Acuity-based Performance Evaluation and Tactical Capacity Planning in Primary Care

Nazanin Aslani

A Thesis
in
the Department
of
Mechanical, Industrial, and Aerospace Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy (Industrial Engineering)
Concordia University
Montréal, Québec, Canada

September 2019

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Nazanin Aslani

Entitled: Acuity-based Performance Evaluation and Tactical Capacity Planning in Primary Care

and submitted in partial fulfillment of the requirements for the degree of

Doctor Of Philosophy (Industrial Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. William Lynch

_____ External Examiner
Dr. Michael Carter

_____ External to Program
Dr. Ketra Schmitt

_____ Examiner
Dr. Ali Akgunduz

_____ Examiner
Dr. Mingyuan Chen

_____ Thesis Supervisor
Dr. Daria Terekhov

Approved by _____
Dr. Ivan Contreras, Graduate Program Director

November 18, 2019

_____
Dr. Amir Asif, Dean
Gina Cody School of Engineering & Computer Science

Abstract

Acuity-based Performance Evaluation and Tactical Capacity Planning in Primary Care

Nazanin Aslani, Ph.D.

Concordia University, 2019

Effective primary care requires timely and equitable access to care for patients as well as efficient and balanced utilization of physician time. Motivated by a family health clinic in Ontario, Canada, this research proposes ways to improve both of these aspects of primary care through tactical capacity planning based on acuity-based performance targets.

First, we propose a new metric based on acuity levels to evaluate timely access to primary care. In Canada, as well as other participant countries in the Organization for Economic Co-operation and Development (OECD), the main metric currently used to evaluate access is the proportion of patients who are able to obtain a same- or next-day appointment. However, not all patients in primary care are urgent and require a same- or next-day appointment. Therefore, accurate evaluation of timely access to primary care should consider the urgency of the patient request. To address this need, we define multiple acuity levels and relative access targets in primary care, akin to the CTAS system in emergency care. Furthermore, current access time evaluation in the province is mostly survey-based, while our evaluation is based on appointment data and hence more objective. Thus, we propose a novel, acuity-based, data-driven approach for evaluation of timely access to primary care.

Second, we develop a deterministic tactical capacity planning (TCP) model to balance workload between weeks for each family physician in the specific primary care clinic in this study. Unbalanced workload among weeks may lead to provider overtime for the weeks with high workload and provider idle time for weeks with low workload. In the proposed TCP model, we incorporate the results from access time evaluation in the first study as

constraints for access time. The proposed TCP model considers 11 appointment types with multiple access targets for each appointment type. The TCP model takes as input a forecast of demand coming from an ARIMA model. We compare the results of the TCP model based on current access time targets as well as targets resulting from our acuity-based metrics. The use of our proposed acuity-based targets leads to allocation of time slots which is more equitable for patients and also improves physician workload balance.

Third, we also propose a robust TCP model based on the cardinality-constrained method to minimize the highest potential physician peak load between weeks. Therefore, the developed robust TCP model enables protection against uncertainty through providing a feasible allocation of capacity for all realizations of demand. The proposed robust TCP model considers two interdependent appointment types (e.g., new patients and follow ups), multiple access time targets for each appointment type and uncertainty in demand for appointments. We conduct a set of experiments to determine how to set the level of robustness based on extra cost and infeasibility probability of a robust solution.

In summary, this dissertation advocates for the definition and subsequent use of acuity-based access time targets for both performance evaluation and capacity allocation in primary care. The resulting performance metrics provide a more detailed view of primary care and lead to not only more equitable access policies but also have the potential to improve physician workload balance when used as input to capacity planning models.

Acknowledgment

Similar to most of researchers, it has been a tough journey for me to complete Ph.D. dissertation. I would like to proudly say I completed this journey with the same motivation and passion towards science that I had before starting my Ph.D. only with the help of love and support from the people around me. Without these helps, I think it can be very easy to lose oneself in this path. The completion of this program and the contributions of this thesis have been influenced by the supporters whom I am about to mention.

First and foremost, I would like to sincerely and deeply thank my supervisor Dr. Daria Terekhov. You have not only been a tremendous mentor for me but a true source of inspiration. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your constant guidance and support during this research and scientific intuition and passion have been priceless. I would also like to thank you for helping me to find my career pathway by enabling me to present my research to other scientists and to hear about their work by attending conferences and workshops and by guiding me to establish new connections.

I would also like to thank my external chair, Professor Michael Carter, for his attention, expertise, and contribution that improved this dissertation. Similar thanks to my other thesis committee members, Dr. Ketra Schmitt, Dr. Ali Akgunduz and Dr. Mingyuan Chen. A big thanks to our graduate program coordinator, Leslie Hosein, who listened to my problems and helped me to figure out how to deal with my problems rather than just keeping them all bottled up inside.

In addition, I would like to thank my lab-mates at data-driven lab, Yingcong Tan, Elaheh Hosseiniiraj, Gerald Potkah and Mahsa Moghaddass for their valuebale inputs in lab practice talks.

My sincere thanks to Health for All (HFA) clinic firstly for their collaboration and giving us the authorization to collect required data for this research. Specifically I would like to

acknowledge Dr. Alan Monavvari for encouraging me to initiate this research based on a real case study at Health for All (HFA) clinic; Dr. Fariborz Fazileh and Dr. Donatus Mutasingwa for their valuable feedback and insights into the realm of primary care booking system; Dr. Stephen Marisette and Dr. Karuna Gupta for their support to realize this research; the Administrative Team Leader, Cathy Teolis, for her patience and help to explain the current appointment types and their specification at HFA. Thank you also to the Program Administrator, Zhanying Shi for providing us some part of the required data for this research.

A special thank to my mother Akram, my father Reza and my brothers Mohammad and Sina, thank you for your love, encouragement and kindness even when we were continents apart. You are the best family one can wish for. To my sister, Saba, thank you for being there when I needed you the most, making me laugh and never leaving my side through tough and ugly times.

Last and most importantly, I am greatly indebted to the person who has been by my side throughout this whole experience, my best friend and love, Fariborz Fazileh who spent sleepless nights, was always my unfailing supporter in the moments when there was no one to answer my queries, and most importantly was my tag-team partner on this journey of life. Without you, none of this would have been possible.

Contribution of authors

This dissertation is presented under the manuscript-based format. It contains three articles that have been accepted for publication or are under review in different journals. These were submitted in the following chronological order. The first article titled "Acuity-based Access Time Evaluation in Primary Care: a Case Study of an Ontario Family Health Team" was accepted to be published in the *Proceedings of the Health Care Systems Engineering Conference 2019 (HCSE'19)* in September 2019. The second manuscript titled "Improving Physician Workload Balance and Timely Access to Primary Care" is intended for submission to the journal *Operations Research for Health Care*. Finally, the third manuscript titled "A Robust Optimization Model for Tactical Capacity Planning in an Outpatient Setting" was submitted for publication to the journal *Health Care Management Science* in September 2019. All three manuscripts are co-authored with Dr. Daria Terekhov who established research guidelines and reviewed the papers before submission. The first manuscript is also co-authored with Dr. Fariborz Fazileh and Dr. Donatus Mutasingwa who were our clinical collaborators representing the Health For All family health team in Markham, Ontario, Canada; they advised on the practical clinical interpretations of the work and reviewed the paper before submission. The third manuscript is also co-authored with Dr. Onur Kuzgunkaya and Dr. Navneet Vidyarthi, who initialized the research of that paper and also reviewed the manuscript before submission. For all of the work presented in this thesis, the author of this thesis acted as the principal researcher with the corresponding duties such as the development of formulations and algorithms, the programming of solution methods and the analysis of computational results along with writing drafts of the papers.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivations

Primary care is the first point of contact between the patient and the healthcare system. Due to the discernible influence of primary care on promoting health and consequently preventing illness and death, strong primary care leads to a strong healthcare system (Starfield et al., 2005) and a healthier population. A strong primary care system ensures that access to care is timely and matches the urgency of patients' health concerns while minimizing costs (Starfield, 2008).

According to the results of the Commonwealth Fund's latest survey (Commonwealth Fund, 2017), Canadian patients are experiencing longer access times compared to other participant countries in the Organization for Economic Co-operation and Development (OECD) (Commonwealth Fund, 2017). Access time is defined as the interval between the arrival of the appointment request and the scheduled time of appointment. Longer access time is due to the scarcity of physician services in Canada compared to most other OECD countries that provide universal health insurance coverage. According to the report of Statistics Canada, in 2017, roughly 4.7 million Canadians aged 12 and older, or approximately 15.8% of Canada's 12 and older population, reported that they did not have a regular medical doctor (Statistics Canada, 2019). Of these, an estimated 2.5 million indicated they did not have one because doctors were not taking new patients, or that doctors were retiring and leaving the area, or simply that no doctors were available where they lived (Globerman et al., 2018). Therefore, reform strategies were developed in the early 2000s to improve timely access to primary care in Ontario through managing the scarce supply of physicians. The main goals behind the reform strategies were having a primary care with a higher level of timely access to care,

health promotion, disease prevention, chronic disease management and setting up teamwork between primary care providers (Hutchison and Glazier, 2013).

1.1.1 Reforms in Ontario Primary Care

With more than 14 million people in 2018, Ontario is the most populous province among the thirteen provinces and territories in Canada (Singh et al., 2019). Ontario has the highest pace among the provinces in Canada in terms of developing reform strategies in primary care (Hutchison and Glazier, 2013). From 2002 to 2007 the Ministry of Health and Long-Term Care in Ontario launched new care delivery and primary care practice models to improve access to care and quality of care in Ontario (Hutchison and Glazier, 2013). The main considered elements in the proposed practice models are team work between groups of family physicians, interdisciplinary collaboration of primary care providers, formal definition of the physician-patient relationship and mixed payment mechanisms (Hutchison et al., 2011). The objectives of these reforms were increasing family physician availability through team practice and improving long-term coordination of care, especially for patients with chronic conditions through patient enrollment with a primary care clinic (Wranik et al., 2019).

Family Practice Models Before Reforms The vast majority of existing family practice models in Ontario before realization of primary care reforms were solo, in which a single family physician runs her/his own clinic. Physicians in solo family practices are funded based on a fee-for-service (FFS) system, where payment to a practice is based on delivered services and type of care (Graber-Naidich, 2015). There also exist a few other practice models such as Community Health Center (CHC) where an interdisciplinary team of providers such as family physicians, nurse practitioners, nutritionist and social workers deliver care to the regions with high proportion of homeless and lower income people (Sweetman and Buckley, 2014). Physicians in a CHC are paid based on *salaried* funding where payment is fixed and is not based on delivered services and number of patients served (Graber-Naidich, 2015). Some

solo family practice and CHCs still exist after the reforms; however, many have converted to new types of family health practices, as described below.

Family Practice Models After Reforms The cornerstone of new family practice models proposed by the primary care reform strategies is patient *rostering* which is considered to be the main factor for improving access to care in primary care in countries such as Canada, Australia, the Netherlands, Norway, New Zealand (Tiagi and Chechulin, 2014). Patient rostering defines a formal relationship between patient and family physician through an official agreement between them. Through this agreement, patients agree to only visit their rostered family physician for primary care services except the times patients are travelling or their needed care is an emergency. In return, the family physician agrees to provide comprehensive care which promotes responsibility of family physicians toward patients (The College of Family Physicians of Canada, 2012). Patient rostering leads to ongoing access to the same family physician over time and results in improving continuity of care for patients (Tiagi and Chechulin, 2014). Some benefits of increasing continuity of care are enhancing patient health status, increasing patient satisfaction, improving coordination of specialist care, reducing emergency department use, and decreasing overall healthcare costs (Christakis et al., 2003; Blewett et al., 2008; Kim et al., 2012; Maarsingh et al., 2016).

The requirements for practicing based on patient rostering models are team work of a group of physicians and providing extended clinical hours (Singh et al., 2019). Collaboration of team of family doctors can alleviate the limitations of family physician availability (Chand et al., 2009). In other words, when a patient from the roster of a family doctor requests an appointment due to a new health concern and the doctor is unavailable, the patient will be seen by the first available doctor in the team. Therefore, family doctor teamwork, or *resource pooling*, results in responding to patient demand variability, reducing access time.

The patient-rostering-based family practice models which were introduced in Ontario in early 2000s are Family Health Group (FHG), Family Health Network (FHN) and Family

Health Organization (FHO). The primary care clinics with FHNs and FHOs practice models can apply to the Ontario Ministry of Health and Long Term Care (MOHLTC) to become a Family Health Team (FHT) (Laberge et al., 2017). These models are described in more detail below:

Family Health Group (FHG): In FHGs a group of three or more family physicians have committed to providing after-hours care in addition to being available during regular hours. For after-hours care requirement, each family physician in an FHG needs to provide one to five times of three-hour clinic sessions in evening or weekend per week. Patient enrollment is *not* mandatory in FHG practice model. Family physicians are paid based on *enhanced fee-for-service* which consists of regular *fee-for-service* for non-enrolled patients and a combination of regular *fee-for-service* and monthly care fee for each enrolled patient.

Family Health Networks (FHN): In FHN, the same as FHG model, groups of three or more family doctors are working as a team to provide care during regular and after hours. However, in contrast to FHG, patient enrollment is mandatory in FHN practice model. If a patient receives care in another clinic, the government reduces the access bonus of the FHN clinic. Physician payments in the FHN model are based on *blended capitation funding* which includes both *capitation funding* and *fee-for-service*. *Capitation funding* which is characterized as annual fixed care fee for visits of enrolled patients as well as 15% of regular fee-for-service, is related to the defined basket of services by government for FHN. However, for services which are not defined in the FHN basket, physicians are paid fully based on regular *fee-for-service* system.

Family Health Organizations (FHO): FHO practice model is very similar to the FHN model but has extended services in its basket, some of which requires more extensive procedures or additional training. An example of extended service in the basket of FHO is *Epistaxis–nasal cauterization* (Ontario Medical Association, 2015) defined as

a procedure to prevent nosebleeds. In this procedure, practitioner makes the inside of nose numb include the usage of chemical swab or an electric current to cauterize the inside of the nose (Healthwise Staff, 2018). Family physicians at FHO receive higher amounts of annual fixed care fee in the *capitation* funding.

Family Health Team (FHT): An FHT is an inter-professional team of health care providers consisting of family doctors, nutritionists, social workers, and other professionals who provide comprehensive care to patients enrolled within the FHT. Clinics associated with an FHT offer different healthcare services, such as pharmacies, diabetes management and mental health services. Each FHT has a different set of services. Primary care clinics based on FHN or FHO practice models can apply to become an FHT. If they are successful in transitioning into an FHT, they can receive funding to bring multidisciplinary care providers to the team.

Table 1.1 presents the comparison between all the existing primary care practice models after reforms.

	Attributes of Primary Care Practice Models			
Practice Model	Resource Pooling	Mandatory Enrollment	Interdisciplinary	Reimbursement based
Solo				***fee-for-service***
CHC	✓		✓	***salaried***
FHG	✓			***enhanced fee-for-service***
FHN	✓	✓		***blended capitation***
FHO	✓	✓		***blended capitation***
FHT	✓	✓	✓	***blended capitation***

Table 1.1: Existing primary care practice models after reforms

In this dissertation, two chapters (Chapter 2 and 3) consider a particular family health team, namely the Health For All team in Markham, Ontario, Canada.

Outcomes of Reforms By 2012, the funding that Ontario spent to support the reform in Ontario was more than $1 billion per year in the new models of primary care which include 200 FHTs providing services to two million provincial residents. Since patient enrollment with a primary care provider in Ontario is voluntary, the growth in such enrollments from 600,000 in 2002 to 9.5 million in February 2011 (72 percent of the provincial population) indicated effectiveness of reform strategies (Hutchison et al., 2011). However, provincial decision-makers were becoming increasingly concerned about the growing cost of the reforms. Moreover, the decision-makers believed the government did not receive the value for money invested since Canadian patients still have longer access time compared to most other participant OECD countries (CIHI, 2018). Moreover, there does not exist any performance measurement approach to evaluate access to primary care following the reform strategy (Marchildon and Hutchison, 2016). Therefore, for having an accurate evaluation of the reform strategy, there is a need for a systematic approach to primary care performance measurement which can provide feedback to decision makers on a regular basis (Haj-Ali et al., 2017). To develop such an approach, there is a need for data regarding analyzing the performance of primary care to track the effectiveness of reform strategy and its relative significant investments (Haj-Ali et al., 2017). In the following, we explain the current approach in Ontario to measure performance of primary care in terms of timely access to care.

1.1.2 Current Performance Evaluation of Ontario Primary Care

Based on the literature in primary care, there exist two general methods to assess timely access to care. One method is based on developing a survey for patients. Another method is based on retrieving data from electronic medical record (EMR) of the clinic (Jones et al., 2003).

Current methods to evaluate performance in Ontario are survey-based. There exist two main validated surveys to evaluate access to care based on patient perceptions, which are Primary Care Assessment Survey (PCAS), and Primary Care Assessment Tool-Short Form

(PCAT-S). The existing surveys in Ontario either use one of these two surveys or a combination of them. For example, Quality and Costs of Primary Care (QUALICOPC) Patient Experiences Survey contains First-Contact Access subscale from PCAT-S survey as well as Organizational Access sub-scale from PCAS survey (Premji et al., 2018). First-Contact Access subscale in PCAT-S survey consists of the following four questions: likelihood of being seen same day; getting advice over the phone when clinic closed, having a phone number to call and likelihood of being seen by doctor during the night. In addition, Organizational Access sub-scale in PCAS survey includes six questions regarding rating doctor's office based on location, hours, usual wait for an appointment, usual wait at the clinic and ability to visit the doctor's office or to speak to doctor by phone (Haggerty et al., 2011).

However, as observed from the above listed questions, survey-based assessment of access to care is subjective, as also alluded to by Haggerty et al. (2011). Furthermore, the urgency of requested appointment is not taken into account (Haggerty et al., 2007).

1.2 Proposed Solutions

The current approach in Ontario for evaluating performance of primary care in terms of timely access to care does not consider the clinical severity of the requested appointment. As it is also mentioned in the study of Haggerty et al. (2007), one of the main measurable terms to evaluate the strength of primary care in Canada is "the ease with which a person can obtain needed care (including advice and support) from the practitioner of choice within a time frame appropriate to the urgency of the problem". Therefore, there is a need to define the notion of *equitable timely access to care* that captures whether or not timely access to care matches the urgency of patients' health concerns. In the following, we are first going to explain the definition of equitable access in primary care. Thereafter, we are going to discuss how to evaluate performance of primary care in terms of equitable timely access to care.

1.2.1 Equity and Equitable Access in Primary Care

Equitable access to primary care services can be analyzed from two perspectives. One perspective is fairness in access to primary care by equitable distribution of family physicians based on geographic and socioeconomic factors (Graber-Naidich, 2015). Inequitable access from this first perspective means family physician distribution is represented by insufficient numbers of physicians in some areas and surpluses in other (Graber-Naidich, 2015).

From the second perspective, which is the focus of this dissertation, equitable access to primary care means fairness in managing access time in terms of clinical equity (urgency or acuity level) (Hador et al., 2000). One of the approaches to achieve equitable access is acuity-based allocation of physician time to requests. Therefore, inequitable access from the second perspective means there is not any prioritization system that patient with greater or more urgent needs receive care ahead of those with less urgent needs (Noseworthy et al., 2003). In Canada, the chance of a patient receiving needed primary care services in a timely manner based on clinical urgency is not guaranteed. This is due to the lack of a reliable tool to evaluate the relative priority of patients before booking their requests (Hador et al., 2000). Therefore, current performance evaluation of timely access to care in Canada does not include any equitable access measure.

1.2.2 Objective Assessment of Primary Care Performance

To provide an objective assessment, access to care should be measured based on appointment system data (Jones et al., 2003). The existing objective assessment methods in the literature to measure access to care are as follows.

Oldham (2001) developed a method to find monthly access score through recording the number of days to the third available routine appointment for each clinician and calculating a median to represent an access score for the specified week. Over a month, the average of four median values was taken to find the monthly access score. In another study, Elwyn et al. (2003) developed another method to find Access Response Index that was developed

as an easy calculable measure of organizational access. This index is derived by counting the number of days until the next available routine appointment, with any clinician, once during every normal working day.

This dissertation is based on the idea that performance of primary care in terms of timely access to care should be assessed objectively and based on data from appointment system. Moreover, this evaluation should present whether or not timely access to primary care is equitable because the demand for primary care exceeds the availability of family physicians, and an increase in access times leads to deterioration in life quality of population (Déry et al., 2019). We recognize that objective assessment requires availability of data, which may not always be possible in a healthcare setting; however, with the advancement of medical health records and appointment systems, we hope there will be more healthcare data available (in this dissertation, we use data from a particular family health team in Markham, Ontario).

In the following section, we are going to focus on the approach which is applied in this dissertation to match family physician time to demand in such a way that access time is equitable.

1.2.3 Tactical Capacity Planning

Tactical capacity planning is a key element of planning and control decisions in healthcare settings, focusing on the allocation of a clinic's resources to appointments of different types. One of the essential elements for addressing physician scarcity and long access times is tactical capacity planning (TCP). TCP involves making tactical-level decisions, i.e., medium-term planning decisions for a group of patients instead of individual ones (Hulshof et al., 2012; Ahmadi-Javid et al., 2017). In TCP, the booking decision regarding the number of reserved time slots for each appointment type is made in advance of demand realization. If demand considered as a known value in TCP model and the realization of data is different from the one expected, the resulting solution may not be feasible. Due to variability and uncertainty in demand for appointments, it is difficult to provide an exact match between the scheduled

physician availability and appointment requests. In this dissertation, we use two approaches for TCP, which are: (1) a forecasting model followed by a deterministic optimization model for balancing the load of one physician in the presence of multiple acuity levels, and (2) a robust optimization model for the case of two appointment types with uncertainty in the demand for one of them.

1.3 Dissertation Overview

There are five chapters in this dissertation.

Chapter 1: Introduction This dissertation begins with the current introductory chapter, which includes the background and practical motivation of our study regarding concerns of the Ministry of Health and Long-Term Care in Ontario (MOHLTC) regarding the effectiveness of the reform strategies in helping to maintain better control over access time for appointments.

Chapter 2: Acuity-based access time evaluation in primary care: a case study of an Ontario clinic Chapter 2 focuses on developing a preliminary measure for equitable access to primary care. The developed measures are translated into five acuity levels in primary care. These five acuity levels prioritize patients based on the access time target relative to the urgency of requested appointment. Thereafter, we propose a scale for objective evaluation of equitable access to primary care based on the data from appointment system (electronic medical record). This study is based on the Health for All family health team located in Markham, Ontario, Canada.

Chapter 3: Improving Physician Workload Balance and Timely Access to Primary Care Chapter 3 presents an optimal equitable capacity reservation plan motivated from the major concerns in the Canadian primary care clinics which are inequitable access and family physician scarcity. The developed model provides equitable access time through prioritizing more urgent patients to get served earlier, based

on the acuity levels and access times defined in Chapter 3. In addition, the developed model addresses physician scarcity through increasing utilization by balancing physician workload between weeks. The model is able to reserve appointment slots in advance based on a forecasting model to determine the demand value to capture variability rather than assuming known demand value. Considering demand forecast provides a data-driven optimal capacity reservation plan. This study is based on the Health for All family health team located in Markham, Ontario, Canada.

Chapter 4: A Robust Optimization Model for Tactical Capacity Planning in an Outpatient Setting Chapter 4 focuses on a robust tactical capacity planning (RTCP) model via cardinality-constrained robust optimization which explicitly considers the number of patients who may need to be scheduled in a subsequent planning horizon. The RTCP model protects against uncertainty in the demand per time period. Due to the presence of the uncertain demand parameter in the right-hand side of two constraints, i.e., individual demand per period and aggregated demand over all the periods in the arrival horizon, formulating the robust tactical planning model required the use of both primal and dual constraints in the robust model. Through employing cardinality-constrained robust optimization, we showed how to control over-conservatism through analyzing the trade-off between the budget of uncertainty and feasibility of the robust plan based on different bounds of the uncertainty set. We also analyzed the price of robustness, i.e., the trade-off between feasibility and cost of robust plan for different budgets of uncertainty and bounds of uncertainty set. The findings of this study provide insight for decision makers how a chosen budget of uncertainty and bounds of uncertainty set impact feasibility and the cost of a tactical capacity plan.

1.4 Summary of Contributions

1. We make a step toward the development of a preliminary framework for prioritization of patients based on their acuity in primary care to evaluate equitable access in Ontario primary care. Following that framework, we are the first to propose an objective assessment tool which is an acuity-based, data-driven indicator for evaluation of timely access to primary care.

2. We develop a novel data-driven TCP approach for an individual physician in primary care for all urgent and multiple non-urgent appointment types (such as chronic disease management and routine appointments), motivated by the main challenges in Canadian primary care which are shortage of family physicians, insufficient appointment slots for urgent requests and lack of equitable timely access to care. This approach consists of a forecasting model to determine the demand and an optimization model that allocates time slots to patient requests of different priority. The developed tactical capacity plan is based on a mixed integer linear model which determines the optimal number of appointment slots to reserve per week for each appointment type during a 12-week planning horizon. The model also tracks the unmet demand for each appointment type per week which should be scheduled in the subsequent planning horizon. Our approach results in developing a weekly planning template for a specific physician in a family health team taking into account the needs and preferences of both patients and family physicians.

3. We provide a multi-objective TCP model for a specific physician in a primary care clinic which balances the physician workload between weeks based on the acuity level of the patient request and prioritizing appointment slots with same/next week access. We therefore address the current limitation of improving access to primary care from the literature through equitable allocation of physician time that prioritizes access based on the acuity level of the patient request rather than considering two general

categories of urgent and non-urgent requests.

4. We apply our proposed framework to the case study of a physician in the Health for All family health team in Markham, Ontario, to determine the capacity reservation plan which can provide workload balance and equitable access distribution. Our empirical results show that appointments can be reserved in such a way that the demand is met within the access targets for each of the acuity levels, and that more urgent demands are prioritized to being served earlier (i.e., same day or same week). In addition, access time distribution for more urgent demand is shifted toward left.

5. We develop a robust TCP model for the case when the demand is uncertain and difficult to forecast. We propose a robust TCP model based on the cardinality-constrained method to minimize the highest potential physician peak load between weeks. Therefore, the developed robust TCP model enables protection against uncertainty through providing a feasible allocation of capacity for all realizations of demand. The proposed robust TCP model considers two interdependent appointment types (e.g., new patients and follow ups), multiple access time targets for each appointment type and uncertainty in demand for first-visit appointments. Our experiments demonstrate how to set the level of robustness based on extra cost and infeasibility probability of a robust solution.

Chapter 2

Acuity-based Access Time Evaluation in Primary Care: a Case Study of an Ontario Family Health Team

Abstract[1] Measuring *access* to primary care is complicated due to a variety of perspectives. In Ontario, Canada, one of the main metrics currently used to evaluate access is the proportion of patients who are able to obtain a same- or next-day appointment with a primary care provider. However, this metric does not accurately reflect patients who do not medically require same- or next-day access. In this study, we demonstrate the need for developing more detailed metrics which capture the urgency of needed care via a case study of an Ontario primary care clinic. Our results show that using the standard metric, the clinic's performance appears unsatisfactory, while using the more detailed acuity-based metrics, the clinic is shown to be performing well for non-urgent requests.

2.1 Introduction

Primary care has been considered as a main element of a high-performing health system from the beginning of the 20th century (Bitton et al., 2017). Hence, it is important to evaluate access to primary care, but this evaluation is difficult as *access* can be viewed and measured in multiple ways (Premji et al., 2018). In Ontario, Canada, the main method for evaluation of access to primary care is through surveys, such as the Commonwealth Fund International Health Policy Survey and the Quality and Costs of Primary Care (QUALICOPC) Patient Experiences Survey (PES). However, these methods have two limitations. First, due to

[1]This chapter is published as a conference paper: Nazanin Aslani, Fariborz Fazileh, Donatus Mutasingwa, and Daria Terekhov. Acuity-based access time evaluation in primary care: a case study of an Ontario clinic. In Proceedings of the 4th International Health Care Systems Engineering Conference (HCSE'19), 2019.

being survey-based, they are influenced by respondent perceptions and biases. Evaluating access only from patient perception can be misleading due to the weak relationship between care accessibility and patient perception of access (Llanwarne et al., 2013). Second, current metrics calculated from survey data, most notably the number of patients who obtain a same-day or next-day appointment, do not consider the urgency of the patient request. However, given the scarcity of healthcare providers in Canada (Globerman et al., 2018), knowing the urgency of the patient could lead to more equitable and effective allocation of available physician time. The need for prioritization of patients in the setting of scarce resources is well-known in medical environments outside of primary care, such as emergency departments (Iserson and Moskop, 2007); it has also recently been examined in the context of non-emergency settings, such as physiotherapy or rehabilitation services (Harding and Taylor, 2013).

Our focus is the evaluation of *access time*, which is defined as the interval between the arrival of an appointment request and the scheduled time of appointment (CIHI, 2018). To address the above limitations, we argue for the evaluation of access time a) through clinic data in order to overcome the potential subjectivity resulting from surveys, and b) based on detailed metrics related to the various patient groups that primary care serves, akin to how emergency care performance is measured through different access time targets for patients of different acuity. While the first argument has already appeared in previous literature, see e.g., Rao et al. (2006), we provide further evidence that there exist discrepancies between performance evaluation from objective data and patient surveys. Our second argument builds on work by Haggerty et al. (2007), whose definition of accessibility considers the appropriateness of access time "to the urgency of the problem", and by Premji (2018), who demonstrates a limitation of the most-prominent metric used for evaluation of Ontario primary care, i.e., the percentage of patients able to obtain a same-day or next-day appointment. Motivated also by the use of prioritization in other medical contexts with scarce resources, we propose to categorize patients in primary care according to the urgency of their request, a proposal

which, to the best of our knowledge, has not been explored in the literature.

Our argument is illustrated by a case study of the Health for All (HFA) clinic, located in Markham, Ontario, Canada. Using a comprehensive data set of patient records from September 2017 to September 2018, we compute both the standard same-day/next-day access metric as well as the proportion of patients obtaining care with access times within targets appropriate to their level of acuity. Our results show a discrepancy between the two evaluation approaches: for non-urgent patients, using the standard metric, the clinic's performance appears unsatisfactory, while using the more detailed metrics, the clinic is shown to be performing well.

2.2 Current Performance Evaluation in Ontario Primary Care

Health Quality Ontario (2018) evaluated performance of primary care in Ontario, Canada, based on the Primary Care Performance Measurement framework, which evaluates nine domains, including *access*. The specific criteria used to evaluate access by Health Quality Ontario (2018) include, among others, *timely access at regular place of care*. Prior to 2017, Health Quality Ontario focused on the percentage of patients with same-/next-day access to a primary care provider, based on the question "The last time you were sick, how quickly could you see *any* doctor, nurse practitioner or physician assistant in this clinic?" In 2017, the latest year for which the performance report is currently available, a distribution of access times is presented, showing that 39.9%, 26.5%, 19.2% and 14.5% had access times of < 2 days, 2-3 days, 4-7 days, and ≥ 8 days, respectively. The percentage of people waiting for ≥ 8 days ranged from 5.6% in the Central West region to 40.7% in the North West. At the same time, 67.6% of the respondent Ontarians reported that their wait for an appointment was "about right", 18.3% said "somewhat too long" and 14.1% said "much too long", with a range of 10.2% (Toronto Central) to 23.6% (North East) in the "much too long" category. Interestingly, the regions with the highest proportion of appointments with high access times are not necessarily the ones with the highest percentage in the "much too

long" wait category. This observation supports our investigation: first, it demonstrates the impact of patient perceptions and expectations; second, it does not capture how many of the appointments were obtained within medically-warranted time frames.

The Quality and Costs of Primary Care (QUALICOPC) Patient Experiences Survey (PES) is a framework for evaluating care quality and outcomes in primary care (Wong et al., 2015). For evaluating timely access to care, QUALICOPC-PES asks: "How many days did you wait for this visit from the time that you tried to make an appointment? For patients who had made an appointment for today's visit: Was it easy to get the appointment? Were you able to arrange an appointment with the doctor as soon as you wanted to?" For data collected between 2013 and winter 2014, 32% (of 1379) respondents said that they obtained a same-/next-day appointment; surprisingly, 87% (of 1536) said they were "able to arrange an appointment as soon as [they] wanted to" (Laberge et al., 2014). These differing statistics again motivate the need for further research into performance evaluation: for instance, were the complaints of patients who did not get a same-/next-day appointment and yet were satisfied less urgent than the ones who were dissatisfied?

The Canadian Institute for Health Information Commonwealth Fund Survey (Canadian Institute for Health Information, 2017) in 2016 reports that "only 43% of Canadians were able to get a same- or next-day appointment at their regular place of care last time they needed medical attention". Furthermore, the study shows the statistic provided for patients who say they could get a same- or next-day appointment for 2016 is 43% while the statistic provided for "primary care physicians who say most (at least 60%) of their patients can get a same- or next-day appointment" is 53% in 2016. The difference between the statistics reported for patients and physicians is a representation of the difference in their perception of who needs same- or next-day appointment as well as the existence of bias in surveys.

2.3 Literature Review

Primary Care Performance Evaluation Jones et al. (2003) state that there exist two main approaches to evaluating primary care performance: appointment-data-based and survey-based. However, to the best of our knowledge, all existing studies in Canada are survey-based. Haggerty et al. (2007) consulted primary healthcare (PHC) experts across Canada to formulate operational definitions of PHC attributes that should be evaluated in the Canadian primary healthcare setting. Importantly, the definition of *first-contact accessibility* as "the ease with which a person can obtain needed care (including advice and support) from the practitioner of choice within a time frame appropriate to the urgency of the problem" received a high level of physician consensus. We highlight in their definition the need to define a "time frame appropriate to the urgency of the problem": for acute patients, the *appropriate* time frame might indeed be same-day or next-day, but for patients requesting a periodic health exam, obtaining an appointment within several weeks of their requests is reasonable. Similarly, by analyzing the data from the QUALICOPC PES, Premji et al. (2018) determined that the same-day/next-day access to primary care indicator does not match patients' perceptions of access to primary care.

Patient Classification in Primary Care The aim of patient classification is to prioritize patients objectively based on equitable criteria to ensure that patients with more urgent needs receive services first (Déry et al., 2019). In the literature, the classification of primary care patients into different types has been considered in the context of improving primary care payment schemes (e.g., Starfield et al. (1991)) as well as improving patient access times. In the latter category, Balasubramanian et al. (2010) study improving access by redesigning a physician panel based on the patients' age and presence of chronic disease; Ozen and Balasubramanian (2013) use the number of simultaneous chronic conditions a patient has to classify patients in primary care and use as a predictor of the number of visits. In the capacity allocation literature, the majority of papers classify patients as urgent and non-

urgent, e.g., Wang and Gupta (2011). However, based on our knowledge none of the existing access-time-focused literature in primary care explores the idea of performance evaluation based on acuity levels or prioritization to provide equitable access time.

2.4 The *Health for All* Clinic Background

The Health for All (HFA) clinic is adjacent to the Markham Stouffville Hospital, located in the City of Markham in the Regional Municipality of York within the Greater Toronto Area of Southern Ontario, Canada. It is located approximately 30 km northeast of Downtown Toronto. The HFA clinic is affiliated with the University of Toronto's Department of Family and Community Medicine. Residents spend their final two years of training with HFA to become family physicians; they see their own patients and go through clinical rotations at the hospital. HFA is a family health team (FHT) – an inter-professional team of health care providers consisting of family doctors, nutritionists, social workers, and other professionals who provide comprehensive care to patients enrolled within the FHT.

Data We use a comprehensive data set from the HFA clinic for the period from September 2017 until September 2018. The data set contains 60682 records listing the provider name, booking date, appointment date, appointment type, primary MD, no show, appointment detail, scheduled time, duration, arrival time and departure time. In order to prepare the data for analysis, we remove extra records and outliers. The records that we remove from consideration are those with negative access time (time of appointment in data set was earlier than the time of booking); without booking or appointment date; with doctor unavailability; home visits; evening and Saturday clinic appointments; and records that were labeled as "deleted" (by the administrators), which (to the best of our knowledge) corresponds to appointments that were rescheduled for later. After removing these records, the remaining data set consists of 39608 records. In order to choose an appropriate outlier labeling method, we considered whether the underlying data is symmetric or skewed (Ben-Gal, 2005). His-

tograms of access times for all appointment types in this study were found to be right-skewed. Therefore the Adjusted Boxplot developed by Hubert and Vandervieren (2008) is applied as an outlier labeling method. After removing the outliers using this method, we are left with 39397 records in the data set.

Current Appointment Types at HFA The appointment classification system at HFA is based on patient complaints, i.e., the reason why the appointment has been requested. When a patients calls the clinic, an administrative clerk asks the patient for the reason of their request, their family doctor, their availability, etc., and suggests a time slot. To aid this process, the HFA clinic currently classifies appointments into 14 types: 12 of these are presented in the first column of Table 2.1, with example conditions given in the second column. Since the focus of this study is on the access time of patients who physically visit the clinic, we do not consider the *Home-visit* appointment type in our analysis. Table 2.1 also omits the *New Patient* category, an appointment for a patient to be introduced to their new family physician. Another category of appointment is referred to as *Blank*. The *Blank* appointment type encompasses appointments such as same-day and same-week requests which are patient-driven (based on the patient perception that they need to be seen by a family physician within the same day or same week of their request). In general, the *Blank* appointment type includes all requests that cannot be considered as one of the other 13 appointment types. Some examples of the conditions typically occurring in this category include seasonal illnesses like flu and sore throat, new issues due to current treatment for an ongoing medical condition, any new symptom a patient experiences that cannot be clearly classified into another category, and prescription renewals. Importantly, requests for an appointment in the *Blank* category correspond to 52% of all HFA appointment requests; anecdotally, it appears that a large proportion of *Blank* appointments request same-day or same-week appointments. However, this hypothesis cannot be confirmed by the current data set due to missing descriptions of patient conditions in the data. In other words, due to

lack of data on which patients request same day or same week appointment, our analysis is limited. In future work, the goal is to develop an exact list of patient conditions and collect data on these conditions and the corresponding appointment time slot that was obtained; obtaining this data will lead to more comprehensive analysis.

App type	**Condition Example**	should be seen within
Follow-up	Soft tissue infection started on an antibiotic	1 week
	Sub-acute abdominal pain with blood work & imaging	2 weeks
	Hypertension with recent medication change	4 weeks
	Thyroid medication dose modification	12 weeks
Injection	Travel medicine injection	1 week
	First visit of a patient who has not started their routine immunization in their infancy	2 weeks
	Intra-articular injection	4 weeks
	Repeat intra-articular injection	12 weeks
Mental Health	Anxiety	2 weeks
	Depression started on new medication	4 weeks
	Mental health condition responded moderately to medication change	12 weeks
First Pre-Natal	Appt requested 1-2 weeks before week 8 of pregnancy	2 weeks
	Appt requested 3-4 weeks before week 8 of pregnancy	4 weeks
Pre-Natal	After week 28 of pregnancy	2 weeks
	Week 12 till week 28 of pregnancy	4 weeks
Well Baby	Baby should be seen in 4m,6m,9m,12m,15m,18m	12 weeks
Pre-Op Assessment	request 1-2 weeks before operation	2 weeks
	request more than 2 weeks before operation	4 weeks
Diabetic Management	Patient should be seen every 3 months	12 weeks
Child Physical	Child should be seen in 2 yr,4 yr,6 yr, 16 yr	12 weeks
Periodic Health Exam	Annual visits for chronic illness and/or health issues	12 weeks
Driver's Physical	Every 5 years if $<$ 46 y/o; 3 years if 46-64 y/o; annually if $>=$ 65 y/o	12 weeks
Blank	Febrile illness	1 day
	Prescription renewal	1 week
	New skin lesion	2 weeks
	Medical forms	4 weeks
	Pap test	12 weeks

Table 2.1: Appointment types, example conditions and proposed access time bounds.

2.5 Proposed Acuity-based Evaluation

As seen from Table 2.1, the complaints for which family medicine clinic appointments are requested vary widely in their nature and urgency. This observation suggests that evaluation of access time which considers the urgency of the patient request would give a more accurate representation of the performance of primary care and lead to more equitable allocation of appointments to patients.

For example, consider patient A who has diabetes and is calling for his/her regular three-month appointment. In this case, as long as the access time of patient A is approximately three months from the previous check-up, the care needs of the patient and risks of adverse health outcomes are appropriately addressed. In contrast, consider patient B who calls for a follow-up appointment for an eight-month-old baby with five days of fever and a possible viral illness diagnosis. Patient B needs to be seen in the clinic on the same day to ensure they are not at risk for significant adverse health outcomes. Table 2.1 provides additional examples of conditions of various urgency – the third column of Table 2.1 gives potential access time upper bounds obtained through discussions with two practitioners from HFA. For the majority of appointments, the access time upper bounds can be defined as the maximum time from the arrival of the patient request until the patient is seen; however, for some periodic appointments such as physicals, or for follow-up appointments, the upper bound is defined as the time between the previous appointment and the next one (e.g., the time between an appointment to resolve an initial complaint and a follow-up appointment to discuss the effectiveness of the care received).

As described in Section 2.2, current methods of primary care performance evaluation in Ontario do not take into account the urgency of the patient request; furthermore, as shown in Section 2.3, a detailed classification of patient urgency in primary care is done for billing purposes or, for the purposes of appointment allocation, is usually limited to two types. Such a performance evaluation approach in primary care is in contrast to performance evaluation in emergency care. In emergency care, the Canadian Triage Acuity Scale (CTAS) is employed

to classify patients into five categories in order to "triage patients according to acuity, risk, and care needs based on their presenting signs and symptoms" and to "ensure that the sickest and highest risk patients are seen first when ED capacity has been exceedeed" (Canadian Association of Emergency Physicians). In addition, ED managers can use CTAS to "capture and analyze ED patient visit based on volume, acuity and by CEDIS presenting complaint" (Canadian Association of Emergency Physicians).

Considering Table 2.1 and using an analogy with the CTAS system in emergency care, we propose a five-level classification of patients in primary care based on their urgency, varying from urgent patients that need to be seen within one day to routine patiens that should be seen within 12 weeks, as shown in Table 2.2. In Table 2.3 we show how the current HFA appointment types map to our proposed acuity levels.

Acuity level	**Description**	**Should be seen within**
1	Same day urgent	1 day
2	Same week urgent	1 week
3	Two weeks non-urgent	2 weeks
4	Four weeks non-urgent	4 weeks
5	Routine	12 weeks

Table 2.2: Proposed acuity levels and corresponding access time targets.

Appointment type	**Acuity level**	**Appointment type**	**Acuity level**
Periodic Health Exam	5	*Injection*	2-5
Child Physical	5	*New Patient*	1-5
Diabetic Management	5	*Pre-Op Assessment*	3,4
Driver's Physical	5	*Pre-Natal*	3,4
First Pre-Natal	3,4	*Well Baby*	5
Follow Up	2-5	*Blank*	1-5
Mental Health	3-5		

Table 2.3: Acuity levels per appointment type at HFA.

Given the acuity levels defined in Table 2.2, we can now evaluate the performance of

primary care with respect to the urgency of the appointment, by calculating the proportion of requests in each acuity category that are scheduled within the proposed access time upper bound (referred to as *access time target*).

2.6 Measuring Timely Access to Care

We first analyze HFA clinic access time performance based on the indicators from the Ontario Ministry of Health and Long-term Care (MOHLTC), followed by performance evaluation based on the acuity levels defined in Table 2.2.

Evaluation based on MOHLTC indicator MOHLTC considers same-/next-day access as one of the major access time indicators (Health Quality Ontario, 2018). Patients at HFA reported "same-day or next-day access when they are sick" as 54% and 52% for 2016 and 2017, respectively. Calculating the same metric from appointment data, we see that only 32% (out of 38367) and 31% (out of 39608) of patients had same-/next-day appointments for 2016/17 and 2017/18 data, suggesting that the surveyed sample consisted of patients who need same-/next-day appointments based on patient perception. In addition, it is not clear whether patients who are over-due to visit for a chronic condition would classify their request as being in the category "when they are sick". However patients who get same-/next-day appointments based on appointment data may not need it. Therefore, in general, there are two approaches of collecting the required data for evaluating performance of the clinic. One approach is a survey, which may not provide an accurate evaluation for two reasons: firstly, due to its subjectivity – it is influenced by an overall favourable perception of patient experience at the clinic, and secondly, due to the difference in the survey sample sizes from the physician and patient perspectives. Furthermore, the values 31% and 32% seem to indicate sub-par performance of the HFA clinic.

In Table 2.4, we present the cumulative percentage of same-/next-day and same-week appointments for each patient class based on our 2017/18 data set.

	Access Time			
Appointment Type	Total #	% same d	% next d	% same w
Periodic Health Exam	2309	10%	13%	14%
Child Physical	910	8%	10%	12%
Diabetic Management	1906	6%	8%	9%
Driver's Physical	38	18%	26%	55%
First Pre-Natal	136	7%	9%	13%
Follow Up	5978	19%	24%	27%
Mental Health	965	10%	13%	16%
Injection	1499	48%	53%	55%
New Patient	1470	13%	17%	18%
Pre-Op Assessment	247	19%	24%	28%
Pre-Natal	708	7%	9%	11%
Well Baby	2711	6%	9%	11%
Blank	20731	37%	42%	46%
All Types	39608	27%	31%	35%

Table 2.4: % of Requested appointment types with same-day/next-day/same-week access time.

Looking at the percentages for the current appointment types, we see that the percentage of same-/next-day visits ranges from 8% (*Diabetic Management*) to 53% (*Injection*), which shows that the HFA clinic does not necessarily reserve time slots with same-/next-day access time for the most urgent requests. Additionally, these results show substantial variability among patient classes.

Evaluation based on acuity-based indicators Figures 2.1–2.3 show access time histograms for three appointment types, *Follow up*, *Diabetic Management* and *Periodic Health Examination*. Importantly, from these histograms we can observe that access time behaviour of different appointment types is quite different. The histogram for *Follow up* shows a zero-inflated distribution with a long right tail that extends beyond 120 days; the histogram for *Diabetic Management* is bi-modal, with peaks at two weeks and 90 days; for the *Periodic Health Exam*, the majority of appointments happen within one month, with a mode of 1 week. In addition to the differences in behaviour among the appointment types, in all three

histograms, we see substantial variability in access times among patients within each appointment type. For *Follow-up* appointments, there are peaks, of diminishing magnitudes, at the end of every week, suggesting the existence of multiple "sub-types" in the *Follow-up* category, that is, patients whose follow-up appointments should be in one week, two weeks, etc. For diabetic appointments, the peak at 90 days matches the suggested interval between two regular visits for a patient with diabetes. For the periodic health exam, the high number of patients being scheduled within one day and one week is particularly surprising, given that these are non-urgent appointments.

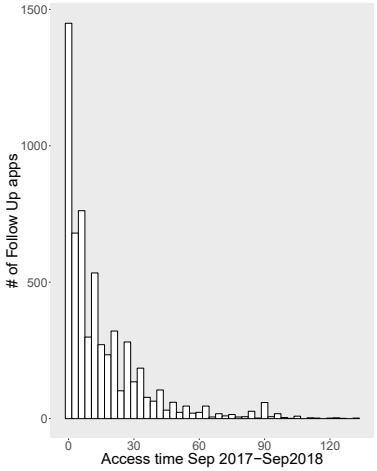


Figure 2.1: *Follow-up* access times (days).

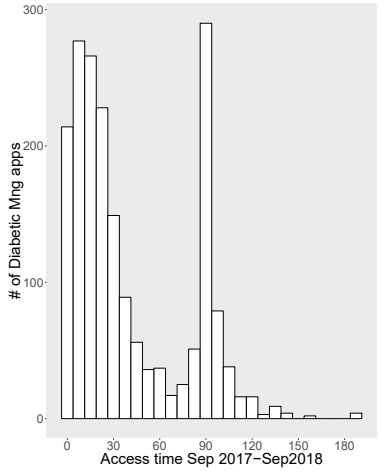


Figure 2.2: *Diabetic Management* access time.

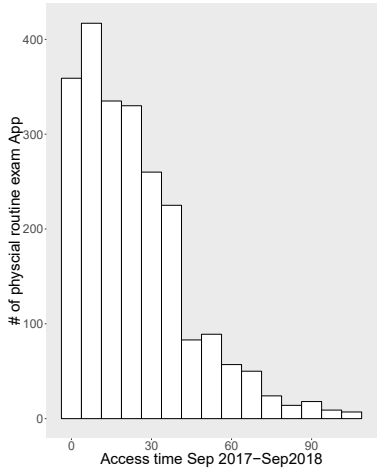


Figure 2.3: *Periodic Health Exam* access time.

Appointment Type	**Range**	**Mean**	**Median**	**Mode**	**% seen within access target**
Periodic Health Exam	0-15	3.44	3	1	98%
Child Physical	0-28	4.05	3	2	95%
Diabetic Management	0-27	5.94	4	2	91%
Well Baby	0-25	4.93	4	1	91%

Table 2.5: Summary of access times (in weeks) for acuity level 5 appointment types at HFA.

We now focus on the appointment types identified in Table 2.3 as having a single acuity

level of 5. For these appointment types, we present additional statistics as well as the evaluation of their access times according to our proposed metric in Table 2.5. In particular, we observe that the majority of patients requesting *Periodic Health Exam*, *Child Physical*, *Diabetic Management* and *Well Baby* appointments were able to obtain them within the suggested upper bounds on access time, demonstrating that HFA performs very well for non-urgent appointments, which is not obvious from standard metrics, and in fact contradicts the conclusion one would make from looking at same-/next-day metrics over all appointment types or even for these specific non-urgent appointment types. Furthermore, we can observe that a large number of acuity level 5 patients obtained a same-/next-day appointment, despite being of low urgency. In a setting with scarce resources, this observation effectively implies that same-/next-day appointment times are not being used effectively by the clinic, and can inform new allocation policies.

Discussion A limitation of our study is that we could not evaluate the performance of the clinic for appointment categories with multiple acuity levels (see Table 2.3) due to the lack of data regarding patient complaints. Furthermore, we note that the acuity level definitions proposed in Table 2.2 constitute a proposal that we expect will be refined by researchers and practitioners in future work. Given these (or refined) acuity level definitions, the list of conditions and corresponding access time targets would require substantial work from clinicians, similarly to the work involved in defining and updating CTAS. We observe that any potential implementation of our proposal in practice immediately raises the question of triage in primary care. However, despite the work required for formalizing a prioritization scheme and triage procedures, doing so may lead to more equitable resource allocation in primary care, which so far remains a setting with limited resources.

2.7 Conclusion

In this paper, we contrasted performance evaluation of primary care using a traditional metric and new metrics via a case study of a primary care clinic, namely the Health for All Clinic in Markham, Ontario, Canada. Inspired by CTAS classification used in emergency departments, we defined acuity levels for the conditions relative to each appointment type in our case study of primary care. Future work needs to focus on developing the exact list of conditions for different acuity levels as well as evaluation of other clinics.

Chapter 3

Improving Physician Workload Balance and Timely Access to Primary Care

Abstract We develop a deterministic tactical capacity planning (TCP) model to balance workload between weeks for an individual family physician in a primary care clinic. Unbalanced workload among weeks may lead to provider overtime for weeks with high workload and provider idle time for weeks with low workload. In the proposed TCP model, we incorporate the results from detailed access time evaluation in a prior study (see Chapter 2) as constraints for access time. The proposed TCP model considers 11 appointment types with multiple access targets for each appointment type. The TCP model takes as input a forecast of demand coming from an ARIMA model. We compare the results of the TCP model based on the distribution of workload among weeks and access time for different appointment types. The results show that the proposed multi-objective model leads to a capacity reservation plan which balances physician workload for a specific physician, provides equitable access to patients, and prioritizes appointment slots with same/next week access.

3.1 Introduction

Primary care is the first point of contact between the patient and the healthcare system. Primary care is responsible for providing a variety of health services such as same-day appointments for patients who perceive their issue as urgent, chronic disease management, routine visits such as periodic health exams, as well as coordination of care to specialists and follow-ups. One of the essential elements to provide timely access to care in the presence of family physician scarcity is tactical capacity planning (TCP). TCP addresses the prob-

lem of allocating available capacity to different patient classes in an outpatient setting such as primary care (Hulshof et al., 2012; Ahmadi-Javid et al., 2017). Efficient TCP ensures that a sufficient number of physician time slots are allocated to both urgent and non-urgent appointment requests while taking physician availability into account.

The task of allocating capacity is complicated due to the need to consider the (potentially conflicting) preferences and needs of patients and family physicians. Patient needs can be expressed by means of *access time*, which is defined as the interval between the arrival of the appointment request and the scheduled time of appointment. To ensure patient satisfaction, TCP should provide equitable access time, which means providing patients with care within access times that are appropriate to their level of urgency. To the best of our knowledge, among the literature focusing on improving access to primary care, Aslani et al. (2019) is the only primary care study which, instead of using just two general categories of urgent and non-urgent, classifies patients based on five acuity levels to provide equitable access (see Chapter 2); this is the study we use to define access time targets in the current paper. On the other hand, due to demand variability between weeks, the workload for family physicians varies substantially and can become unbalanced. The presence of unbalanced workload among weeks may in turn lead to overtime for the weeks with higher workload and idle time for the weeks with lower workload. Thus, balancing physician workload between weeks is necessary to ensure higher levels of physician satisfaction.

In this paper, we consider both aspects, proposing an approach which starts by forecasting the demand for patients with different acuity levels, followed by solving an optimization model which considers the access time targets of patients of different acuity while balancing physician workload. Our approach is illustrated through the case study of a primary care organization in Ontario, Canada, namely the Health for All (HFA) family health team which is located in the City of Markham and is a teaching unit affiliated with the University of Toronto's Department of Family and Community Medicine.

Assumptions As mentioned in Chapter 1, in a family health team setting, a team of family physicians are working together in an effort to provide more available time slots for patients who prefer to be seen within the same or next day of their request. However, due continuity of care concerns, it is still preferable, both from the patient and the provider perspectives, for the physicians to mainly see patients from their own roster. Thus, in this study, we focus on determining the optimal number of reserved time slots for a single physician, i.e., determining how much capacity should be reserved for each patient type in order for the physician to be able to take care of her/his roster appropriately; in fact, if this goal can be achieved, the resource pooling among physicians would be needed only occasionally. Therefore, we assume that at HFA a developed TCP model for a family physician is independent of other family physicians at HFA. Future work will require looking at resource pooling and the effect of resource pooling on the tactical capacity plan.

An additional assumption of this study is that same-day and same-week appointment types are considered as one category since the time unit used for modelling is a week, and the allocation of slots is weekly. In future work, we will develop a daily TCP model which is based on the developed weekly TCP model in this study to consider same-day and same-week appointment types as two separate categories.

We also assume that a forecasting approach that results in a specific forecasted demand value is accurate enough for us to use deterministic optimization models. In the future, we can consider the effect of noise or uncertainty in demand within the optimization models.

Contributions The specific contributions of this work are:

- We develop a novel data-driven TCP approach for an individual physician in primary care motivated by the main challenges in Canadian primary care which are shortage of family physicians, insufficient appointment slots for urgent requests and lack of equitable timely access to care. This approach consists of a forecasting model to determine the demand and an optimization model that allocates time slots to patient

requests of different priority. The developed tactical capacity plan is based on a mixed integer linear model which determines the optimal number of appointment slots to reserve per week for each appointment type during a 12-week planning horizon. The model also tracks the unmet demand for each appointment type per week which should be scheduled in the subsequent planning horizon.

- We provide a multi-objective TCP model for a specific physician in a primary care clinic which balances the physician workload between weeks based on the acuity level of the patient request and prioritizing appointment slots with same/next week access. We therefore address the current limitation of improving access to primary care from the literature through equitable allocation of physician time that prioritizes access based on the acuity level of the patient request rather than considering two general categories of urgent and non-urgent requests.

- We apply our proposed framework to the case study of a physician in the Health for All family health team in Markham, Ontario, to determine the capacity reservation plan which can provide workload balance and equitable access distribution. Our empirical results show that appointments can be reserved in such a way that the demand is met within the access targets for each of the acuity levels, and that more urgent demands are prioritized to being served earlier (i.e., same day or same week). In addition, access time distribution for more urgent demand is shifted toward left.

The paper is organized as follows. Section 3.2 describes recent work in tactical capacity planning in primary care to improve access. Section 3.4 describes the problem and steps to provide an optimal equitable tactical capacity plan in primary care that meets the needs of both patients and providers. Three optimization models are proposed. The first model focuses only on balancing workload based on the acuity level of appointments. The second model aims to ensure an individual physician's workload balance and equitable access time for patients. Finally, the third model is multi-objective and aims to achieve workload balance,

equitable access and an equitable workload distribution. In Section 3.7, we conduct an extensive set of experiments and discuss the results. Section 3.8 concludes the paper.

3.2 Literature Review

The literature review covers the literature on three topics: patient classification, tactical capacity planning in outpatient setting and tactical capacity planning in primary care.

3.2.1 Patient Classification

Papers focusing on planning and scheduling in outpatient settings classify patients based on features such as service time variability, no-show rate, patient preferences, priority levels, first-time or returning patients, pre-scheduled or same-day patients, and combination of appointments for a specific treatment (Wang and Gupta, 2011; Ahmadi-Javid et al., 2017). These studies can be divided into two categories based on two performance indicators: waiting time (direct waiting time) and access time (indirect waiting time). The terms that are used in this document are waiting time and access time. Waiting time is the difference between a patient's scheduled time of appointment and the time when the patient is actually served by the service provider. Access time is the interval between the arrival of the appointment request and the scheduled time of appointment (Gupta and Denton, 2008). The literature with access time as an indicator mainly used priority levels, first-time or returning patients, pre-scheduled or same-day patients, and combination of appointments for a specific treatment as the features for patient classification (Dobson et al., 2011; Gupta and Wang, 2008; Ahmadi-Javid et al., 2017). In this study we focus on the impact of reserved capacity through classifying patients based on the defined priority levels in chapter 2 for primary care based on our knowledge the literature with the focus on tactical capacity planning in primary care only classifies patients into the two categories of urgent and routine appointment types. However, considering more detailed patient prioritization will lead to more equitable allocation of physician time (Déry et al., 2019). The literature with the focus on tactical

capacity planning in outpatient setting are explained in the next section.

3.2.2 Tactical Capacity Planning in Outpatient Setting

As already mentioned in Section 3.1, TCP is a key element of planning and control decisions in outpatient settings. TCP focuses on how to allocate the available capacity of a clinic's resources among different patient classes (Hulshof et al., 2012; Ahmadi-Javid et al., 2017). Table 3.1 presents some of the recent studies with the focus on TCP in outpatient settings.

reference	Attributes of TCP for multiple patient groups				
	outpatient clinic setting	TCP	patient group	equitable access	physician preference
Qu et al. (2007)	primary care	static	urgent&pre-scheduled	–	–
Gupta and Wang (2008)	primary care	dynamic	urgent&pre-scheduled	–	–
Patrick et al. (2008)	diagnostic	dynamic	priority levels	✓	–
Qu et al. (2011)	primary care	static	urgent&pre-scheduled	–	–
Qu et al. (2012)	primary care	static	urgent&pre-scheduled	–	✓
Balasubramanian et al. (2012)	primary care	static	urgent&pre-scheduled	–	–
Hulshof et al. (2013)	integrated care	static	multiple appt	✓	–
Qu et al. (2013)	women's clinic	static	specific equipment	–	✓
Balasubramanian et al. (2014)	primary care	dynamic	urgent&pre-scheduled	–	–
Gocgun and Puterman (2014)	chemotherapy	dynamic	priority levels	✓	–
Nguyen et al. (2015)	urology	static	first visit, revisit	✓	–
Wiesche et al. (2017)	primary care	static	urgent &pre-scheduled	–	✓

Table 3.1: Summary of literature in tactical capacity planning for multiple patient groups in outpatient setting

Generally, TCP approaches are either static or dynamic. Static TCP provides a tactical

plan based on the demand over a few months through deterministic allocation of capacity to different patient groups (Hulshof et al., 2013). However, dynamic approaches result in plans through considering daily variability in demand for capacity allocation (Vermeulen et al., 2009). Static TCP is a prerequisite for an effective dynamic TCP. In other words, static TCP provides a tactical rule at the beginning of the planning horizon by determining an optimal number of reserved time slots for each patient class (Hulshof et al., 2013). Dynamic TCP adjusts the optimal number of reserved time slots for each patient class in response to the variability in demand or/and resource availability in each period of the planning horizon (Hulshof et al., 2016).

As presented in Table 3.1, the TCP models developed in some of these studies considered equitable access. In the current literature, to achieve equitable access patients are classified either based on priority levels, or access target(s) are defined for a specific percentile of patients in a particular patient class (Patrick et al., 2008; Hulshof et al., 2013; Gocgun and Puterman, 2014; Nguyen et al., 2015). As an example of classifying patients based on acuity levels, Patrick et al. (2008) classified patients based on multiple priority levels to allocate capacity of CT-Scans based on the urgency of the request. Patrick et al. (2008) considered an access target for each priority level and defined class-specific costs for late booking while the relative access target for a patient class is not achieved. Therefore, the developed TCP model in Patrick et al. (2008) aims to book patients in a particular urgency class prior to a specific target date, which is similar to what we propose in this study.

Gocgun and Puterman (2014) extended the developed model in Patrick et al. (2008) by defining priority levels through considering both access target and tolerance limit. For example, the tolerance limit (2,0) corresponds to a tolerance limit of 2 days before and 0 day after the specific access target date. The reason is that there are some requests which cannot be booked too early due to the required interval between patient visits for the treatment process. As future work, we plan to extend the defined patient classification for primary care in Chapter 2 based on our inspiration from the study of Gocgun and Puterman (2014); we

further discuss this idea in Chapter 5.

In terms of defining access target(s) for a particular patient class, Hulshof et al. (2013) developed a TCP model for an integrated outpatient setting in which a patient should visit more than one outpatient clinic and therefore book multiple appointments to complete her/his treatment. Hulshof et al. (2013) considered each outpatient clinic as a stage and defined a specific access target to provide equitable access for patients in each stage. Nguyen et al. (2015) classified patients based on the attribute of re-entry systems. This re-entry system considers the fact that any first visit may lead to multiple revisits in an outpatient setting. Nguyen et al. (2015) developed a TCP model with equitable access through restricting the probability distribution of access time for first visit patient class, and limiting the variation of access time for revisit patient type.

As it is presented in Table 3.1 the literature with the focus on primary care did not incorporate equitable access to develop a TCP model. In Section 3.2.3, we elaborate on the studies with the focus on developing TCP model in primary care.

The last attribute that is presented in Table 3.1 is physician preference, which is considered in only a few studies. There exist two main metrics in the literature to calculate physician preferences: the average number of consulted patients in each period and the variation in the number of consulted patients between periods in the planning horizon. We elaborate more on physician preferences in Section 3.2.4.

3.2.3 Tactical Capacity Planning in Primary Care

TCP in primary care mainly considers two appointment types: same-day appointments, which are booked as calls come during the workday and pre-scheduled appointments, which are booked in advance of a given workday. The main challenge which is addressed in the literature is providing an optimal percentage of same-day and pre-scheduled appointments. Some studies (Qu et al., 2007, 2011; Dobson et al., 2011) focus on finding an optimal number of same-day and pre-scheduled appointments which is the fixed for different days of a week.

Other studies (Balasubramanian et al., 2014; Wiesche et al., 2017) consider demand variation for different days of a week and therefore the optimal number of time slots reserved for same-day and same-week varies in a daily basis.

In terms of fixed number of reserved time slots for same day and pre-scheduled appointments, Qu et al. (2007) describe the conditions for the optimal percentage of same-day appointments and propose an analytical formulation to find the percentage of same-day to maximize the expected number of patients seen. The presented formulation is based on the ratio of average demand for same-day appointments to provider capacity and the ratio of the no-shows for the prescheduled and same-day appointments. Gupta and Wang (2008) consider the effect of patient choice on developing a TCP model in a primary-care clinic where patients may have preference for physician and date of service. Patients are divided into those that request same-day service and those that seek an advanced appointment. Although a penalty function is included to penalize the clinic if it cannot meet the request of a patient, the model is not designed to track patient access time. Qu et al. (2011) address the challenge of lack of enough time slot for same-day appointment due to pre-scheduled appointments by proposing two time horizons for open access scheduling. If the correlation coefficient of historical requests for pre-scheduled and open appointments in each session is high and positive, two time horizons for open access scheduling is advised. Qu et al. (2012) extend the preceding analytical model of Qu et al. (2007) by reducing variability in addition to increasing the average number of patients seen through developing a mean variance model.

For the studies with the focus on varied number of same-day and pre-scheduled appointments per day, Balasubramanian et al. (2014) present a mathematical model for dynamic allocation of same-day demands into multiple physicians in a primary care. This allocation decision is made in presence of dynamic arrivals of same-day requests over a workday, incomplete information for same-day demand and presence of pre-scheduled demand. This study shows the policy regarding the location of time slots assigned to prescheduled appointments

significantly impact the number of time slots allocated to same-day appointments. Through evaluating different policies, the one which books prescheduled appointments in two blocks of early morning and early afternoon works very well. Balasubramanian et al. (2014) assumes fixed number of time slots for pre-scheduled appointments and only find optimal number of same-day appointment which varies over different days of a week. However, Wiesche et al. (2017) additionally determine the optimal capacity for pre-scheduled appointment on each day, due to varying patient requests throughout the weekday. Wiesche et al. (2017) include both patient and provider preferences in their studies. The measure to address patient preference is number of patients with same-day access and provider preference is time slot utilization.

These papers provide significant insights for tactical capacity planning in primary care through taking into account only same-day and prescheduled appointment types with the main goal of increasing number of reserved time slots for same-day requests. However, prescheduled appointment types include a broad spectrum of conditions such as chronic disease (i.e. diabetes, hypertension), mental health and pre-natal, which have different acuity levels and access time targets. In this study, we propose tactical capacity planning for both urgent and multiple types of non-urgent appointments. Moreover, in previous studies, the demand for non-urgent appointments is daily based and should be scheduled within a week (Balasubramanian et al., 2014; Wiesche et al., 2017). However in this study, the demand for non-urgent appointments should be scheduled within 2-12 weeks. Therefore, we consider a 12-week planning horizon. Due to the presence of longer planning horizon, we provide a weekly number of time slots that should be reserved for each appointment type.

3.2.4 Physician Preferences in Primary Care

According to the existing literature, physician preferences in primary care are mainly determined based on the two factors of physician revenue and utilization (Qu et al., 2012; Wiesche et al., 2017). Based on our knowledge, there are only two studies with the focus on

physician preferences in developing TCP models in primary care. Current studies used two main metrics to evaluate physician preferences which are the average number of patients consulted in each session, as well as the variation in the number of consulted patients between clinic sessions (Qu et al., 2012; Wiesche et al., 2017). Variation in the number of consulted patients results in an unbalanced workload among weeks and leads to provider overtime for weeks with high workload and provider idle time for weeks with low workload (Qu et al., 2012). Physician preferences should be evaluated in association with patient needs (Wiesche et al., 2017). For example, considering only physician preferences for high utilization will lead to consulting a large number of patients per day that provides higher revenue. However, it might result in longer patient access as well as insufficient capacity for urgent requests (Wiesche et al., 2017).

In this study, we consider variation in the number of consulted patients as a metric for physician preferences. However, consideration of physician preference in association with patient needs in this study is different from previous studies since patient needs are defined based on five acuity levels. In particular, access targets for appointments with acuity levels 1 and 2 are less than one week; thus, we aim to evenly distribute the appointments of acuity levels 1 or 2 among the weeks of the planning horizon. On the other hand, access targets for appointments with acuity level 3, 4 and 5 exceed one week. Therefore, the flexibility in access time of these appointments may lead to variability in the scheduled appointments among weeks. Apart from flexibility, appointments with lower acuity place a heavy burden on a physician's workload since acuity and physician workload do not have a direct relationship. Low acuity appointments are usually scheduled for conditions that are chronic in nature; for such conditions, to make a diagnosis, prescription or recommendation, the physician has to spend a longer time reviewing a patient's medical history, as compared to an acute condition. For example, for diabetes, the physician needs to review what the patient has experienced over three months and elicit relevant information from it, versus in the case of, e.g., a patient's appointment due to a common cold, when the patient's experienced symptoms have been

going on for days (most likely up to a maximum of one week), even if the patient has some chronic condition in addition to the cold. From a physician's perspective, appointments with lower acuity take more energy so they prefer to have a limit on the number of such appointments per week. Provided with such a limit, a physician schedule will have an even distribution of appointments with lower acuity in between weeks which will improve his/her workload balance. The advantage of having a balanced physician workload from the perspective of the clinic is its consistency due to the same total number of appointments per week which makes operational level planning more convenient.

3.3 Problem Description

A capacity reservation plan determines the number of appointment slots that should be reserved per appointment type in each week for a specific physician (Ahmadi-Javid et al., 2017). Figure 3.1 depicts the schematic view for the basic structure of capacity reservation plan in primary care. The filled circles represent the weeks. We define the first S weeks as the *arrival horizon*, and we define an index set $\mathcal{S} = \{1, \dots, S\}$. It is assumed that $D_{p,t}$ requests for appointment type p arrive at the start of each week t in the *arrival horizon* $\mathcal{S}$. In this study we consider multiple appointment types and each appointment type may have single or multiple acuity levels. Therefore, we define the index set $\mathcal{K} = \{1, \dots, 5\}$ for the defined acuity levels in primary care from our prior study (Aslani et al., 2019). We introduce the index set $\mathcal{P}_\kappa$ for all appointment types with acuity level κ. Let the set $\mathcal{P}$ include all appointment types with all acuity levels which means $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2 \cup \mathcal{P}_3 \cup \mathcal{P}_4 \cup \mathcal{P}_5$. It should be mentioned that Figure 3.1 does not represent the details regarding appointment types and their relative acuity levels. A desired level of demand for each appointment type should be met and scheduled in T weeks. Due to the variation in access time of patients with different acuity levels, we defined the term *planning horizon* which is an extension of *arrival horizon* and we introduce an index set $\mathcal{H} = \{1, \dots, T\}$. All the given weeks in the *planning horizon* has the fixed physician time availability of $\{\phi_1, \dots, \phi_T\}$. Our goal in this study is to develop

a *capacity reservation plan* for the demand of appointment type with acuity levels from the defined set $\mathcal{P}_\kappa$ for a specific family physician to address two general goals. One objective is minimizing the imbalance in the physician workload between weeks based on acuity level of appointment. The second objective is prioritizing requested appointment in such a way to provide equitable timely access to care based on acuity level of appointments.

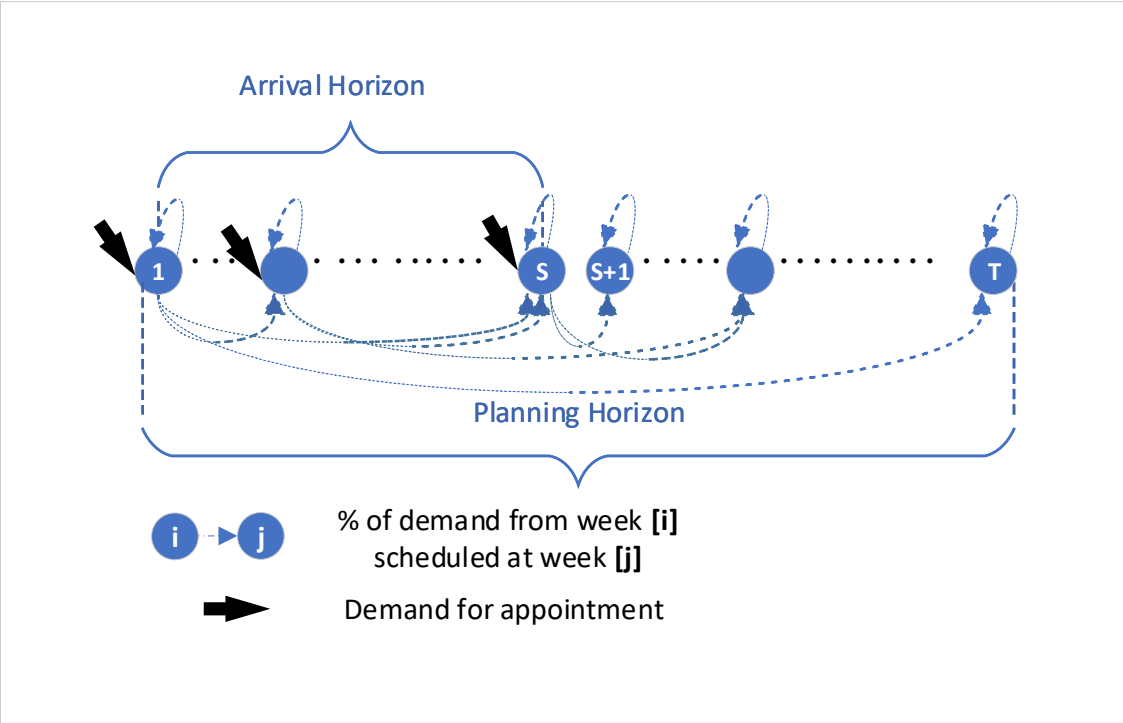


Figure 3.1: Schematic view of arrival and planning horizon, adapted from Figure 3 by (Nguyen et al., 2015)

3.4 Overall Approach

An optimal capacity reservation plan takes into account both the needs of patients and the preferences of primary care providers. We present an optimization model for capacity reservation for both urgent and non-urgent appointments in primary care to deliver care in a timely manner. The novelty of this study is developing an efficient linear optimization approach that tackles the main concern in the Canadian primary care clinics: developing a tactical capacity plan to achieve equitable access based on the acuity level of patient needs

in such a way that physician time is utilized efficiently. The optimization model distributes the reserved appointment slots in a manner to balance physician workload between weeks of a planning horizon. Therefore, appointment slots for patients with lower acuity level are reserved on weeks when the demand high-acuity appointments is lower.

To be able to reserve appointment slots in advance, we should find the realized weekly demand, $D_{p,t}$ for each appointment type. Therefore, we develop an individual forecast model for each appointment type to predict the demand. Considering the demand forecast provides a data-driven optimal capacity reservation plan which is more accurate than considering a known demand value due to capturing variability. To ensure of obtaining equitable access time, we incorporate the access targets developed by Aslani et al. (2019) (see Chapter 2) as access time constraints for each appointment type. Since family physicians at HFA clinic prefer to take care of their own patients (rosters, panel), we focus on a single family physician. In the following we introduce the basic structure we consider to develop a tactical capacity plan. Thereafter, we elaborate the steps to forecast demand and finally we describe how we develop a model to find an optimal capacity reservation plan.

3.5 Forecasting Model

We apply time series analysis to forecast demand for all appointment types. To do so, we first pre-process data for all appointment types to get weekly demand value for each appointment type. Thereafter, we test time series stationarity. Finally, we fit multiple forecast models and select the best model based on various goodness-of-fit metrics. By pre-processing, we mean deleting the extra information that we do not consider in our model, finding and removing the outliers, as well as reformatting the required data in such a way as to get weekly demand value for each appointment type.

To build a forecast model, the time series should be stationary which means the mean, variance and covariance of the series should not be a function of time (Box et al., 2015; Abraham and Ledolter, 2009). To evaluate stationarity of time series, we apply augmented

Dickey Fuller test (ADF) which is a statistical test with null hypothesis of "the tested time series is not stationary". Thereafter, we propose multiple forecast models to the stationary time series based on ARIMA or seasonal ARIMA analysis method (Holt, 1957; Winters, 1960). ARIMA model is described by the three factors of (p, d, q) where 'p' represents order of the autoregressive model; 'd' represents the order of difference; and 'q' represents the order of moving average model. Seasonal ARIMA model is explained by (p, d, q)(P, D, Q)[S] in which (P, D, Q) represents the seasonal part and S represents the length of the seasonality. In this study we compare goodness-of-fit based on Akaike information criterion (AIC). AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model. In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model. In other words, AIC deals with both the risk of overfitting and the risk of underfitting. In addition, we compared accuracy of models based on Mean Absolute Error (MAE). MAE is applied in terms of comparing forecast models based on single time series.

3.6 Optimization Models

This section develops a modeling approach to determine the optimal reserved capacity for each appointment type in the determined planning horizon. The challenges of fixed availability of a family physician as well as equitable access time for multiple types of urgent and non-urgent appointment types are addressed. The details of the optimization model vary for the designed objective based on the complexity of the addressed challenges. In the following we provide three variations of optimal capacity allocation model: 1) optimal capacity reservation to balance workload, 2) optimal capacity reservation to provide balanced workload and equitable access, and 3) optimal capacity reservation to provide balanced workload and equitable access distribution. In particular, model 1 allocates physician time to different appointment types in a way to minimize the disbalance in the physician workload of

family physician between weeks based on acuity level of appointment. In model 2, along with providing physician workload balance, we also ensure equitable timely access through incorporating the obtained access targets from our prior study (Aslani et al., 2019) (Chapter 2) as constraints for access time. In model 3, in addition to the features in model 2, more urgent requests are prioritized to have lower access time. Therefore, model 3 provides both equitable timely access to care which is based on acuity level of requested appointment type individually as well as equitable distribution of access time which is based on prioritizing appointment types by comparing their acuity levels.

3.6.1 Model 1: Optimal Capacity Reservation Plan to Balance Workload

Due to the fixed weekly availability of family physicians in the clinic and from our experience through collaborating with HFA clinic, family physicians prefer their workload be as consistent as possible between weeks. Inconsistent distribution of workload among weeks may lead to family physician overtime for the weeks with high workload and provider idle time for the weeks with low workload. Therefore, in this study we define the term "workload balance" as a representation of the family physician preferences. We say that a workload is *fully balanced* when the total number of planned appointments for the physician is the same among weeks of the planning horizon. We measure the weekly workload of a family physician by calculating the total number of reserved appointment slots to serve demand for appointment types with different acuity levels. To this end, we define a decision variable $x_{p,j}$ as the number of time slots reserved for appointment type p in week j.

Let α_p be a desired level of demand for each appointment type p that should be met and scheduled in T weeks, i.e., we want to ensure $\sum_{j\in\mathcal{H}} x_{p,j} \geq \alpha_p \sum_{t\in\mathcal{H}} D_{p,t}$. We define the number of unmet requests for appointment type p at the start of week j as $w_{p,j}$, so that $\sum_{j=t}^{T} x_{p,j} + w_{p,t} = D_{p,t}$. Due to the fact that the required number of time slots to serve demand varies based on the requested appointment type, we introduce the parameter τ_p to link the required number of appointment slots to the required number of 15-minute

time slots. Let τ_p be the required number of time slots to serve appointment type p. As an example, the *Periodic health exam* appointment type in HFA requires two 15-minute time slots ($\tau_{periodic-health-exam} = 2$) versus the *Well baby* appointment at HFA requires one 15-minute time slot ($\tau_{Well-baby} = 1$). To match weekly family physician availability to the total number of reserved appointment slots, $\sum_{p \in \mathcal{P}} \tau_p x_{p,j} \leq \phi_j$ has to hold true for all $j \in \mathcal{H}$. We also define γ_p which is a no-show rate for appointment type p similarly to Qu and Shi (2009). Due to the presence of acuity level, "workload balance" should be defined further in detail which is explained in the following.

Dependence of Workload Balance on Acuity Level The definition of workload balance for the appointments that should be scheduled within a same day/week (acuity levels 1 & 2) is different from the appointments with other acuity levels. This is due to the definition of appointments with acuity levels 1 & 2 that requires sufficient number of reserved time slots in each week to avoid any risk to patient health (see Chapter 2). In other words, having an insufficient number of reserved time slots for an appointment in this class cannot be "cancelled out" with excessive reserved time slots of other appointments. Therefore, a fully balanced workload for appointments with acuity level 1 or 2 means the same number of each appointment type is scheduled each week.

On the other hand, a fully balanced workload for appointments with acuity levels 3, 4 and 5 means that the total number of reserved time slots for these appointment types is the same among weeks. In other words, the excessive number of reserved slots for an appointment type can cover for an insufficient number of other appointment types in the same week.

The developed formulation in this study to measure workload is based on calculating the total number of reserved appointment slots to serve demand for appointment types. Thereafter, the value for workload balance is obtained by finding the absolute value of deviation of number of reserved appointment slots between any two weeks in the planning horizon. This absolute value is calculated for any two weeks in the planning horizon. In

the following we explain the difference between workload measurement for different acuity levels.

Formulation of measuring workload for appointments with acuity level 3 & 4 & 5

In case of having appointments with acuity levels 3 & 4 & 5, we calculate the absolute value for deviation of number of reserved appointment slots for all appointment types between any two weeks in the planning horizon because there is no need to have a consistent number of weekly reserved appointment slots for each appointment type with acuity levels 3 & 4 & 5. The following equation represents the difference between the total number of planned appointment slots with acuity levels 3 & 4 & 5 for a physician among weeks of the planning horizon:

$$\sum_{j=1}^{T-1} \sum_{j'=j+1}^{T} \mid \sum_{p \in \mathcal{P}_3 \cup \mathcal{P}_4 \cup \mathcal{P}_5} (1-\gamma_p)\tau_p(x_{p,j} - x_{p,j'}) \mid \tag{3.1}$$

As it is seen, Equation (3.1) is nonlinear. Therefore, we introduce two new variables to make it linear. One variable is $u_{j,j'}$ for the cases that deviation of all the reserved appointment slots for all the appointment types between weeks j and $j^{'}$ is positive. The other variable is $o_{j,j'}$ for the cases that the deviation of all the reserved appointment slots for all the appointment types between weeks j and $j^{'}$ is negative.

$$u_{j,j'} = \begin{cases} \sum_{p \in \mathcal{P}_3 \cup \mathcal{P}_4 \cup \mathcal{P}_5} (1-\gamma_p)\tau_p(x_{p,j} - x_{p,j'}), & \text{if } \sum_{p \in \mathcal{P}_3 \cup \mathcal{P}_4 \cup \mathcal{P}_5} (1-\gamma_p)\tau_p(x_{p,j} - x_{p,j'}) \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

$$o_{j,j'} = \begin{cases} -\sum_{p \in \mathcal{P}_3 \cup \mathcal{P}_4 \cup \mathcal{P}_5} (1-\gamma_p)\tau_p(x_{p,j} - x_{p,j'}), & \text{if } \sum_{p \in \mathcal{P}_3 \cup \mathcal{P}_4 \cup \mathcal{P}_5} (1-\gamma_p)\tau_p(x_{p,j} - x_{p,j'}) \leq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

As a result, the linear format of workload measure for appointments with acuity levels 3 & 4 & 5 is as follows:

$$\sum_{j=1}^{T-1} \sum_{j^{'}=j+1}^{T} (u_{j,j^{'}} + o_{j,j^{'}}). \tag{3.4}$$

Formulation of measuring workload for appointments with acuity level 1 & 2

In case of having appointments with acuity levels 1 & 2, we calculate the absolute value for deviation of number of reserved appointment slots for each appointment type in this category individually between any two weeks in the planning horizon because it is needed to have consistent number of weekly reserved appointment slots for each appointment type with acuity levels 1 & 2 due to their urgency. The following equation represents the difference between total number of planned appointment slots with acuity levels 3 & 2 for a physician among weeks of the planning horizon:

$$\sum_{j=1}^{T-1} \sum_{j^{'}=j+1}^{T} \sum_{p \in \mathcal{P}_1 \cup \mathcal{P}_2} \mid (1-\gamma_p)\tau_p(x_{p,j} - x_{p,j^{'}}) \mid . \tag{3.5}$$

Equation (3.5) is nonlinear. Therefore, we introduce two new variables to make it linear. One variable is $v^{p}_{j,j^{'}}$ for the cases that deviation of all the reserved appointment slots for appointment type p between weeks j and $j^{'}$ is positive. The other variable is $l^{p}_{j,j^{'}}$ for the cases that the deviation of all the reserved appointment slots for appointment type p between weeks j and $j^{'}$ is negative.

$$v^{p}_{j,j^{'}} = \begin{cases} (1-\gamma_p)\tau_p(x_{p,j} - x_{p,j^{'}}), & \text{if } (1-\gamma_p)\tau_p(x_{p,j} - x_{p,j^{'}}) \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

$$l^{p}_{j,j^{'}} = \begin{cases} -(1-\gamma_p)\tau_p(x_{p,j} - x_{p,j^{'}}), & \text{if } (1-\gamma_p)\tau_p(x_{p,j} - x_{p,j^{'}}) \leq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.7}$$

As a result, the linear format of workload measure for appointments with acuity levels 1 & 2 is as follows:

$$\sum_{j=1}^{T-1} \sum_{j^{'}=j+1}^{T} \sum_{p \in \mathcal{P}_1 \cup \mathcal{P}_2} (v^{p}_{j,j^{'}} + l^{p}_{j,j^{'}}) \tag{3.8}$$

Therefore, the idea of allocating physician time to different appointment types (or reserving appointment slots for different appointment types) to minimize unbalanced workload of the family physician is modeled as below:

$$Min \quad \sum_{j=1}^{S-1} \sum_{j^{'}=j+1}^{S} \sum_{p \in \mathcal{P}_1 \cup \mathcal{P}_2} (v^{p}_{j,j^{'}} + l^{p}_{j,j^{'}}) + \sum_{j=1}^{T-1} \sum_{j^{'}=j+1}^{T} (u_{j,j^{'}} + o_{j,j^{'}}) \tag{3.9}$$

$$v^{p}_{j,j^{'}} - (1-\gamma_p)\tau_p(x_{p,j} - x_{p,j^{'}}) \geq 0 \quad \forall p \in \mathcal{P}_1 \cup \mathcal{P}_2 \quad \forall j, j^{'} \in \mathcal{H} \tag{3.10}$$

$$l^{p}_{j,j^{'}} + (1-\gamma_p)\tau_p(x_{p,j} - x_{p,j^{'}}) \geq 0 \quad \forall p \in \mathcal{P}_1 \cup \mathcal{P}_2 \quad \forall j, j^{'} \in \mathcal{H} \tag{3.11}$$

$$u_{j,j^{'}} - \sum_{p \in \mathcal{P}_3 \cup \mathcal{P}_4 \cup \mathcal{P}_5} (1-\gamma_p)\tau_p(x_{p,j} - x_{p,j^{'}}) \geq 0 \quad \forall j, j^{'} \in \mathcal{H} \tag{3.12}$$

$$o_{j,j^{'}} + \sum_{p \in \mathcal{P}_3 \cup \mathcal{P}_4 \cup \mathcal{P}_5} (1-\gamma_p)\tau_p(x_{p,j} - x_{p,j^{'}}) \geq 0 \quad \forall j, j^{'} \in \mathcal{H} \tag{3.13}$$

$$\sum_{j \in \mathcal{S}} x_{p,j} \geq \alpha_p \sum_{t \in S} D_{p,t} \quad \forall p \in \mathcal{P}_1 \cup \mathcal{P}_2 \tag{3.14}$$

$$\sum_{j \in \mathcal{H}} x_{p,j} \geq \alpha_p \sum_{t \in S} D_{p,t} \quad \forall p \in \mathcal{P}_3 \cup \mathcal{P}_4 \cup \mathcal{P}_5 \tag{3.15}$$

$$\sum_{j \in \mathcal{S}} x_{p,j} = \sum_{t \in \mathcal{S}} (D_{p,t} - w_{p,t}) \quad \forall p \in \mathcal{P}_1 \cup \mathcal{P}_2 \tag{3.16}$$

$$\sum_{j \in \mathcal{H}} x_{p,j} = \sum_{t \in \mathcal{S}} (D_{p,t} - w_{p,t}) \quad \forall p \in \mathcal{P}_3 \cup \mathcal{P}_4 \cup \mathcal{P}_5 \tag{3.17}$$

$$\sum_{j=t}^{T} x_{p,j} + w_{p,t} \geq D_{p,t} \quad \forall p \in \mathcal{P} \quad \forall t \in \mathcal{H} \tag{3.18}$$

$$\sum_{p \in \mathcal{P}} \tau_p x_{p,j} \leq \phi_j \quad \forall j \in \mathcal{H} \tag{3.19}$$

$$x_{p,j}, w_{p,j}, v^{p}_{j,j'}, l^{p}_{j,j'} \in \mathbb{Z} \quad \forall p \in \mathcal{P} \quad j, j^{'} \in \mathcal{H} \quad \forall n \in \mathcal{N} \tag{3.20}$$

$$u_{j,j'}, o_{j,j'} \geq 0 \quad \forall \quad j, j^{'} \in \mathcal{H} \quad \forall n \in \mathcal{N} \tag{3.21}$$

The designed model allocates physician time to different appointment types in a way to minimize the disbalance in the physician workload of family physician between weeks (3.9). Constraint (3.10) calculated positive deviation of total reserved appointment slots with any access time between any two weeks for a specific appointment type in class of acuity level 1 & 2. Constraint (3.11) calculated negative deviation of total reserved appointment slots with any access time between any two weeks for a specific appointment type in class of acuity level 1 & 2. Constraint (3.12) calculated positive deviation of total reserved appointment slots with any access time for all the appointment types in class of acuity level 3 & 4 & 5 between any two weeks. Constraint (3.13) calculated negative deviation of total reserved appointment slots with any access time for all the appointment types in class of acuity level 3 & 4 & 5 between any two weeks.

Constraints (3.14) and (3.17) ensure that the desired service level, α_p, of serving patients with different acuity within their access target is achieved. Constraints (3.16) - (3.18) track the unmet demand for appointment types with different acuity levels in each week. To match weekly family physician availability to total number of reserved appointment slot equation (3.19) has to hold true. The domains of the decision variables are described in (3.20) and (3.21).

The results support the decision maker to know for a specific family physician how many patients with a specific appointment type they allow to schedule each week while the physician preference is met. Through running the model, we get poor results of workload balance (shown and discussed further in Section 3.7.3.1). This is due to the fact that the model does not have any information regarding the access time for each appointment type. The former decision variable $x_{p,j}$ is not suitable when it comes to reflecting patient influence on the planning decision. In addition, the model does not consider any priority to allocate

appointment slots to demand. However, as it is mentioned in Aslani et al. (2019) (Chapter 2), allocation of physician time should be based on the acuity of requested appointment. In other words, allocation of physician time to demand should provide equitable access. In the next section, we modify the proposed basic model in Equations (3.9)-(3.21) in such a way that we ensure the reserved time slot realized after arrival of demand request. Therefore, model 1 is considered as a relaxed version of the developed model in the next section (model 2). As a result, model 1 provides a lower bound for optimal physician workload balance in model 2 because model 2 also address equitable timely access to care.

3.6.2 Model 2: Optimal Capacity Reservation Plan to Provide Balanced Workload & Equitable Access

Ministry of Health and Long Term Care in Ontario (MOHLTC) requires primary care clinics to maintain better control over access time for appointments and report performance of the clinic. As a result, allocation of physician time to demand should also consider access time. Therefore, we introduce a decision variable $x^n_{p,j}$, which is the number of reserved appointment slots to serve appointment type p in week j with access time of n weeks. As a result the formulation of workload balance (3.9) should be updated based on the new defined decision variable $x^n_{p,j}$.

Updated formulation of measuring workload for appointments with acuity level 3 & 4 & 5

The following equation represents the difference between total number of planned appointment slots with acuity levels 3 & 4 & 5 for a physician among weeks of the planning horizon:

$$\sum_{j=1}^{T-1} \sum_{j'=j+1}^{T} \Big| \sum_{n\in\mathcal{N}} \sum_{p\in\mathcal{P}_3\cup\mathcal{P}_4\cup\mathcal{P}_5} (1-\gamma_p)\tau_p(x^n_{p,j} - x^n_{p,j'}) \Big| . \tag{3.22}$$

To make Equation (3.22) linear we will take the same process as the previous section. Therefore, we introduce two new variables to make it linear. One variable is $u_{j,j'}$ for the cases that

deviation of all the reserved appointment slots for all the appointment types between weeks j and $j^{'}$ is positive. The other variable is $o_{j,j'}$ for the cases that the deviation of all the reserved appointment slots for all the appointment types between weeks j and $j^{'}$ is negative.

$$u^*_{j,j'} = \begin{cases} \sum_{p\in\mathcal{P}_3\cup\mathcal{P}_4\cup\mathcal{P}_5}(1-\gamma_p)\tau_p(\sum_{n=0}^{j-1} x^n_{p,j} - \sum_{n=0}^{j'-1} x^n_{p,j'}), & \text{if } \sum_{n\in\mathcal{N}}\sum_{p\in\mathcal{P}_3\cup\mathcal{P}_4\cup\mathcal{P}_5} \\ & (1-\gamma_p)\tau_p(x^n_{p,j} - x^n_{p,j'}) \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.23}$$

$$o^*_{j,j'} = \begin{cases} -\sum_{p\in\mathcal{P}_3\cup\mathcal{P}_4\cup\mathcal{P}_5}(1-\gamma_p)\tau_p(\sum_{n=0}^{j-1} x^n_{p,j} - \sum_{n=0}^{j'-1} x^n_{p,j'}), & \text{if } \sum_{n\in\mathcal{N}}\sum_{p\in\mathcal{P}_3\cup\mathcal{P}_4\cup\mathcal{P}_5} \\ & (1-\gamma_p)\tau_p(x^n_{p,j} - x^n_{p,j'}) \leq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.24}$$

As a result, the linear format of workload measure for appointments with acuity levels 3 & 4 & 5 is as follows:

$$\sum_{j=1}^{T-1}\sum_{j'=j+1}^{T}(u^*_{j,j'} + o^*_{j,j'}) \tag{3.25}$$

Formulation of measuring workload for appointments with acuity level 1 & 2

In case of having appointments with acuity levels 1 & 2, we calculate the absolute value for deviation of number of reserved appointment slots for each appointment type in this category individually between any two weeks in the planning horizon because it is needed to have consistent number of weekly reserved appointment slots for each appointment type with acuity levels 1 & 2.

$$\sum_{n\in\mathcal{N}}\sum_{j=1}^{T-1}\sum_{j'=j+1}^{T}\sum_{p\in\mathcal{P}_1\cup\mathcal{P}_2} \mid (1-\gamma_p)\tau_p(x^n_{p,j} - x^n_{p,j'}) \mid \tag{3.26}$$

As it is seen, Equation (3.26) is non-linear. Therefore, we introduce two new variables to make it linaer. One variable is $v^{p}_{j,j^{'}}$ for the cases that deviation of all the reserved appointment slots for appointment type p between weeks j and $j^{'}$ is positive. The other variable is $l^{p}_{j,j^{'}}$ for the cases that the deviation of all the reserved appointment slots for appointment type p between weeks j and $j^{'}$ is negative.

$$v^{p}_{j,j^{'}} = \begin{cases} (1-\gamma_p)\tau_p(\sum_{n=0}^{j-1} x^{n}_{p,j} - \sum_{n=0}^{j^{'}-1} x^{n}_{p,j^{'}}), & \text{if } \sum_{n\in\mathcal{N}}(1-\gamma_p)\tau_p(x^{n}_{p,j} - x^{n}_{p,j^{'}}) \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.27}$$

$$l^{p*}_{j,j^{'}} = \begin{cases} -(1-\gamma_p)\tau_p(\sum_{n=0}^{j-1} x^{n}_{p,j} - \sum_{n=0}^{j^{'}-1} x^{n}_{p,j^{'}}), & \text{if } \sum_{n\in\mathcal{N}}(1-\gamma_p)\tau_p(x^{n}_{p,j} - x^{n}_{p,j^{'}}) \leq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.28}$$

As a result, the linear format of workload measure for appointments with acuity levels 1 & 2 is as follows:

$$\sum_{j=1}^{T-1} \sum_{j^{'}=j+1}^{T} \sum_{p\in\mathcal{P}_1\cup\mathcal{P}_2} (v^{p}_{j,j^{'}} + l^{p}_{j,j^{'}}) \tag{3.29}$$

In addition access time should be equitable to match the acuity of the requested appointment. To ensure equitable allocation of physician time to demand for appointment types, we incorporate the obtained access targets from our prior study Aslani et al. (2019) (Chapter 2) as constraints for access time. Due to the presence of multiple access targets for some appointment types, we define an access rule through considering multiple access constraints for each appointment type which could control the distribution of access within the relative access targets for each appointment type. To this end, we introduce $\alpha_{p,k}$ as a desired level of demand for appointment type p with acuity level κ. Therefore, we included access time into the capacity reservation model through incorporating access time into the decision variable ($x^{n}_{p,j}$) and service level ($\alpha_{p,k}$). Therefore, the developed model to address the trade-off

between the physician workload balance and patient equitable access time is as follows:

$$Min \quad (\sum_{j=1}^{S-1}\sum_{j^{'}=j+1}^{S}\sum_{p\in\mathcal{P}_1\cup\mathcal{P}_2}(v^{p}_{j,j^{'}}+l^{p}_{j,j^{'}})+(\sum_{j=1}^{T-1}\sum_{j^{'}=j+1}^{T}u^{*}_{j,j^{'}}+o^{*}_{j,j^{'}}) \tag{3.30}$$

$$v^{p}_{j,j^{'}}-(1-\gamma_p)\tau_p(\sum_{n=0}^{j-1}x^{n}_{p,j}-\sum_{n=0}^{j^{'}-1}x^{n}_{p,j^{'}})\geq 0 \quad \forall p\in\mathcal{P}_1\cup\mathcal{P}_2 \quad \forall j,j^{'}\in\mathcal{H} \tag{3.31}$$

$$l^{p}_{j,j^{'}}+(1-\gamma_p)\tau_p(\sum_{n=0}^{j-1}x^{n}_{p,j}-\sum_{n=0}^{j^{'}-1}x^{n}_{p,j^{'}})\geq 0 \quad \forall p\in\mathcal{P}_1\cup\mathcal{P}_2 \quad \forall j,j^{'}\in\mathcal{H} \tag{3.32}$$

$$u^{*}_{j,j^{'}}-\sum_{p\in\mathcal{P}_3\cup\mathcal{P}_4\cup\mathcal{P}_5}(1-\gamma_p)\tau_p(\sum_{n=0}^{j-1}x^{n}_{p,j}-\sum_{n=0}^{j^{'}-1}x^{n}_{p,j^{'}})\geq 0 \quad \forall j,j^{'}\in\mathcal{H} \tag{3.33}$$

$$o^{*}_{j,j^{'}}+\sum_{p\in\mathcal{P}_3\cup\mathcal{P}_4\cup\mathcal{P}_5}(1-\gamma_p)\tau_p(\sum_{n=0}^{j-1}x^{n}_{p,j}-\sum_{n=0}^{j^{'}-1}x^{n}_{p,j^{'}})\geq 0 \quad \forall j,j^{'}\in\mathcal{H} \tag{3.34}$$

$$\sum_{j\in\mathcal{S}}\sum_{n=0}^{a}x^{n}_{p,j}\geq\alpha_{p,k}\sum_{t\in\mathcal{S}}D_{p,t} \quad \forall p\in\mathcal{P}_1\cup\mathcal{P}_2 \quad \forall k\in\mathcal{K} \quad \forall a\in\mathcal{N} \tag{3.35}$$

$$\sum_{j\in\mathcal{H}}\sum_{n=0}^{a}x^{n}_{p,j}\geq\alpha_{p,k}\sum_{t\in\mathcal{S}}D_{p,t} \quad \forall p\in\mathcal{P}_3\cup\mathcal{P}_4\cup\mathcal{P}_5 \quad \forall k\in\mathcal{K} \quad \forall a\in\mathcal{N} \tag{3.36}$$

$$\sum_{j=t}^{S}x^{j-t}_{p,j}+w_{p,t}=D_{p,t} \quad \forall p\in\mathcal{P}_1\cup\mathcal{P}_2 \quad \forall t\in\mathcal{S} \tag{3.37}$$

$$\sum_{j=t}^{T}x^{j-t}_{p,j}+w_{p,t}=D_{p,t} \quad \forall p\in\mathcal{P}_3\cup\mathcal{P}_4\cup\mathcal{P}_5 \quad \forall t\in\mathcal{S} \tag{3.38}$$

$$\sum_{n\in\mathcal{N}}\sum_{p\in\mathcal{P}}\tau_p x^{n}_{p,j}\leq\phi_j \quad \forall j\in\mathcal{H} \tag{3.39}$$

$$x^{n}_{p,j},w_{p,j},v^{p}_{j,j^{'}},l^{p}_{j,j^{'}}\in\mathbb{Z} \quad \forall p\in\mathcal{P} \quad j,j^{'}\in\mathcal{H} \quad \forall n\in\mathcal{N} \tag{3.40}$$

$$u^{*}_{j,j^{'}},o^{*}_{j,j^{'}}\geq 0 \quad \forall \quad j,j^{'}\in\mathcal{H} \quad \forall n\in\mathcal{N} \tag{3.41}$$

Objective function (3.30) minimizes the deviation of total reserved appointment slots between weeks to minimize unbalanced workload for a specific physician. The deviation is

measured differently for acuity levels of 1 & 2 compared to the other acuity levels. The deviation is calculated individually for all appointment types with acuity levels of 1 and 2. However, for the appointment types with acuity levels of 3, 4 & 5, the deviation includes all the appointment types. Constraint (3.31) calculated positive deviation of total reserved appointment slots with any access time between any two weeks for a specific appointment type in class of acuity level 1 & 2. Constraint (3.32) calculated negative deviation of total reserved appointment slots with any access time between any two weeks for a specific appointment type in class of acuity level 1 & 2. Constraint (3.33) calculated positive deviation of total reserved appointment slots with any access time for all the appointment types in class of acuity level 3 & 4 & 5 between any two weeks. Constraint (3.34) calculated negative deviation of total reserved appointment slots with any access time for all the appointment types in class of acuity level 3 & 4 & 5 between any two weeks. Constraints (3.35) and (3.36) ensure that the desired service level, $\alpha_{p,\kappa}$, of serving patients with different acuity within their access target is achieved. Constraints (3.37) and (3.38) ensure appointments are scheduled after demand arrival and unmet demand for appointment types with different acuity levels in each week are tracked. Constraint (3.39) ensures physician availability in each week matches total reserved appointment slots ($\tau_p x^n_{p,j}$) for all appointment types. Constraints (3.40) and (3.41) describe the domains for the decision variables.

The results of the proposed model in Equations (3.30)-(3.41) provides the reserved capacity for each appointment type per week for a specific physician while both physician and patient preferences are met. Thus, the output of the model in Equations (3.30)-(3.41) supports the decision makers to make complex trade-offs between meeting care accessibility targets for patients and balancing physicians' work load between weeks. In this model, equitable access is addressed through planning appointment within the determined access target relative to the acuity of appointment. Therefore, the model in Equations (3.30)-(3.41) ensures patient access time does not exceed their acuity level. However, there can exist multiple equivalent solutions that meet all access targets and therefore Model 2 does not achieve

full prioritization of appointments based on urgency. For example, if there exists an available same-week time slot, the model might assign this slot to an acuity 4 appointment even though there may be acuity 2 and 3 demand, as long as the access targets are met.

3.6.3 Model 3: Optimal Capacity Reservation Plan to Provide Balanced Workload & Equitable Access Distribution

To have equitable distribution of access time, we minimize the weighted maximum number of total reserved appointment slots with same week, next week and two week access for acuity 5. Therefore, we introduce three new decision variables of z_p^0, z_p^1 and z_p^2 which are defined as the maximum number of total reserved appointment slots to serve appointments with acuity 5 with same week, next week and two weeks access time. The new model is multi-objective. We define ω_1 as the weight for minimizing unbalanced workload and ω_2, ω_3 and ω_4 as the weights for minimizing inequitable distribution of access time. In particular, ω_2 is the weight relative to minimizing same-week access allocation to appointments with acuity level 5; ω_3 is the weight relative to minimizing next week access allocation to appointments with acuity level 5; and ω_4 is the weight relative to minimizing two week access allocation to appointments with acuity level 5. The updated model is as follows:

$$\begin{aligned} Min \quad \omega_1 \Big(\sum_{j=1}^{S-1} \sum_{j'=j+1}^{S} \sum_{p \in \mathcal{P}_1 \cup \mathcal{P}_2} (v_{j,j'}^p + l_{j,j'}^p + \sum_{j=1}^{T-1} \sum_{j'=j+1}^{T} u_{j,j'}^* + o_{j,j'}^*) + \\ \omega_2 \sum_{p \in \mathcal{P}_5} z_p^0 + \omega_3 \sum_{p \in \mathcal{P}_5} z_p^1 + \omega_4 \sum_{p \in \mathcal{P}_5} z_p^2 \end{aligned} \tag{3.42}$$

$$\text{s.t.} \qquad (3.31) - (3.34) \tag{3.43}$$

$$\sum_{j \in \mathcal{S}} x_{p,j}^0 \leq z_p^0 \quad \forall p \in \mathcal{P}_5 \tag{3.44}$$

$$\sum_{j \in \mathcal{S}} x_{p,j}^1 \leq z_p^1 \quad \forall p \in \mathcal{P}_5 \tag{3.45}$$

$$\sum_{j \in \mathcal{S}} x_{p,j}^2 \leq z_p^2 \quad \forall p \in \mathcal{P}_5 \tag{3.46}$$

$$\text{s.t.} \qquad (3.35) - (3.39) \tag{3.47}$$

$$x_{p,j}^n, w_{p,j}, v_{j,j'}^p, l_{j,j'}^p \in \mathbb{Z} \quad \forall p \in \mathcal{P} \quad j, j' \in \mathcal{H} \quad \forall n \in \mathcal{N} \tag{3.48}$$

$$u_{j,j'}^*, o_{j,j'}^* \geq 0 \quad \forall \quad j, j' \in \mathcal{H} \quad \forall n \in \mathcal{N} \tag{3.49}$$

Objective function (3.42) include two parts. One part aims to minimize unbalanced physician workload which is similar to the objective function of the proposed model in section 3.6.2 and the other part focus on minimizing inequitable distribution of access time. Constraints (3.31)-(3.34) which are obtained from the proposed model in section 3.6.2 calculate positive and negative deviations of total reserved appointment slots with any access time between any two weeks in the planning horizon for appointment with different acuity levels. Constraint (3.44) finds the total number of reserved appointment slots with same week access time for appointment with acuity level 5. Constraint (3.45) determines the total number of reserved appointment slots with next week access time for appointment with acuity level 5. Constraint (3.46) shows the total number of reserved appointment slots with two weeks access time for appointment with acuity level 5. Constraints (3.31)-(3.34) are also obtained from the proposed model in section 3.6.2. Constraints (3.35) and (3.36) ensure that the desired service level, $\alpha_{p,\kappa}$, of serving patients with different acuity within their access target is achieved. Constraint (3.37) and (3.38) ensure appointments are scheduled after demand arrival and unmet demand for any appointment type in each week is tracked. Constraint (3.39) ensures physician availability in each week matches total reserved appointment slots ($\tau_p x_{p,j}^n$) for all appointment types. Constraints (3.48) and (3.49) describe the domains for the decision variables.

3.7 Experimental Results

We evaluate the above models in the context of the Health for All family health team (further referred to as HFA) located in Markham, Ontario, Canada. HFA is a teaching unit affiliated with the University of Toronto's Department of Family and Community Medicine. The team of family physicians at HFA includes 12 faculty and 24 residents. Planning and scheduling decisions at HFA are made by five administrators who must make complex trade-offs between meeting care accessibility targets for both urgent and non-urgent patients and balancing physicians' workload between weeks in the absence of planning/scheduling software. The task of planning is even more challenging in the presence of multiple non-urgent appointment types and access time targets for each appointment.

3.7.1 Data

Administrative data was retrieved from the comprehensive data set from HFA for the period from September 2017 until September 2018. The data set contains 60,682 records listing the provider name, booking date, appointment date, appointment type, primary MD, whether the patient was a no-show, appointment detail, scheduled time, duration, arrival time and departure time. In order to prepare the data for analysis, we remove extra records and outliers. The records that we remove from consideration are those with negative access time (i.e., when the time of appointment in the data set was before the time of booking); with no booking or appointment date; with doctor unavailability; home visits; evening and Saturday clinic appointments; and records that were labeled as "deleted", which generally corresponded to an appointment that was rescheduled for later.

In this study, we assume *arrival horizon* is $\mathcal{S} = \{1, \dots, 10\}$ and the *planning horizon* is $\mathcal{H} = \{1, \dots, 12\}$. All the given weeks in the planning horizon have fixed physician time availability of $\{30, 30, 60, 60, 60, 60, 60, 60, 60, 60, 30, 30\}$, for each of the weeks respectively. We intentionally consider lower capacity in the first and last two weeks of the planning horizon to reflect a more realistic, dynamic setting which would include multiple planning

horizons. Through considering lower physician availability in the first two weeks, we assume that we do not have an empty system and some proportion of appointment slots in the first two weeks are already reserved for the requests from the previous planning horizon. In addition, by considering lower physician availability in the last two weeks, we reserve some appointment slots for the requests with acuity levels 1 and 2 in the next arrival horizon.

The number of appointment slots that should be reserved for the next planning horizon is calculated based on the values in Tables 3.4 and 3.7. As presented in Table 3.2 appointment types with acuity levels 1 and 2 are *Blank*, *Follow-up* and *Injection*. Table 3.4 shows that 68% of requests for *Blank* have acuity level 1 and 2, 52% of requests for *Follow-up* have acuity level 2 and 71% of requests for *Injection* have acuity level 2. Moreover, as presented in Table 3.7 (which shows the forecasted demand over the arrival horizon for a particular family physician at HFA) the maximum demand for *Blank* is 31, for *Follow-up* is 14 and for *Injection* is 3. Therefore, we calculate the value of reserved time slots for the next planning horizon as $(0.68 \times 31) + (0.52 \times 14) + (0.71 \times 3) = 30.48$. Based on this approximation, in this study we reserve 30 time slots for the arriving requests in the subsequent planning horizon; however, sensitivity analysis based on changing this value should be conducted in future work. The acuity level of appointments is obtained from the study presented by Aslani et al. (2019) (Chapter 2). Table 3.2 shows the acuity levels for all 11 appointment types at HFA. Table 3.3 describes the corresponding access targets for the defined acuity levels at HFA.

Appointment type	**Acuity level**	**Appointment type**	**Acuity level**
Periodic Health Exam	5	*Injection*	2-5
Child Physical	5	*New patient*	5
Diabetic Management	5	*Pre-Op Assessment*	3,4
Mental Health	3-5	*Pre-Natal*	3,4
Blank	1-5	*Well Baby*	5
Follow-up	2-5		

Table 3.2: Acuity levels per appointment type at HFA.

Acuity level	Description	Should be seen within
1	Same day urgent	same week
2	Same week urgent	1 week
3	Two weeks non-urgent	2 weeks
4	Four weeks non-urgent	4 weeks
5	Routine	12 weeks

Table 3.3: Proposed acuity levels and corresponding access time targets.

Due to the fact that some appointment types have multiple acuity levels, we define $\alpha_{p,k}$ as the proportion of demand for appointment type p with acuity level κ. Therefore, for an appointment type with multiple acuity levels, $\alpha_{p,k}$ defines proportion of demand that should be meet within the access target of acuity level κ. Table 3.4 presents the values we assume for $\alpha_{p,k}$, based on approximation of percentages falling into the respective categories in the data; more accurate calculation of $\alpha_{p,k}$ would require knowing the true conditions of the patients in the dataset.

	Acuity level				
	1	2	3	4	5
Periodic Health Exam	-	-	-	-	93%
Child Physical	-	-	-	-	93%
Diabetic Management	-	-	-	-	93%
Follow-up	-	52%	67%	84%	99%
Mental Health	-	-	48%	85%	99%
Injection	-	71%	79%	90%	99%
New Patient	-	-	-	-	93%
Pre-Op Assessment	-	-	76%	99%	-
Pre-Natal	-	-	52%	99%	-
Well Baby	-	-	-	-	93%
Blank	46%	68%	79%	95%	99%

Table 3.4: Desired level of demand per appointment type per acuity level

Some of the current appointment types include patients of multiple acuity levels – again, due to lack of data on the actual patient conditions, we have to make an assumption. In

particular, we classify the appointments into acuity levels as shown in Table 3.5, with some appointment types corresponding to two acuity levels.

Acuity level	Appointment type
1, 2	*Blank*, *Follow-up*, *Injection*
3, 4	*Mental Health*, *Pre-Op Assessment*, *Pre-Natal*
5	*Well baby*, *New Patient*, *Diabetic Management*, *Periodic Health Exam*, *Child Physical*

Table 3.5: Approximation of acuity levels for appointment types

3.7.2 Forecasting Demand for Appointment Types

As stated in Section 3.5, we use ARIMA modelling for forecasting the demand for various appointment types. Here we first provide an example for how we apply the forecasting approach to the *Blank* appointment type. Implementation and analysis of forecasting models was done using R Studio Version 1.1.456 (RStudio Team, 2016) with the *dplyr* (Wickham et al., 2019), the *ggplot2* (Wickham, 2016), the *tseries* (Trapletti and Hornik, 2018), the *univOutl* (D'Orazio, 2018), and the *robustbase* (Maechler et al., 2018) packages.

Demonstration of the Forecasting Approach for *Blank* Appointment Type We took the following three steps to pre-process demand data for the *Blank* appointment type: 1) counting demand value for each booking date, 2) aggregating daily demand value into weekly, and 3) testing time series stationarity. Through applying the ADF test, we found out the time series is non-stationary and we therefore applied differencing. In Figure 3.2, the first figure from left presents the original time series of *Blank* demand which is non-stationary and has a non-constant variance. Therefore, we took the logarithm of the data in order to stabilize the variance. Thereafter, due to the presence of seasonality we took first-order seasonal difference of data with lag 8. We then applied the ADF test again, which this time showed that the time series is stationary. In Figure 3.2, the first figure from right presents stationary time series for *Blank* demand. Figure 3.3 presents the examination of the

diagnostics of ARIMA (1,0,0)(0,1,2)[7] which shows all the residual assumptions (normality and independence) are satisfied. The residuals are uncorrelated because the Box-Ljung plot shows all p-values are significantly above the 0.05 level, indicating that the residuals are white noise. However, the normality assumption is not completely satisfied because the QQ-plot displays the residuals that do not fully follow the QQ-line. Therefore, we propose another ARIMA model which can verify normality assumption.

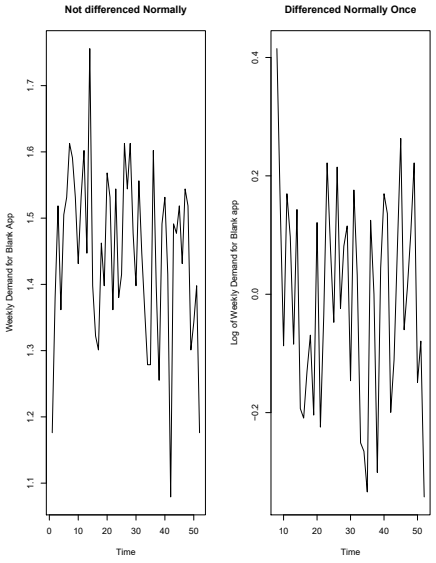


Figure 3.2: Non-stationary & stationary time series for *Blank* appointment type

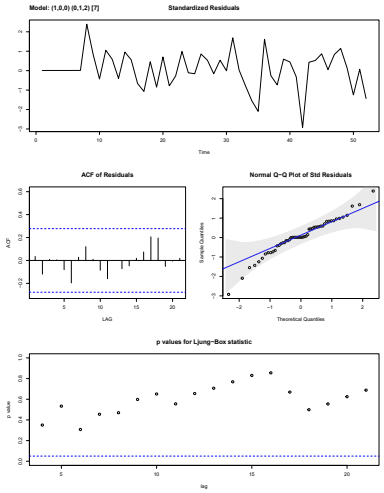


Figure 3.3: Residual Diagnostic ARIMA (1,0,0)(0,1,2)[7]

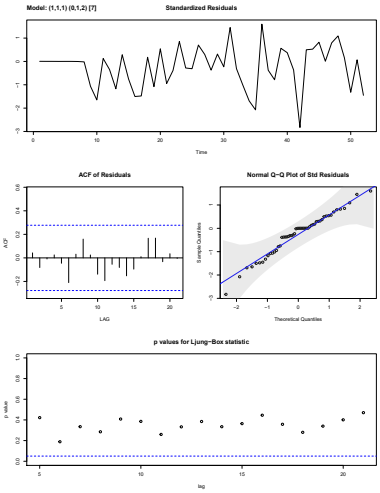


Figure 3.4: Residual Diagnostic ARIMA (1,1,1)(0,1,2)[7]

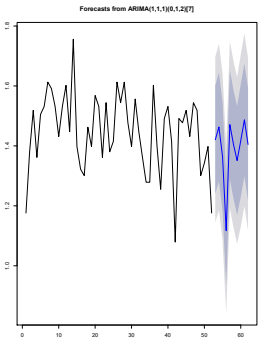
The new suggested model is ARIMA (1,1,1)(0,1,2)[7]. Figure 3.4 presents the examination of the diagnostics of this model which shows all the residual assumptions (normality and independence) are satisfied. The residuals are uncorrelated because the Box-Ljung plot shows all p-values are significantly above the 0.05 level, indicating that the residuals are white noise. In addition, the normality assumption is also satisfied because the QQ-plot displays the residuals follow the QQ-line.

Goodness of Fit and Forecast Accuracy For *Blank* Appointment Type Table 3.6 presents the comparisons between the proposed ARIMA model for *Blank* appointment type based on goodness of fit in terms of AIC and based on accuracy in terms of MAE.

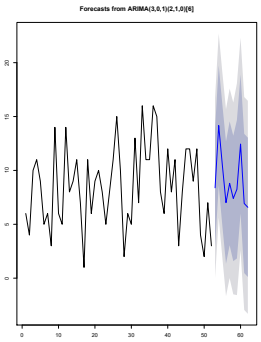
Blank	**ARIMA model**	**AIC**	**MAE**
	ARIMA (3,1,0)(0,1,1)[7]	-20	0.054
	ARIMA (3,1,1)(0,1,1)[7]	-19.36	0.051
	ARIMA (2,1,1)(0,1,1)[7]	-20.29	0.053
	ARIMA(1,1,1)(2,1,0)[7]	-21.29	0.042
	ARIMA (1,1,1)(2,1,1)[7]	-19.77	0.041
	ARIMA (1,1,1)(0,1,2)[7]	-23.35	0.032

Table 3.6: Goodness of fit and forecast accuracy for *Blank*

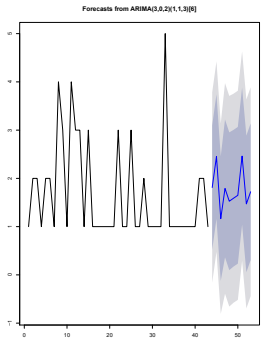
The model with the lowest AIC and MAE is selected. Therefore ARIMA (1,1,1)(0,1,2)[7] is selected to forecast weekly demand for *Blank* appointment type. Figure 3.5 presents forecasts for all the 11 appointment types.



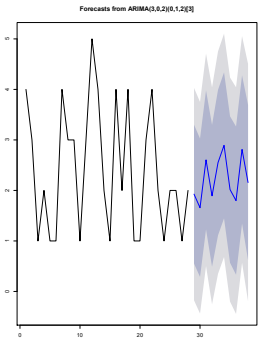
(a) *Blank*



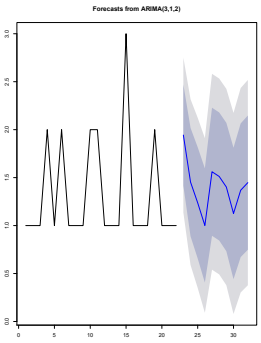
(b) *Follow-up*



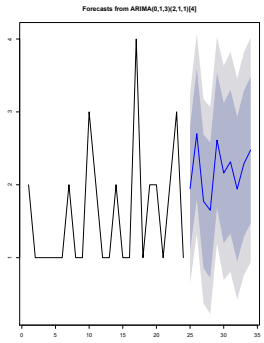
(c) *Injection*



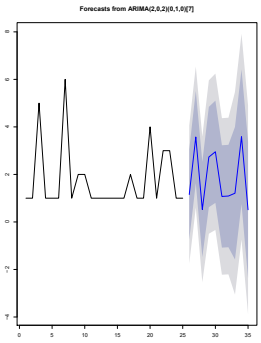
(d) *Mental Health*



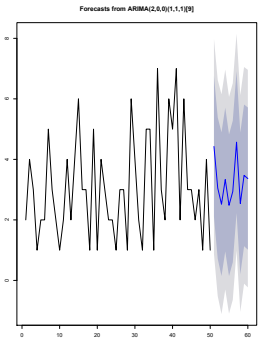
(e) *Pre-op Assess*



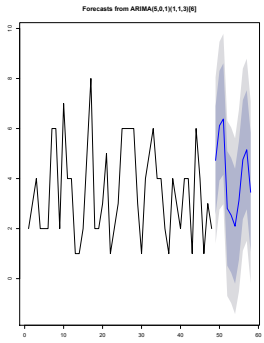
(f) *Pre-Natal*



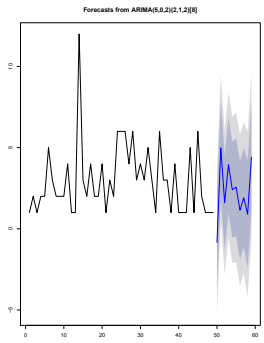
(g) *Child Physical*



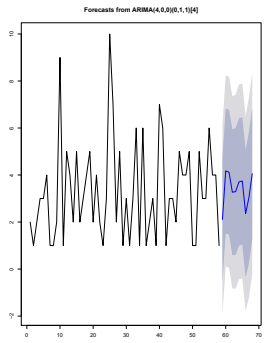
(h) *Well Baby*



(i) *Diabetic Mng*



(j) *New Patient*



(k) *Annual Phys*

Figure 3.5: Forecast for different appointment types

Based on the forecast, the demand for the next 10 weeks, specifically September 2, 2018 to November 24, 2018 for each appointment type is presented in Table 3.7 as follows:

	Weeks									
	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}
Periodic Health Exam	2	4	4	5	3	5	2	3	3	2
Child Physical	1	4	1	3	3	1	1	1	4	1
Diabetic Management	5	6	6	3	3	2	3	5	5	3
Follow-up	10	10	11	14	12	10	5	7	8	7
Mental Health	2	2	3	2	3	3	2	2	3	2
Injection	2	3	1	2	2	2	2	3	2	2
New Patient	0	5	2	4	2	3	1	2	1	4
Pre-Op Assessment	2	1	1	1	2	2	2	1	1	2
Pre-Natal	2	3	2	2	3	2	2	2	2	3
Well Baby	4	3	2	3	2	3	5	3	4	3
Blank	26	29	23	13	30	25	22	26	31	25

Table 3.7: Forecast of demand for appointment types at HFA

3.7.3 Comparison of Capacity Reservation Plans

In this section, we compare the obtained results for capacity reservation plans from the three proposed models based on workload distribution, access time distribution and unmet demand. We refer to the obtained results from models 1,2 and 3 as capacity reservation plan 1, capacity reservation plan 2 and capacity reservation plan 3 respectively.

3.7.3.1 Workload Distribution

Figure 3.6 presents the workload distribution based on the model proposed in Section 3.6.1 in which access time is not considered.

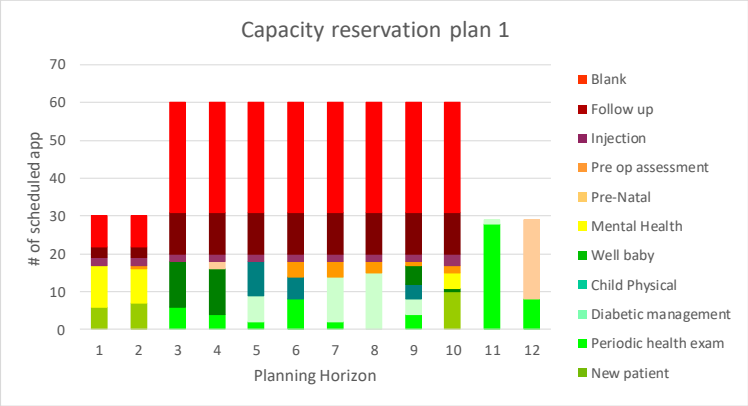


Figure 3.6: Workload Distribution Over the Planning Horizon for Capacity Reservation Plan 1.

As observed in Figure 3.6, appointment slots are reserved for *New Patient* in the 1^{st} week of planning horizon. However, as it is seen in Table 3.7, there is not any request for *New Patient* in the 1^{st} week of planning horizon. There exist the same trend for other appointment types such as *Mental Health* and *Well Baby*. It means there is a possibility that capacity reservation plan 1 becomes infeasible due to the mismatch in allocation of physician time to the demand forecast. Figure 3.6 shows appointment slots for *New Patient* which has acuity 5 are reserved in the 1^{st} week. However, the first two weeks of *planning horizon* should be mainly dedicated to more urgent appointment types (mainly to acuity level 1 & 2). The reason is that the decision variable $x_{p,j}$ cannot link the day that patient request arrive and scheduling date. Therefore, Model 1 does not prioritize patients in terms of access time.

Figure 3.7 presents workload distribution based on the proposed model in section 3.6.2 in which access time is considered.

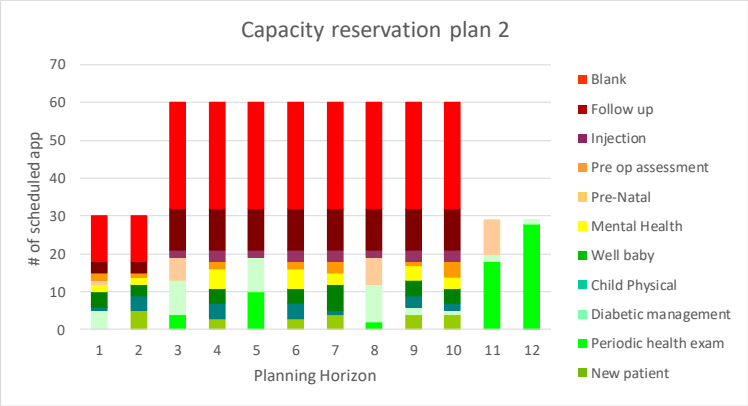


Figure 3.7: Workload Distribution Over the Planning Horizon for Capacity Reservation Plan 2.

As observed in Figure 3.7, appointments are distributed in such a way that demand are met within the relative access target to the acuity level of appointment. Figure 3.7 presents appointment slots in the first two weeks of *planning horizon* is also allocated to appointment with acuity level 5. Therefore the model only consider the demand met within the relative access target to the demand acuity. However the proposed model in section 3.6.2 does not prioritize more urgent demand to be served earlier.

Workload distribution based on the proposed model in section 3.6.3 is presented in Figure 3.8. Figure 3.8 shows appointments are reserved in a way that demand are met within the relative access target to the acuity level of appointment and more urgent demand are prioritized to be served earlier.

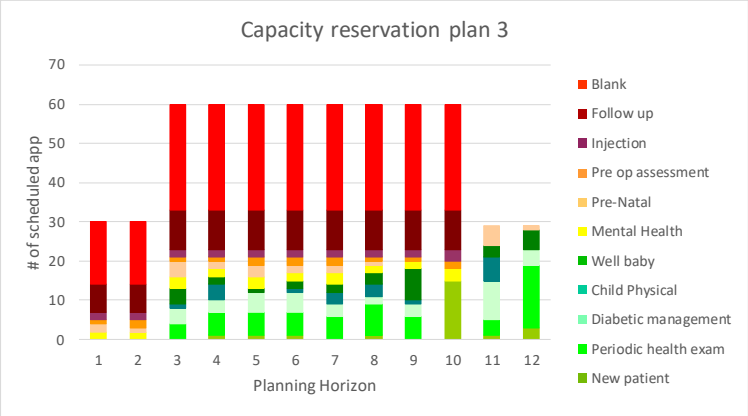


Figure 3.8: Workload Distribution Over the Planning Horizon for Capacity Reservation Plan 3.

As it is observed in Figure 3.8, appointment slots are reserved after realization of demand. In addition, the reserved appointment slots in the first two weeks is not assigned to appointment type with acuity level 5 which proves the model prioritize demand based on the relative acuity level. The main proportion of reserved appointment slots in the first two weeks is assigned to *Blank*, *Follow-up* and *Injection* (more urgent requests). In addition, distribution of workload between *Blank*, *Follow-up* and *Injection* appointment slots are in a balanced way between week 1-10 of *planning horizon*.

Table 3.8 presents compares the resulted values for the workload balance and optimality gap between the three proposed capacity reservation plans. Plan 1 provides a lower bound for capacity reservation plan 2 since it is relaxed version of plan 2.

	Capacity reservation plan		
	1	2	3
Workload balance	634.22	645.02	662.1
Optimality gap	0.01%	0.72%	0.01%
Run time	1 hr	1 hr	7 min
Solution status code	11	11	102

Table 3.8: Comparison of the three proposed capacity reservation plans

CPLEX concludes feasibility and optimality of solution in terms of "solution status code". The "solution status code" for capacity reservation plan 1 & 2 is equal to 11 which means

"Aborted due to a time limit". The "solution status code" for capacity reservation plan 3 is equal to 102 which means "Optimal solution within epgap or epagap tolerance found" (ILOG, 2002).

3.7.3.2 Access Time Distribution

In this section, we compared the obtained access time distribution from the proposed models in sections 3.6.2 and 3.6.3 based on acuity level of appointments obtained from Aslani et al. (2019).

Figure 3.9 provides the comparison for access time distribution of appointments with acuity level 5. As it is observed, the resulted access time distributions for model 3 is more shifted toward right. It can be interpreted, model 3 does not prioritize appointment types with acuity 5 to get appointment slots with same/next week access time.

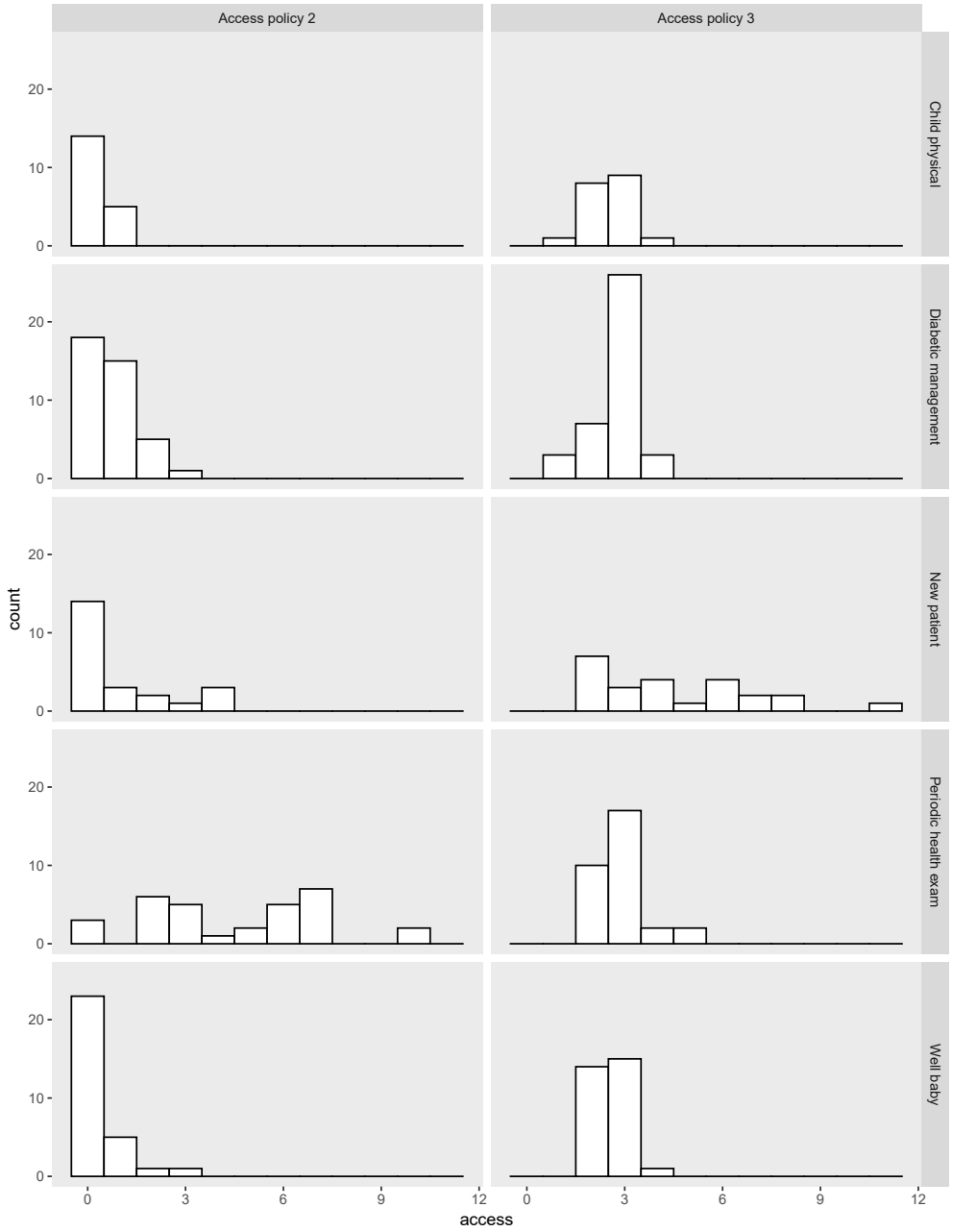


Figure 3.9: Access time distribution for appointments with mainly acuity level 5

Figure 3.10 provides the comparison for access time distribution of appointments with acuity levels 3 & 4. As it is observed, the resulted access time distributions for model 3 is more shifted toward left. In other word, the proportion of appointments with same-week access are increased in model 3. It can be interpreted as model 3 prioritize appointments

with acuity levels 3 & 4 to get same week appointment slots.

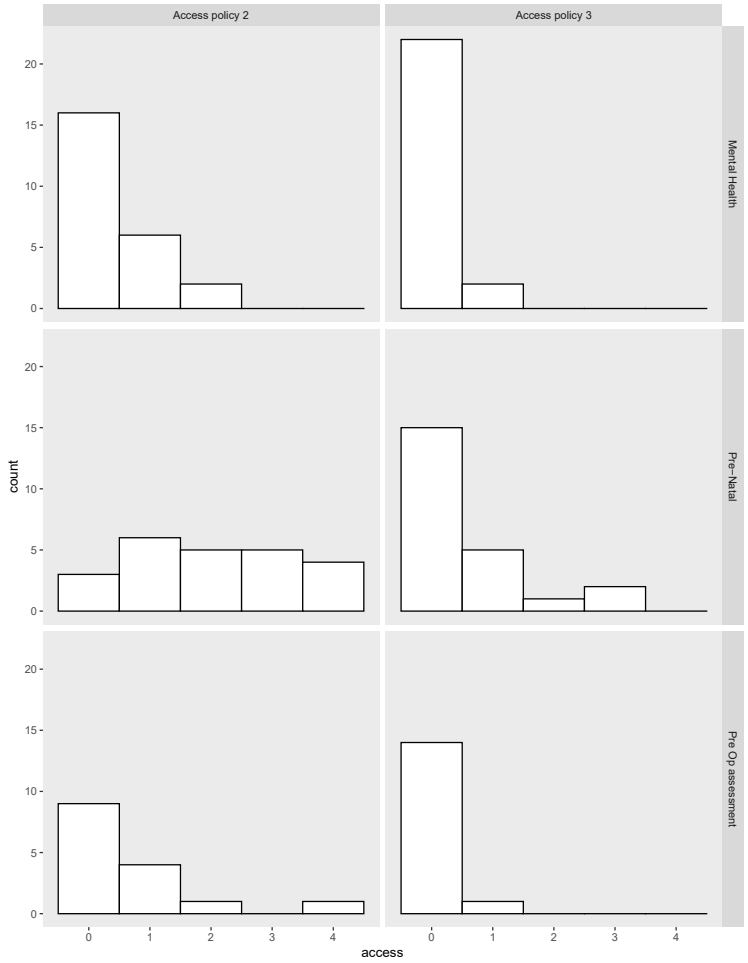


Figure 3.10: Access time distribution for appointments with mainly acuity level 3 & 4

Figure 3.11 provides the comparison for access time distribution of appointments that mainly include acuity levels 1,2 & 3. As it is observed, the resulted access time distributions for model 3 is more shifted toward left. In other word, the proportion of appointments with same/next week access are increased in model 3. It can be interpreted as model 3 prioritize appointments with acuity levels 1,2 & 3 to get same/next week appointment slots.

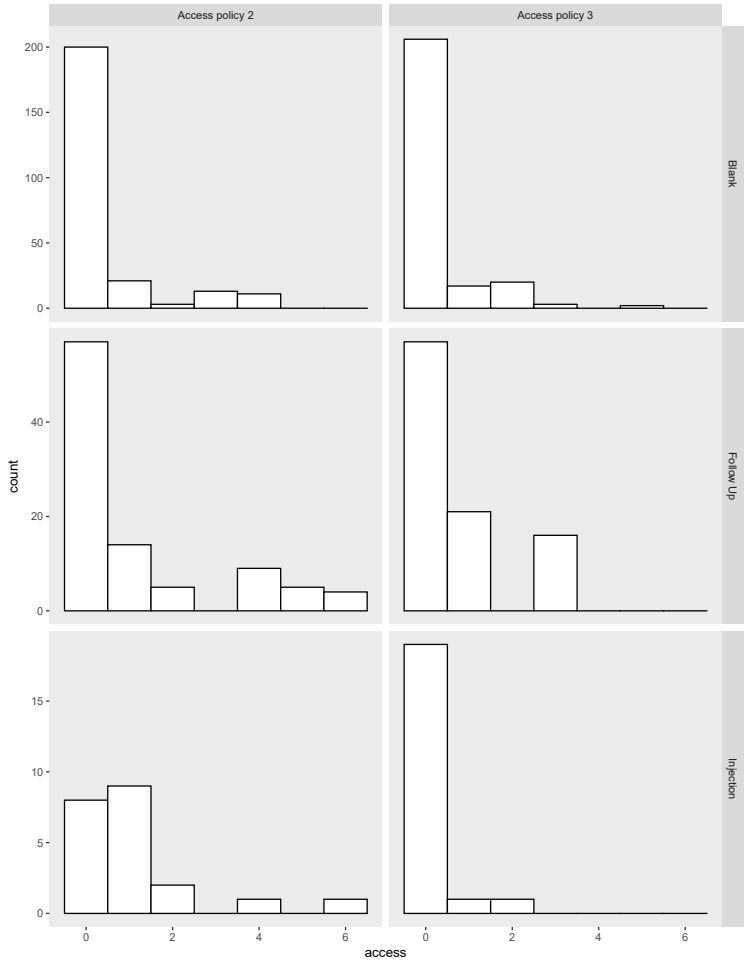


Figure 3.11: Access policy for appointments with mainly acuity level 1 & 2

One of the main messages through comparing access time distribution between capacity reservation plans 2 and 3 is the shift in same/next week allocation to different acuity levels. Therefore we present a comprehensive comparison in terms of same week allocation in Tables 3.9.

Table 3.9 compares same week allocation for different acuity levels between capacity reservation plans 2 & 3. For example to calculate percentage of same week allocation for appointments that are approximated to have acuity level 1 and/or 2, we consider the relevant appointment types from Table 3.5.

	Appointments with mainly acuity level		
# same week allocation	1,2	3,4	5
Model 2	73%(265/363)	45%(28/62)	49%(72/142)
Model 3	78%(282/363)	82%(51/62)	0

Table 3.9: Evaluating same week allocation based on capacity reservation plans 2 & 3

Same week allocation for appointment with: acuity level 5 decreases to zero, acuity level 3,4 increase by 45%, and acuity level 1,2 increase by 9%. The results show mainly prioritize more urgent request to get same week allocation. Same week allocation for appointment with acuity levels 3,4 is higher than acuity levels 1,2 since workload should be balanced in *arrival horizon* for acuity level 1 & 2.

Another message through comparing access time distribution is the gap between access target and realized access time for appointments with different acuity levels. To calculate this gap, we use $\alpha_{p,k}$, which is defined as the proportion of demand for appointment type p with all the feasible acuity level κ. Therefore we analyzed the gap between $\alpha_{p,k}$ and the realized number of reserved appointments with access time relative to acuity level κ. Table 3.10 presents the realized access time for each appointment types in capacity reservation plan 2.

	Acuity level				
	1	2	3	4	5
Periodic Health Exam	-	-	-	-	94%
Child Physical	-	95%	-	-	-
Diabetic Management	-	-	-	95%	-
Follow-up	-	76%	81%	90%	100%
Mental Health	-	100%	-	-	-
Injection	-	81%	90%	90%	100%
New Patient	-	-	-	96%	-
Pre-Op Assessment	-	-	93%	100%	-
Pre-Natal	-	-	61%	100%	-
Well Baby	-	-	-	94%	-
Blank	80%	88%	90%	99%	-

Table 3.10: Realized access time in capacity reservation plan 2

Table 3.11 presents the realized access time for each appointment types in capacity reservation plan 3.

	Acuity level				
	1	2	3	4	5
Periodic Health Exam	-	-	-	-	94%
Child Physical	-	-	-	95%	-
Diabetic Management	-		-	95%	-
Follow-up	-	83%	83%	100%	-
Mental Health	-	100%	-	-	-
Injection	-	95%	100%	-	-
New Patient	-	-	-	-	96%
Pre-Op Assessment	-	100%	-	-	-
Pre-Natal	-	-	91%	100%	-
Well Baby	-	-	-	94%	-
Blank	82%	89%	97%	98%	99%

Table 3.11: Realized access time in capacity reservation plan 3

3.7.3.3 Unmet Demand

We also compare the three proposed capacity reservation planes based on the magnitude and distribution of unmet demand. The magnitude of unmet demand for proposed plans is presented in Table 3.12.

Appointment types	**Total unmet demand**		
	Capacity reservation plan 1	Capacity reservation plan 2	Capacity reservation plan 3
Periodic health exam	2	2	2
Child Physical	1	1	1
Diabetic Management	2	2	2
Follow-up	0	0	0
Mental Health	0	0	0
Injection	0	0	0
New Patient	1	1	1
Pre-Op Assessment	0	0	0
Pre-Natal	0	0	0
Well Baby	2	2	2
Blank	2	2	2

Table 3.12: Total unmet demand for capacity reservation plans 1,2,3

As we see, the magnitude of unmet demand is the same between all the three planes. The distribution of unmet demand in the three proposed planes is presented in Figure 3.12 - 3.14. The highest peak for unmet demand between week is 6 in capacity reservation plan 1, 4 in capacity reservation plan 2 and 3 in capacity reservation plan 3. In other word, unmet demand in model 3 is more distributed which is the result of having most balanced workload distribution between the three proposed plans.

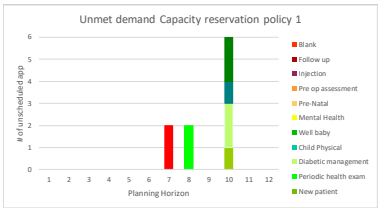


Figure 3.12: Unscheduled demand in plan 1

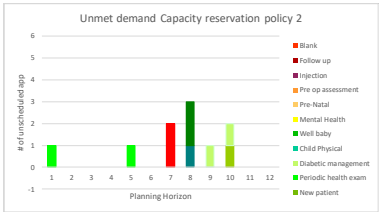


Figure 3.13: Unscheduled demand in plan 2

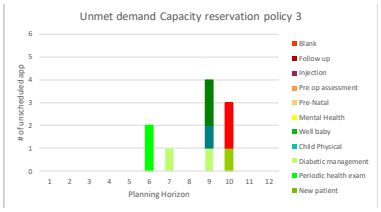


Figure 3.14: Unscheduled demand in plan 3

3.8 Conclusion

In this study, we develop an optimal capacity reservation plan in the presence of multiple non-urgent appointment types and multiple access time targets that takes into account the needs of both patients and primary care providers. The novelty of this study is developing an efficient linear optimization approach that tackles the main concern in the Canadian primary care clinics which are: shortage of family physicians, insufficient appointment slots for urgent request and lack of equitable criteria. The developed TCP model has multi-objective which balance physician workload for a specific physician between weeks based on acuity level of appointment, provide equitable access and prioritize appointment slots with same/next week access. The optimization model distribute the reserved appointment slots in a manner to balance physician workload between weeks and achieve equitable access . Therefore appointment slots for patients with lower acuity level are reserved on weeks with lower patient demand for higher acuity level. To be able to reserve appointment slots in advance, we develop an individual forecast model for each appointment type to predict the demand. Considering demand forecast provides data driven optimal capacity reservation plan which is more accurate due to capturing variability. This study improve access to primary care through equitable allocation of physician time that prioritize patient access based on acuity level of patient request obtained from Aslani et al. (2019) rather than considering two general category of urgent and non-urgent requests. The developed model in this study, incorporate the obtained access targets from Aslani et al. (2019) as an access time constraint for each appointment type. The developed TCP model can also deal with dynamic aspect of multiple planning horizon through tracking patients who may need to be scheduled in the subsequent planning horizon and considering the demand from previous and next planning horizon to determine phyisicin weekly availability. To make this reservation plan as a decision making support for schedulers, as a future work, we should also develop another optimization model to distribute the obtained weekly reserved time slots between half-days of a week that the family physician is available.

Chapter 4

A Robust Optimization Model for Tactical Capacity Planning in an Outpatient Setting

Abstract[1] Tactical capacity planning is a key element of planning and control decisions in healthcare settings, focusing on the allocation of a clinic's resources to appointments of different types. One of the most scarce resources in healthcare is physician time. Due to uncertainty in demand for appointments, it is difficult to provide an exact match between the scheduled physician availability and appointment requests. Our study, therefore, uses cardinality-constrained robust optimization to develop tactical capacity plans which are robust against uncertainty, providing a feasible allocation of capacity for all (or the majority of) realizations of demand. Our approach takes into account multiple appointment types and multiple access time targets. We experimentally evaluate our approach and its practical implications under different levels of conservatism. We show that we can guarantee 100% feasibility of the robust tactical capacity plan while not being fully conservative, which will lead to the clinic saving money while being able to meet demand despite uncertainty. We also show how the robust model helps us to identify the critical time periods (weeks) which contribute to worst case physician peak load, which could be valuable to decision-makers.

4.1 Introduction

According to the 2018 Canadian Institute of Health Information report (CIHI, 2018), physician services in Canada are scarce compared to most other participant countries in the

[1]This chapter is submitted to the *Health Care Management Science* journal as: Nazanin Aslani, Onur Kuzgunkaya, Navneet Vidyarthi, Daria Terekhov. A Robust Optimization Model for Tactical Capacity Planning in an Outpatient Setting, Submitted September 2019.

Organization for Economic Co-operation and Development (OECD). Therefore, Canadian patients have longer access time, defined as the interval between the arrival of the appointment request and the scheduled time of appointment (CIHI, 2018). One of the essential elements for addressing physician scarcity and long access times is tactical capacity planning (TCP). TCP involves making tactical-level decisions, i.e., medium-term planning decisions for a group of patients instead of individual ones (Ahmadi-Javid et al., 2017). TCP deals with two related questions: resource allocation and capacity design. The aim of resource allocation is the allocation of known capacity (e.g., total number of available physician hours) to different patient classes (Hulshof et al., 2013). The focus of capacity design, on the other hand, is to determine the resource capacity required to meet certain performance targets (e.g., access time targets) (Nguyen et al., 2015). In practice, the two questions are inter-dependent, since resource allocation is based on the available capacity while capacity requirements are influenced by the effectiveness of resource allocation.

In this paper, we focus on the capacity design problem combined with resource allocation in order to address the concerns of physician scarcity and long access times in outpatient settings, i.e., those in which care delivery happens without overnight hospitalization. We study an outpatient setting with the following characteristics: two appointment types, corresponding to appointments for new patients and follow-up visits; dependence between appointment types, since the number of follow-ups depends on the number of new patients; and access time targets for each appointment type. This setting was first described and studied by Nguyen et al. (2015) in the context of an outpatient clinic of an urology department. This study can also be applied for primary care setting in which appointments are classified into the two categories of new patient (new complaint) and follow-ups. Nguyen et al. (2015) developed a deterministic model for finding the total required physician time and its allocation to each patient type to meet access time targets. Our work builds on Nguyen et al.'s in two ways. First, we modify their model in order to control the total number of appointments scheduled in the given planning horizon. Second, more importantly, we address uncertainty in

demand by developing a robust optimization model which determines the required physician capacity and distributes it among patient types and is robust against uncertainty, providing a feasible allocation of capacity for all (or the majority of) realizations of demand. Thus, the contributions of this paper are focused on extending the previous work in order to make the resulting models more applicable in practice. Specifically,

- We address two limitations of an outpatient tactical capacity planning approach from the literature (Nguyen et al., 2015): we extend their model to deal with demand uncertainty and explicitly consider the number of patients who may need to be scheduled in the subsequent planning horizon.

- We develop a robust TCP model based on the cardinality -constrained method of Bertsimas and Sim (2003). We make a general assumption that there is uncertainty in new patient demand, without knowledge of the exact time period in which uncertainty occurs.

- We conduct an extensive set of experiments to determine the level of robustness based on cost and infeasibility probability of a robust solution. We also use our approach to identify the most critical time periods in the planning horizon.

The paper is organized as follows. Section 4.2 describes recent work in robust optimization in healthcare planning and scheduling. Section 4.3 describes the problem, presents the deterministic model of Nguyen et al. (2015) and discusses their limitations. Section 4.4 focuses on the development of two cardinality-constrained robust optimization models to address demand uncertainty. The first model optimizes the maximum required physician capacity over the planning horizon, while the second is a multi-objective model which allows to evaluate the trade-off between capacity and the number of patients left unscheduled in the current time period. In Section 4.5, we conduct an extensive set of experiments and discuss the results. In section 4.6, we discuss the three key observations from our study. Section 4.7 concludes the paper.

4.2 Robust Optimization in Healthcare Planning and Scheduling

One of the major challenges in the development of planning and scheduling models for healthcare environments is data uncertainty. If data uncertainty is overlooked during the model development process, and the realization of data is different from the one expected, the resulting solution may not be feasible. This issue can be addressed by providing a robust solution which remains feasible when (specific) model parameters deviate from their nominal values via robust optimization (RO) (Ben-Tal et al., 2009; Soyster, 1973). In this approach, the distribution of the uncertain parameter is unknown but its value is assumed to belong to an *uncertainty set*. The quality of a robust approach is evaluated based on two criteria: remaining feasible despite changing parameter values and the cost of doing so. The cost of a robust solution is attributed to potential over-conservatism and is measured by evaluating a trade-off between the robustness and the optimal objective value. Robust optimization, in contrast to other approaches to addressing uncertainty such as stochastic and chance-constrained programming, does not require knowledge of the probability distribution of uncertain parameters which in fact may not be available in practice (Birge and Louveaux, 2011; Henrion, 2004). Nguyen et al. (2018) extended their previous paper (Nguyen et al., 2015) to address demand uncertainty through formulating a stochastic linear optimization model with chance constraints, assuming either full or partial knowledge of the uncertainty in each period of the planning horizon. In contrast, in our paper, we do not assume any knowledge of the probability distribution and develop a robust tactical capacity planning model based on a particular robust optimization method: cardinality-constrained RO.

In cardinality-constrained robust optimization over-conservatism is avoided through defining a polyhedral uncertainty set in which the level of conservatism is controlled through defining *budget of uncertainty* (Jalilvand-Nejad et al., 2016). If the budget of uncertainty is zero, the decision-maker is not conservative at all. However, if the budget of uncertainty is equal to the number of constraints in which the uncertain parameter exists, the decision maker has the highest level of conservatism; the decision-maker can prevent being overly con-

servative by defining a budget of uncertainty between zero and highest level of conservatism where the cardinality of the parameters permitted to change is constrained (Bertsimas and Sim, 2003, 2004). Cardinality-constrained RO has the advantages of providing a computationally tractable model as well as simplicity that makes it simplicity that is appealing to decision-makers; it has therefore been successfully applied in many areas, including logistics and production systems (Hazır and Dolgui, 2013), tactical planning in supply chains (Sanei Bajgiran et al., 2017) and healthcare management problems (Addis et al., 2015).

Cardinality-constrained RO in healthcare has mainly been applied in inpatient settings. For example, Denton et al. (2010) apply cardinality-constrained RO for allocating surgery blocks to operating rooms to deal with uncertainty in surgery duration. A surgery block is defined as one or more consecutive surgeries which are performed by a specific surgeon in the same operating room during an eight-hour period. Denton et al. (2010) apply cardinality-constrained RO to deal with the mentioned uncertainty due to the limitation in data availability for surgery durations as well as the capability of decision makers to provide reasonable estimates for the lower and upper bounds for surgery durations. The objective of their model is to allocate surgery blocks to operating rooms to minimize the worst possible over-time cost for all realizations of surgery block durations within the defined range. Tang and Wang (2015) develop an RO model for allocating operating room time to different sub-specialty of surgeries to deal with uncertainty in demand for surgeries, considering both elective and emergency cases. The developed RO model is based on implementor/adversary algorithm of Bienstock (2007), which decomposes the original problem into a master problem (implementor) and a sub-problem (adversary). The sub-problem chooses a value for uncertain demand which deviates from its nominal value and controls the maximum number of uncertain demands based on the cardinality-constrained RO method. The master problem minimizes the revenue loss for the shortage of operating rooms as well as a penalty cost for idleness of operating rooms for the generated demand from the sub-problem. Geranmayeh (2015) develops a cardinality-constrained RO model to allocate blocks of operating rooms to each surgeon to

deal with uncertainty in the number of referrals to the inpatient ward. The number of referrals cannot be estimated via an average because doing so would force the surgeon to have same number of inpatients per surgery block. However, the decision for referral can be made only after the surgery is done. Geranmayeh (2015) applies cardinality-constrained RO since the distribution for the number of referrals is not known, but the range can be estimated. The objective of the developed RO model is protecting the hospital against congestion in the inpatient ward due to the highest possible number of referrals to the inpatient unit.

On the contrary, cardinality-constrained RO has been rarely applied in *outpatient* settings. We found only two such papers in outpatient settings. Pour (2016) develops a cardinality-constrained RO model to schedule patients from a radiotherapy waiting list to deal with uncertainty in treatment duration and treatment time. The goal of their model is maximizing the number of scheduled patients from the waiting list for the worst possible scenario of treatment durations. Mirahmadi Shalamzari (2018) applies cardinality-constrained RO method to develop an admission planning model for multi-priority independent patients for magnetic resonance imaging (MRI) exam in the presence of patient arrival uncertainty. The aim of their model is to minimize weighted patient waiting time for the highest possible demand for MRI exams. Among the studies described above, Mirahmadi Shalamzari (2018) is the only one that focuses on uncertainty that arises in the right-hand side parameter of a constraint. In our study, similar to Mirahmadi Shalamzari (2018), we consider demand uncertainty that happens in the right-hand side of two constraints. However, in our study, patient types are *dependent* and we explicitly consider demand uncertainty only in one appointment type which implicitly impacts the demand for the other type of appointment. Furthermore, in the current study we control the level of conservatism for both the total demand in the planning horizon as well as the demand in a single period.

4.3 Problem Background

In this section, we describe the problem of interest, the previous work of Nguyen et al. (2015) and some of the limitations of that work. Throughout most of the paper, we adopt and, where necessary, extend the notation of Nguyen et al. (2015).

4.3.1 Problem Description

This study focuses on the same outpatient clinic setting as described in Nguyen et al. (2015). This clinic follows a re-entry appointment system based on which patient appointments are classified into the two general categories of first-visit (FV) and re-visit (RV) patients. The outpatient clinic is looking for an appointment system which can guarantee the availability of physician capacity to match the uncertain patient demand to reduce the difficulty of accessing care. In this study, care accessibility is evaluated based on access time, which is the interval between the arrival of the request and the scheduled appointment. It is assumed that f_i requests for FV appointments arrive at each period i of the *arrival horizon*, $\mathcal{S} = \{1, \dots, S\}$. The requests for both FV and RV appointments should be scheduled in the planning horizon, $\mathcal{T} = \{1, \dots, T\}$, which is defined as the extension of arrival horizon and should be long enough to cover the maximum access time allowed for the last arriving FV and RV requests from the arrival horizon.

Furthermore, the two appointment types, FV and RV, are dependent: FV patients may need to revisit the outpatient clinic for completing their care, therefore at their next visit they are considered as RV patients. If FV patients make their appointment requests after the end of the current arrival horizon, S, they will be considered as FV patients for the next arrival horizon. In addition, RV patients who cannot be scheduled in the current planning horizon will be postponed to the next planning horizon, which is modeled by scheduling them in period $T+1$. Therefore, we define index set $\mathcal{T}' = \{1, \dots, T+1\}$. In this study, we denote the total number of postponed RV patients ("scheduled" at $T+1$) as Ψ and assume that these become pre-scheduled RV patients for the next planning horizon. Unlike in previous

work on this problem, we treat Ψ as a parameter that controls the trade-off between the number of patients seen in the current planning horizon (for which the access targets are enforced) and the number of patients postponed until the next planning horizon. Figure 1 represents the schematic view for the re-entry appointment system explained above.

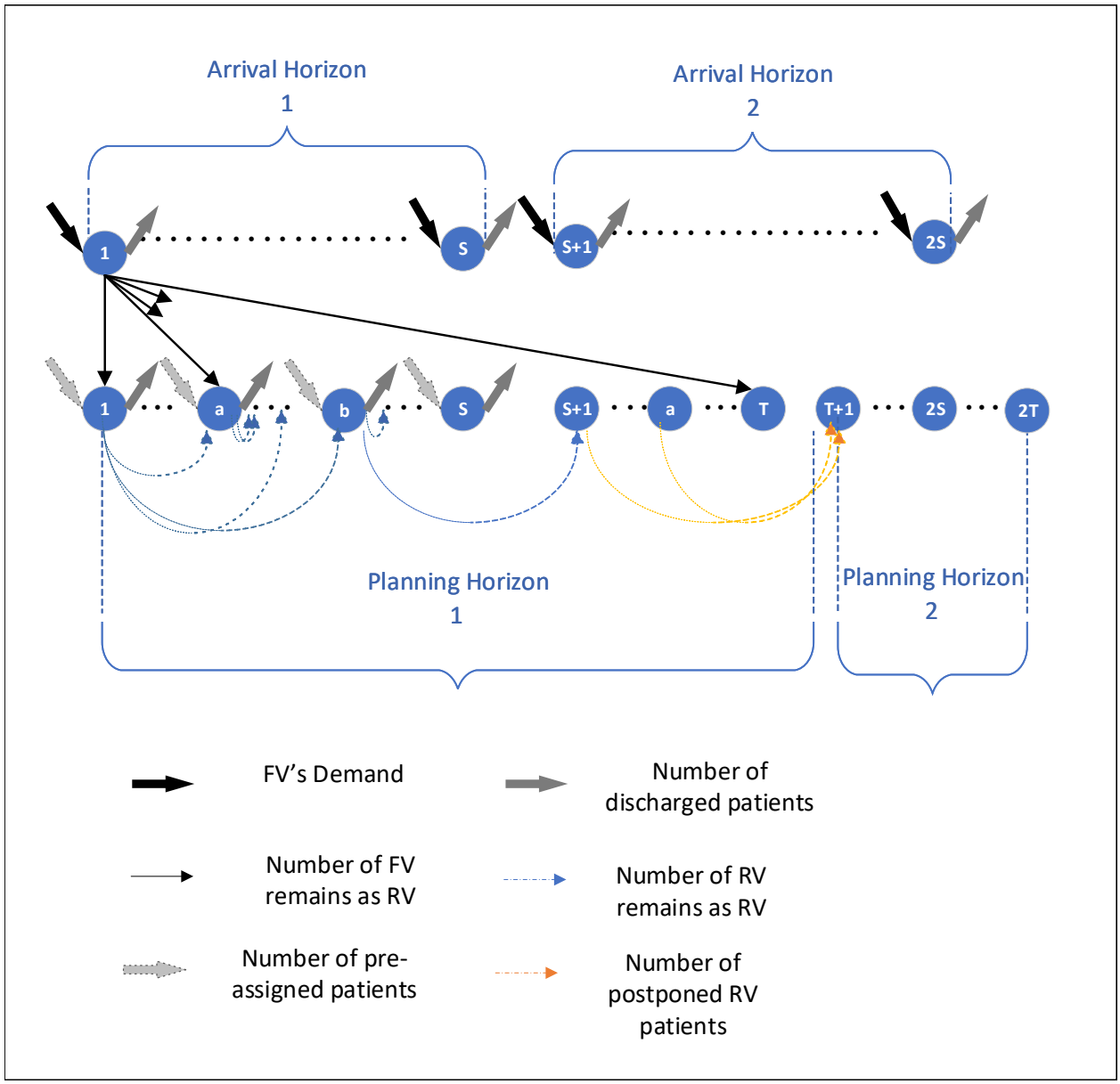


Figure 4.1: Schematic view of re-entry appointment system, adapted from Figure 3 of the paper by Nguyen et al. (2015).

The focus of the problem is therefore on planning the required physician time for FV and RV patients that arrive in the arrival horizon so as to meet the access time targets.

4.3.2 Deterministic Tactical Capacity Model of Nguyen et al. (2015)

Nguyen et al. (2015) presented a mixed-integer programming model for the deterministic version of the above problem. The goal of our deterministic tactical capacity planning (DTCP) model is minimizing the maximum required physician time between weeks subject to ensuring weekly demand for appointments is satisfied and access time targets are met. Maximum required physician time can also be seen as physicians' peak load. The mathematical formulation of the DTCP model is presented in equations (4.1)–(4.22). The model is based on a network flow model with two sets of nodes, one set for arrival periods of FV patient requests and the other set for the scheduled periods for RV patient requests. There are four general sets of constraints: conservation of flow between FV nodes, conservation of flow between RV nodes, access time targets and finally required capacity for each patient type. The notation for the model is defined as follows.

Decision Variables

$z_{i,j}$ Number of FV patients who make a request in the i^{th} period and have their appointment scheduled in the j^{th} period.

$x_{i,j}$ The number of FV patients who make a request in the i^{th} period and have appointment in the j^{th} period, and still remain in the system as RV patients after their appointment in the j^{th} period.

$y_{i,j}$ The number of RV patients who have an appointment in the i^{th} period and have their next appointment in the j^{th} period, and still remain as RV patients after the j^{th} period.

d_j^r The number of RV patients who are discharged after their appointment in period j.

C_i^f, C_i^r, C_i Capacity in the i^{th} period (minutes) for FV, RV, and both patient types, respectively.

q The maximum required capacity per period (minutes).

Parameters

u_m,u_p, u_{100}	Appointment lead-time targets (number of time periods) for median, p^{th} percentile ($0 < p < 1$), and 100^{th} percentile of FV appointment requests (f_i), respectively.
$[a, b]$	Range of RV appointment access time target (number of time periods).
$\bar{a}$	Mean RV appointment access time target (number of time periods).
τ^f,τ^r	Consultation times for FV and RV patients (minutes). These times are specified by the doctors.
α,β	Discharge rates for FV and RV patients, respectively ($0 < \alpha, \beta < 1$).
r_j^f,r_j^r	Number of pre-scheduled FV and RV patients with appointments in period j who still remain as RV patients after their appointments.

Sets

Z	Set of all $z_{i,j} : j - i \geq 0$.
L^m, L^p, L^{100}	Set of all $z_{i,j} \in Z$ that have $j - i \leq u_m$ and set of all $z_{i,j} \in Z$ that have $j - i \leq u_p$ and set of all $z_{i,j} \in Z$ that have $j - i \leq u_{100}$, respectively.

The original formulation of Nguyen et al. (2015) for the DTCP problem is presented below.

$$Min \quad q \tag{4.1}$$

$$\text{s.t.} \qquad q \geq C_j \quad \forall j \in \mathcal{T} \tag{4.2}$$

$$\sum_{j=i}^{T} z_{i,j} = f_i \quad \forall i \in \mathcal{S} \tag{4.3}$$

$$\sum_{j=i}^{T} z_{i,j} = 0 \quad \forall i \in \mathcal{T} \backslash \mathcal{S} \tag{4.4}$$

$$x_{i,j} - (1 - \alpha) z_{i,j} = 0 \quad \forall i \in \mathcal{S}, \forall j \in \mathcal{T} \tag{4.5}$$

$$(r_j^f + r_j^r + \sum_{i=1}^{j} x_{i,j} + \sum_{i=1}^{j} y_{i,j}) - (d_j^r + \sum_{i=j}^{T} y_{j,i}) = 0 \quad \forall j \in \mathcal{S} \tag{4.6}$$

$$(r_j^f + r_j^r + \sum_{i=1}^{j} x_{i,j} + \sum_{i=1}^{j} y_{i,j}) - (d_j^r + \sum_{i=j}^{T+1} y_{j,i}) = 0 \quad \forall j \in \mathcal{T} \backslash \mathcal{S} \tag{4.7}$$

$$d_j^r - \beta(r_j^f + r_j^r + \sum_{i=1}^{j} y_{i,j} + \sum_{i=1}^{j} x_{i,j}) = 0 \quad \forall j \in \mathcal{S} \tag{4.8}$$

$$y_{i,j} = 0 \quad \forall j - i < a \quad \forall i \in \mathcal{T}, \forall j \in \mathcal{T}^{'} \tag{4.9}$$

$$y_{i,j} = 0 \quad \forall j - i > b \quad \forall i \in \mathcal{T}, \forall, j \in \mathcal{T}^{'} \tag{4.10}$$

$$y_{i,j} = 0 \quad \forall j \geq T + 1 \quad \forall i \in \mathcal{S} \tag{4.11}$$

$$\sum_{i=S+1}^{T} d_i^r = 0 \tag{4.12}$$

$$\sum_{z_{i,j} \in L^m} z_{i,j} \geq (1/2 \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{T}} z_{i,j}) + 1 \tag{4.13}$$

$$\sum_{z_{i,j} \in L^p} z_{i,j} \geq p \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{T}} z_{i,j} \tag{4.14}$$

$$\sum_{z_{i,j} \in L^{100}} z_{i,j} = \sum_{i \in \mathcal{S}} f_i \tag{4.15}$$

$$z_{i,j} = 0 \;\; \forall i, j \in \mathcal{T} : j - i \geq u_{100} + 1 \tag{4.16}$$

$$z_{i,j} = 0 \;\; \forall i, j \in \mathcal{T} : j - i < 0 \tag{4.17}$$

$$\sum_{j=1}^{T} \sum_{i=1}^{j} (j - i) y_{i,j} - \bar{a} \sum_{j=1}^{T} \sum_{i=1}^{j} y_{i,j} \leq 0 \tag{4.18}$$

$$C_j^f - (\tau^f r_j^f + \tau^f \sum_{i=1}^{j} z_{i,j}) = 0 \quad \forall j \in \mathcal{T} \tag{4.19}$$

$$C_j^r - (\tau^r r_j^r + \tau^r \sum_{i=1}^{j} y_{i,j}) = 0 \quad \forall j \in \mathcal{T} \tag{4.20}$$

$$C_j - (C_j^f + C_j^r) = 0 \quad \forall j \in \mathcal{T} \tag{4.21}$$

$$z_{i,j}, x_{i,j}, y_{i,j'}, C_j^f, C_j^r, d_i^f, d_j^r \geq 0 \quad \forall i \in \mathcal{T}, \forall j \in \mathcal{T}, , \forall j^{'} \in \mathcal{T}^{'} \tag{4.22}$$

The DTCP model minimizes the maximum required capacity (physician time). Constraint (4.2) defines a decision variable (q) for the maximum required capacity which should be greater than the total required physician time for both patient types. The first set of constraints which is the conservation flow at FV nodes is modeled in constraints (4.3) and (4.5). Constraint (4.3) assures FV patient demand over each period of arrival horizon is fully covered. Constraint (4.5) determines the proportion of scheduled FV patients who need to revisit the clinic in the future (in other words, the number of FV patients who will be considered as RV patients for their next visit).

The second set of constraints which is the conservation flow at RV nodes is modeled in constraints (4.6) and (4.8). Constraint (4.6) presents the balance between inflow into and outflow from each RV nodes. The inflow into a RV node includes the total number of pre-scheduled RV and FV patients as well as total number of RV patients with an appointment at the RV node. The outflow from a RV node include the total number of RV patients who will be discharged plus the number of RV patients who will revisit again after their appointment at the RV node. Constraint (4.8) shows the number of scheduled RV patients whose care will be completed and discharged after their appointment at the RV node.

The third set of constraints which is designed for controlling the access time of FV and RV patients is specified from constraint (4.9) to constraint (4.18). The targets for RV patients accessibility are designed based on defining a restricted range as well as mean access time. Constraints (4.9) and (4.10) assures the access time of the scheduled RV patients' appointments belongs to the restricted range of $[a, b]$. Constraint (4.11) forces RV patients' appointments to be scheduled before the last date of planning horizon. Constraint (4.12) prohibits the discharge of RV patients if their appointment is made after the arrival horizon. This model restricts the access time of FV patients through controlling the distribution of their access time. The three constraints of (4.13), (4.14) and (4.15) assures FV patients' appointments accessibility based on defining three targets for median, p^{th} percentile and 100^{th} percentile of FV patient requests. Constraint (4.18) specifies mean access time for

RV patients' appointment meets the designed mean access time target. The last set of constraints regarding the required capacity for each patient type and total required capacity are presented in the constraints (4.19), (4.20) and (4.21).

4.3.3 Modifications of Nguyen et al.'s Model

In this study, we modify the adjusted model of Nguyen et al. (2015) to control the number of scheduled RV patients in each planning horizon which prevents congestion of suspended RV patients at time period $T+1$, which we denote by Ψ. The reason for this congestion is the increase in the value of Ψ as the level of uncertainty increases. Therefore, we control Ψ to serve a strategically agreed number of patients at each planning horizon which is one the main objectives of TCP (Hulshof et al., 2013). In the original DTCP model, constraint (4.3) ensures the total number of FV patients that should be served is equal to FV demand. Since RV patients are a consequence of FV patients, there is currently no constraint regarding the strategic number of RV patients that should be served. Therefore, when the demand for FV patients becomes uncertain, we would like to be sure the extra RV patients resulting from uncertain FV demand are also scheduled in the current planning horizon instead of being postponed to the period $T+1$ to be scheduled in the next planning horizon. We therefore propose to add the constraint (4.23):

$$\sum_{i \in \mathcal{T}} y_{i,T+1} \leq \Psi. \tag{4.23}$$

In this case, we need to solve the TCP problem in two stages. In the first stage, we find the value for Ψ and in the second stage we find the robust solution. In practice there may be more than one way to determine Ψ, depending on the clinic's goals and availability of data. In this paper, we find this value by the solving the DTCP model (4.1)–(4.22) for a particular demand scenario (the one used by Nguyen et al. (2015)) and then insert the obtained value for Ψ in the right-hand side of equation 4.23. In practice, this number should be determined

in consultation with the clinic; an alternative approach is discussed in Section 4.6.

4.4 Robust Tactical Capacity Planning Model

The DTCP model in this study is designed based on the single value of demand for appointment in each week. However, the clinic may face periods of high congestion due to uncertainty in weekly demand. Therefore, we cannot ensure care accessibility through considering a single value for weekly demand for appointments. We need to design a TCP model in such a way as to accommodate different ranges of demand. Among the operations research methods, we choose robust optimization (RO) since it will allow us to minimize the required physician time due to realization of worst case scenario for the uncertain demand which belongs to a fixed interval. The interval in which the uncertain parameter varies is defined as an uncertainty set. Applying RO results in the smallest required physician time that will be feasible regardless of what realization of demand uncertainty occurs in the uncertainty set. The RTCP model decides the required physician time for each of the defined appointment type for the worst case realization of demand to achieve the designated targets for the access time of each appointment type. The RTCP model in this study will address this goal by minimizing the maximum required physician time for the worst case realization of demand between weeks. This is equivalent to minimizing the highest potential physicians' peak load. In this section, we develop the robust counterpart of the model presented in the previous section to develop a robust tactical capacity planning (RTCP) model. The TCP model is affected by uncertainty in the parameter f_i which is the number of FV patient requests in the i^{th} period. We use the deterministic demand values from Nguyen et al. (2015) as the nominal demand values in this study.

4.4.1 Uncertainty Set

We assume that uncertainty in f_i generally happens in the arrival horizon without the knowledge of the specific periods of the arrival horizon that are affected. This assumption

matches reality since we cannot know in advance in which week the uncertainty will be manifested. We assume that demand at each period is modeled as a non-negative, bounded random variable $\tilde{f}_i, i \in S$, which is independent of $\tilde{f}_j$, $j \in S$, $j \neq i$. To incorporate the uncertainty we focus on the case when the number of FV patient requests in the arrival horizon at each period exceeds its nominal value, since our model does not have any penalties for idleness. Specifically, we assume that each uncertain parameter, $\tilde{f}_i$, takes values in the box uncertainty set of $[0, \bar{f}_i + \hat{f}_i]$, where $\bar{f}_i$ is known as the *nominal* value and $\hat{f}_i$ is the maximum deviation of the uncertain parameter from its nominal value. Recall that the uncertain parameter $\tilde{f}_i$ appears in the following constraints of the original model:

$$\sum_{j=i}^{T} z_{i,j} = \tilde{f}_i \quad \forall i \in S \tag{4.24}$$

$$\sum_{z_{i,j} \in L^{100}} z_{i,j} = \sum_{i \in \mathcal{S}} \tilde{f}_i \tag{4.25}$$

4.4.2 Budget of Uncertainty

We define a budget of uncertainty based on the assumption that uncertainty generally happens in the arrival horizon without the knowledge of the specific period(s) affected. Specifically, the budget of uncertainty Γ takes values in $[0, s]$, where s (number of periods in the arrival horizon) is equal to the maximum number of periods which can incorporate uncertainty over the arrival horizon. In other words, the budget of uncertainty (Γ) is the degree of freedom that the scheduler can consider regarding number of periods in the arrival horizon s in which the number of FV patient requests (f_i) deviates from its nominal values ($\Gamma = s$ will be the worst case).

4.4.3 Robust Formulation of Constraint (4.15)

Protection function aims to guarantee feasibility of the constraint with uncertain parameters through adding the following function:

$$\eta(\Gamma) = \max_{\{P\cup\{t\} \mid P\subseteq S, |P|=\lfloor\Gamma\rfloor, t\in S\setminus P\}} \{\sum_{i\in P} \hat{f}_i + (\Gamma - \lfloor\Gamma\rfloor)\hat{f}_t\}. \tag{4.26}$$

Let κ_i be the proportion of deviation of uncertain parameter $\tilde{f}_i$ from the the nominal value $\bar{f}_i$ towards the maximum weekly demand of $\bar{f}_i + \hat{f}_i$, with $0 \leq \kappa_i \leq 1, \forall i \in S$. Before deriving the robust equivalent, we rewrite constraint (4.15), as follows:

$$\sum_{z_{i,j}\in L^{100}} z_{i,j} = \sum_{i\in S} \bar{f}_i + \sum_{i\in S} \kappa_i \hat{f}_i \tag{4.27}$$

Below, the linear equivalent for the non-linear protection function η is provided. The objective function of the following problem is maximizing the proportion of demand uncertainty in each period of arrival horizon:

$$\text{Maximize} \qquad \sum_{i\in S} (\kappa_i \hat{f}_i) \tag{4.28}$$

$$\text{s.t.} \qquad \sum_{i\in S} \kappa_i \leq \Gamma \tag{4.29}$$

$$0 \leq \kappa_i \leq 1 \qquad \forall i \in S. \tag{4.30}$$

Due to the fact that the presented equivalent model is feasible and bounded, by strong duality theorem, its dual model is also feasible and bounded. The dual model of the above linear program is derived by introducing the dual variables λ (constraint (4.29)) and μ_i (constraint (4.30)):

$$\text{Minimize} \qquad \lambda\Gamma + \sum_{i\in S} \mu_i \tag{4.31}$$

$$\text{s.t.} \qquad \lambda + \mu_i \geq \hat{f}_i \qquad \forall i \in S \tag{4.32}$$

$$\lambda \geq 0, \mu_i \geq 0 \quad \forall i \in S. \tag{4.33}$$

Therefore, the linear robust formulation of constraint (4.15) is as follows:

$$\sum_{z_{i,j} \in L^{100}} z_{i,j} = (\sum_{i \in S} \bar{f}_i) + \lambda\Gamma + \sum_{i \in S} \mu_i, \tag{4.34}$$

$$\lambda + \mu_i \geq \hat{f}_i \qquad \forall i \in S, \tag{4.35}$$

$$\lambda \geq 0, \mu_i \geq 0, \qquad \forall i \in S. \tag{4.36}$$

4.4.4 Robust Reformulation of Constraint (4.3)

Constraint (4.3) will be protected against uncertainty if we can obtain the value for κ_i (proportion of deviation of $\tilde{f}_i$ from its nominal value in each period of arrival horizon). This can be realized by including dual feasibility and strong duality conditions as follows:

$$\sum_{j=i}^{T} z_{i,j} = \bar{f}_i + \kappa_i \hat{f}_i \quad \forall i \in S, \tag{4.37}$$

$$\sum_{i \in S} \kappa_i \hat{f}_i = \lambda\Gamma + \sum_{i \in S} \mu_i, \tag{4.38}$$

$$\sum_{i \in S} \kappa_i \leq \Gamma, \tag{4.39}$$

$$0 \leq \kappa_i \leq 1 \quad \forall i \in S, \tag{4.40}$$

$$\lambda + \mu_i \geq \hat{f}_i \quad \forall i \in S, \tag{4.41}$$

$$\lambda \geq 0, \mu_i \geq 0, \quad \forall i \in S. \tag{4.42}$$

We are not aware of other work that leverages the above primal-dual relationship for modeling uncertainty in a particular time period when the budget of uncertainty is defined for the entire

time horizon.

4.4.5 RTCP Models

We develop an RTCP model based on the chosen approach to control the number of RV patients postponed to the next planning horizon discussed in Section 4.3.3. The RTCP model based on considering a fixed number of postponed RV patients is as follows:

$$\min \ (4.1), \tag{4.43}$$

$$\text{subject to} \quad (4.2) \tag{4.44}$$

$$\sum_{j=i}^{T} z_{i,j} = \bar{f}_i + \kappa_i \hat{f}_i, \quad \forall i \in S, \tag{4.45}$$

$$(4.5) - (4.14), \tag{4.46}$$

$$\sum_{z_{i,j} \in L^{100}} z_{i,j} = (\sum_{i \in S} \bar{f}_i) + \lambda\Gamma + \sum_{i \in S} \mu_i, \tag{4.47}$$

$$\lambda + \mu_i \geq \hat{f}_i \quad \forall i \in S, \tag{4.48}$$

$$\sum_{i \in S} \kappa_i \hat{f}_i = \lambda\Gamma + \sum_{i \in S} \mu_i, \tag{4.49}$$

$$\sum_{i \in S} \kappa_i \leq \Gamma, \tag{4.50}$$

$$0 \leq \kappa_i \leq 1 \quad \forall i \in S, \tag{4.51}$$

$$\lambda \geq 0, \ \ \mu_i \geq 0, \quad \forall i \in S, \tag{4.52}$$

$$(4.18) - (4.23). \tag{4.53}$$

In the following, we will provide a numerical study to determine the level of budget of uncertainty to protect tactical capacity planning against demand uncertainty.

4.5 Experimental Results

The RO approach of Bertsimas and Sim (2003) hedges against uncertainty such that the proposed solution is feasible for a given budget of uncertainty. The issue is that setting the level for the budget of uncertainty has to be done prior to the realization of uncertainty. Thus,

we need to answer the following two questions: 1) How to ensure the feasibility of RTCP under all demand realizations? 2) How to find a balance between the level of protection of robust TCP and additional cost of robust TCP? In order to answer these questions, we will provide some numerical results for the implementation of both the deterministic and the robust model in IBM ILOG CPLEX Optimization Studio V12.6.2.0 on Lenovo ThinkPad X260 Corei7 2.60 GHz.

4.5.1 Data

In order to answer the two questions, we set up experiments in which $\bar{f}_i$ is based on the data from Nguyen et al. (2015). The data set used by Nguyen et al. (2015) include the weekly demand of FV patients over 52 weeks in *arrival horizon* which should be scheduled in 82 weeks in *planning horizon*. The data set includes available physician time in minutes in each week of the planning horizon. For our experiments, we take the nominal demand values from Appendix A2.2 of Nguyen (2014); these values are the actual realizations of demand from year 2009 and 2010 in an outpatient (urology) clinic from Tan Tock Seng Hospital in Singapore. To generate uncertain data we should define an appropriate interval as an uncertainty set for each uncertain FV demand in each week of planning horizon. Usually the minimum and the maximum observed data are used as the lower and the upper bounds of an uncertainty set (Jalilvand-Nejad et al., 2016). Due to the fact that uncertain parameter is in the right-hand of constraint, we only considered maximum observed demand to find the maximum deviation from nominal demand value. In this study, we calculate $\bar{f}_i$ and $\hat{f}_i$ in three ways, based on yearly, seasonal, and monthly maximum values:

- **Yearly:** We use the maximum and minimum realized FV requests over the *arrival horizon* (52 weeks) from Nguyen et al.'s data.
  - Let $f_{max}^{(52)} = \max_{i\in\{1,\dots,52\}} f_i$ and $f_{min}^{(52)} = \min_{i\in\{1,\dots,52\}} f_i$.
  - To calculate the nominal value, we set $\bar{f}_i^{(52)} = \frac{f_{max}^{(52)} + f_{min}^{(52)}}{2}$ for all $i = 1, \dots 52$.

- To calculate the maximum deviation from the nominal, we find $\hat{f}_i^{(52)} = f_{max}^{(52)} - \bar{f}_i$ for all $i = 1, \ldots, 52$.

- **Seasonal:** We use the maximum and minimum realized FV requests over each 13-week period from Nguyen et al.'s data.
  - Let $h_{max,j}^{(13)} = \max_{i \in \{1+13j,\ldots,13+13j\}} f_i$ and $h_{min,j}^{(13)} = \min_{i \in \{1+13j,\ldots,13+13j\}} f_i$ for each season $j = 0, \ldots, 3$.
  - To calculate the nominal value, we find $\bar{h}_j^{(13)} = \frac{h_{max,j}^{(13)} + h_{min,j}^{(13)}}{2}$ for $j = 0, \ldots, 3$.
  - To calculate the maximum deviation from the nominal, we find $\hat{h}_j^{(13)} = h_{max,j}^{(13)} - \bar{h}_j^{(13)}$ for all $j = 0, \ldots, 3$.
  - We set $\bar{f}_i^{(13)} = \bar{h}_j^{(13)}$ and $\hat{f}_i^{(13)} = \hat{h}_j^{(13)}$ whenever $i \in \{1 + 13j, \ldots, 13 + 13j\}$ for $j = 0, \ldots, 3$.

- **Monthly:** We use the maximum and minimum realized FV requests over each 4-week period from Nguyen et al.'s data.
  - Let $h_{max,j}^{(4)} = \max_{i \in \{1+4j,\ldots,4+4j\}} f_i$ for $j = 0, \ldots, 12$ and $h_{min,j}^{(4)} = \min_{i \in \{1+4j,\ldots,4+4j\}} f_i$ for $j = 0, \ldots, 12$.
  - To calculate the nominal value, we find $\bar{h}_j^{(4)} = \frac{h_{max,j}^{(4)} + h_{min,j}^{(4)}}{2}$ for $j = 0, \ldots, 12$.
  - To calculate the maximum deviation from the nominal, we find $\hat{h}_j^{(4)} = h_{max,j}^{(4)} - \bar{h}_j^{(4)}$ for all $j = 0, \ldots, 12$.
  - We set $\bar{f}_i^{(4)} = \bar{h}_j^{(4)}$ and $\hat{f}_i^{(4)} = \hat{h}_j^{(4)}$ whenever $i \in \{1+4j, \ldots, 4+4j\}$ for $j = 0, \ldots, 12$.

Table 4.1 presents the bound values for the three types of uncertainty sets defined above. Due to the fact that the uncertain parameter is in the right-hand side of the constraint, we consider only the maximum observed demand to find the maximum deviation from nominal demand value. We employ the uncertainty set bounds to randomly generate demand scenarios. For example, to generate one instance of a problem based on uncertainty set $[\bar{f}_i^{(52)}, \bar{f}_i^{(52)} + \hat{f}_i^{(52)}]$,

we draw a random sample from Uniform$[\bar{f}_i^{(52)}, \bar{f}_i^{(52)} + \hat{f}_i^{(52)}]$ for each i. We note that the above three methods of calculating uncertainty set bounds induce different levels of variability, as the interval $[\bar{f}_i^{(52)}, \bar{f}_i^{(52)} + \hat{f}_i^{(52)}]$ is wider than $[\bar{f}_i^{(13)}, \bar{f}_i^{(13)} + \hat{f}_i^{(13)}]$, which is in turn wider than $[\bar{f}_i^{(4)}, \bar{f}_i^{(4)} + \hat{f}_i^{(4)}]$, for each $i = 1, \ldots, 52$.

Table 4.1: Values of annual, seasonal and monthly based uncertainty sets.

Weeks			**Uncertainty set**		
Year	Season	4weeks	Bounds/ year	Bounds/ season	Bounds/4 weeks
w1-w52	w1-w13	w1-w4	$[181 \pm 67]$	$[162 \pm 47]$	$[157 \pm 43]$
		w5-w8			$[186 \pm 23]$
		w9-w12			$[186 \pm 22]$
		w13-w16			$[171 \pm 17]$
	w14-w26	w17-w20		$[189 \pm 38]$	$[185 \pm 34]$
		w21-w24			$[194 \pm 13]$
		w25-w28			$[180 \pm 47]$
	w27-w39	w29-w32		$[190 \pm 58]$	$[181 \pm 34]$
		w33-w36			$[218 \pm 30]$
		w37-w40			$[198 \pm 47]$
	w40-w52	w41-w44		$[200 \pm 45]$	$[213 \pm 32]$
		w45-w48			$[189 \pm 17]$
		w49-w52			$[186 \pm 31]$

4.5.2 Results

In this section, we provide the results of the experiment for the RTCP model presented in Section 4.4.5, which includes two stages. In the first stage, we run the DTCP model with the demand scenario used by Nguyen et al. (2015) to find the value for Ψ (number of RV patients postponed to $T+1$). We find this number to be 12929 patients; we set $\Psi = 12929$ for all experiments. In practice, this number should be determined in consultation with the clinic; an alternative approach is discussed in Section 4.6.

The two main goals of this experiment are analyzing the probability of infeasibility of the robust optimal solution to control over-conservatism and evaluating the price of robustness,

which is defined as the trade-off between cost and feasibility of the robust optimal solution.

Trade-off Between Infeasibility Probability of Robust Optimal Solution and Level of Conservatism To calculate the empirical infeasibility probability of RO solution, we evaluate the feasibility of robust optimal solution for different level of robustness (conservatism) in the presence of randomly generated demand. To do so, we apply Monte-Carlo simulation. Therefore, for all the values of budget of uncertainty from [0, 52], we randomly generate 200 scenarios for uncertain demand from the uniform distributions (**Sanei Bajgiran et al., 2017**; **Mirahmadi Shalamzari, 2018**) of the three defined uncertainty sets $[\bar{f}_i^{(52)}, \bar{f}_i^{(52)} + \hat{f}_i^{(52)}]$, $[\bar{f}_i^{(13)}, \bar{f}_i^{(13)} + \hat{f}_i^{(13)}]$, and $[\bar{f}_i^{(4)}, \bar{f}_i^{(4)} + \hat{f}_i^{(4)}]$, for each i. The generated scenarios are the simulated values of demand. The proposed RTCP model becomes infeasible if the available physician time is not sufficient to meet generated random demand. In other words, for each demand scenario the optimal solution corresponding to physician capacity from the RTCP model is inserted as a parameter into the DTCP model where f_i is replaced by generated demand scenarios. Thereafter, we calculate the empirical probability of RO solution infeasibility by dividing the number of infeasible instances by 200 (the total number of simulated demand scenarios for each uncertainty set type). Therefore, we solve the deterministic model $200 \times 52 \times 3 = 31200$ times to find the trade-off between infeasibility probability and level of conservatism (robustness).

Figure 4.2 represents the impact of the budget of uncertainty on the magnitude of infeasibility for the robust solution for the 200 generated demand scenarios for the three types of uncertainty sets. Figure 4.2 shows that the budget of uncertainty required to ensure feasibility is 35, 28, and 24 for the 52-week, 13-week and 4-week uncertainty sets.

In addition to the budget of uncertainty, another element that could control over-conservatism in practice is the subset of data (which in our case corresponds to some period of time) used to determine the bounds of the uncertainty set. We considered three periods over which the bounds were estimated from the data, i.e., 52 weeks, 13 weeks and 4 weeks.

Since $\hat{f}_i^{(52)} \geq \hat{f}_i^{(13)}$ and $\hat{f}_i^{(52)} \geq \hat{f}_i^{(4)}$ for every i, we see that a higher budget of uncertainty is required to ensure feasibility in the case when the uncertainty set is estimated based on all 52 weeks. If the variation of demand is, for instance, highly seasonal (e.g., substantially lower demand values in the summer as opposed to winter), then the results based on $\hat{f}_i^{(52)}$ will be unnecessarily conservative.

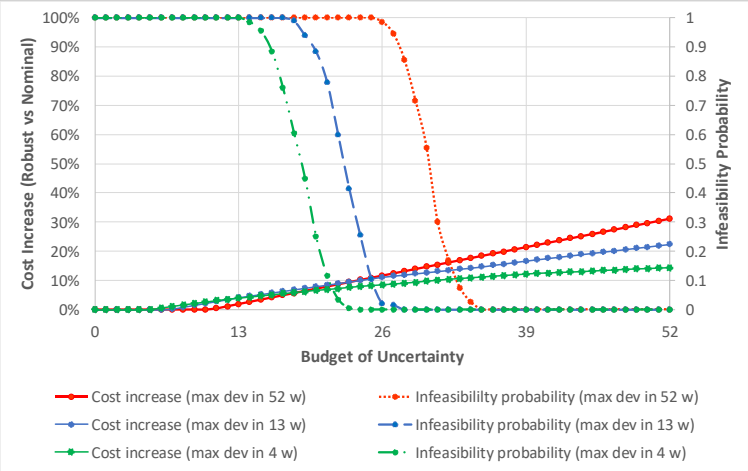


Figure 4.2: Cost of Robustness and Infeasibility Probability for RO Solution.

Price of Robustness The price of robustness presents the trade-off between the additional cost and the feasibility of the robust solution for different budgets of uncertainty. The extra cost of robust solution is calculated as $\frac{q^R - q^N}{q^N}$, where q^R is the objective value of robust optimal solution and q^N is the objective value for nominal optimal solution. Figure 4.2 presents the price of robustness in terms of different budgets of uncertainty.

We can obtain from Figure 4.2 the penalty of being fully confident about feasibility of robust solution. It can be seen that when demand variability is based on the uncertainty set from narrower interval (i.e., uncertainty sets based on monthly deviation), the penalty imposed by the robust model is at most 8.05%. However, when demand variability is based on the uncertainty set from wider interval, the system should pay a higher penalty to assure that all patient demand can be scheduled, i.e., 11.91% based on seasonal variation and

18.42% based on yearly variation. Furthermore, we can again interpret the above results as follows: if the variation of demand is in reality highly seasonal (e.g., substantially lower demand values in the summer as opposed to winter) but the uncertainty set was estimated based on 52-week data (i.e., using $\bar{f}_i^{(52)}$ and $\hat{f}_i^{(52)}$), then an unnecessary cost of about 6.51% is incurred.

Robust Solution vs. Worst-Case Solution We compare the objective value of the robust problem (q^{RO}) for the budget of uncertainty for which the infeasibility probability becomes 0 with that of the worst case deterministic model (q^{wc}). Table 4.2 shows that q^{wc} - $q^{RO}_{avg} \geq 0$ for all the three types of uncertainty sets, implying that the robust solution protects fully against uncertainty at a lower cost than the worst-case solution.

	Uncertainty Set Interval		
Objective value	4w	13w	52w
q^{wc}	130.262	139.290	149.330
q^{RO}	122.996	127.397	134.800

Table 4.2: Robust TCP vs. worst-case TCP performance for various uncertainty sets.

Identification of the Critical Time Periods that Contribute to the Worst-Case Physician Peak Load The objective of this analysis is finding the periods where realization of demand uncertainty will lead to the worst possible maximum physician peak load. The solution generated from the RO model provides a schedule that performs well even when the realized demand $\tilde{f}_i$ is greater than its nominal value $\bar{f}_i$. Having the knowledge of which time periods are the critical ones can help decision makers in scheduling patients. In particular, we can identify the critical weeks in the planning horizon that lead to the worst-case physician peak load by finding the value of κ_i in constraint (4.43). As defined in Section 4.4.5, κ_i is the proportion of deviation of $\tilde{f}_i$ from its nominal value $\bar{f}_i$ in each period of arrival horizon. Therefore, the periods with non-zero κ_i values are the critical ones.

Figures 4.3–4.5 show the critical time periods for the 4-week-based uncertainty set. When the budget of uncertainty Γ is equal to 4, the selected critical weeks by the RTCP model are weeks 25, 28, 37 and 38. We think that this behaviour is due to the fact that the maximum $\hat{f}_i^{(4)}$ for all $i = 1, \ldots, 52$ (see Table 4.1) is 48, which is realized over weeks 25–28 and 37–40. For $\Gamma = 24$, six of the 4-week intervals with highest $\hat{f}_i^{(4)}$ are selected, which are weeks 1–4, 17–21, 25–28, 29–32, 37–40 and 41–44. For $\Gamma = 35$, the additional selected critical weeks compared to $\Gamma = 24$ are weeks 6–8, 33–36 and 49–52.

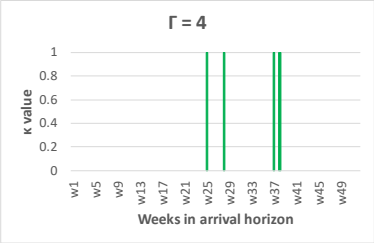


Figure 4.3: κ Values for 4-Week-Based Uncertainty Set & $\Gamma = 4$.

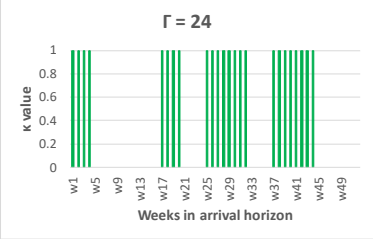


Figure 4.4: κ Values for 4-Week-Based Uncertainty Set & $\Gamma = 24$.

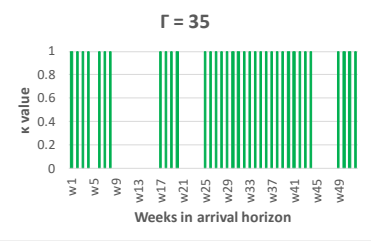


Figure 4.5: κ Values for 4-Week-Based Uncertainty Set & $\Gamma = 35$.

Figures 4.6–4.8 show the critical time periods for the 13-week-based uncertainty set. When $\Gamma = 4$, the critical weeks selected by the RTCP model are 27, 29, 34, 35 and 37. The values of κ for weeks 35 and 37 are fractional. The reason for choosing these weeks as critical is that the maximum $\hat{f}_i^{(13)}$ for all $i = 1, \ldots, 52$ (see Table 4.1) is 58, which is realized over one season in weeks 27–39. For $\Gamma = 24$, the critical weeks are chosen from the two seasons with highest $\hat{f}_i^{(13)}$. Since 13 is not a factor of 24, κ values for some weeks are fractional. For $\Gamma = 35$, the critical weeks are from all the weeks in the two seasons with highest $\hat{f}_i^{(13)}$ and nine weeks of the season with third highest $\hat{f}_i^{(13)}$. As observed in Figure 4.8, all the κ values are integer, although 13 is not a factor of 35. We conjecture that this behaviour is due to the fact that for seasonal-based uncertainty set, infeasibility probability of the robust solution becomes zero for $\Gamma = 28 < 35$, and that therefore the model does not need to be

as "careful" in choosing the critical periods once the probability of infeasibility becomes 0; similarly, we suspect there are multiple equivalent solutions with different κ values after the budget of uncertainty is high enough to ensure no infeasibility.

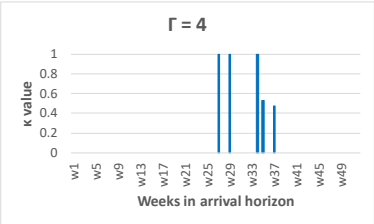


Figure 4.6: κ values for seasonal based uncertainty set & $\Gamma = 4$.

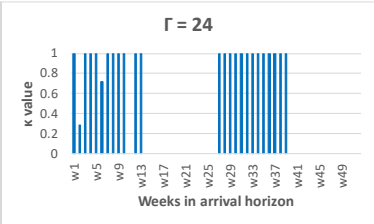


Figure 4.7: κ Values for 13-Week-Based Uncertainty Set & $\Gamma = 24$.

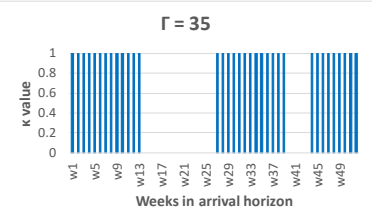


Figure 4.8: κ Values for 13-Week-Based Uncertainty Set & $\Gamma = 35$.

Figures 4.9–4.11 depict the critical time periods for 52-week-based uncertainty set. Due to the fact $\hat{f}_i^{(52)}$ for all $i = 1, \ldots, 52$ is the same, the selected critical weeks are not based on the week with highest possible demand variation. As seen in Figure 4.9, for $\Gamma = 4$ the weeks toward the end of arrival horizon are chosen due to the fewer remaining periods to allocate available physician time to demand as well as limited physician time availability in each period. Thereafter, for $\Gamma = 24$, the weeks from the beginning of arrival horizon are also chosen, in addition to the weeks at the end of arrival horizon.

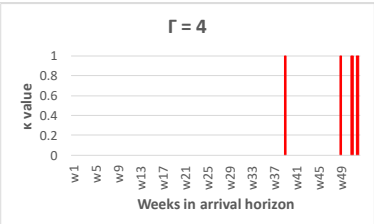


Figure 4.9: κ values for yearly based uncertainty set & $\Gamma = 4$.

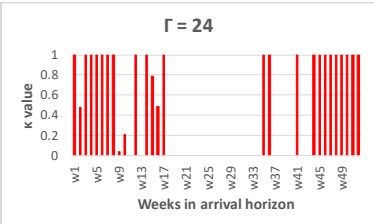


Figure 4.10: κ values for yearly based uncertainty set & $\Gamma = 24$.

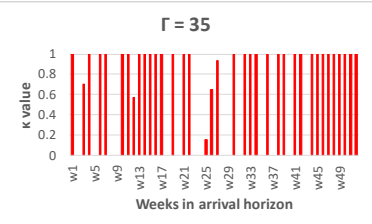


Figure 4.11: κ values for yearly based uncertainty set & $\Gamma = 35$.

As observed from Figures 4.3–4.11, identifying the critical weeks that lead to worst-case

physician workload is non-trivial and would be very challenging without solving a robust optimization model.

Frequency of critical time periods that leads to worst-case physician peak load Figures 4.12–4.14 depict the frequency of κ values over the weeks in the arrival horizon for the three defined uncertainty sets. Having knowledge of the most critical weeks can provide some direction to schedulers to avoid congestion.

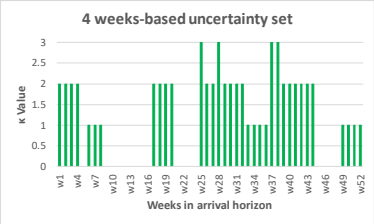


Figure 4.12: Frequency of κ for 4-Week-Based Uncertainty Set & $\Gamma = 4, 24, 35$.

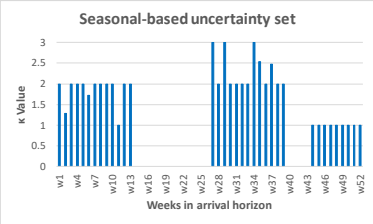


Figure 4.13: Frequency of κ for 13-Week-Based Uncertainty Set & $\Gamma = 4, 24, 35$.

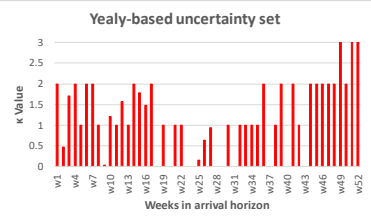


Figure 4.14: Frequency of κ for 52-Week-Based Uncertainty Set & $\Gamma = 4, 24, 35$.

4.6 Discussion

Here, we further discuss three key observations from our study, namely the need to explicitly model the existence of a subsequent planning horizon even in a static model, the importance of considering the effect of uncertainty in particular time periods, and the necessity for careful translation of data into uncertainty sets.

Consideration of the Next Planning Horizon First, in developing a static model for TCP, we need to explicitly consider the number of patients that may have to be postponed to the next planning horizon, since in reality the static model will be used in a dynamic setting. Above, we have shown that if a guideline for the maximum number of patients that could be postponed to the next planning horizon is either known or can be estimated from

past data, then we recommend the use of our RTCP model. It is possible, however, that such a guideline is not available or cannot be easily estimated from the data. In this case, we recommend considering a multi-objective RTCP model which should be calibrated via changing objective function weights to meet the clinic's goals. In particular, we can modify the objective function of both the deterministic model (4.1)–(4.23) and the robust model (4.43)–(4.53) to $\min h_1 q + h_2 \tau^r \sum_{i \in \mathcal{T}} y_{i,T+1}$, while removing the constraint (4.23). The two terms h_1 and h_2 would represent the penalty cost associated with each unit of extra required physician time and each extra RV patient postponed to period $T+1$, respectively. Thus, in this model Ψ is a variable and one can evaluate the trade-off between the maximum capacity for the current time period and Ψ.

Timing of Worst-Case Realizations Second, we have shown, as expected, that if uncertainty in demand is not considered, then the calculated physician capacity is an underestimate of what is actually needed to achieve the required access time targets. We note that in our problem, the consideration of the worst-case realization of demand is done with respect to time; that is, our budget of uncertainty controls how many periods in the arrival horizon are subject to uncertainty, and the solution is based on identifying those periods in which achieving the highest potential demand would have the most impact on the maximum capacity required. Our results show that identifying such time periods without a robust model would have been difficult, as the determination of the timing of the worst-case realizations is dependent on the definition of the uncertainty sets and the budget of uncertainty, and since the allocation of appointments is a complex process that requires consideration of access time targets. For instance, the worst-case realizations for $\Gamma = 4$ with yearly-based uncertainty sets occurs at the end of the arrival horizon (see Figure 4.9), since there is a small remaining number of periods in that planning horizon (and there is a constraint on how many patients can be allocated to the subsequent one); yet at other times, such as for $\Gamma = 24$ with yearly-based uncertainty sets (see Figure 4.10), the worst-case realizations can

in addition occur in the beginning of the arrival horizon since such realizations have an effect on the entire planning horizon.

Translation of Data into Uncertainty Sets The third observation is that the step of translating available data into an uncertainty set is important and should be carefully considered as it can lead to substantially different results and costs. In fact, the translation step can influence the true conservatism of a solution – an over-estimate or under-estimate in the uncertainty set definition will lead to an over-estimate or an under-estimate, respectively, of the resulting capacity and cost. For example, if the variation of demand is in reality highly seasonal (e.g., substantially lower demand values in the summer as opposed to winter) but the uncertainty set is estimated based on 52 weeks of data, then in some of our experiments an unnecessary cost of about 6.51% is incurred. Therefore, it is important to understand the given data prior to the development of the uncertainty set as well as to validate the definitions of the uncertainty set with the stakeholders prior to, as well as following, the optimization process.

4.7 Conclusion

In this study, we developed a robust tactical capacity planning model via cardinality constrained robust optimization which explicitly considers the number of patients who may need to be scheduled in a subsequent planning horizon. The robust tactical capacity planning model protects against uncertainty in the demand per time period. Due to the presence of the uncertain demand parameter in the right-hand side of two constraints, i.e., individual demand per period and aggregated demand over all the periods in the arrival horizon, formulating the robust tactical planning model required the use of both primal and dual constraints in the robust model. By employing cardinality-constrained robust optimization, we showed how to control over-conservatism through analyzing the trade-off between the budget of uncertainty and feasibility of the robust plan based on different bounds of the un-

certainty set. We also analyzed the price of robustness, i.e., the trade-off between feasibility and cost of robust plan for different budgets of uncertainty and bounds of uncertainty set. The findings of this study provides an insight for decision makers how a chosen budget of uncertainty and bounds of uncertainty set impact feasibility and the cost of tactical capacity plan. We conducted a set of experiments to compute the feasibility and cost of robust solutions for different budgets of uncertainty and different methods of estimating the uncertainty sets. We showed that we can guarantee 100% feasibility of a robust tactical capacity plan while not being fully conservative, which will lead to cost savings for the clinic while being able to meet demand despite uncertainty. We also show how the robust model helps us to identify the critical time periods (weeks) which contribute to worst case physician peak load. Having the knowledge of critical time periods gives insight to the decision-makers about how to avoid congestion due to demand uncertainty.

Chapter 5

Conclusions and Future Work

In this final chapter, we summarize the work presented in the previous chapters, re-state the major contributions of this dissertation and state directions for future work.

5.1 Summary and Contributions

The thesis focuses on a subject which has gained a lot of recent interest and funding investments, specifically, the organization and delivery of primary care. This subject has been discussed and analyzed mainly from the health policy but not from the operations research perspective. In this thesis, on the contrary, we focus on the use of operations research methods, including forecasting and optimization, to improve access to primary care. The major contributions of this study can be classified into two parts.

Part 1: How should primary care access performance be evaluated?

- We make a step toward the development of a preliminary framework for prioritization of patients based on their acuity in primary care with the goal of taking *equity* into account in performance evaluation.
- We are the first to propose an acuity-based, data driven metric for evaluation of timely access to primary care that leads to more equitable and efficient allocation of physician time.

Part 2: Given a representative performance metric, how can we improve access to primary care through

- Being the first to develop a capacity reservation plan for an individual physician in primary care for all urgent and multiple non-urgent appointment types (such

as chronic disease management and routine appointments) to address the trade-off between providing equitable access and efficient utilization of physician time. This approach consists of a forecasting model to determine the demand and an optimization model that allocates time slots to patient requests of different priority. The developed tactical capacity plan is based on an mixed integer linear model which determines the optimal number of appointment slots to reserve per week for each appointment type during a 12-week planning horizon. The model also tracks the unmet demand for each appointment type per week which should be scheduled in the subsequent planning horizon. Our approach results in developing a weekly planning template for a specific physician in a family health team taking into account the needs and preferences of both patients and family physicians.

- Developing a robust TCP model based on the cardinality-constrained method to minimize the highest potential physician peak load between weeks for the case when we may not be able to forecast demand accurately. Therefore, the developed robust TCP model enables protection against uncertainty through providing a feasible allocation of capacity for all realizations of demand. The proposed robust TCP model considers two interdependent appointment types (e.g., new patients and follow ups), multiple access time targets for each appointment type and uncertainty in demand for appointments. We conduct a set of experiments to determine how to set the level of robustness based on extra cost and infeasibility probability of a robust solution

5.2 Future work

Future work stemming from this thesis can be divided into research extensions and implementation plan.

5.2.1 Research extensions

We plan to extend our research as follows:

- Developing weekly planning template for all the care providers in an interdisciplinary primary care setting that includes resource pooling

    1. Optimal # of time slots reserved based on the demand forecast and data driven access time distribution for each appointment type/week for a specific care provider for a determined planning horizon

    2. Assigning an exact time slot per half day of a specific physician to each appointment type.

- Developing online booking system based on the designed weekly planning template for all the care providers.

- Extending the current definition of acuity levels in this study based on both access target and tolerance limit (Gocgun and Puterman, 2014). As an example for some cases of follow-up appointment type such as modification in medication dose, the appointment should not be scheduled too early for having enough time for the medication to start having an effect on patient health status. [1]

- Meeting patient preference by developing a multi-agent scheduling system. Each party (family physicians, patients) represents an agent and the developed scheduling system makes these scheduling agents to match slack time in physicians' and patients' schedules. In other words, each agent is aware of the preferences and limitations of its owner. The objective of a multi-agent scheduling system is to design the agents and the interaction rules between agents to achieve an effective schedule (Vermeulen et al., 2007). Each agent has the ownership of their own calendars and the agents exchange information among each other with the goal of finding an open time slot to schedule an

[1] Thanks to Dr. Michael Carter for bringing into my attention the necessity of considering tolerance limit.

appointment. Finally, agents will negotiate multiple available time slots based on the designed interaction rules to meet both patient and physician preferences (Crawford and Veloso, 2004).[2]

5.2.2 Implementation plan

The focus of the implementation plan is on developing policies and tools for implementation of equitable scheduling of appointments that takes into account the acuity level of patient requests, described below in terms of a three-stage plan.

- **Stage 1:** Acuity-based access time evaluation at the Health for All (HFA) clinic.
    1. Assigning acuity levels for patient conditions
        (a) Meet (potentially multiple times) with physicians to present and discuss the plan of action and preliminary findings.
        (b) Survey physicians regarding willingness to participate, as well as their current perceptions regarding patient conditions and corresponding acuity levels.
        (c) Perform an experiment: for each patient visit during this period, the participating physicians define a condition and assign an acuity level for all the visits.
        (d) Analyze the data collected from the physicians: all the defined conditions and the assigned acuity levels.
        (e) Synthesize the data in order to define a preliminary patient classification framework for HFA.
        (f) Survey physicians regarding the appropriateness and accuracy of the synthesized results and make any required changes to the evaluation/classification framework.

[2]Thanks to Dr. Ketra Schmitt for bringing the idea of multi-agent scheduling systems to my attention.

2. Develop and validate a checklist for administrative staff to determine the urgency of the current request (note: currently, the administrative staff already does this but our proposal is to formalize this approach through standardized questions and a checklist, combined with our acuity-based framework for primary care).
3. Train the administrative staff regarding the defined acuity levels and appointment types.
4. Perform a pilot study to see the challenges for schedulers to assign acuity levels based on the defined appointment types.
5. Meet with physicians to discuss the revision of appointment types, questions and checklists based on any challenges reported by the administrative staff with regards to the new system.
6. Finalize the list of patient conditions, the corresponding acuity level and the questions/checklist to be employed by the administrative staff.

Stage 2: Development of appointment scheduling policies.

1. Meet (potentially multiple times) with physicians to present and discuss characteristics of good and bad appointment scheduling policies.
2. Propose multiple candidate scheduling policies
   - Propose "rule of thumb" policies based on discussions with physicians.
   - Develop and test optimization models to develop "optimal" policies.
   - Translate solutions to optimization models into implementable policies.
3. Build a simulation model that will allow for evaluation of the effectiveness of various appointment scheduling policies on past data.
4. For each candidate policy determined in step 2 of the current stage, the simulation will determine the percentage of time that access time targets are met for each acuity class defined during Stage 1, as well as on how balanced

the resulting schedules would be for physicians.

5. The top policies will be presented to the clinical and administrative staff of the clinic, who will evaluate the adequacy of the policies and the feasibility of their implementation in practice. Any new suggestions or adjustments will be again validated using simulation. Finally, the best implementable policy will be chosen.
6. Train administrative staff to implement the proposed policy in practice.
7. Collect data from the pilot test and evaluate true effectiveness; go back to a previous step (redesign of the policy, consultation with staff, etc.) if an improvement in access times is not obtained.

Stage 3: Identify a plan for periodic review of the effectiveness of the policies and continuous improvement.

5.3 Conclusion

Motivated by the major concern in the Canadian primary care clinics which is timely and equitable access to care due to family physicians scarcity, the central thesis of this dissertation is based on the two following questions: 1) How should primary care access performance be evaluated? and 2) Given a representative performance metric, how to improve access to primary care thorough operations research methods, including forecasting and optimization.

The first manuscript addresses how to evaluate primary care access performance. It focuses on developing a data-driven assessment tool to objectively evaluate equitable access to care. In particular, we develop a preliminary guideline to translate clinical severity of requested appointments into five acuity levels in primary care. These five acuity levels prioritize patients based on the relative access time target to the clinical severity of requested appointment.

The second and third manuscripts address how to improve access through developing tactical capacity planning (TCP) models. The second manuscript presents the contributions

of a TCP model for a specific physician in a family health team which results in a number of reserved physician time slots per week for all the 11 considered appointment types before demand realization based on demand forecast. The developed decision model is considered as a tool to provide equitable timely access through considering the developed guideline in the first manuscript. The developed TCP model in the second manuscript improve access for the optimal physician workload balance between weeks. The third manuscript presents the contribution of a TCP model when demand is uncertain to improve access for the optimal worst case physician peak load.

Bibliography

Bovas Abraham and Johannes Ledolter. *Statistical methods for forecasting*, volume 234. John Wiley & Sons, 2009.

Bernardetta Addis, Giuliana Carello, Andrea Grosso, Ettore Lanzarone, Sara Mattia, and Elena Tànfani. Handling uncertainty in health care management using the cardinality-constrained approach: Advantages and remarks. *Operations Research for Health Care*, 4: 1–4, 2015.

Amir Ahmadi-Javid, Zahra Jalali, and Kenneth J Klassen. Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, 258(1):3–34, 2017.

Nazanin Aslani, Fariborz Fazileh, Donatus Mutasingwa, and Daria Terekhov. Acuity-based access time evaluation in primary care: a case study of an Ontario clinic. In *Proceedings of the 4th International Health Care Systems Engineering Conference (HCSE'19)*, 2019.

Hari Balasubramanian, Ritesh Banerjee, Brian Denton, James Naessens, and James Stahl. Improving clinical access and continuity through physician panel redesign. *Journal of General Internal Medicine*, 25(10):1109–1115, 2010.

Hari Balasubramanian, Ana Muriel, and Liang Wang. The impact of provider flexibility and capacity allocation on the performance of primary care practices. *Flexible Services and Manufacturing Journal*, 24(4):422–447, 2012.

Hari Balasubramanian, Sebastian Biehl, Longjie Dai, and Ana Muriel. Dynamic allocation of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments. *Health care management science*, 17(1):31–48, 2014.

Irad Ben-Gal. Outlier detection. In *Data Mining and Knowledge Discovery Handbook*, pages 131–146. Springer, 2005.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

D Bertsimas and M Sim. Robust discrete optimization and network flows. *Mathematical programming*, 98(1):49–71, 2003.

Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations research*, 52(1): 35–53, 2004.

Daniel Bienstock. Histogram models for robust portfolio optimization. *Journal of computational finance*, 11(1):1, 2007.

John R Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

Asaf Bitton, Hannah L Ratcliffe, Jeremy H Veillard, Daniel H Kress, Shannon Barkley, Meredith Kimball, Federica Secci, Ethan Wong, Lopa Basu, Chelsea Taylor, et al. Primary health care as a foundation for strengthening health systems in low-and middle-income countries. *Journal of General Internal Medicine*, 32(5):566–571, 2017.

Lynn A Blewett, Pamela Jo Johnson, Brian Lee, and Peter B Scal. When a usual source of care and usual provider matter: adult prevention and screening services. *Journal of general internal medicine*, 23(9):1354, 2008.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Canadian Association of Emergency Physicians. The Canadian Triage & Acuity Scale (CTAS). http://ctas-phctas.ca/?page_id=17. Accessed March 2nd, 2019.

Canadian Institute for Health Information. Commonwealth fund survey 2016: Chartbook. Technical report, 2017.

Suresh Chand, Herbert Moskowitz, John B Norris, Steve Shade, and Deanna R Willis. Improving patient flow at an outpatient clinic: study of sources of variability and improvement factors. *Health care management science*, 12(3):325–340, 2009.

Dimitri A Christakis, Jeffrey A Wright, Frederick J Zimmerman, Alta L Bassett, and Frederick A Connell. Continuity of care is associated with well-coordinated care. *Ambulatory Pediatrics*, 3(2):82–86, 2003.

CIHI. National health expenditure trends, 1975 to 2018. *Canadian Institute for Health Information*, 2018.

Commonwealth Fund. International profiles of health care systems. Technical report, 2017.

Elisabeth Crawford and Manuela Veloso. Opportunities for learning in multi-agent meeting scheduling. In *Proceedings of the AAAI 2004 Symposium on Artificial Multiagent Learning, Washington, DC*, 2004.

Brian T Denton, Andrew J Miller, Hari J Balasubramanian, and Todd R Huschka. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations research*, 58(4-part-1):802–816, 2010.

Julien Déry, Angel Ruiz, François Routhier, Marie-Pierre Gagnon, André Côté, Daoud Ait-Kadi, Valérie Bélanger, Simon Deslauriers, and Marie-Eve Lamontagne. Patient prioritization tools and their effectiveness in non-emergency healthcare services: a systematic review protocol. *Systematic Reviews*, 8(1):78, 2019.

Gregory Dobson, Sameer Hasija, and Edieal J Pinker. Reserving capacity for urgent patients in primary care. *Production and Operations Management*, 20(3):456–473, 2011.

Marcello D'Orazio. *univOutl: Detection of Univariate Outliers*, 2018. URL https://CRAN.R-project.org/package=univOutl. R package version 0.1-4.

Glyn Elwyn, Wendy Jones, Melody Rhydderch, and Peter Edwards. Developing a measure of patient access to primary care: the access response index (aros). *Journal of evaluation in clinical practice*, 9(1):33–37, 2003.

Shirin Geranmayeh. *Optimizing surgical scheduling through integer programming and robust optimization*. PhD thesis, Université d'Ottawa/University of Ottawa, 2015.

Steven Globerman, Bacchus Barua, and Sazid Hasan. The supply of physicians in Canada: Projections and assessment. http://www.fraserinstitute.org, 2018.

Yasin Gocgun and Martin L Puterman. Dynamic scheduling with due dates and time windows: an application to chemotherapy patient appointment booking. *Health care management science*, 17(1):60–76, 2014.

Anna Graber-Naidich. *Operations Research Methodologies to Improve the Quality, Accessibilty and Equity of Primary Care*. PhD thesis, Department of Mechanical and Industrial Engineering, University of Toronto, 2015.

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

Diwakar Gupta and Lei Wang. Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research*, 56(3):576–592, 2008.

David C Hador, Steering Committee of the Western Canada Waiting List Project, et al. Setting priorities for waiting lists: defining our terms. *Cmaj*, 163(7):857–860, 2000.

Jeannie Haggerty, Fred Burge, Jean-Frédéric Lévesque, David Gass, Raynald Pineault, Marie-Dominique Beaulieu, and Darcy Santor. Operational definitions of attributes of primary health care: consensus among Canadian experts. *The Annals of Family Medicine*, 5(4):336–344, 2007.

Jeannie L Haggerty, Jean-Frédéric Lévesque, Darcy A Santor, Frederick Burge, Christine Beaulieu, Fatima Bouharaoui, Marie-Dominique Beaulieu, Raynald Pineault, and David Gass. Accessibility from the patient perspective: comparison of primary healthcare evaluation instruments. *Healthcare Policy*, 7(Spec Issue):94, 2011.

Wissam Haj-Ali, Brian Hutchison, Primary Care Performance Measurement Steering Committee, et al. Establishing a primary care performance measurement framework for ontario. *Healthcare Policy*, 12(3):66, 2017.

Katherine Harding and Nicholas Taylor. Triage in non-emergency services. In *Patient Flow: Reducing Delay in Healthcare Delivery*, pages 229–250. Springer, 2013.

Öncü Hazır and Alexandre Dolgui. Assembly line balancing under uncertainty: Robust optimization models and exact solution method. *Computers & Industrial Engineering*, 65 (2):261–267, 2013.

Health Quality Ontario. Measuring up 2018. Technical report, 2018.

Healthwise Staff. Nose cautery for nosebleeds: What to expect at home. https://myhealth.alberta.ca/Health/aftercareinformation/pages/conditions.aspx?hwid=abp6135, 2018. Accessed: 2019.

René Henrion. Introduction to chance-constrained programming. *Tutorial paper for the Stochastic Programming Community home page*, 2004.

CC Holt. Forecasting seasonals and trends by exponentially weighted moving averages, onr memorandum (vol. 52), pittsburgh, pa: Carnegie institute of technology. *Available from the Engineering Library, University of Texas at Austin*, 1957.

Mia Hubert and Ellen Vandervieren. An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, 52(12):5186–5201, 2008.

Peter JH Hulshof, Nikky Kortbeek, Richard J Boucherie, Erwin W Hans, and Piet JM Bakker. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in or/ms. *Health systems*, 1(2):129–175, 2012.

Peter JH Hulshof, Richard J Boucherie, Erwin W Hans, and Johann L Hurink. Tactical resource allocation and elective patient admission planning in care processes. *Health care management science*, 16(2):152–166, 2013.

Peter JH Hulshof, Martijn RK Mes, Richard J Boucherie, and Erwin W Hans. Patient admission planning using approximate dynamic programming. *Flexible services and manufacturing journal*, 28(1-2):30–61, 2016.

Brian Hutchison and Richard Glazier. Ontario's primary care reforms have transformed the local care landscape, but a plan is needed for ongoing improvement. *Health affairs*, 32(4): 695–703, 2013.

Brian Hutchison, JEAN-FREDERIC LEVESQUE, Erin Strumpf, and Natalie Coyle. Primary health care in canada: systems in motion. *The Milbank Quarterly*, 89(2):256–288, 2011.

ILOG. Solution status codes. https://www.tu-chemnitz.de/mathematik/discrete/manuals/cplex/doc/refman/html/appendixB.html, 2002.

Kenneth V Iserson and John C Moskop. Triage in medicine, part i: concept, history, and types. *Annals of Emergency Medicine*, 49(3):275–281, 2007.

Amir Jalilvand-Nejad, Rasoul Shafaei, and Hamid Shahriari. Robust optimization under correlated polyhedral uncertainty set. *Computers & Industrial Engineering*, 92:82–94, 2016.

Wendy Jones, Glyn Elwyn, Peter Edwards, Adrian Edwards, Melody Emmerson, and Richard Hibbs. Measuring access to primary care appointments: a review of methods. *BMC Family Practice*, 4(1):8, 2003.

Min Young Kim, Ju Heon Kim, Il-Kwon Choi, In Hong Hwang, and Soo Young Kim. Effects of having usual source of care on preventive services and chronic disease control: a systematic review. *Korean journal of family medicine*, 33(6):336, 2012.

Maude Laberge, Jocelyn Pang, Kevin Walker, Sabrina Wong, William Hogg, and Walter P. et al. Wodchis. QUALICOPC (Quality and Costs of Primary Care) Canada: A focus on the aspects of primary care most highly rated by current patients of primary care practices. 2014.

Maude Laberge, Walter P Wodchis, Jan Barnsley, and Audrey Laporte. Costs of health care across primary care models in ontario. *BMC health services research*, 17(1):511, 2017.

Nadia R Llanwarne, Gary A Abel, Marc N Elliott, Charlotte AM Paddison, Georgios Lyratzopoulos, John L Campbell, and Martin Roland. Relationship between clinical quality and patient experience: analysis of data from the english quality and outcomes framework and the national GP patient survey. *The Annals of Family Medicine*, 11(5):467–472, 2013.

Otto R Maarsingh, Ykeda Henry, Peter M van de Ven, and Dorly JH Deeg. Continuity of care in primary care and association with survival in older people: a 17-year prospective cohort study. *Br J Gen Pract*, 66(649):e531–e539, 2016.

Martin Maechler, Christophe Croux Peter Rousseeuw, Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Tobias Verbeke, Manuel Koller, "L. T.") Conceicao c("Eduardo", and Maria Anna di Palma. *robustbase: Basic Robust Statistics*, 2018. URL http://CRAN.R-project.org/package=robustbase. R package version 0.93-3.

Gregory P Marchildon and Brian Hutchison. Primary care in ontario, canada: New proposals after 15 years of reform. *Health Policy*, 120(7):732–738, 2016.

Akram Mirahmadi Shalamzari. Robust multi-class multi-period scheduling of MRI services with wait time targets. Master's thesis, University of Waterloo, 2018.

T.-B.-T Nguyen, A.-I Sivakumar, and S.-C Graves. A network flow approach for tactical resource planning in outpatient clinics. *Health care management science*, 18(2):124–136, 2015.

Thi Thu Ba Nguyen. *Modelling, analysis, and optimization in resource planning for outpatient clinics*. PhD thesis, Nanyang Technological University, 2014.

Thu Ba T Nguyen, Appa Iyer Sivakumar, and Stephen C Graves. Capacity planning with demand uncertainty for outpatient clinics. *European Journal of Operational Research*, 267 (1):338–348, 2018.

TW Noseworthy, JJ McGurran, DC Hadorn, and Steering Committee of the Western Canada Waiting List Project. Waiting for scheduled services in canada: development of priority-setting scoring systems. *Journal of evaluation in clinical practice*, 9(1):23–31, 2003.

John Oldham. *Advanced Access in primary care*. National Primary Care Development Team Manchester, 2001.

Ontario Medical Association. section on general and family practice: Diagnostic codes. https://londonreferral.files.wordpress.com/2018/02/2015-common-family-practice-codes.pdf, 2015. Accessed: 2019.

Asli Ozen and Hari Balasubramanian. The impact of case mix on timely access to appointments in a primary care group practice. *Health care management science*, 16(2):101–118, 2013.

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

Farnaz Haji Pour. *Robust radiotherapy appointment scheduling*. PhD thesis, Concordia University Montreal, Quebec, Canada, 2016.

Kamila Premji. *In our rush to offer "McMedicine", do we even know what patients really want?*, 2018. https://healthydebate.ca/opinions/same-day-access-family-doctor.

Kamila Premji, Bridget L Ryan, William E Hogg, and Walter P Wodchis. Patients' perceptions of access to primary care: Analysis of the QUALICOPC patient experiences survey. *Canadian Family Physician*, 64(3):212–220, 2018.

Xiuli Qu and Jing Shi. Effect of two-level provider capacities on the performance of open access clinics. *Health care management science*, 12(1):99, 2009.

Xiuli Qu, Ronald L Rardin, Julie Ann S Williams, and Deanna R Willis. Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European Journal of Operational Research*, 183(2):812–826, 2007.

Xiuli Qu, Ronald L Rardin, and Julie Ann S Williams. Single versus hybrid time horizons for open access scheduling. *Computers & Industrial Engineering*, 60(1):56–65, 2011.

Xiuli Qu, Ronald L Rardin, and Julie Ann S Williams. A mean–variance model to optimize the fixed versus open appointment percentages in open access scheduling systems. *Decision Support Systems*, 53(3):554–564, 2012.

Xiuli Qu, Yidong Peng, Nan Kong, and Jing Shi. A two-phase approach to scheduling multi-category outpatient appointments–a case study of a women's clinic. *Health care management science*, 16(3):197–216, 2013.

Mala Rao, Aileen Clarke, Colin Sanderson, and Richard Hammersley. Patients' own assessments of quality of primary care compared with objective records based measures of technical quality of care: cross sectional study. *BMJ*, 333(7557):19, 2006.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2016. URL http://www.rstudio.com/.

Omid Sanei Bajgiran, Masoumeh Kazemi Zanjani, and Mustapha Nourelfath. Forest harvesting planning under uncertainty: a cardinality-constrained approach. *International Journal of Production Research*, 55(7):1914–1929, 2017.

Jatinderpreet Singh, Simone Dahrouge, and Michael E Green. The impact of the adoption of a patient rostering model on primary care access and continuity of care in urban family practices in ontario, canada. *BMC family practice*, 20(1):52, 2019.

Allen L Soyster. Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations research*, 21(5):1154–1157, 1973.

Barbara Starfield. Refocusing the system. *New England Journal of Medicine*, 359(20): 2087–2091, 2008.

Barbara Starfield, Jonathan Weiner, Laura Mumford, and Donald Steinwachs. Ambulatory care groups: a categorization of diagnoses for research and management. *Health Services Research*, 26(1):53, 1991.

Barbara Starfield, Leiyu Shi, and James Macinko. Contribution of primary care to health systems and health. *The milbank quarterly*, 83(3):457–502, 2005.

Statistics Canada. Primary health care providers, 2017. https://www150.statcan.gc.ca/n1/pub/82-625-x/2019001/article/00001-eng.pdf, 2019.

Arthur Sweetman and Gioia Buckley. Ontario's experiment with primary care reform. *SPP Research Paper*, (7-11), 2014.

Jiafu Tang and Yu Wang. An adjustable robust optimisation method for elective and emergency surgery capacity allocation with demand uncertainty. *International Journal of Production Research*, 53(24):7317–7328, 2015.

The College of Family Physicians of Canada. Best Advice: Patient rostering in family practice, 2012.

Raaj Tiagi and Yuriy Chechulin. The effect of rostering with a patient enrolment model on emergency department utilization. *Healthcare Policy*, 9(4):105, 2014.

Adrian Trapletti and Kurt Hornik. *tseries: Time Series Analysis and Computational Finance*, 2018. URL https://CRAN.R-project.org/package=tseries. R package version 0.10-45.

Ivan Vermeulen, Sander Bohte, Koye Somefun, and Han La Poutré. Multi-agent pareto appointment exchanging in hospital patient scheduling. *Service Oriented Computing and Applications*, 1(3):185–196, 2007.

Ivan B Vermeulen, Sander M Bohte, Sylvia G Elkhuizen, Han Lameris, Piet JM Bakker, and Han La Poutré. Adaptive resource allocation for efficient patient scheduling. *Artificial intelligence in medicine*, 46(1):67–80, 2009.

Wen-Ya Wang and Diwakar Gupta. Adaptive appointment systems with patient preferences. *Manufacturing & Service Operations Management*, 13(3):373–389, 2011.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2019. URL https://CRAN.R-project.org/package=dplyr. R package version 0.8.3.

Lara Wiesche, Matthias Schacht, and Brigitte Werners. Strategies for interday appointment scheduling in primary care. *Health care management science*, 20(3):403–418, 2017.

Peter R Winters. Forecasting sales by exponentially weighted moving averages. *Management science*, 6(3):324–342, 1960.

Sabrina T Wong, Leena W Chau, William Hogg, Gary F Teare, Baukje Miedema, Mylaine Breton, Kris Aubrey-Bassler, Alan Katz, Fred Burge, Antoine Boivin, et al. An international cross-sectional survey on the quality and costs of primary care (QUALICO-PC): recruitment and data collection of places delivering primary care across Canada. *BMC Family Practice*, 16(1):20, 2015.

Wiesława Dominika Wranik, Sheri Price, Susan M Haydt, Jeanette Edwards, Krista Hatfield, Julie Weir, and Nicole Doria. Implications of interprofessional primary care team characteristics for health services and patient health outcomes: A systematic review with narrative synthesis. *Health Policy*, 2019.