	360.0	304.0	264.0
Overall Weekly EPA	100 Rooms	0.994	0.998	0.992	0.995
	8 Rooms	0.986	0.996	0.986	0.990
Average number of simulated clients seen per week	100 Rooms	351.3	352.5	351.3	351.6
	8 Rooms	348.1	351.9	348.0	349.8
Average number of simulated clients seen within an hour of arrival, per week	100 Rooms	329.1	344.9	319	333.3
	8 Rooms	323.1	343.1	314.9	330.0
Average number of simulated clients leaving unseen, per week	100 Rooms	1.99	0.6	2.8	1.6
	8 Rooms	5.1	1.3	5.1	3.3

Table 6.4: Staffing levels and simulated performance measures for the original and three optimized schedules, without incorporating staff breaks. For each schedule, the total weekly staffing levels are shown for MOAs at the front desk, RNs, and MRPs. However, the joint and sequential simulation optimization applications only optimized the staffing for RNs and MRPs. Each optimization application was run independently for each day, and was based on simulation EPA estimates from 5,000 simulated days with access to an arbitrary large number of exam rooms (100 rooms). The performance measures shown for each schedule were estimated from 85,000 simulated weeks under two numbers of simulated exam rooms: an arbitrary large capacity (100 exam rooms) and the current capacity (8 exam rooms). Client-centred performance metrics include the average number of simulated clients per week who receive MRP or RN care, who receive MRP or RN care within an hour of arrival, or who leave without receiving care.

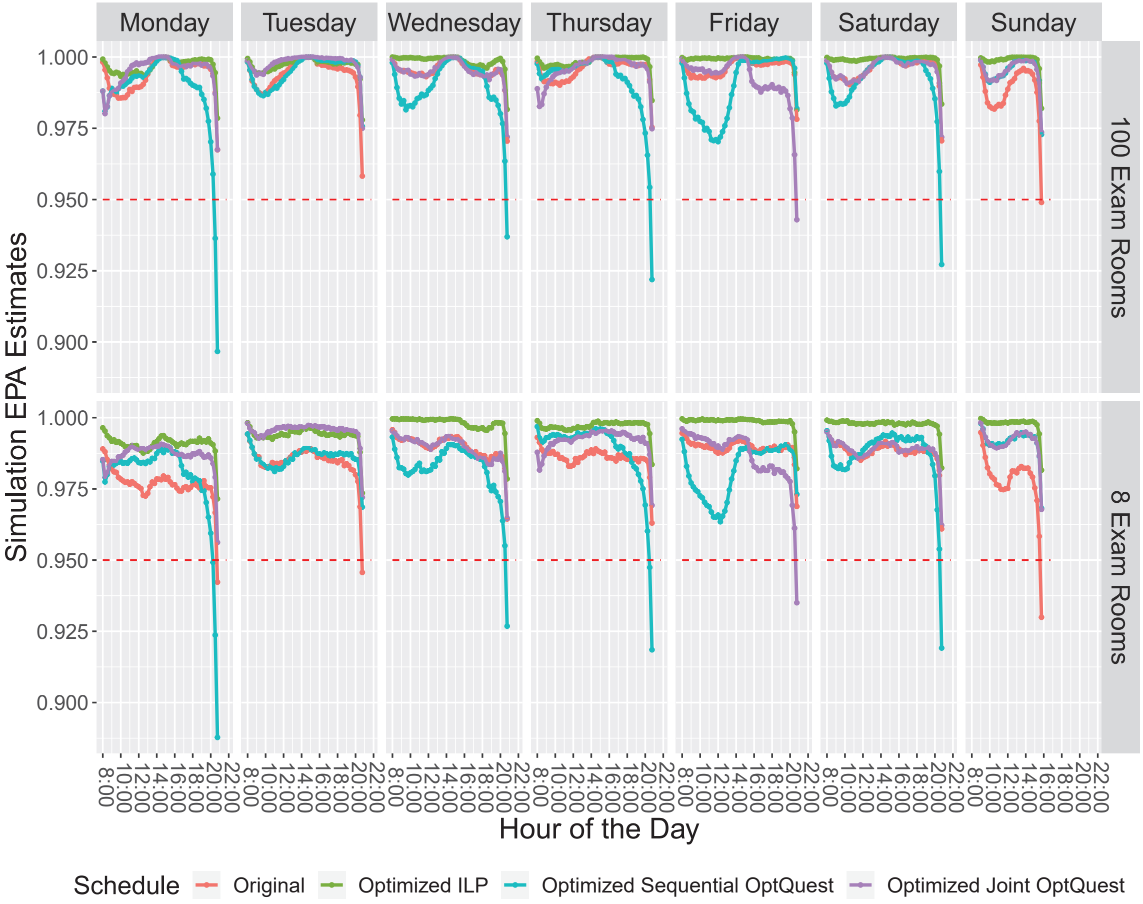

Figure 6.6: Comparison of simulation EPA estimates for the original and optimized schedules, without consideration for breaks. EPA estimates are shown for each 15-minute of the day, calculated based on client outcomes in 50,000 simulated days. Two sets of EPA estimates are shown, one with an arbitrary large number of exam rooms (100 rooms), and one with the current number of exam rooms at the UPCC (8 rooms).

Overall, the ILP optimized schedule has the best estimated client-centred performance measures compared to the original and other optimized schedules, with consistently high simulation EPA values throughout the week (See Figure 6.6), almost no simulated clients who leave without being seen, and the highest number of clients who are seen within an hour of arrival (see Table 6.4). However, the ILP optimized schedule also has the highest total staffing of these options, with a net total staffing increase of 151 staff hours per week compared to the original schedule. On the other hand, both OptQuest optimized schedules use less staff than the ILP schedule, but do not consistently meet EPA targets (Figure 6.6. The joint OptQuest schedule has only 7 more MRP shifts per week than the original schedule and 2 less RN shifts, and reduces the number of simulation clients who leave without being seen by 34% (1.8 clients per week), and increase the number of clients seen within an hour by 2.1% (6.9 simulation clients per week)—under the current number of exam rooms.

Table 6.5 compares the number of simulation runs used in the three approaches to shift-based staffing without breaks. Solving the deterministic ILP formulation required no further simulation runs beyond the simulation required for ISA, all under 10 simulation runs. Each daily ILP optimization was solved independently with Python[3] using the Gurobi solver[4], and all computation times were less than a tenth of a second[5]. In comparison, simulation optimization with the OptQuest meta-heuristic required at least 300 simulation calls for Monday through Saturday, with sequential optimization requiring a total of at least 450 simulation calls per weekday. Each simulation call involves 5,000 simulated days and can run in around one minute[6]. OptQuest runs were performed in parallel over at least 8 cores.

[3] `https://www.python.org/`

[4] `https://www.gurobi.com/`

[5] Computation times were evaluated on a 2.3 GHz quad-core Intel core i7 processor.

[6] Computation times were evaluated on a 2.3 GHz quad-core Intel core i7 processor

Weekday	Optimization Approach			
	ILP + ISA	Sequential OptQuest		Joint OptQuest
	(RNs & MRPs)	RNs	MRPs	(RNs & MRPs)
Monday	11	221	242	312
Tuesday	10	244	220	329
Wednesday	8	232	230	354
Thursday	10	221	237	311
Friday	16	227	231	357
Saturday	12	238	233	323
Sunday	10	5	5	25

Table 6.5: Number of simulation calls used for each shift-based optimization approach applied to a set of shifts without breaks considered. ILP schedule optimization does not use any simulation runs beyond those used in the ISA, which was run using 500,000 simulated days per simulation call. For both the joint and sequential simulation optimization approaches, performance measures were estimated using 5,000 simulated days per simulation call. OptQuest simulation optimization runs until the meta-heuristic determines that no significant improvements are made, or a maximum of 500 simulation calls has been reached.

6.3.2 Shifts with Breaks Considered

To address an important operational factor in staff scheduling, I also performed optimization of shift-based staffing using a set of shift options that incorporate staff breaks. I used the same three approaches from Subsection 6.3.1—namely ISA and ILP, sequential OptQuest, and joint OptQuest—to optimize weekly shift-based staffing that incorporates breaks. I used the ILP formulation in equation (4.8) to find the minimum number of RNs and MRPs on each of the shifts and break combinations in Table 4.3 needed to cover the small-interval staffing requirements yielded from ISA (shown in Figure 6.3). I used shift options without breaks for ILP optimization of front desk MOA shifts, since the breaks for MOAs at the front desk are covered by the MOAs who work in the clinical pod. I also applied the OptQuest meta-heuristic to perform simulation optimization of RN and MRP staffing, both sequentially and jointly. Each application of simulation optimization used the original staff schedule as a starting point, and had a timeout limit of 500 simulation calls. Each simulation call uses 5,000 simulated days to estimate EPA values in each 15-minute time interval and assumes an arbitrary large number of exam rooms (100 rooms). The ILP and simulation optimization approaches were all solved independently for each day of the week, with Sunday shifts restricted to day shifts to reflect the UPCC opening hours. Figure 6.7 displays the three resulting optimized schedules along with the original schedule. For each schedule, the total staffing levels and overall simulation performance measures are compared in Table 6.6, with time-dependent EPA estimates for each schedule shown in Figure 6.8. Table 6.7 compares the number of simulation runs used in each optimization approach.

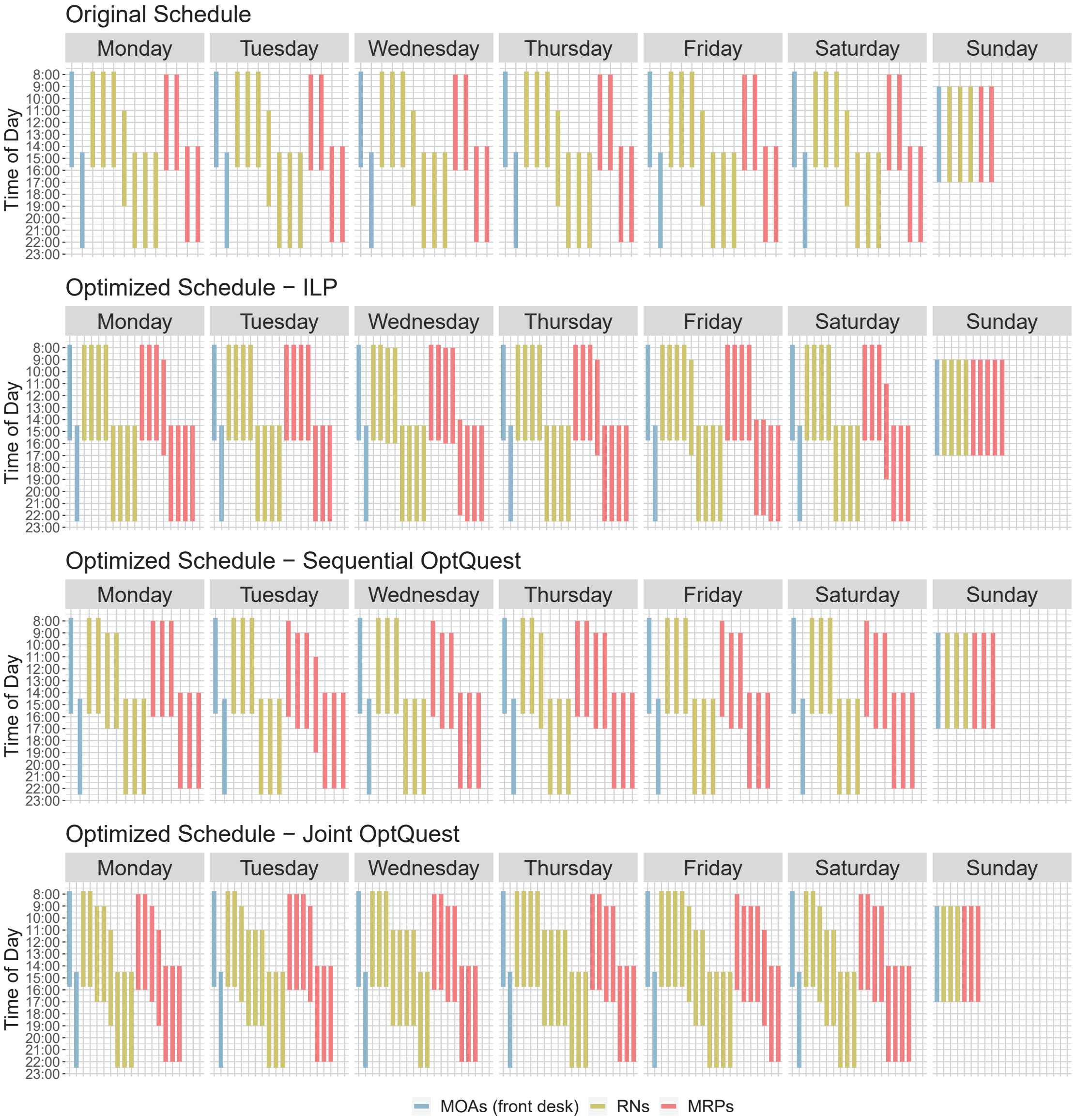

Figure 6.7: Original and optimized shift-based staffing for a scenario that considers staff breaks.

Metric		Original Schedule	Optimized Schedules		
			ILP	Sequential OptQuest	Joint OptQuest
Number of Shifts	MOAs	13	13	13	13
	RNs	45	52	40	59
	MRPs	26	50	41	46
Total Active Staff Hours	MOAs	99.50	99.50	99.50	99.50
	RNs	301.50	353.50	267.00	399.75
	MRPs	182.00	342.75	287.00	322.00
Overall Weekly EPA	100 Rooms	0.984	0.998	0.991	0.996
	8 Rooms	0.965	0.996	0.985	0.994
Average number of simulated clients seen per week	100 Rooms	347.8	352.6	350.0	352.0
	8 Rooms	340.9	351.8	347.9	351.0
Average number of simulated clients seen within an hour of arrival, per week	100 Rooms	283.5	344.9	311.8	340.3
	8 Rooms	272.3	343.6	307.6	338.5
Average number of simulated clients leaving unseen, per week	100 Rooms	5.5	0.7	3.1	1.2
	8 Rooms	12.4	1.3	5.23	2.1

Table 6.6: Staffing levels and simulated performance measures for the original and three optimized schedules, which all incorporate staff breaks for RNs and MRPs. For each schedule, the total weekly staffing levels are shown for MOAs at the front desk, RNs, and MRPs. Active staff hours exclude breaks and shift time outside opening hours. Simulation optimization (both joint and sequential) only optimized for the shift-based staffing of RNs and MRPs. The performance measures shown were estimated from 50,000 simulated weeks under two numbers of simulated exam rooms: an arbitrary large number (100 exam rooms) and the current capacity (8 exam rooms). Client-centred performance metrics include the average number of simulated clients per week who receive MRP or RN care, who receive MRP or RN care within an hour of arrival, or who leave without receiving care.

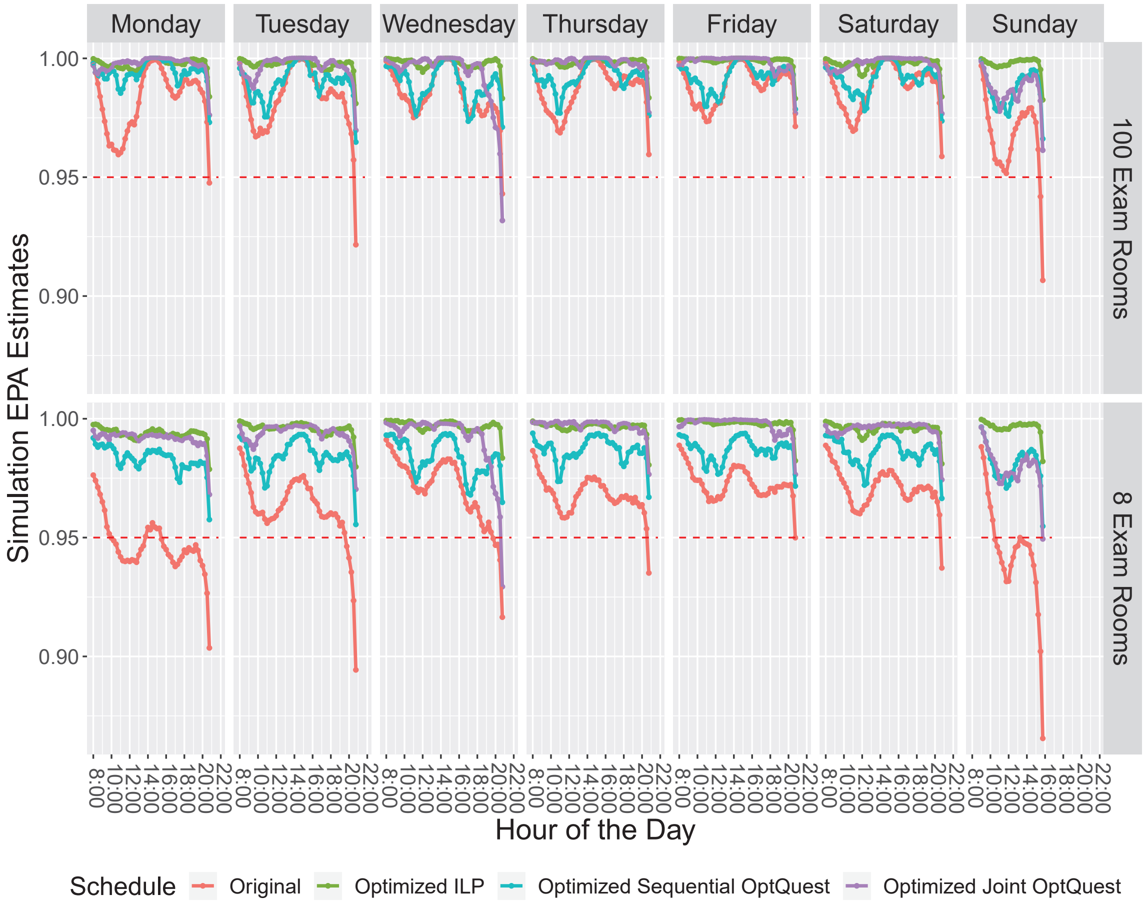

Figure 6.8: Comparison of simulation EPA estimates for the original and optimized schedules, with consideration for breaks. The red dashed line represents the target EPA value. EPA estimates are calculated as the mean number of simulated clients receiving RN or MRP treatment divided by the mean number of simulated client arrivals, broken up into 15-minute intervals based on the arrival time at the UPCC. The shown estimates are made using 50,000 simulated days, and under either an arbitrarily large number of exam rooms (100 rooms) or the current number of exam rooms at the UPCC (8 rooms).

In Figure 6.5, the original schedule typically staffs 3 RNs and 2 MRPs on each morning and evening shift on Mondays through Saturdays, with an additional RN on an afternoon shift to provide break coverage. The optimized schedule from ILP increases staffing levels by having 4 RNs and 3-4 MRPs on each morning and evening shift on Mondays through Saturdays, with an additional day or afternoon shift for MRPs on Monday, Thursday, and Saturday, and 4 RNs and MRPs on the Sundays. The optimized schedule obtained by sequentially applying OptQuest sightly decreases RN staffing levels from the original schedule by removing the afternoon RN shift from six days of the week, although keeping a day shift on Mondays. The sequential OptQuest results typically have 6-7 MRP shifts scheduled per day on Monday through Saturday. The optimized schedule obtained from applying OptQuest jointly (to both RNs and MRPs) has substantially higher staffing levels than the other schedules, with 8-11 RN shifts and 7-8 MRP shifts on Mondays through Saturdays.

Overall, the ILP optimized schedule has the best estimated client-centred performance measures compared to the original and other optimized schedules, with consistently high simulation EPA values throughout the week (See Figure 6.8), almost no simulated clients who leave without being seen, and the highest number of clients who are seen within an hour of arrival (see Table 6.6). However, the ILP optimized schedule also has the highest total staffing levels, with a net staffing increase of 212.75 active staff hours per week compared to the original schedule. On the other hand, sequential simulation optimization identified a schedule that meets the client access targets and uses 142.25 less active staff hours than the ILP schedule. Compared to the original schedule, the sequential OptQuest schedule reduces RN staffing by around 35 hours and increases MRP staffing by 105 hours; in the simulation, these changes reduces the number of clients who leave without being seen by 58% (7 more clients seen per week), and increase the number of clients seen within an hour by 13% (35 more clients per week), under the current number of exam rooms. However, under the optimization scenario of 100 exam rooms, the original staff schedule meets the access targets on Thursday, Friday, and Saturday; indicating that the additional staffing used in the sequential OptQuest solution—which was found using a smaller number of simulation runs per call than the post-optimization evaluation—may not be necessary. Furthermore, the schedule found by joint simulation optimization uses substantially more staff hours per week than the sequential simulation optimization results, with 133 more RN hours and 35 more MRP hours.

Table 6.7 compares the number of simulation runs used in the three approaches to shift-based staffing without breaks. Solving the deterministic ILP formulation required no further simulation runs beyond the simulation required for ISA, all under 20 simulation runs. Each daily ILP optimization was solved independently with Python[7] using the Gurobi

[7] https://www.python.org/

Weekday	Optimization Approach			
	ILP + ISA	Sequential OptQuest		Joint OptQuest
	(RNs & MRPs)	RNs	MRPs	(RNs & MRPs)
Monday	11	360	339	318
Tuesday	10	355	376	318
Wednesday	8	328	289	311
Thursday	10	375	284	304
Friday	16	351	297	334
Saturday	12	366	281	330
Sunday	10	78	80	272

Table 6.7: Number of simulation calls used for each shift-based optimization approach applied to a set of shifts that consider breaks. ILP schedule optimization does not use any simulation runs beyond those used in the ISA, which was run using 500,000 simulated days per simulation call. For both the joint and sequential simulation optimization approaches, performance measures were estimated using 5,000 simulated days per simulation call. OptQuest simulation optimization runs until the meta-heuristic determines that no significant improvements are made, or a maximum of 500 simulation calls has been reached.

solver[8], and all computation times were less than a tenth of a second[9]. In comparison, simulation optimization with the OptQuest meta-heuristic required at least 300 simulation calls for Monday through Saturday, with sequential optimization requiring a total of at least 600 simulation calls per weekday. Each simulation call involves 5,000 simulated days and can run in around one minute minutes[10]. Simulation runs for OptQuest optimization were performed in parallel over at least 8 cores.

[8] `https://www.gurobi.com/`

[9] Computation times were evaluated on a 2.3 GHz quad-core Intel core i7 processor

[10] Computation times were evaluated on a 2.3 GHz quad-core Intel core i7 processor

Chapter 7

Discussion and Conclusions

7.1 Discussion of Results

I developed a new approach to optimizing team-based staffing for an urgent and primary care centre in Vancouver, Canada. My approach determines minimum staffing levels for each healthcare discipline, based on meeting client-centred access targets in a queueing model representation of the urgent care stream. I used a network queueing model and simulation implementation to represent the stochastic flow of clients through different components of urgent care at the centre, based on the operational profile of the UPCC and client visit data. I built a discrete event simulation (DES) that estimates client-centred access indicators for urgent care and incorporates the sequential and interactive nature of team-based care. I then embedded the simulation into staffing optimization procedures to identify both small-interval staffing requirements and shift-based staffing levels needed to meet access targets. My optimization analysis focuses on the expected proportion of model clients who receive care, as opposed to leaving the clinic without being seen due to an extended wait.

I identified small-interval staffing requirements by extending the iterative staffing algorithm (ISA) [44, 52] to consider multiple staff types and expected access proportions. I applied the ISA framework to determine the staffing levels needed to meet access targets in each 15-minute interval by iterating over simulation output and sequentially considering each staffing discipline. The resulting small-interval staffing requirements fluctuate during the day and week in order to meet modeled client demand, without needing to incorporate shift options. In my results, ISA staffing levels follow some of the trends in client arrival rates; higher daily staffing is typically reflective of higher daily arrivals and some intraday staffing peaks occur during or after arrival rate peaks. However, the ISA results also demonstrate that UPCC opening and closing can have a bigger effect than arrival rates on expected percentage access, and highlight a consequence of optimizing solely on whether or not clients leave without being seen.

My analysis compares two different approaches to determine shift-based staffing requirements for the urgent care stream at the UPCC. Firstly, I applied an integer linear programming (ILP) formulation to minimize the staffing level on each shift such that the ISA small-interval requirements are satisfied [42, 45]. Secondly, I used simulation optimization to select shift-based staffing that satisfy simulation client access targets in each 15-minute interval—measured directly for each combination of shifts—by using the OptQuest metaheuristic. The resulting schedules for both approaches recommend that physician staffing could be increased from the original schedule in order to better meet targets on client access. In particular, The OptQuest results show that by adding 15 physician shifts per week, simulation client access increased by around 7 clients seen per week, which reduced the number of simulation clients leaving without being seen by 58%, and also increased the number of clients seen within an hour by 13% (35 clients). Increased client access to care can improve client outcomes, reduce emergency department visits, boost health equity, and balance staff workloads. Because the simulation slightly overestimates the number of clients seen compared to historic data, and because additional work—including administrative duties and calling clients regarding test results—is not captured, these staffing results underestimate the true staffing requirements at the UPCC. The original UPCC staffing levels were chosen based on average demand projections made before the clinic began operating. In comparison, my approach makes staffing recommendations based on UPCC data and modeled client-centred outcomes that directly correspond to client access to care. To my knowledge, no other studies optimize team-based primary or urgent care using time-dependent client access measures.

All three staffing optimization procedures in my analysis (ISA/ILP, sequential OptQuest, and joint OptQuest) were performed using simulations with an arbitrarily large number of exam rooms, and the resulting schedules do not always meet access targets under the current exam room capacity. This highlights the impact of limited rooms on client access, and suggests that strategies are needed to maximize the utilization of physical space.

In this application, ILP was the most computationally efficient approach to shift-based staffing, since it did not require any further simulation calls beyond those needed to identify the ISA small-interval requirements, which was under 20 simulation calls for each day. The ILP formulation was solved independently for each staffing discipline and day of the week, and it quickly determined staffing for a range of shift options that incorporated timing options for staff breaks. In comparison, simulation optimization was substantially more computationally intensive, often exceeding hundreds of simulation calls for each day. However, the ILP approach has been known to lead to over staffing [45, 83], and in my application, simulation optimization was more accurate, since it identified a schedule that addresses simulation access targets and uses less staff than the ILP solution. However, this accuracy decreased when simulation optimization was performed to jointly optimize for both nurse and physician staffing, under which a local optima was returned. While sequential simulation optimization for each staff type produced the smallest estimated staffing requirements, it also used the most simulation calls, and on some days the results do not meet access targets since the number of runs per simulation call was lower in the optimization procedure than comparison.

The trade-off between accuracy and efficiency of these optimization approaches highlight the challenge of finding useful simulation optimization results under the increased number of variables needed to consider team-based care and multiple shift options that incorporate the impact of breaks. To my knowledge, no other studies optimize small-interval or shift-based staffing for multiple types of staff by using time-dependent expected proportions of client-centred outcomes.

7.2 Limitations and Future Work

Ongoing collaboration with the UPCC will support further extension and implementation of the analysis in this thesis. Future work will extend the analysis to optimize staffing for social workers and longitudinal care services at the UPCC, which will open later this year. For these services, the stochastic modeling of booked appointments can extend to consider KPIs for the indirect wait time between appointment booking and the next available appointment slot. The simulation for both urgent and longitudinal care can be used to explore different workflow procedures at the UPCC, including strategies to optimize the use of limited physical space.

The analysis in this thesis focuses on a sole performance measure, but the approach generalizes to include other client-centred access indicators. Future analysis could incorporate modeled client wait times, which are an important part of primary care access [19]. Further work could incorporate the impact of client acuity levels on client access and could use acuity specific access targets, including models where priority accumulates over time [36, 157].

The staffing optimization results in this thesis motivate further investigation into more accurate and efficient hybrid simulation optimization techniques, which could incorporate the impact of limited exam room capacity. The optimization objective in my analysis minimized total staffing hours and some simulation optimization results demonstrated a substitution between staff types. Future work could incorporate objectives that weigh the relative staffing costs for each healthcare discipline. Additional analysis could explore the sensitivity of the staffing results to different model parameter choices and inform robust staffing optimization. The model can be combined with parameter projections to provide long-term capacity planning.

7.3 Conclusions

The analysis in this thesis provide a new approach to optimize staffing based on client-centred access indicators in a team-based urgent care context. By combining embedded simulation with optimization searches, my analysis quantifies the interplay between staffing levels, team-based care interaction, and client access to care, which is an acknowledged need in primary care [153]. The results make staffing recommendations that can improve care access by maintaining targets throughout the day and week. The approach in this thesis can be applied to other urgent care centres [137] or walk-in clinics [29].

To address team-based staffing, my analysis extends stabilization techniques [44] to incorporate multiple staff types and observation-based performance measures. The work in this thesis contributes to the larger body of work on staffing optimization and capacity planning. This approach is more broadly applicable in settings outside of urgent care that provide interdisciplinary services, for example in emergency departments [44, 155], collaborative emergency centres [38, 165], intensive case management [141], and the organizational structure of corporate work teams [158].

In the BC context, additional urgent and primary care centres are being established as part of a provincial initiative to increase healthcare and reduce low-acuity emergency department admissions [134, 78]. My approach can inform the successful implementation of team-based urgent and primary care care and contribute to improving access to community-based healthcare. Ongoing collaboration will incorporate this analysis into a learning health system that supports continuous operational improvement [56, 124].

Bibliography

[1] Registered nurses scope of practice: Standards, limits, conditions. Technical report, The British Columbia College of Nurses and Midwives, 2020. Available at: `https://www.bccnm.ca/Documents/standards_practice/rn/RN_ScopeofPractice.pdf`.

[2] Salah Aguir, Fikri Karaesmen, O. Zeynep Akşin, and Fabrice Chauvet. The impact of retrials on call center performance. *OR Spectr*, 26(3):353–376, 2004.

[3] Amir Ahmadi-Javid, Zahra Jalali, and Kenneth J. Klassen. Outpatient appointment systems in healthcare: A review of optimization studies. *Eur J Oper Res*, 258(1):3–34, 2017.

[4] Mohamed A. Ahmed and Talal M. Alkhamis. Simulation optimization for an emergency department healthcare unit in Kuwait. *Eur J Oper Res*, 198(3):936–942, 2009.

[5] Uwe Aickelin and Kathryn A. Dowsland. Exploiting problem structure in a genetic algorithm approach to a nurse rostering problem. *J Sched*, 3(3):139–153, 2000.

[6] Uwe Aickelin and Kathryn A. Dowsland. An indirect genetic algorithm for a nurse-scheduling problem. *Comput Oper Res*, 31(5):761–778, 2004.

[7] Anthony J. Alessandra, Ted E. Grazman, Ravi Parameswaran, and Ugur Yavas. Using simulation in hospital planning. *Simul*, 30(2):62–67, 1978.

[8] Mor Armony and Avishai Mandelbaum. Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. *Oper Res*, 59(1):50–65, 2011.

[9] Júlíus Atlason, Marina A. Epelman, and Shane G. Henderson. Call center staffing with simulation and cutting plane methods. *Ann Oper Res*, 127(1):333–358, 2004.

[10] Júlíus Atlason, Marina A. Epelman, and Shane G Henderson. Optimizing call center staffing using simulation and analytic center cutting-plane methods. *Manage Sci*, 54(2):295–309, 2008.

[11] Turgut Aykin. Optimal shift scheduling with multiple break windows. *Manage Sci*, 42(4):591–602, 1996.

[12] M. N. Azaiez and S. S. Al Sharif. A 0-1 goal programming model for nurse scheduling. *Comput Oper Res*, 32(3):491–507, 2005.

[13] Jonathan F. Bard and Hadi W. Purnomo. Incremental changes in the workforce to accommodate changes in demand. *Health Care Manag Sci*, 9:71–85, 2006.

[14] Stephen E. Bechtold and Larry W. Jacobs. Implicit modeling of flexible break assignments in optimal shift scheduling. *Manage Sci*, 36(11):1339–1351, 1990.

[15] Ilham Berrada, Jacques A. Ferland, and Philippe Michelon. A multi-objective approach to nurse scheduling with both hard and soft constraints. *Soc Econ Plann Sci*, 30(3):183–193, 1996.

[16] Marco Better and Fred Glover. Selecting project portfolios by optimizing simulations. *Eng Econ*, 51(2):81–97, 2006.

[17] Papiya Bhattacharjee and Pradip Kumar Ray. Simulation modelling and analysis of appointment system performance for multiple classes of patients in a hospital: A case study. *Oper Res Health Care*, 8:71–84, 2016.

[18] Rachel Floersheim Boaz. Manpower utilization by subsidized family planning clinics: An economic criterion for determining the professional skill-mix. *J Hum Resour*, 7(2):191–207, 1972.

[19] Thomas Bodenheimer, Amireh Ghorob, Rachel Willard-Grace, and Kevin Grumbach. The 10 building blocks of high-performing primary care. *Ann Fam Med*, 12(2):166–171, 2014.

[20] Lawrence Brown, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, and Linda Zhao. Statistical analysis of a telephone call center:A queueing-science perspective. *J Am Stat Assoc*, 100(469):36–50, 2005.

[21] Jens O. Brunner, Jonathan F. Bard, and Rainer Kolisch. Midterm scheduling of physicians with flexible shifts using branch and price. *IIE Trans*, 43(2):84–109, 2010.

[22] Elwood S. Buffa, Michael J. Cosgrove, and Bill J. Luce. An integrated work shift scheduling system. *Decis Sci*, 7(4):620–630, 1976.

[23] Edmund K. Burke, Peter Cowling, Patrick De Causmaecker, and Greet Vanden Berghe. A memetic approach to the nurse rostering problem. *Appl Intell*, 15:199–214, 2001.

[24] Edmund K. Burke, Timothy Curtois, Gerhard Post, Rong Qu, and Bart Veltman. A hybrid heuristic ordering and variable neighbourhood search for the nurse rostering problem. *Eur J Oper Res*, 188(2):330–341, 2008.

[25] Edmund K. Burke, Patrick De Causmaecker, Greet Vanden Berghe, and Hendrik Van Landeghem. The state of the art of nurse rostering. *J Sched*, 7:441–499, 2004.

[26] Arnold Buss and Ahmed Al Rowaei. A comparison of the accuracy of discrete event and discrete time. In Björn Johansson, Sanjay Jain, Jairo Montoya-Torres, Joe Hugan, and Enver Yücesan, editors, *Proceedings of the 2010 Winter Simulation Conference, Dec 5–8, Baltimore, MD, USA*, pages 1468–1477. ACM, 2010.

[27] Ignacio Castillo, Tarja Joro, and Yong Yue Li. Workforce scheduling with multiple objectives. *Eur J Oper Res*, 196(1):162–170, 2009.

[28] Tugba Çayirli, Pinar Dursun, and Evrim Didem Güneş. An integrated analysis of capacity allocation and patient scheduling in presence of seasonal walk-ins. *Flex Serv Manuf J*, 31:524–561, 2019.

[29] Tugba Çayirli and Evrim Didem Güneş. Outpatient appointment scheduling in presence of seasonal walk-ins. *J Oper Res Soc*, 65(4):512–531, 2014.

[30] Tugba Çayirli and Emre Veral. Outpatient scheduling in health care: A review of literature. *Prod Oper Manag*, 12(4):519–549, 2003.

[31] Tugba Çayirli, Emre Veral, and Harry Rosen. Designing appointment scheduling systems for ambulatory care services. *Health Care Manag Sci*, 9:47–58, 2006.

[32] Tugba Çayirli, Emre Veral, and Harry Rosen. Assessment of patient classification in appointment system design. *Prod Oper Manag*, 17(3):338–353, 2008.

[33] Tugba Çayirli, Kum Khiong Yang, and Ser Aik Quek. A universal appointment rule in the presence of no-shows and walk-ins. *Prod Oper Manag*, 21(4):682–697, 2012.

[34] B. Cheang, H. Li, A. Lim, and B. Rodrigues. Nurse rostering problems—a bibliographic survey. *Eur J Oper Res*, 151(3):447–460, 2003.

[35] Hong Chen and David D. Yao. *Single Station Queues*, volume 46 of *Applications of Mathematics: Stochastic Modelling and Applied Probability*, chapter 6, pages 125–155. Springer, New York, 2001.

[36] Marta Cildoz, Amaia Ibarra, and Fermin Mallor. Accumulating priority queues versus pure priority queues for managing patients in emergency departments. *Oper Res Health Care*, 23:100224, 2019.

[37] Gordon M. Clark. Use of Polya distributions in approximate solutions to nonstationary M/M/s queues. *Commun ACM*, 24(4):206–217, 1981.

[38] Alison M. Coates. Sustaining rural access to emergency care through collaborative emergency centres in Nova Scotia. *Health Reform Obs*, 7(2):Article 2, 2019.

[39] Albert Corominas and Amaia Lusa. LETRIS: Staffing service systems by means of simulation. *J Ind Eng Manag*, 5(2):285–296, 2012.

[40] Stefan Creemers, Mieke Defraeye, and Inneke Van Nieuwenhuyse. G-RAND: A phase-type approximation for the nonstationary G(t)/G(t)/s(t)+G(t) queue. *Perform Eval*, 80:102–123, 2014.

[41] Federico Della Croce and Fabio Salassa. A variable neighborhood search based matheuristic for nurse rostering problems. *Ann Oper Res*, 218:185–199, 2014.

[42] George B. Dantzig. Letter to the editor—A comment on Edie's "Traffic delays at toll booths". *J Oper Res Soc Am*, 2(3):339–341, 1954.

[43] Jimmie L. Davis, William A. Massey, and Ward Whitt. Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Manage Sci*, 41(6):1107–1116, 1995.

[44] Mieke Defraeye and Inneke Van Nieuwenhuyse. Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm. *Decis Support Syst*, 54(4):1558–1567, 2013.

[45] Mieke Defraeye and Inneke Van Nieuwenhuyse. Personnel scheduling in queues with time-varying arrival rates: Applications of simulation-optimization. In Gabriella Dellino and Carlo Meloni, editors, *Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, volume 59 of *Operations Research/Computer Science Interfaces Series*, chapter 9, pages 203–223. Springer, New York, NY, 2015.

[46] Mieke Defraeye and Inneke Van Nieuwenhuyse. Staffing and scheduling under non-stationary demand for service: A literature review. *Omega*, 58:4–25, 2016.

[47] Kathryn A. Dowsland. Nurse scheduling with tabu search and strategic oscillation. *Eur J Oper Res*, 106(2):393–407, 1998.

[48] Stephen G. Eick, William A. Massey, and Ward Whitt. The physics of the $M_t/G/\infty$ queue. *Oper Res*, 41(4):731–742, 1993.

[49] Agner Krarup Erlang. Sandsynlighedsregning og telefonsamtaler. *Nyt Tidsskrift for Matematik*, 20(B):33–39, 1909.

[50] A. T. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier. Staff scheduling and rostering: A review of applications, methods and models. *Eur J Oper Res*, 153(1):3–27, 2004.

[51] Pamela Fayerman. New Vancouver urgent care centre already risks relocation for condo development. *The Vancouver Sun*, 2019.

[52] Zohar Feldman, Avishai Mandelbaum, William A. Massey, and Ward Whitt. Staffing of time-varying queues to achieve time-stable performance. *Manage Sci*, 54(2):324–338, 2008.

[53] Tian Feng, Xing Qing-hua, Wang San-tao, and Zhang Wei. Simulation and optimization of vehicle scheduling in flight logistic support process based on arena. In *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, volume 15, pages V15–446–V15–449, 2010.

[54] Daniela Fishbein, Siddhartha Nambiar, Kendall McKenzie, Maria Mayorga, Kristen Miller, Kevin Tran, Laura Schubel, Joseph Agor, Tracy Kim, and Muge Capan. Objective measures of workload in healthcare: A narrative review. *Int J Health Care Qual Assur*, 33(1):1–17, 2020.

[55] Lori S. Franz, Hope M. Baker, G. Keong Leong, and Terry R. Rakes. A mathematical model for scheduling and staffing multiclinic health regions. *Eur J Oper Res*, 41(3):277–289, 1989.

[56] Lawrence M. Friedman, Curt D. Furberg, David L. DeMets, David M. Reboussin, and Christopher B. Granger. *Fundamentals of clinical trials*. Springer, 2015.

[57] Ofer Garnet, Avishai Mandelbaum, and Martin Reiman. Designing a call center with impatient customers. *Manuf Serv Oper Manag*, 4(3):208–227.

[58] Michel Gendreau and Jean-Yves Potvin. *Handbook of Metaheuristics*, volume 272 of *International Series in Operations Research & Management Science*, page vii. Springer, 3 edition, 2019.

[59] Michel Gendreau and Jean-Yves Potvin. Tabu search. In Michel Gendreau and Jean-Yves Potvin, editors, *Handbook of Metaheuristics*, volume 272 of *International Series in Operations Research & Management Science*, chapter 2, pages 37–55. Springer, 2019.

[60] Amireh Ghorob and Thomas Bodenheimer. Sharing the care to improve access to primary care. *N Engl J Med*, 366:1955–1957, 2012.

[61] Fred Glover, Manuel Laguna, and Rafael Martí. Scatter search. In *Advances in Evolutionary Computation: Theory and Applications*, Natural Computing Series, pages 519–537. Springer, 2003.

[62] Ashley Goodman, Kim Fleming, Nicole Markwick, Tracey Morrison, Louise Lagimodiere, Thomas Kerr, and Western Aboriginal Harm Reduction Society. "They treated me like crap and I know it was because I was Native": The healthcare experiences of Aboriginal peoples living in Vancouver's inner city. *Soc Sci Med*, 178:87–94, 2017.

[63] Linda V. Green and Peter J. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Manage Sci*, 37(1):84–97, 1991.

[64] Linda V. Green and Peter J. Kolesar. On the accuracy of the simple peak hour approximation for Markovian queues. *Manage Sci*, 41(8):1353–1370, 1995.

[65] Linda V. Green and Peter J. Kolesar. The lagged PSA for estimating peak congestion in multiserver Markovian queues with periodic arrival rates. *Manage Sci*, 43(1):80–87, 1997.

[66] Linda V. Green, Peter J. Kolesar, and João Soares. Improving the SIPP approach for staffing service systems that have cyclic demands. *Oper Res*, 49(4):549–564, 2001.

[67] Linda V. Green, Peter J. Kolesar, and João Soares. An improved heuristic for staffing telephone call centers with limited operating hours. *Prod Oper Manag*, 12(1):46–61, 2003.

[68] Linda V. Green, Peter J. Kolesar, and Anthony Svoronos. Some effects of nonstationarity on multiserver Markovian queueing systems. *Oper Res*, 39(3):502–511, 1991.

[69] Linda V. Green, Peter J. Kolesar, and Ward Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Prod Oper Manag*, 16(1):13–39, 2007.

[70] Linda V. Green and Sergei Savin. Reducing delays for medical appointments: A queueing approach. *Oper Res*, 56(6):1526–1538, 2008.

[71] Linda V. Green, Sergei Savin, and Mark Murray. Providing timely access to care: What is the right patient panel size? *Jt Comm J Qual Patient Saf*, 33(4):211–218, 2007.

[72] Linda V. Green and João Soares. Note—computing time-dependent waiting time probabilities in M(t)/M/s(t) queuing systems. *Manuf Serv Oper Manag*, 9(1):54–61, 2007.

[73] Linda V. Green, João Soares, James F. Giglio, and Robert A. Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Acad Emerg Med*, 13(1):61–68, 2006.

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

[75] Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Oper Res*, 29(3):567–588, 1981.

[76] Pierre Hansen, Nenad Mladenović, Jack Brimberg, and José A. Moreno Pérez. Variable neighborhood search. In Michel Gendreau and Jean-Yves Potvin, editors, *Handbook of Metaheuristics*, volume 272 of *International Series in Operations Research & Management Science*, chapter 3, pages 57–97. Springer, 2019.

[77] Mor Harchol-Balter. Phase-type distributions and matrix-analytic methods. In *Performance Modeling and Design of Computer Systems : Queueing Theory in Action*, chapter 21, pages 359–379. Cambridge University Press, 2013.

[78] Jen Homwood. B.C. government's primary health-care strategy focuses on faster, team-based care. *BC Gov News*, 2018.

[79] Anthony A. Hudgins, Jack L. Graves, Betty W. Abbott, Ellen Rhone Blair, Catherine Meyers, and Paula Van Ness. Issues in family planning clinic management. *Fam Community Health*, 5(1):47–59, 1982.

[80] Peter J. H. Hulshof, Nikky Koortbeek, Richard J. Boucherie, Erwin W. Hans, and Piet J. M. Bakker. Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health Syst*, 1(2):129–175, 2012.

[81] Rudy Hung. Hospital nurse scheduling. *J Nurs Adm*, 25(7):21–23, 1995.

[82] Armann Ingolfsson, Elvira Akhmetshina, Susan Budge, Yongyue Li, and Xudong Wu. A survey and experimental comparison of service-level-approximation methods for nonstationary M(t)/M/s(t) queueing systems with exhaustive discipline. *INFORMS J Comput*, 19(2):201–204, 2007.

[83] Armann Ingolfsson, Fernanda Campello, Xudong Wu, and Edgar Cabral. Combining integer programming and the randomization method to schedule employees. *Eur J Oper Res*, 202(1):153–163, 2010.

[84] Armann Ingolfsson, Md Amanul Haque, and Alex Umnikov. Accounting for time-varying queueing effects in workforce scheduling. *Eur J Oper Res*, 139(3):585–597, 2002.

[85] Navid Izady and David J. Worthington. Approximate analysis of non-stationary loss queues and networks of loss queues with general service time distributions. *Eur J Oper Res*, 213(3):498–508, 2011.

[86] Sheldon H. Jacobson, Shane N. Hall, and James R. Swisher. Discrete-event simulation of health care systems. In Randolph Hall, editor, *Patient Flow*, volume 206 of *International Series in Operations Research Management Science*, chapter 12, pages 273–309. Springer, New York, NY, 2 edition, 2013.

[87] D. L. Jagerman. Nonstationary blocking in telephone traffic. *The Bell System Technol J*, 54(3):625–661, 1975.

[88] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *Linear Regression*, chapter 3. Springer, 2 edition, 2013.

[89] Brigitte Jaumard, Frédéric Semet, and Tsevi Vovor. A generalized linear programming model for nurse scheduling. *Eur J Oper Res*, 107(1):1–18, 1998.

[90] Denise Jaworsky, Anne Gadermann, Arnaud Duhoux, Trudy E. Naismith, Monica Norena, Matthew J. To, Stephen W., and Anita Palepu. Residential stability reduces unmet health care needs and emergency department utilization among a cohort of homeless and vulnerably housed persons in Canada. *J Urban Health*, 93(4):666–681, 2016.

[91] Otis B. Jennings, Avishai Mandelbaum, William A. Massey, and Ward Whitt. Server staffing to meet time-varying demand. *Manage Sci*, 42(10):1383–1394, 1996.

[92] J. B. Jun, S. H. Jacobson, and J. R. Swisher. Application of discrete-event simulation in health care clinics: A survey. *J Oper Res Soc*, 50(2):109–123, 1999.

[93] Seifedine Kadry, Aremn Bagdasaryan, and Mohammad Kadhum. Simulation and analysis of staff scheduling in hospitality management. In *2017 7th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO)*, pages 1–6, 2017.

[94] Edward L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*, 53(282):457–481, 1958.

[95] Jun Woo Kim and Sung Ho Ha. Consecutive staffing solution using simulation in the contact center. *Ind Manag Data Syst*, 110(5):718–730, 2010.

[96] Donald Ervin Knuth. *Random Numbers*, chapter 3. Addison-Wesley Professional, 3 edition, 1997.

[97] Howard K. Koh and James J. O'Connell. Improving health care for homeless people. *JAMA Forum*, 316(24):2586–2587, 2016.

[98] Qingxia Kong, Shan Li, Nan Liu, Chung-Piaw Teo, and Zhenzhen Yan. Appointment scheduling under time-dependent patient no-show behavior. *Manage Sci*, 66(8):3480–3500, 2020.

[99] Nikky Kortbeek, Maartje E. Zonderland, Aleida Braaksma, Ingrid M. H. Vliegen, Richard J. Boucherie, Nelly Litvak, and Erwin W. Hans. Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals. *Perform Eval*, 80:5–26, 2014.

[100] Harold Joseph Kushner. *Heavy Traffic Analysis of Controlled Queueing and Communication Networks*, volume 47 of *Applications of Mathematics: Stochastic Modelling and Applied Probability*, chapter 1. Springer, 2001.

[101] J. Mauricio Lach and Ricardo M. Vázquez. Simulation model of the telemedicine program. In R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, editors, *Proceedings of the 36th Conference on Winter Simulation, Dec 5 – 8, Washington, DC, USA*, pages 2012–2017. ACM, 2004.

[102] Manuel Laguna. Optquest: Optimization of complex systems. Technical report, Opt-Tek Systems, Inc., 2011.

[103] Manuel Laguna and Rafael Martí. *Scatter Search: Methodology and Implementations in C*, volume 24 of *Operations Research/Computer Science Interfaces Series*, chapter 1. Springer, 2003.

[104] Eyitayo Lambo. An optimization-simulation model of a rural health center in Nigeria. *Interfaces*, 13(3):29–35, 1983.

[105] Sangbok Lee and Yuehwern Yih. Analysis of an open-access scheduling system in outpatient clinics: A simulation study. *SIMUL*, 86(8–9):503–518, 2010.

[106] Nan Liu and Thomas D'Aunno. The productivity and cost-efficiency of models for involving nurse practitioners in primary care: A perspective from queueing analysis. *Health Serv Res*, 47(2):594–613, 2012.

[107] Nan Liu, Stacey R. Finkelstein, and Lusine Poghosyan. A new model for nurse practitioner utilization in primary care: Increased efficiency and implications. *Health Care Manage Rev*, 39(1):10–20, 2014.

[108] Nan Liu and Serhan Ziya. Panel size and overbooking decisions for appointment-based services under patient no-shows. *Prod Oper Manag*, 23(12):2209–2223.

[109] Ran Liu and Xiaolan Xie. Physician staffing for emergency departments with time-varying demand. *INFORMS J Comput*, 30(3):588–607, 2018.

[110] Yunan Liu and Ward Whitt. Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with abandonment. *Queueing Syst*, 67:145–182, 2011.

[111] Yunan Liu and Ward Whitt. A network of time-varying many-server fluid queues with customer abandonment. *Oper Res*, 59(4):835–846, 2011.

[112] Yunan Liu and Ward Whitt. The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Syst*, 71(4):405–444, 2012.

[113] Yunan Liu and Ward Whitt. A many-server fluid limit for the $G_t/GI/s_t+GI$ queueing model experiencing periods of overloading. *Oper Res Lett*, 40(5):307–312, 2012.

[114] Yunan Liu and Ward Whitt. Many-server heavy-traffic limit for queues with time-varying parameters. *Ann Appl Probab*, 24(1):378–421, 2014.

[115] Christine Loignon, Catherine Hudon, Émilie Goulet, Sophie Boyer, Marianne De Laat, Nathalie Fournier, Cristina Grabovschi, and Paula Bush. Perceived barriers to healthcare for persons living in poverty in Quebec, Canada: the EQUIhealThY project. *Int J Equity Health*, 14:art 4, 2015.

[116] Avi Mandelbaum and William A. Massey. Strong approximations for time-dependent queues. *Math Oper Res*, 20(1):33–64, 1995.

[117] Avi Mandelbaum, William A. Massey, and Martin I. Reiman. Strong approximations for Markovian service networks. *Queueing Syst*, 30(1):149–201, 1998.

[118] Avi Mandelbaum, William A. Massey, Martin I. Reiman, Alexander Stolyar, and Brian Rider. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommun Syst*, 21(2):149–171, 2002.

[119] Avishai Mandelbaum and Sergey Zeltyn. Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. In Dieter Spath and Klaus-Peter Fähnrich, editors, *Advances in Services Innovations*, chapter 2, pages 17–45. Springer-Verlag Berlin Heidelberg, 2007.

[120] Avishai Mandelbaum and Sergey Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centres. *Oper Res*, 57(5):1189–1205, 2009.

[121] William A. Massey and Ward Whitt. An analysis of the modified offered-load approximation for the nonstationary Erlang loss model. *Ann Appl Probab*, 4(4):1145–1160, 1994.

[122] Jyotiprasad Medhi. Queueing systems: General concepts. In *Stochastic Models in Queueing Theory*, chapter 2. Academic Press, 1 edition, 1991.

[123] Jyotiprasad Medhi. Stochastic processes. In *Stochastic Models in Queueing Theory*, chapter 1. Academic Press, 1 edition, 1991.

[124] Matthew Menear, Marc-André Blanchette, Olivier Demers-Payette, and Denis Roy. A framework for value-creating learning health systems. *Health Res Policy Sys*, 17:Article 79, 2019.

[125] Harvey H. Millar and Mona Kiragu. Cyclic and non-cyclic scheduling of 12 h shift nurses by network programming. *Eur J Oper Res*, 104(3):582–592, 1998.

[126] Holmes E. Miller, William P. Pierskalla, and Gustave J. Rath. Nurse scheduling using mathematical programming. *Oper Res*, 24(5):857–870, 1976.

[127] Shyam L. Moondra. An L.P. model for work force scheduling for banks. *J Bank Res.*, pages 299–301, 1976.

[128] Arup K. Mukherjee. A simulation model for management of operations in the pharmacy of a hospital. *SIMUL*, 56(2):91–103, 1991.

[129] Mark Murray, Mike Davies, and Barbara Boushon. Panel size: How many patients can one doctor manage? *Fam Pract Manag*, 14(4):44–51, 2007.

[130] Mark Murray and C. Tantau. Same-day appointments: Exploding the access paradigm. *Fam Pract Manag*, 7(8):45–50, 2000.

[131] Barry L. Nelson and Michael R. Taaffe. The $[\mathrm{Ph}_t/\mathrm{Ph}_t/\infty]^K$ queueing system: Part II—The multiclass network. *INFORMS J Comput*, 16(3):275–283, 2004.

[132] G. F. Newell. Queues with time-dependent arrival rates I—The transition through saturation. *J Appl Probab*, 5(2):436–451, 1968.

[133] Canadian Association of Emergency Physicians. The Canadian Triage and Acuity Scale Combined Adult/Paediatric Education Program. `http://ctas-phctas.ca/wp-content/uploads/2018/05/participant_manual_v2.5b_november_2013_0.pdf`, 2013. Accessed: 2021-09-23.

[134] British Columbia Ministry of Health. 2020/21 – 2022/23 Service plan. `https://www.bcbudget.gov.bc.ca/2020/sp/pdf/ministry/hlth.pdf`, 2020. Accessed: 2021-09-22.

[135] Vera Z. Osidach and Michael C. Fu. Computer simulation of a mobile examination center. In Stephen E. Chick, Paul J. Sánchez, David Ferrin, and Douglas J. Morrice, editors, *Proceedings of the 35th Conference on Winter Simulation, Dec 7–10, New Orleans, LA, USA*, pages 1868–1875. ACM, 2003.

[136] Asli Ozen and Hari Balasubramanian. The impact of case mix on timely access to appointments in a primary care group practice. *Health Care Manag Sci*, 16:101–118, 2013.

[137] Jorge Pacheco, Cristóbal Cuadrado, and María Soledad Martínez-Gutiérrez. Urgent care centres reduce emergency department and primary care same-day visits: A natural experiment. *Health Policy Plan*, 34(3):170–177, 2019.

[138] D. Parr and J. M. Thompson. Solving the multi-objective nurse scheduling problem with a weighted cost function. *Ann Oper Res*, 155:279–288, 2007.

[139] Lisa Patvivatsiri. A simulation-based approach for optimal nurse scheduling in an emergency department. Master's thesis, Virginia Polytechnic Institute and State University, 2003.

[140] Lisa Patvivatsiri, Barbara M. P. Fraticelli, and C. Patrick Koelling. A simulation-based approach for optimal nurse scheduling in an emergency department. In *IIE Annual Conference Proceedings*. Institute of Industrial and Systems Engineers (IISE), 2006.

[141] Bernadette Pauly, Corrine Lowen, Kathleen Perkin, Nicole Jackson, and Centre for Addictions Research of British Columbia. Intensive case management team model of care: Standards and guidelines. Technical report, British Columbia Ministry of Health, Victoria, 2014. Available at: `http://www.llbc.leg.bc.ca/public/pubdocs/bcdocs2020_2/715254/715254_icmt_standards.pdf`.

[142] Yidong Peng, Xiuli Qu, and Jing Shi. A hybrid simulation and genetic algorithm approach to determine the optimal scheduling templates for open access clinics admitting walk-in patients. *Comput Ind Eng*, 72:282–296, 2014.

[143] Sanja Petrovic and Greet Vanden Berghe. A comparison of two approaches to nurse rostering problems. *Ann Oper Res*, 194:365–384, 2012.

[144] Provincial Health Services Authority. Towards reducing health inequities: A health system approach to chronic disease prevention. A discussion paper. Vancouver, BC: Population & Public Health, Provincial Health Services Authority, 2011.

[145] Erfan Rahimian, Kerem Akartunalı, and John Levine. A hybrid integer programming and variable neighbourhood search algorithm to solve nurse rostering problems. *Eur J Oper Res*, 258(2):411–423, 2017.

[146] Monia Rekik, Jean-François Cordeau, and François Soumis. Implicit shift scheduling with multiple breaks and work stretch duration restrictions. *J Sched*, 13:49–75, 2010.

[147] Ahmad Ridley, Michael Fu, and William A. Massey. Customer relations management: Call center operations: Fluid approximations for a priority call center with time-varying arrivals. In Stephen E. Chick, Paul J. Sánchez, David Ferrin, and Douglas J. Morrice, editors, *Proceedings of the 35th Conference on Winter Simulation, Dec 7–10, New Orleans, LA, USA*, pages 1817–1823. ACM, 2003.

[148] Lori E. Ross, Simone Vigod, Jessica Wishart, Myera Waese, Jason Dean Spence, Jason Oliver, Jennifer Chambers, Scott Anderson, and Roslyn Shields. Barriers and facilitators to primary care for people with mental health and/or substance use issues: A qualitative study. *BMC Fam Pract*, 16:art 135, 2015.

[149] Sheldon M. Ross. *Simulation.* Elsevier Science & Technology, 5 edition, 2012.

[150] Michael H. Rothkopf and Shmuel S. Oren. A closure approximation for the nonstationary M/M/s queue. *Manage Sci*, 25(6):522–534, 1979.

[151] Elizabeth Sandoval, Sandy Smith, James Walter, Sarah-Anne Henning Schuman, Mary Pat Olson, Rebecca Striefler, Stephen Brown, and John Hickner. A comparison of frequent and infrequent visitors to an urban emergency department. *J Emerg Med*, 38(2):115–121, 2010.

[152] Matthias Schacht. Improving same-day access in primary care: Optimal reconfiguration of appointment system setups. *Oper Res Health Care*, 18:119–134, 2018.

[153] Ali Rafik Shukor, Sandra Edelman, Dean Brown, and Cheryl Rivard. Developing community-based primary health care for complex and vulnerable populations in the Vancouver coastal health region: HealthConnection clinic. *Perm J*, 22:18–010, 2018.

[154] Ali Rafik Shukor, Ronald Joe, Gabriela Sincraian, Niek Klazinga1, and Dionne Sofia Kringos. A multi-sourced data analytics approach to measuring and assessing biopsychosocial complexity: The Vancouver community analytics tool complexity module (VCAT-CM). *Community Ment Health J*, 55:1326–1343, 2019.

[155] David Sinreich and Ola Jabali. Staggered work shifts: A way to downsize and restructure an emergency department workforce yet maintain current operational performance. *Health Care Manag Sci*, 10:293–308, 2007.

[156] Lukas A. J. Stalpers and Edward L. Kaplan. Edward L. Kaplan and the Kaplan-Meier survival curve. *Br J Hist Math*, 33(2):109–135, 2018.

[157] Peter Stanford, David A.and Taylor and Ilze Ziedins. Waiting time distributions in the accumulating priority queue. *Queueing Syst*, (77):297–330, 2014.

[158] Eric Sundstrom, Kenneth P. De Meuse, and David Futrell. Work teams: Applications and effectiveness. *Am Psychol*, 45(2):120–133, 1990.

[159] M. R. Taaffe and K.L. Ong. Approximating nonstationary Ph(t)/M(t)/s/c queueing systems. *Ann Oper Res*, 8(1):103–116, 1987.

[160] Boon Aik Tan, Aldas Gubaras, and Nipa Phojanamongkolkij. Simulation study of Dreyer urgent care facility. In E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, editors, *Proceedings of the 33rd Conference on Winter Simulation, Dec 8–11, San Diego, CA, USA*, pages 1922–1927. ACM, 2002.

[161] Gary M. Thompson. Improved implicit optimal modeling of the labor shift scheduling problem. *Manage Sci*, 41(4):595–607, 1995.

[162] Seyda Topaloglu. A shift scheduling model for employees with different seniority levels and an application in healthcare. *Eur J Oper Res*, 198(3):943–957, 2009.

[163] Debbie Travers. Triage: How long does it take? How long should it take? *J Emerg Nurs*, 25(3):238–240, 1999.

[164] Feray Tunçalp, Evrim Güneş, and Lerzan Örmeci. Modeling strategic walk-in patients in appointment systems: Equilibrium behavior and capacity allocation. Available at SSRN: `https://ssrn.com/abstract=3687717` or `http://dx.doi.org/10.2139/ssrn.3687717`, 2020.

[165] Peter Vanberkel and J. R. Wing. Mixing scheduled patients with walk-in patients: A simulation optimization approach. In *ORAHS 2018 Conference: Connected Care. 44th International Conference of the EURO Working Group on Operational Research Applied to Health Services, Jul 29–Aug 3, Oslo, Norway*, pages 50–51, 2018.

[166] Neal A. Vanselow, Molla S. Donaldson, and Karl D. Yordy. A new definition of primary care. *JAMA*, 273(3):192, 1995.

[167] Shan Wang, Nan Liu, and Guohua Wan. Managing appointment-based services in the presence of walk-in customers. *Manage Sci*, 66(2):667–686, 2020.

[168] Yanchao Wang, Warren L. Hare, L. Vertesi, and Alexander R. Rutherford. Using simulation to model and optimize acute care access in relation to hospital bed count and bed distribution. *J Simul*, 5(2):101–110, 2011.

[169] Darrell Whitley. Next generation genetic algorithms: A user's guide and tutorial. In Michel Gendreau and Jean-Yves Potvin, editors, *Handbook of Metaheuristics*, volume 272 of *International Series in Operations Research & Management Science*, chapter 8, pages 245–274. Springer, 2019.

[170] Ward Whitt. Staffing a call center with uncertain arrival rate and absenteeism. *Prod Oper Manag*, 15(1):88–102, 2006.

[171] Toni I. Wickert, Pieter Smet, and Greet Vanden Berghe. The nurse rerostering problem: Strategies for reconstructing disrupted schedules. *Comput Oper Res*, 104:319–337, 2019.

[172] Lena Wolbeck, Natalia Kliewer, and Inês Marques. Fair shift change penalization scheme for nurse rescheduling problems. *Eur J Oper Res*, 284(3):1121–1135, 2020.

[173] Jinn-Yi Yeh and Wen-Shan Lin. Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department. *Expert Syst Appl*, 32(4):1073–1083, 2007.

[174] Christos Zacharias and Mor Armony. Joint panel sizing and appointment scheduling in outpatient care. *Manage Sci*, 63(11):3978–3997, 2017.

[175] Huimin Zhou. Optimization of physician panel design in primary care. Master's thesis, Binghamton University, Binghamton, NY, 7 2018.

[176] Samantha L. Zimmerman, Alan Bi, Trevor Dallow, Alexander R. Rutherford, Tamon Stephen, Cameron Bye, David Hall, Andrew Day, Nicole Latham, and Krisztina Vasarhelyi. Optimising nurse schedules at a community health centre. *Oper Res Health Care*, 30:100308, 2021.

[177] Samantha L. Zimmerman, Alexander R. Rutherford, Alexa van der Waall, Monica Norena, and Peter Dodek. A queuing model for ventilator capacity management during the COVID-19 pandemic. Available at: `https://doi.org/10.1101/2021.03.17.21253488`, 2021.

Appendix A

Extended Iterative Staffing Algorithm

This appendix section describes the details of the sequential-EPA-ISA approach, which I introduce to solve equation (4.4) and find the minimum small-interval staffing for multiple staff disciplines so that small-interval EPA targets are met. The sequential-EPA-ISA approach extends the ISA framework of Defraeye and Van Niuewenhuyse [44] to address multiple staff disciplines in a care network and access targets on a performance measure for the expected proportion of clients receiving service, which is referred to in this thesis as EPA and defined in Section 4.3.

The pseudo-code in algorithms 1 and 2 describe phase I and phase II of the presented EPA-ISA approach, which extend the two phase ISA framework of Defraeye and Van Niuewenhuyse [44] to address staffing requirements for a single staff type in a service network using EPA constraints. Both phases of the EPA-ISA algorithm are then applied sequentially to each staff discipline in the sequential-EPA-ISA approach 3. See Subsection 4.4.1 for a higher level discussion of the ISA framework and the extensions made in the EPA-ISA and sequential-EPA-ISA algorithms.

The presented EPA-ISA phase I and II algorithms optimize staffing levels for a single staffing discipline m that provides set of associated services K_m in a queue network. In the phase I algorithm 1, an initial staffing requirement guess for staff m is made using the product of the daily average arrival rate $\bar{\lambda}_k$ multiplied by the average service time $1/\mu_k$, which is added together for each relevant service k. The algorithm then iteratively updates the staffing requirements for discipline m in each staffing interval according to simulation EPA estimates. For each staffing interval i_h with number h, staffing levels are updated according to estimated EPA values in an associated set of measurement intervals with indices in $G_{h,k}$ for each service. The set $G_{h,k}$ is defined based on a specified number of extra time intervals η_k for each service chosen according to equation (4.5). For each staffing interval and service, an amplification factor $A_{m,i_h,k}$ is used, which is greater than 1 if the minimum EPA in the associated measure intervals is below the targets, and less than one if the EPA targets are satisfied. Staffing in each interval is scaled up or down according to the maximum associated amplification factor across each relevant service. If the associated EPA targets are satisfied, then staffing can be scaled down; whereas staffing is increased if the EPA targets are not satisfied in any associated interval for some relevant service. In each phase I iteration, if the simulated solution meets all targets, then the current optimal required staffing S^*_m and cost c^*_m are updated to the current solution and current cost c^j_m; otherwise, the current iteration index j is added to a set of infeasible indices B_m. Here the variable $w_{m,i}$ is used to reflect the staffing cost of discipline m in interval i. Lastly, phase I checks three stopping criteria, namely whether a staffing solution has been repeated, whether the squared coefficient of variation of EPA values SCV^j_k are small and alternate between iterations, or whether a maximum number of iterations $MaxIter1$ has been reached.

Phase II of the EPA-ISA algorithm 2 draws on most of the variables used in the phase I algorithm. Since the solution(s) found in phase I do not necessarily satisfy all the EPA constraints, phase II incrementally increases the staffing levels for each solution to address unsatisfactory EPA estimates—under the same set $G_{h,k}$ of measurement indices for each interval. For each infeasible staffing iterate indexed in set B_m, phase II draws on the EPA estimates from phase I to increase the staffing requirement by one in each unsatisfactory interval. If the updated staffing cost c^l_m is less than the current optimal, then phase II will reassess the simulation EPA estimates for the new staffing level and continue to increment as needed until either the EPA targets are satisfies, the staffing cost c^l_m is not longer advantageous, or a maximum number of iteration $MaxIter2$ is reached. To increase efficiency, phase II considers each infeasible solutions in order of increasing maximum deviation from the staffing targets in variable o^l_m; when this metric is the same for multiple staffing solutions, the algorithm uses the lowest cost schedule to break ties. Here the pre-calculated cost c^l_m already incorporates the first staffing increment in phase II.

The sequential-EPA-ISA approach in Algorithm 3 loops over each staffing discipline in order of client interaction at the centre, and sequentially applies EPA-ISA phases I and II. After each iteration, the resulting optimal staffing vector S^*_m for staff discipline m is incorporated into an overall staffing matrix, that is used as input for subsequent iterations.

Algorithm 1 EPA-ISA Phase I (m, $\tilde{S}^{(-m)}$)

```
Initialize staffing $S^0_{m,i} \leftarrow \lceil \sum_{k \in K_m} \frac{\bar{\lambda}_k}{\mu_k} \rceil \quad \forall\, i \in I$ and $S^0_{m',i} \leftarrow \tilde{S}_{m',i} \quad \forall\, m' \in M \setminus m,\, i \in I$
Initialize iteration counter $j \leftarrow 0$
Initialize optimal cost $c^*_m \leftarrow \infty$ and infeasible set $B_m \leftarrow \emptyset$
Initialize flag $STOP \leftarrow FALSE$
Choose lag factor $\eta_k \leftarrow \lceil \frac{\sigma_k}{\Delta} \rceil, \quad \forall k \in K_m$
Set interval assignment $G_{h,k} \leftarrow \Big\{ \max\left(0, h - \eta_k\right),\, ...,\, \min\left(h, |I_c|\right) \Big\} \quad \forall i_h \in I_s,\, k \in K_m$
while $STOP == FALSE$ do
    Simulate the queueing network to estimate $\widehat{EPA}^{Sj}_{k,k}(i)$, $\forall\, k \in K_m$, $i \in I$
    Update the staffing requirements:
    $A_{m,i_h,k} \leftarrow 1 + \frac{\max_{g \in G_{h,k}} \{1 - \widehat{EPA}^{Sj}_{k,k}(i_g)\} - \alpha_{k,k}}{2(j+2)\alpha_{k,k}}, \quad \forall i_h \in I_s,\, k \in K_m$
    $A_{m,i} = \max_{k \in K_m} \{A_{m,i,k}\}, \quad \forall i \in I_s$
    $S^{j+i}_{m,i} \leftarrow \begin{cases} \lceil S^j_{m,i} A^j_{m,i} \rceil & A^j_{m,i} \geq 1 \\ \max(\lfloor S^j_{m,i} A^j_{m,i} \rfloor, 1) & A^j_{m,i} < 1 \end{cases} \qquad \forall i \in I_s$
    $S^{j+i}_{m',i} = S^j_{m\prime,i} \quad \forall\, m' \in M \setminus m,\, i \in I$
    Update staffing costs and feasibility:
    $c^j_m = \leftarrow \sum_{i_h \in I_s} \Big\{ w_{m,i_h} s^j_{m,i_h} + w_{m,i_h} \mathbb{1}\Big(\exists k \in K_m, g \in G_{h,k} : (1 - \widehat{EPA}^{Sl}_{k,k}(i)) > \alpha_{k,k}\Big)\Big\}$
    if $(1 - \widehat{EPA}^{Sj}_{k,k}(i)) < \alpha_{k,k}) \quad \forall k \in K_m,\, i \in I_c$ then
        if $c^j_m < c^*_m$ then
            $c^*_m \leftarrow c^j_m$ and $S^*_m \leftarrow S^j_m$
        end if
    else
        Add $j$ to $B_m$
        $o^l_m \leftarrow \max_{i \in I_s, k \in K_m} \left\{ \frac{(1 - \widehat{EPA}^{Sj}_{k,k}(i)) - \alpha_{k,k}}{\alpha_{k,k}} \right\}$
    end if
    Evaluate stopping criteria:
    $SCV^j_k \leftarrow$ squared coefficient of variation of $\max_{g \in G_{h,k}} \{1 - \widehat{EPA}^{Sj}_{k,k}(i_g)\}$ over $i_h \in I_s$, $\forall k \in K_m$
    if $\exists\, l < j+1 \,:\, S^l_{m,i} = S^{j+1}_{m,i} \quad \forall i \in I_s$ then
        $STOP \leftarrow TRUE$
    else if $j >= 2$ and $SCV^j_k < 1 \forall k \in K_m$ and $SCV^j_k$ alternates $\forall k \in K_m$ then
        $STOP \leftarrow TRUE$
    else if j>= MaxIter1 then
        $STOP \leftarrow TRUE$
    end if
    $j \leftarrow j + 1$
end while
```

Algorithm 2 EPA-ISA Phase II (m)

```
Input: All stored variables from phase I
for $l \in B_m$, in order of increasing $o^l_m$ and using $c^l_m$ to break ties do
    $S^l_{m,i} \leftarrow \begin{cases} S^l_{m,i} + 1 & if\ (1 - \widehat{EPA^{S^l}_{k,k}}(i_g)) > \alpha_{k,k})\ \exists g \in G_{h,k},\ \exists k \in K_m \\ S^l_{m,i} & otherwise \end{cases} \qquad \forall i \in I_s$
    $q \leftarrow 0$
    while $c^l_m < c^*_m$ and $q < MaxIter2$ do
        Simulate the queueing network to estimate EPA for the new staffing matrix $S^l$
        if $\exists k \in K_m,\ i \in I_c : (1 - \widehat{EPA^{S^l}_k}(i)) > \alpha_k)$ then
            $S^l_{m,i} \leftarrow \begin{cases} S^l_{m,i} + 1 & if\ \exists k \in K_m\ \exists g \in G_{h,k}\ (1 - \widehat{EPA^{S^l}_{k,k}}(i_g)) > \alpha_{k,k}) \\ S^l_{m,i} & otherwise \end{cases} \qquad \forall i \in I_s$
            $c^l_m \leftarrow \sum_{i \in I} w_{m,i} s^l_{m,i}$
        else if $c^l_m < c^*_m$ then
            $c^*_m \leftarrow c^l_m$
            $S^*_m \leftarrow S^l_m$
        end if
        $q \leftarrow q + 1$
    end while
end for

Output: $S^*_m \leftarrow S^l_m$
```

Algorithm 3 Sequential-EPA-ISA

```
Initialize staffing matrix $S_{m,i} \leftarrow \lceil \sum_{k \in K_m} \frac{\bar{\lambda}_k}{\mu_k} \rceil \quad \forall\, i \in I$ and $\forall\, m \in M$
for $m \in M$, by order of client interaction do
    Run EPA-ISA phase I ($m$, $S^{(-m)}$)
    Run EPA-ISA phase II ($m$) to get $S^*_m$
    $S_{m,i} \leftarrow S^*_m$
end for
```

Appendix B

Adjustment for Overlapping Service Times

This appendix section describes an approach to adjust service time estimates to accounting for overlapping records for the same physician. The basic steps of this approach are described in Algorithm 4. The core adjustment in Line 5 reduces service times during each interval of overlap, where the magnitude of reduction is equal to the interval duration split evenly by the number of clients recorded as receiving treatment during that time. In other words, for every interval that a physician is recorded as multi-tasking to providing care for multiple clients the algorithm assumes that they divide their time equally between those clients. By adjusting service times, this approach approximates the time spent in the provision of treatment for each individual and is a simplification of both human behavior and ambiguous data logging.

Algorithm 4 Modifying service time estimates for overlapping care.

```
1: Create an initial service time estimate for each client visit, calculated as the length of
   the interval between each seen and departure time
2: for Each individual provider $p$ and date $d$ do
3:    Construct a time series for the number of clients $b$ recorded as being seen by provider $p$
      at time $t$ on day $d$.
4:    for Each time interval $\tilde{t}$ where there is a constant non-zero number of clients $b$ being
      seen by provider $p$. do
5:       Reduce the estimated service time by $T/b$ for each of these clients, where $T$ is the
         length of time interval $\tilde{t}$.
6:    end for
7: end for
```