Resource Allocation Models in Healthcare Decision Making

by

Abdelhalim Hiassat

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Management Sciences

Waterloo, Ontario, Canada, 2017

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	MEHMET A. BEGEN Associate Professor
Co-Supervisor	F. SAFA ERENAY Assistant Professor
Co-Supervisor	OSMAN Y. OZALTIN Assistant Professor
Internal Member	JAMES H. BOOKBINDER Professor
Internal Member	QI-MING HE Professor
Internal-external Member	ALI ELKAMEL Professor

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

We present models for allocating limited healthcare resources efficiently among target populations in order to maximize society's welfare and/or minimize the expected costs. In general, this thesis is composed of two major parts.

Firstly, we formulate a novel uncapacitated fixed-charge location problem which considers the preferences of customers and the reliability of facilities simultaneously. A central planner selects facility locations from a set of candidate sites to minimize the total cost of opening facilities and providing service. Each customer has a strict preference order over a subset of the candidate sites, and uses her most preferred available facility. If that facility fails due to a disruptive event, the customer attends her next preferred available facility. This model bridges the gap between the location models that consider the preferences of customers and the ones that consider the reliability of facilities. It applies to many healthcare settings, such as preventive care clinics, senior centers, and disaster response centers. In such situations, patient (or customer) preferences vary significantly. Therefore, there could be a large number of subgroups within the population depending on their preferences of potential facility sites. In practice, solving problems with large numbers of population subgroups is very important to increase granularity when considering diverse preferences of several different customer types. We develop a Lagrangian branch-and-bound algorithm and a branch-and-cut algorithm. We also propose valid inequalities to tighten the LP relaxation of the model. Our numerical experiments show that the proposed solution algorithms are efficient, and can be applied to problems with extremely large numbers of customers.

Secondly, we study the allocation of colorectal cancer (CRC) screening resources among individuals in a population. CRC can be early-detected, and even prevented, by undergoing periodic cancer screenings via colonoscopy. Current guidelines are based on existing medical evidence, and do not explicitly consider (i) all possible alternative screening policies, and (ii) the effect of limited capacity of colonoscopy screening on the economic feasibility of the screening program. We consider the problem of allocating limited colonoscopy capacity for CRC screening and surveillance to a population composed of patients of different risk groups based on risk factors including age, CRC history, etc. We develop a mixed inte-

ger program that maximizes the quality-adjusted life years for a given patient population considering the population's demographics, CRC progression dynamics, and relevant constraints on the system capacity and the screening program effectiveness. We show that the current guidelines are not always optimal. In general, when screening capacity is high, the optimal screening programs recommend higher screening rates than the current guidelines, and the optimal screening policies change with age and gender. This shows the significance of incorporating screening capacity into the decisions of optimal screening policies.

Acknowledgements

Firstly, I would like to express my immense gratitude to my two co-supervisors. Dr. Fatih Safa Erenay has been my biggest support and source of motivation. His vast knowledge and high intelligence are surpassed only by his kindness and courtesy. Dr. Erenay's door was always open whenever I had a question or needed help. While I wish him better students in the future, I certainly could not have imagined having a better supervisor for my doctoral studies. Dr. Osman Ozaltin has been my model scholar for his exceptional knowledge and unparalleled robustness. His clear vision and humbling questions have significantly improved this thesis. Despite being away for the better part of my studies, he always made himself available for my never-ending questions. Both supervisors valued learning, and allowed me to make mistakes. For that, I am ever grateful.

The members of the Examining Committee have improved this thesis and made it more organized and easier to read through their comments and suggestions. I am grateful for their time and effort spent in reading the manuscript, and for showing up for an exam in August.

The courses I have taken (or audited) at the University of Waterloo have been instrumental in my development and in shaping this thesis. I would like to thank the instructors of these courses for making hard concepts easy to understand and introducing me to new topics and ideas: Hossein Abouee Mehrizi, Steve Drekic, Samir Elhedhli, Bon Koo, Osman Ozaltin, Frank Safayeni, Anindya Sen, Levent Tuncel.

I also would like to thank my group-mates who helped in challenging my ideas and generating new ones. The long and frequent talks with Mustafa, Najmaddin, Bahar, Gizem, Onur, Tagi, and Burak have sharpened my understanding and enhanced my knowledge. I am thankful for the joyful times and the anxiety about future plans our meetings have brought. I had enjoyed a practically private office on campus thanks to my amazing office-mate, Ata. When he was around, which was not too often, he was constantly considerate and understanding.

I had the opportunity to be an instructor and a teaching assistant for a number of courses during my studies. I was inspired by my students every time I was in class. I

have learned a lot about myself and about the subjects that I was teaching. Thanks to all of my students, instructors, and teaching assistants that I worked with for making my job interesting and possible. Thanks to Dr. Samir Elhedhli for nominating me to be an instructor, which has both slowed my research progress and significantly enriched my experience.

I am grateful for the support of the staff of the Department of Management Sciences at the University of Waterloo. A special thanks to Wendy Fleming, Lisa Hendel, Shelley Vossen, and Kathy Tytko for all the administrative help throughout the years.

My friends in Waterloo made my life slightly less boring. Tarek has made my first days less scary. Thanks to my roommates Selva, Arty, and Burak for being so neat and clean that I did not want to try my luck a fourth time. Thanks to Gizem and Onur for the delicious food and car rides. Thanks to Ibrahim and Tagi for organizing football games. Also, thanks to Khaled for being a food and cooking partner, a FIFA opponent, and an immigration advisor. I also want to thank my friends in Jordan, UAE, USA, and elsewhere for staying in touch despite the long distance and torturous time zones.

Last but definitely not least, thanks to my family for keeping me sane and teaching me everyday what unconditional love is. Being away from them is always the hardest challenge. I would have never been here without their support and unlimited belief in me. A man cannot wish for more.

Dedication

She always kept chocolate in her room. My brothers realized she would always give if I were the one asking. I always feared the time she wouldn't.

For the trick that never failed... to my mother.

He rarely expressed his feelings. It was that extra tight hug when I was back home from the airport that told everything. It was an untold secret.

For the promise still kept hours before his passing... to my late father.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CB-LBB	Callbacks Implementation of Lagrangian Branch-and-Bound.
CRC	Colorectal Cancer.
CT	Computed Tomography.
DCBE	Double-Contrast Barium Enema.
FIT	Fecal Immunochemical Test.
FOBT	Fecal Occult Blood Test.
FSIG	Flexible sigmoidoscopy.
GDP	Gross Domestic Product.
LBB	Lagrangian Branch-and-Bound.
MDP	Markov Decision Process.
MISCAN	MIcrosimulation SCreening ANalysis.
OR	Operations Research.
OR/MS	Operations Research/Management Sciences.
POMDP	Partially Observable Markov Decision Process.

PQ-LBB	Priority Queue Implementation of Lagrangian Branch-and-Bound.
QALYs	Quality Adjusted Life Years.
RSR	Relative Survival Rate.
RUFLO	Reliable Uncapacitated Fixed-charge Location Problem with Order.
RUFLO-R	Reformulation of RUFLO Using the Proposed Valid Inequality.
SEER	Surveillance, Epidemiology, and End Results Program.
SPLPO	Simple Plant Location Problem with Order.
Stack-LBB	Stack Queue Implementation of Lagrangian Branch-and-Bound.
UFLP	Uncapacitated Fixed-Charge Location Problem.
USPSTF	US Preventive Services Task Force.

List of Symbols

Chapter 2

$\mathcal{I}$	Set of customers, $i \in \mathcal{I} = \{1, \ldots, I\}$,
$\mathcal{J}$	Set of candidate facilities, $j \in \mathcal{J} = \{1, \ldots, J\}$, (facility J is the dummy facility),
$\mathcal{R}$	Set of backup levels, $r \in \mathcal{R} = \{1, \ldots, R\}$,
$\mathcal{L}$	A partition of customers such that $\bigcup_{\ell \in \mathcal{L}} \mathcal{I}_\ell = \mathcal{I}$,
f_j	Fixed cost for opening and operating location $j \in \mathcal{J}$,
q_j	Probability of failure of location $j \in \mathcal{J}$, where $0 \leq q_j \leq 1$,
η_i	Demand for customer $i \in \mathcal{I}$,
d_{ij}	Cost of serving customer $i \in \mathcal{I}$ from location $j \in \mathcal{J}$,
X_j	$\begin{cases} 1 & \text{if facility } j \in \mathcal{J} \text{ is open;} \\ 0 & \text{otherwise,} \end{cases}$
Y_{ijr}	$\begin{cases} 1 & \text{if customer } i \in \mathcal{I} \text{ is assigned to facility } j \in \mathcal{J} \text{ at level } r \in \mathcal{R}; \\ 0 & \text{otherwise,} \end{cases}$
P_{ijr}	Probability that customer $i \in \mathcal{I}$ is served by facility $j \in \mathcal{J}$ at level $r \in \mathcal{R}$,
ϕ	Penalty cost for not serving customer $i \in \mathcal{I}$,
b	Preference multiplier, $b \in [0, 1]$,
h	Neighborhood search distance parameter,
$h(i, j)$	The order of facility $j \in \mathcal{J}$ in the preference list of customer $i \in \mathcal{I}$.

Chapter 3 MDP Model

$\mathcal{I}$	The set of disease progression states, indexed by i,
$\mathcal{R}$	The set of risk levels, indexed by R, $\mathcal{R} := \{LR, HR, PC, UCT, D\}$,
$\mathcal{O}$	The set of observations, indexed by o, $\mathcal{O} := \{T\text{-}, P+, C+, SD\}$,
$\mathcal{A}$	The set of action vectors, indexed by $a(t)$,
$\mathcal{T}$	The set of time periods, indexed by t,
$\mathcal{S}$	The set of health states, indexed by s,
$\mathcal{X}$	The set of system states, indexed by $X(t)$,
s	The health state vector, $s := \{s_0, s_1, \ldots, s_M\}$,
s_i	The core health state, where $i \in \{0, 1, \ldots, M\}$,
$X_i(t)$	Number of individuals in state $s_i, i \in \{0, \ldots, M\}$ at time $t \in \mathcal{T}$,
$X(t)$	A vector of $X_i(t)$ for all $i \in \{0, \ldots, M\}$,
$\overline{s_i}$	The set of health states that leads to health state s_i,
$\underline{s_i}$	The set of health states that health state s_i leads to,
$\overline{X_i(t)}$	Random variable that represents the number of individuals transitioning into health state s_i during the interval $[t, t + \Delta t]$,
$\underline{X_i(t)}$	Random variable that represents the number of individuals transitioning from health state s_i during the interval $[t, t + \Delta t]$,
$u_{ij}(t)$	Random variable that represents the number of individuals transitioning from health state s_i to health state s_j during the interval $[t, t + \Delta t]$,
$\mathbf{u}(t)$	A vector of $u_{ij}(t)$ for all $i \in \{0, \ldots, M\}$,
t^{max}	The duration (in years) that the model is run for,
$a_i(t)$	Action taken at time t, which is the proportion of individuals in core health state s_i to undergo colonoscopy,
$a^R(t)$	Action taken at time t, which is the proportion of R risk level individuals to undergo colonoscopy, where Equations (3.3b)-(3.3d) apply,
$a(t)$	Action vector at time $t \in \mathcal{T}$, or $a(t) = \{a^{LR}(t), a^{HR}(t), a^{PC}(t)\} \in \mathcal{A}$, and $t \in \mathcal{T}$,
$\hat{a}$	Treatment given to an individual patient; either undergo colonoscopy, or do nothing. $\hat{a} \in \{dn, cl\}$,

L^{max} The CRC screening capacity limit,

$p_t(s_j|s_i, \hat{a}, o)$ The probability that an individual patient will be in core health state s_j in year $t+1$ given that the patient is in core health state s_i, treatment $\hat{a} \in \{dn, cl\}$ is selected, and screening result $o \in \mathcal{O}$ is observed in year t, where $s_j \in \underline{s_i}$,

$q(s_i, \hat{a}, o, s_j)$ The expected reward (in QALYs) of individual patient for going from core health state s_i at time t to core health state $s_j, s_j \in \underline{s_i}, i \in \{0, \ldots, M\}$ at time $t+1$ when treatment $\hat{a} \in \{dn, cl\}$ is taken and observation $o \in \mathcal{O}$ is seen,

$g_t(s_j|s_i, \hat{a})$ The probability that a patient will be in core health state $s_j \in \underline{s_i}$ in year $t+1$ given that the patient is in core health state s_i and treatment $\hat{a} \in \{dn, cl\}$ is selected in year t,

$P_{u_{ij}(t)}(c|a_i(t))$ The probability that the c individuals would move from core health state s_i at time t to core health state $s_j \in \underline{s_i}$ at time $t+1$ when action $a_i(t)$ is performed,

$r_t(c|s_i, a_i(t))$ The immediate reward of transitioning c patients from core health state s_i to core health state s_j after action $a_i(t)$ is taken,

$P_t(X', X, a(t))$ The probability of going from system state X at time $t \in \mathcal{T}$ to system state s' at time $t+1$ when action $a(t) \in \mathcal{A}$ is taken,

$\hat{u}_{ij}(t)$ A realizations of $u_{ij}(t), s_j \in \underline{s_i}$,

$\hat{\mathbf{u}}(X', X)$ The vector of $\hat{u}_{ij}$ such that the transition from system state X at time t to health state X' at time $t+1$ is feasible,

$\hat{\mathbf{U}}(X', X)$ The set of all $\hat{\mathbf{u}}(X', X)$ vectors,

$V_t^*(X)$ The maximum expected TQALYs from for a system at state X in year t to year t^{max},

$r_{t^{max}}(X)$ Terminal reward,

λ Discount factor.

Chapter 3 MIP Model

$\mathcal{I}$ The set of disease progression states, indexed by i,

$\mathcal{J}$ The set of age groups, indexed by j,

$\mathcal{K}$	The set of genders, indexed by k,
$\mathcal{R}$	The set of risk levels, indexed by R, $\mathcal{R} := \{LR, HR, PC, UCT, D\}$,
$\mathcal{O}$	The set of observations, indexed by o, $\mathcal{O} := \{T\text{-}, P+, C+, SD\}$,
$\mathcal{A}$	The set of all action vectors a_t,
$\mathcal{H}$	The set of all history states, indexed by h,
$\mathcal{T}$	The set of time periods, indexed by t, $\mathcal{T} = \{0, 1, \ldots, t^{max}\}$, where t^{max} represents the last time epoch for which the model is run,
$\mathcal{L}$	The set of possible policies,
$s^i_{j,k,h}$	The health state defined by disease progression stage $i \in \mathcal{I}$, age group $j \in \mathcal{J}$, gender $k \in \mathcal{K}$, and disease history $h \in \mathcal{H}$,
$X^i_{j,k,h}(t)$	The number of people in the health state defined at time t by disease progression stage $i \in \mathcal{I}$, age group $j \in \mathcal{J}$, gender $k \in \mathcal{K}$, and disease history $h \in \mathcal{H}$,
$\widetilde{X}^i_{j,k,h}(t)$	The adjusted $X^i_{j,k,h}(t)$ after aging,
$a^R_{j,k,h}(t)$	The action of individuals in state $s^i_{j,k,h}$ to either undergo colonoscopy or not at time t. Since disease progression is unobservable within each risk level, the index i does not appear here,
a_t	The vector of all actions $a^R_{j,k,h}(t)$ at time t, $a_t = \left\{a^R_{j,k,h}(t)\right\}\ \forall\ R, j, k,$ and h. Thus, $a_t \in \mathcal{A}$,
$\hat{a}$	The type of treatment a subgroup is subjected to, $\hat{a} \in \{cl, dn\}$,
$f\left(o \vert s^i_{j,k,h}, \hat{a}\right)$	The rate of observing observation $o \in \mathcal{O}$ at time t when action $\hat{a} \in \{cl, dn\}$ is taken on state $s^i_{j,k,h}$,
$p\left(s^{i'}_{j',k',h'} \vert s^i_{j,k,h}, \hat{a}, o\right)$	The rate at which individuals will be in state $s^{i'}_{i',j',k'}$ given that they are in state $s^i_{j,k,h}$, action $\hat{a} \in \{cl, dn\}$ is taken, and screening result o is observed,
$q\left(s^{i'}_{j',k',h'}, s^i_{j,k,h}, \hat{a}, o\right)$	Immediate rewards (in expected QALYs) for all individuals going from state $s^i_{j,k,h}$ to state $s^{i'}_{j',k',h'}$ given action $\hat{a}$ and screening result o is observed,
$q_{t^{max}}(s^i_{j,k,h})$	Terminal reward for individuals in state $s^i_{j,k,h}$ at the last time period, t^{max},
L^{max}	The capacity limit for colonoscopy resource available.

υ_j	Rate of aging to age group j from an immediate predecessor age group $\underline{j}$,
$\theta^R_{j,k,h}$	Compliance rate for age individuals in states $s^i_{j,k,h}$ $\forall i$,
τ_P	Sensitivity of colonoscopy to polyps,
τ_C	Sensitivity of colonoscopy to cancer,
$\omega_{\hat{a}}$	Probability of CRC self-detection given action $\hat{a}$,
$\rho^{i,i'}_j$	Lesion progression rate from states i to state i',
$\delta^{i,j}_{\hat{a},o}$	Rate of mortality in state i given treatment $\hat{a}$ and observation o.
$\gamma_{i,j}$	Rate of completion within year t of treatment initiated at state i,
$\underline{h}$	An immediate predecessor of h (e.g., if $h = 1$ then $\underline{h} = 0$),
$\underline{j}$	An immediate predecessor of j (e.g., if $j = 1$ then $\underline{j} = 0$),
h_0	1 if $h = 0$, 0 otherwise,
λ_t	Discount factor in year t,
$q^i_{J,h,k}$	Terminal reward (QALYs after age J) for state $s^i_{j,k,h}$,
d_C, d_{CT}, d_{UCT}	Disutility of undetected CRC, CRC treatment, and being in the UCT state,
$d_{poly}(cl)$	Disutility of undergoing colonoscopy with polypectomy,
$d_{\neg poly}(cl)$	Disutility of undergoing colonoscopy without polypectomy,
$\kappa^{i,j}_{\hat{a},o}$	Probability of immediate mortality from screening complications,
κ^j_{UCT}	Probability of immediate mortality from treatment at age j.
m_ℓ	Binary variable set to one if policy $\ell \in \mathcal{L}$ is selected, zero otherwise.

Chapter 1

Introduction

The resource allocation models discussed in this thesis are part of a wider umbrella of Operations Research/Management Sciences (OR/MS) techniques. This chapter starts with a quick overview of OR/MS applications in healthcare industries. This is followed by a section focused on some OR/MS applications in resource allocation in healthcare. Then, the target problems of this thesis are introduced, as well as the research questions this thesis aims to answer.

1.1 Preface

Healthcare industry represents approximately 15% of the Gross Domestic Product (GDP) of the United States (The Economist, 2004), and about 16% of GDP or $2.1 trillion in 2006 even though 50 million Americans do not have health insurance and another 25 million remain underinsured (Catlin et al., 2008). US national health spending is expected to account for 20% of GDP or $4 trillion in 2015 (Dobrzykowski, 2012). Corresponding figures in Canada indicate that health spending was $214.9 billion in 2014, a $4.5 billion increase from the year before. This represent 11% of Canada's GDP in 2014. Health expenditure, on average, accounts for about 40% of provincial/territorial government budgets (CIHI, 2015b). The growth of healthcare spending can be attributed to the increase in life expectancy, new government policies, and improvements in the service quality.

Healthcare professionals are required to carry out their tasks in an effective and efficient manner. Healthcare institutions face new challenges such as increased complexity of processes, the need for efficient utilization of resources, increased pressure to improve the quality of services, and the need to control the workload of healthcare personnel (De Vries et al., 1999). This is where optimization models and tools are most useful.

Operations Research (OR) has been used considerably in healthcare decision making. Early applications include nurse staffing and operations room scheduling, and applications are increasing rapidly (Brailsford and Vissers, 2011). Healthcare has become a major industry, with large number of workers in healthcare organizations and consumers of healthcare services. OR is being utilized recently more to address day-to-day hospital management, resource-constrained operations, or treatment planning aspects in healthcare (Royston, 2009). Key healthcare optimization issues include service planning, resource scheduling, logistics, medical therapeutics, disease diagnosis, and preventive care (Rais and Viana, 2011). Several articles discuss the OR applications in healthcare (Brailsford et al., 2009; Cayirli and Veral, 2003; Dobrzykowski et al., 2014; Fakhimi and Propert, 2013; Gupta and Denton, 2008; Hulshof et al., 2012; Jun et al., 1999).

Many problems faced by OR researchers in healthcare are not analytically different from problems in other industries. However, healthcare delivery systems have quite unique characteristics. Some of these are: the possibilities of death or low quality of remaining life, the difficulty in measuring quality and value of outcomes, the sharing of decisions among several decision makers (physicians, nurses, and administrators), third party payment mechanisms for diagnoses and treatments, and the concept of healthcare access as a right of citizens in society (Pierskalla and Brailer, 1994). Jarrett (1998) indicates a reluctance to implement supply chain management principles in healthcare operations. He attributes this reluctance to healthcare organizations' emphasis on the differences in their operations and the vitality of the services they offer compared to manufacturing industries. Research has shown that implementing supply chain management concepts can reduce costs (Poulin, 2003), while increasing the quality of services as labor productivity is improved (Baltacioglu et al., 2007).

Healthcare applications are modeled and solved using OR tools available to other industries. However, the characteristics of healthcare industry dictate the usage of these

methodologies. Carter (2002) provides a brief insight into the methodologies used in healthcare. Simulation is a popular choice since most healthcare queuing problems are too complex to be analyzed theoretically. Günal and Pidd (2010) provide a recent review on simulation models in healthcare. One of the major issues in healthcare is waiting times (e.g., wait lists for transplants, waiting at the emergency room, etc.). Simulation offers a great tool to visualize local decisions, and the effect of different scenarios on the whole system. One major drawback of simulation is the difficulty of collecting data. Practical and ethical issues arise when it comes to measuring input data related to patients. Also, since caregivers provide service to multiple patients at the same time, it is hard to measure the time needed for each patient.

Linear, goal, and integer programming have been used in a number of applications including facility and staff scheduling, budget allocation, and case-mix management. One major obstacle in using these models is that doctors, not administration, eventually decide what the hospital does. They are generally more concerned about the patient care than they are about the hospital's case-mix issues. Systems dynamics models are used in areas like AIDS epidemic modeling. The large number of stakeholders (governments, public-health agencies, and healthcare providers) requires input from many directions to allocate limited resources. System dynamics models are suited for such environments. However, these models require further work to model overall epidemic control strategy, and to improve the usefulness of outcomes. Queuing models are used to find and improve waiting times. In healthcare perspective, as the queues increase, people either look elsewhere, their health states worsen, or perhaps they die waiting. This is an important aspect to consider. Finally, quality management is used most in pharmaceutical industry. Other fields are behind in terms of applying statistical tools to monitor and control quality. This is mainly because of the reluctance of the medical community to acknowledge and report errors and problems.

Hans et al. (2012) suggest a two dimensional framework for healthcare planning and control that spans four hierarchical levels of control and four managerial areas. The four levels of control are: (1) Strategic, which addresses structural decision making, and it involves dimensioning and development of the healthcare delivery process, (2) Tactical, which involves the organization of operations and execution of that delivery process, and (3 and 4) Operational (online and offline), which involves the short term decision planning

regarding the execution of the healthcare delivery process.

The managerial areas in the framework of Hans et al. (2012) are: (1) Medical planning, which comprises decision making by clinicians regarding medical protocols, treatments and diagnoses, (2) Resource capacity planning, which addresses the dimensioning, planning , scheduling, monitoring, and control of *renewable* resources. These include staff, equipment and facilities (bed linen, sterile instruments, physical therapy equipment), (3) Material planning, which addresses the acquisition, storage, and distribution of all *consumable* resources such as suture materials, blood, bandages, and food, and (4) Financial planning, which addresses how an organization should manage its costs and revenue to achieve its objectives, given the current and future circumstances. The applications discussed in this thesis lie at the intersection of resource capacity planning on the managerial dimension, and at the strategic level on the hierarchy of control. Other applications with similar characteristics include case-mix planning (the volume and composition of patient groups that an ambulatory facility serves), capacity dimensioning, and workforce planning.

1.2 Resource Allocation Models in Healthcare

We now give examples of resource allocation models in healthcare. This is not intended to be a comprehensive analysis. A full review of literature in this area is beyond the scope of this work. Interested readers are referred to the excellent review by Rais and Viana (2011).

1.2.1 Facility Location

When deciding on the number and locations of facilities to open and operate, the decision maker needs to balance between customer satisfaction and associated costs. On the one hand, opening too many facilities increases customer satisfaction and setup costs while reducing the traveling costs. On the other hand, opening fewer facilities decreases customer satisfaction and setup cost while increasing the traveling cost. This type of setting is known in literature as the facility location problem.

Location problems are characterized by four components (ReVelle and Eiselt, 2005), namely: (1) customers, who are already located at points or on routes, (2) facilities that will be located, (3) a space in which customers and facilities are located, and (4) a metric that indicates distances or times between customers and facilities. The problem of locating facilities and allocating customers to them is in the core business of many industries. A logistics company must locate warehouses, an industrial firm must locate assembly plants, and a government must locate new hospitals, care centers, and schools, etc. Location problems are not new to OR researchers and practitioners. A wide range of models has been explored and solved. Formulations range in complexity from simple linear, single-stage, single-product, uncapacitated, deterministic models to non-linear probabilistic models. Algorithms include, among others, local search and mathematical programming-based approaches (Klose and Drexl, 2005). Recent reviews of the literature include Klose and Drexl (2005), where they classify and review different types of location models (including continuous and network location models), Revelle et al. (2008), who review discrete location modeling, and Melo et al. (2009) who review facility location in the context of supply chain management.

It is worth mentioning that some project management applications can be modeled as facility location models. For example, vendor selection problems can be modeled as location models, as discussed in Current and Weber (1994), Demirtas and Üstün (2008), and Jayaraman et al. (1999).

Location problems in healthcare have much in common with those in other industries which have a geographically dispersed customer base that requires easy-to-access quality facilities while the cost is as low as possible. Healthcare facilities are different, however, because they may be subject to national and international control regulations and standards such as maximum response time of emergency vehicles. Daskin and Dean (2005) give a review of location models in healthcare.

Pierskalla and Brailer (1994) differentiate location (or siting) problems in healthcare into five categories. The first category is the regionalization, which is sought to improve the cost or quality of a healthcare system through more effective distribution of services, such as determining the number and distribution of CT scanners in a given region. Regionalization problems are either optimal clustering problems or resource allocation problems. The

second category is the locating or removal of a single facility, such as an acute care hospital which needs to be geographically close to its customer base, and major consideration is given to the current location of similar facilities or institutions in the region. The third category is the location of ambulatory neighborhood clinics, which are used primarily for touring outpatient medical care and for preventive care. Proximity to patients is also an important criterion in the decision making process, as well as network linkage and structure to other institutions in the region. The fourth category is the location of specialized long-term care facilities, where the primary criteria for these locations are costs (of site acquisitions, construction, and operation), rather than closeness to customers. The fifth category is the siting of emergency medical services (EMS) where the primary criterion is the speed of response. Speed of response includes distance to the problem occurrence location, together with the distance from the occurrence location to the treatment facility.

Different location models in healthcare deal with different aspects of location-allocation decisions. Li et al. (2002) study the impact of strategic decisions on community hospitals in the US, while Griffin et al. (2008) use statistical techniques to estimate the demand for community health centers. Gu et al. (2010) model preventive healthcare facilities given a requirement of minimum patients to retain accreditation. Syam and Côté (2010) develop a model for specialized healthcare clinics that has a minimum service requirement. Mahar et al. (2011) investigate the effect of pooling of specialized services in a multi-hospital setting. Mestre et al. (2015) also consider multi-hospital networks but with uncertain parameters. These models provide great insights into how unique the healthcare location models are compared to other location models. These models, however, differ significantly from the models discussed in this thesis in terms of objectives, parameters, and/or scope.

Uncertainty may arise from many factors in location models. Owen and Daskin (1998) provide a review on the facility location research addressing uncertainty in some of the system's parameters, including travel times, construction costs, and demand quantities. There are, however, other sources of uncertainty that are not discussed in their review. One of which arises from damages to facilities which cause disruptions in the allocation decisions. Models that consider this possibility and account for it are named reliable models. The models discussed in this thesis are reliable models in the sense that it is assumed that facilities may fail (become unavailable).

The location models in this thesis incorporate the concepts of reliability into the healthcare setting. Moreover, they also account for customer choices and preferences.

1.2.2 Cancer Screening

According to the American Cancer Society, about 1,688,780 new cancer cases are expected to be diagnosed, and about 600,920 are expected to die of cancer in the United States in 2017 (American Cancer Society, 2017). Furthermore, the Canadian Cancer Society estimated that 206,200 Canadians will develop cancer and 80,800 will die of cancer in 2017 (Canadian Cancer Society, 2017). Cancer is the leading cause of death (about 30% of all deaths) in Canada, and the second most common cause of death (about 25% of all deaths) in the US, exceeded only by heart diseases.

Screening for cancer is an important weapon in the fight against cancer. Colorectal cancer (CRC), for example, mostly originates from benign growths on the inner surface of the colon and rectum (Loeve et al., 2004). Thus, detecting suspicious tissues and removing them before they become malignant is an effective method for the prevention of cancer, and can have significant impact on the patient's health.

The American Cancer Society currently provides screening guidelines for cancers of the breast, cervix, colorectum, endometrium, lung, and prostate, and general recommendations for a cancer-related component of a periodic checkup to examine the thyroid, oral cavity, skin, lymph nodes, testicles, and ovaries.

Cancer screening is among the common preventive healthcare programs, which also include flu shots, blood tests, and anti-smoking advice. Zhang et al. (2009) categorize preventive healthcare programs into three groups based on their objectives: (1) primary prevention aims at reducing the likelihood of diseases in people with no symptoms, for example, by immunizations of healthy children, (2) secondary prevention aims at identifying and treating low-risk people, for example, detecting colorectal polyps before their transition to cancerous lesions, and (3) tertiary prevention aims at treating symptomatic patients in an effort to decrease complications, for example, sugar control in a diabetic person.

Preventive healthcare is inherently different from healthcare for acute problems, and current healthcare systems worldwide fall remarkably short (Zhang et al., 2009). Only 5% of the $1.4 trillion spent on direct health care in the United States goes to preventive health measures and the promotion of general health (Falkenheimer, 2004). Cohen et al. (2008) compare selected preventive measures and treatments and conclude that preventive services, in general, are no more and no less likely to save money than treatments. They note, however, that screening for colorectal cancer reduce mortality either at low cost or at a cost savings (Ness et al., 2000).

It is economically infeasible to screen every individual in the population very often. In fact, a screening procedure itself has its own health risks. Esserman et al. (2009) suggest that screening may be increasing the burden of low-risk cancers without significantly reducing the burden of more aggressively growing cancers, and therefore, not resulting in the reduction in cancer mortality. This suggests that a trade-off is required whenever an optimal screening policy is planned.

Alagoz et al. (2011) and Pierskalla and Brailer (1994) provide reviews on OR models used for cancer screening, while Stevenson (1995) summarizes statistical models of planning and evaluation of cancer screening. Heidenberger (1996) provides a review of quantitative studies that aim at determining the best screening strategy to be used.

1.2.3 Other Applications

Some other significant applications of operations research models in healthcare are presented here. This, however, is not intended to be a comprehensive list. Such a comprehensive review is beyond the scope of this thesis.

Staffing

Staffing is part of patient scheduling, which involves setting the timetable to match patients with caregivers. The time of appointment, the length of time between appointments, the specific type of caregiver who will be responsible for treating patients, and the physical space that will be required to deliver the necessary treatment may all be involved in making

scheduling decisions. The goal is to ensure the maximum utilization of personnel and facility resources and patient flow without incurring additional costs or excessive patient waiting.

A number of studies have addressed the problem of bottlenecks in healthcare clinics by scheduling staff to meet patient demand. Alessandra and Grazman (1978) vary staff patterns to accommodate patient arrival rate. They recommend distributing the current morning appointment patients to the afternoon shift, while keeping the staffing and arrival rate the same. Chan et al. (2002) use integer programming and discrete-event simulation to study a medical records department to determine the optimal staff schedule and understand the workflow. Klafehn et al. (1989) address the linkage between patient flow and the number of staff available in an emergency department. They conclude that moving one nurse from the regular emergency area to a triage position reduces patient waiting lines and patient waiting times. Butler et al. (1996) report significant savings in nurse staffing as well as length of stay of chemotherapy patients at a hospital in Detroit by applying OR tools.

Bed Requirement

It is important for the hospital or clinic to decide how many beds are needed to meet the demand, while maintaining reasonable bed utilization rates. Most bed planning simulation models in the literature attempt to overcome bed shortages or policies that lead to patient misplacement, bumping, or rejection (Jun et al., 1999).

Lowery and Martin (1991) study the the critical care bed requirement. They consider the interrelationships between different hospital units and demonstrate improvements in their methodologies over previous models. Lane et al. (2000) use system dynamics simulation to show that reductions in dedicated emergency bed capacities for patients admitted from an emergency room may increase cancellation rates for elective treatments, rather than increasing waiting times. They conclude that looking at one performance measure in the system can be misleading. Harrison et al. (2005) suggest a simulation model for stochastic bed occupancy problem. Akkerman and Knip (2004) show that the number of beds could be reduced in a cardiac surgery center if recovering patients are transferred

once they no longer require the center's specialized care services.

Moreover, Berman et al. (2007) study the situations where some emergency rooms reach their capacity limit, and therefore, signal to the ambulance dispatch to redirect to another hospital. Their model does not incorporate patients' preferences, which might be related to availability of staff and/or equipment.

Resources for Disease Prevention

One rule for resource allocation suggests that resources be allocated to interventions in an increasing order of their incremental cost-effectiveness ratios. The incremental cost-effectiveness ratio is defined as the total incremental cost associated with an intervention, divided by the total incremental benefits of the intervention. However, this approach may not account for nonlinear scaling of interventions, may ignore nonlinear epidemic growth, and may not capture potential interactions between interventions (Zaric and Brandeau, 2001). Epidemics tend to follow nonlinear growth curves, and incremental investment in an epidemic control program may not yield constant reductions in the chance of disease transmission (Brandeau et al., 2003). Linear and integer programming models for healthcare resource allocation problems have been proposed (e.g., Earnshaw et al., 2002; Earnshaw and Dennett, 2003; Epstein et al., 2007; Stinnett and Paltiel, 1996; Van Zon and Kommer, 1999).

Simple epidemic models with single population have been analyzed using control theory (e.g., Blount et al., 1997; Müller, 1998). The goal here is to determine optimal control over time (e.g., the optimal vaccination rate). Another setting includes allocation of resources among multiple populations with the goal of eradication of the disease or optimization of some function of the equilibrium state of the epidemic (e.g., May and Anderson, 1984; Zaric and Brandeau, 2002).

Zaric and Brandeau (2001) analyze the optimal allocation of investment funds to HIV prevention programs to maximize life years saved by estimation of a production function relating the investment to change in risky behavior. They note that the effectiveness of a particular intervention may depend on the population to which it is targeted (e.g., a high-

risk vs. a low-risk group), the amount already invested in the intervention, and the level of investment in other HIV prevention programs (e.g., television ads to increase awareness).

Vaccine Allocation: influenza

Influenza is a highly contagious disease. Each year 5-15% of the population is infected with influenza resulting in around 3-5 million sever cases and 250,000 - 500,000 deaths worldwide (CDC, 2014). The annual burden of influenza epidemics on the US economy extends to $87.1 billion in 2003, factoring in the cost of medical treatments and working day losses (Molinari et al., 2007). There are many common intervention methods for mitigating the effects of pandemic influenza including social distancing strategies (e.g., school closure, quarantine, isolation), public health measures (e.g., improved hygiene, respiratory protection), and using vaccination or prophylaxis with antiviral medications to reduce the susceptibility of individuals against influenza virus (Chao et al., 2010). Among them, immunization with a well-matched vaccine provides the most efficient and durable response (Talbot et al., 2013).

Medlock and Galvani (2009) develop a compartmental model to determine the optimal age-specific allocation of vaccine stocks for mitigating an influenza pandemic in the US population, based on different outcome measures including number of infections, mortality, and economic cost. Uribe-Sánchez et al. (2011) use simulation to optimize the allocation of the influenza intervention resources over multiple regions. The goal is to minimize the adverse effect of the pandemic given the budget limitations.

Organ Transplantation

Human organs are very scarce resources. Candidates for organ transplantation are placed on waiting lists. In the United States, as of 2013, these lists had approximately 58,000 candidate patients for kidney transplant, 13,000 for liver transplant, 2,500 for heart transplant, and 1,300 for lung transplant (OPTN, 2014). In Canada, the numbers as of 2013 are 3,200 for kidney transplant, 700 for liver transplant, 160 for heart transplant, 300 for lung transplant (CIHI, 2015a). At the same time, some available organs end up being

wasted. Each year about 18% of kidneys, 10% of livers, and 5% of lungs are discarded in the United States. This shows the importance of having an effective allocation system in place. Moreover, the transplantation process is costly. As indicated in Akan et al. (2012), the base cost of a liver transplant is around US$450,000, and around US$1 million when the costs of surgery and medication are factored in. These figures show the significance of improving the efficiency of donated organ usage.

Several researchers model the accept/decline model for organ transplantation (e.g., Alagoz et al., 2004, 2007; Sandikci et al., 2008). Some authors provide simulation to demonstrate and compare different allocation policies (e.g., Bertsimas et al., 2013; Shechter et al., 2005). Kong et al. (2010) develop a set-partitioning model for a liver allocation system that takes into account the geographic composition of donors and candidates.

1.3 Target Problems

In this thesis, two main problems are discussed; a reliable facility location problem with customer preferences, and an allocation of limited cancer screening resources among individuals in a population.

In the first problem, a central authority is looking to locate facilities (e.g., hospitals, warehouses, etc.) among a set of candidate locations. The aim is to decide how many facilities to open, where to open them, and how to allocate demand (patients, customers, etc.) to them. Adding to the complexity of the problems, the patients have preferences over which facility to be assigned to. This means that each patient, when faced with a choice between two available facilities A and B, an order is present (say, B is preferred over A), and must be considered in the allocation decision. Moreover, it will be assumed that there is a possibility of failure for each facility. In case of failure, the patients originally assigned to the failed facility need to be re-allocated to the next preferred available facility.

For the location problem, the development of the model is discussed, and a Lagrangian relaxation scheme is developed and embedded within a branch-and-bound structure. The methodology exploits the special characteristics of the model to arrive at an efficient solution. Three different implementation strategies are discussed and tested. Extensive numer-

ical results are shown. Later on, a reformulation and more advanced solution methodologies are presented in which extremely large datasets are solved. A developed branch-and-bound scheme as well as a branch-and-cut technique are presented and tested to verify their effectiveness. Moreover, a constraint is proposed to significantly tightens the LP relaxation of the formulation.

The aim of the cancer screening resources problem is to suggest an optimal screening policy such that the welfare of the society is maximized, according to some quality measure (e.g., quality-adjusted life years). The decision maker is required to allocate limited screening capacity among a population of individuals. The disease progression for each individual patient follows a stochastic process of its own. The problem is formally introduced, and two modeling approaches are presented. The mixed-integer program approach is solvable using commercial software and the computing power of a personal computer. Insights on the recommended policies are shown and recommendations for future lines of research are discussed.

1.4 Research Questions

The main question of this thesis is how to allocate or assign resources in various healthcare settings so that the society welfare is maximized (based on some welfare measure). To answer this question, two main applications are considered. Within each application, a number of questions are answered.

The first application is a reliable facility location model with customer preferences. This model attempts to answer the following questions:

- How many facilities to open, given the possibility of failure of some (or all) facilities?
- Where to locate open facilities?
- How many customers should be assigned to each open facility?
- Which customers to assign to each facility?

- If a facility randomly fails, to which facility are customers re-assigned?
- What is the total expected cost of opening the desired facilities?
- How can such models be solved efficiently?
- What are the possible implementation strategies for the solution methodology?
- How do different implementation strategies compare with respect to time needed to converge?
- How can the solution methodology be improved to allow for extremely large instances?

The other application is allocating CRC screening capacity to individuals in a population. The developed framework is an attempt to answer the following questions:

- What is the optimal CRC screening policy for a population with varying health states, given limited screening capacity?
- What are the shortcoming of existing CRC screening guidelines?
- What is the effect of key system parameters on the overall resource allocation recommendations?

1.5 Thesis Outline

This thesis is organized as follows. Chapter 2 presents the reliable facility location model with customer preference. The motivation, description, and formulation of the model are presented. Three different implementation strategies are introduced and thoroughly tested. Next, a formulation with fewer variables is presented, and enhanced solving techniques for the model are discussed. Comprehensive testing is shown to highlight the superiority of the solution methods and their ability to solve extremely large instances, as well as the

effectiveness of the constraints which aim to tighten the LP relaxation. Directions for future research are then discussed.

Chapter 3 then formally introduces an analytical framework for modeling a colorectal cancer screening problem for a representative population. The limited screening resources are considered. A Markov Decision Process model is first given, together with the challenges associated with it. In a following section, a mixed integer program is discussed and solved to extract optimal screening policies. Finally, possible directions for extended models are presented.

A summary and a highlight on the conclusions are shown in Chapter 4. The lessons and insights learned from the different problems and models of this thesis are mentioned as well as recommendations on future research work.

1.6 Connection of Models

The connection between the models presented in this thesis is best viewed by considering a case study. In CRC screening setting, there are two main phases of designing a system to allocate available screening resources among individuals in the population. The first phase deals with finding the optimal locations at which service is provided and allocate patients to these facilities. To do so, it is essential to know the preferences of patients. The second phase utilizes the available capacity and recommends screening guidelines that aim at improving the health status of the whole population.

The decisions of the first phase are considered in Chapter 2, where two models are presented. The two models consider the preferences of patients when deciding on the number and location of service facilities to have in the system. Unlike the Preliminary Model, the Modified Model allows patients to completely control their choices.

Once the locations are decided on, the second phase of this case study deals with allocating available resources of screening among the population in order to maximize the society's welfare. Two modeling approaches are introduced in Chapter 3. The MDP model is first discussed. However, since the MDP model is found to be hard to solve, the MIP

model is introduced with additional factors like age and personal history. The MIP model can be solved for reasonably-sized instances. The results of the MIP model constitute optimal policies of CRC screening.

In conclusion, this thesis considers and solves the two phases of a typical preventative healthcare application. Incorporating the two phases into a single model is theoretically possible, and may provide great insights. However, doing so is beyond the scope of this thesis.

Chapter 2

Reliable Facility Location Model with Customer Preference

In this chapter, we introduce the facility location model, where it is assumed that each facility has a probability of failure, and customers order available facilities according to their preferences. The following section motivates the problem, provides a review of related literature, and discusses the contributions. Section 2.2 introduces the preliminary model, which is solved in Section 2.3. The numerical results of the preliminary model are shown in Section 2.4. Then, Section 2.5 discusses the modified model, which is solved and tested in sections 2.6 and 2.7, respectively.

2.1 Introduction

Several studies report that firms can save millions through redesigning their distribution systems by determining the optimal number, capacity, and location of their facilities (Camm et al., 1997; Teo and Shu, 2004). The classical facility location models aim to balance the cost of opening facilities and logistics/service costs considering spatial, budgetary, service-quality related constraints (Klose and Drexl, 2005).

Facility location is a vital strategic decision in the design of supply chains and service

networks. In classical facility location models, a central planner setups perfectly-reliable facilities and makes customer allocations in order to balance the cost of opening facilities against the cost of providing service (Daskin, 1995). Such classical models, however, are not applicable if customers patronize the facility of their choice. For instance, in the context of preventative healthcare, patients' choices of care provider depend on both quality and accessibility (Haase and Müller, 2015). That is, patients may not patronize the closest or cheapest health facility to seek better care (Baldwin et al., 2008; Charlton et al., 2015).

Furthermore, in real life, facilities may fail to serve customers due to several types of disruptions including natural disasters, road/weather conditions, unplanned maintenance breaks, and not having the necessary capacity or expertise (Snyder et al., 2016).

The aim of this chapter is to propose a novel uncapacitated fixed-charge location problem which considers the preferences of customers and the reliability of facilities simultaneously. Although there are facility location models in the literature that consider customer preferences and reliability of facilities separately, studies incorporating both of these aspects are limited (Herrera et al., 2008).

A central planner selects facility locations from a set of candidate sites to minimize the total cost of opening facilities and providing service to customers. The central planner does not allocate customers to constructed facilities. Rather, each customer has a strict preference order over the candidate sites, and patronizes her most preferred available facility. If that facility fails due to a disruptive event, the customer attends her next preferred available facility. If none of the available facilities is a preferred location for a customer, then she does not seek service and incurs a disutility. The proposed model bridges the gap between the location models that consider the preferences of customers and the ones that consider the reliability of facilities.

The proposed model will be referred to as the Reliable Uncapacitated Fixed-charge Location Problem with Order (RUFLO). We formulate a preliminary version of the RUFLO in Section 2.2, and develop a Lagrangian branch-and-bound procedure to solve it efficiently in Section 2.3. We implement the procedure in three different ways, and conduct extensive numerical analysis in Section 2.4. The results show that the proposed algorithm is able to solve small and medium instances efficiently. After that, a modified version of the RUFLO

model is discussed in Section 2.5. In Section 2.6, we develop a Lagrangian branch-and-bound algorithm and a branch-and-cut algorithm to solve a strengthened reformulation of the RUFLO model. We also propose a neighborhood search method to generate upper bounds from feasible solutions. Our numerical experiments in Section 2.7 show that the proposed solution algorithms are efficient, and can be applied to problems with extremely large number of customers. This is an important contribution because solving real-life location problems may require considering large number of different customer types regarding their preferences. We conclude the chapter in Section 2.9 with final remarks and potential future research directions.

2.1.1 Review of Related Literature

Hanjoul and Peeters (1987) first introduced the so-called Simple Plant Location Problem with Order (SPLPO) to consider customers' preferences when locating facilities. They assume that each customer has a known preference order over the candidate sites, and attends her most preferred available facility. Cánovas et al. (2007) strengthen the formulation of SPLPO with valid inequalities. Hansen et al. (2004) present a bilevel location model to consider customer preferences. At the upper level, a set of facilities is selected, whereas at the lower level, customers attend open facilities according to their preferences. Camacho-Vallejo et al. (2014) and Marić et al. (2012) develop heuristic methods to solve this bilevel formulation.

Vasilev et al. (2009) present new lower bounds for the SPLPO by introducing valid inequalities and show improvements in the linear relaxation and integrality gap. Vasilyev and Klimentova (2010) add valid inequalities related to the preferences as a single-level integer linear program. Other papers discuss the bilevel p-median problem by considering customers' preferences, including Aksen et al. (2013). Lee and Lee (2012) present a facility location problem with covering constraints and preferences as a mixed integer program, and propose a heuristic based on Lagrangian relaxation.

Zhang et al. (2012c) propose an optimal-choice model and a probabilistic-choice model for locating preventive health care facilities. In the first model, each patient attends the most attractive facility similar to the SPLPO. The second model assumes that a patient

may patronize each facility with a certain probability, which is modeled by a multinomial logit function increasing with the attractiveness of the facility. Haase and Müller (2015) present a mixed-integer formulation for this probabilistic-choice model. Verter and Zhang (2015) give a detailed discussion on the location models for preventative healthcare facilities. Ishii et al. (2007) present a fuzzy modeling structure for the facility location problem with customer preferences. They represent the satisfaction degree of the customer based on the distance to the facility site. Their objective is to find the site of the facility which maximizes the minimal satisfaction degree among all demand points and maximizes the preferences of the site. Our study differs from the aforementioned studies in the literature as we consider the reliability of facilities in addition to the preferences of customers.

It is assumed that open facilities are always operational in classical facility location models. Due to the fast growing awareness of service sustainability and reliability, an increasing number of studies have incorporated uncertain environmental and social factors into facility location decisions (Baron et al., 2011; Chen et al., 2014; Mestre et al., 2015). In principle, when a facility fails, it cannot provide the intended service, and consequently, customers who originally are assigned to that facility need to be forwarded (reassigned) to other facilities.

There is a big literature on the reliable supply chain. Snyder and Daskin (2005) study the reliable facility location problem in which customers are assigned to a number of backup facilities. They formulate a p-median problem and an Uncapacitated Fixed-Charge Location Problem (UFLP) for selecting facility locations. Snyder and Daskin (2005) assume that all locations have identical and independent failure probabilities. Cui et al. (2010) present a model for the reliable UFLP with site-specific failure probabilities. They propose a continuum approximation and also formulate a mixed-integer program which is solved by Lagrangian decomposition. Aboolian et al. (2013) extend the model in Cui et al. (2010) by relaxing the limit on the number of backup facilities, and develop an efficient search-and-cut algorithm. Shen et al. (2011) formulate the reliable UFLP as a two-stage stochastic program, and then as a nonlinear integer program. Their stochastic programming model can capture the dependence among site-specific failure probabilities. Furthermore, Lu et al. (2015) also present a model that allows disruptions to be correlated, and apply distributionally robust optimization to minimize the expected cost under the worst-case

distribution. We refer the reader to Snyder et al. (2016) for a comprehensive review of facility location models with disruption.

Peng et al. (2011) take a different modeling approach, and use the p-robustness criterion (Snyder and Daskin, 2006) to explicitly bound the cost in disruption scenarios with the objective of minimizing the nominal cost, that is, the cost when no disruptions occur. Lu et al. (2015) present a model that allows disruptions to be correlated with an uncertain joint distribution, and apply distributionally robust optimization to minimize the expected cost under the worst-case distribution for given marginal disruption probabilities. Li and Ouyang (2010) develop a continuum approximation approach to reliable facility location design under correlated probabilistic disruptions. In a similar fashion, Lim et al. (2013) use a stylized continuous location model to investigate the impact of misestimating the disruption probability in the presence of correlated disruptions and finite capacity.

As with other location models, researchers try to incorporate other levels of the supply chain decisions into a single integrated model. In the context of reliable supply chains, Chen et al. (2011) incorporate location and inventory costs, and formulate the model as a mixed integer program. Their model is solved using Lagrangian relaxation. Ahmadi-Javid and Seddighi (2013) consider location-routing under various risk scenarios. Moreover, Qi et al. (2010) present a location-inventory-routing model with random supply disruptions at either the supplier or retailer. Their model is formulated as a nonlinear integer program. The objective function is approximated to make the model easier to analyze, and it is solved using a Lagrangian relaxation approach embedded in a branch-and-bound procedure. Garcia-Herreros et al. (2014) extend that work to capacitated distribution centers and to multiple commodities.

In another direction of research, facilities can be subjected to hardening (or fortification), at an additional cost, to make them more reliable. Lim et al. (2010) propose two types of facilities; one that is unreliable (has a probability of failure), and another that is reliable but more expensive. Li et al. (2013) build on that and incorporate a fortification budget constraint. These models are solved using Lagrangian relaxation-based algorithms. Moreover, Scaparra and Church (2008) assume that attacks (or damages) can occur to only a subset of locations, and hence, resources for fortification are used accordingly.

Considering the preferences of customers and the reliability of facilities might be imperative in facility network design (An et al., 2013; Wagner et al., 2010; Zhang et al., 2009) especially when customers choose a facility to attend based on their preferences (which may be based on distance, quality, familiarity, etc.), and open facilities face disruption risk (Akgün et al., 2015; Teng et al., 2014; Verma and Gaukler, 2015; Zhang et al., 2009). In this chapter, we also consider the reliability of facilities with independent and different failure probabilities. However, we do not assign customers to their closest facilities. Instead, we let each customer attend her most preferred facility as long as it remains operational. Less preferred facilities work as a backup, and serve that customer only if all more preferred facilities have failed. The main aim is to open facilities at a subset of the potential sites in such a way that each customer is assigned up to a certain number of facilities in the order of preference, and that the total cost of opening facilities plus the expected service cost is minimized. This problem will be referred to as the reliable uncapacitated fixed-charge location problem with order (RUFLO). It bridges the gap between the facility location models that consider customers' preferences and those that consider reliability of candidate sites.

We propose a novel mixed-integer programming formulation for the RUFLO. Unlike, Cui et al. (2010) and Daskin (1995), who employed Lagrangian relaxation to solve the RUFLP, and unlike Aboolian et al. (2013), who proposed a search-and-cut algorithm, we develop an efficient decomposition method through a split variable reformulation. This proposed approach can easily be adapted to solve the p-median-based version of our model.

2.1.2 Motivation

To demonstrate the importance of incorporating customer preferences, we look at the optimal solution for the 49-node US map (`USmap49`) dataset with two levels of reliability as reported in Cui et al. (2010). Column (1) of Table 2.1 shows the five locations open according to the optimal solution of their model. The *Dummy* facility will be discussed later. However, for now, it can be thought of as lost demand due to high service cost. The customers are then assigned to the nearest open facility as their first level of assignment. This assignment is shown in column (2). The failure cost of a facility is the extra cost

endured by the system due to a failure in that facility (with all other open facilities still available). For example, if the facility in Sacramento, CA fails, a total of $\$838,308$ will be endured by the system due to re-allocation of the %19 of demand originally assigned to it. The failure costs of all locations are shown in column (3) where re-allocation is also based on distance. The Dummy facility does not fail. Hence, it has no failure cost. Alternatively, in column (4) we show a scenario in which the same locations are open. However, we assume random preferences for each customer, and assign customers based on their preferences. This is equivalent of opening locations based on the model of Cui et al. (2010), but customers choose the location of service based on their preferences. This result in changing the allocation decisions for the first level as it is apparent in column (4). Similarly, in this case, the failure costs, column (5), are also different. If a facility fails, the customer will go to the next available facility on their list. The difference in costs is significant and upward of 100%, as shown in column (6).

To conclude, the values in column (5) represent the *actual failure cost* if customer indeed have preferences, and these preferences were not taken into consideration while optimizing the model. Obviously, the values of column (5) can never be less than those in column (3) since the model of Cui et al. (2010) always assign customers to the nearest facility. The assignments for the second level in both cases are omitted for clarity. However, since $|\mathcal{R}| = 2$, all customers not assigned to the Dummy in the first level, will be assigned to it in the second level. In particular, 100% of demand would go to the Dummy at the second level in the case of no consideration for preferences, while 60% of the demand would be assigned to the Dummy in the case of preferences. It is important to note that the overall costs (i.e., the objective function) of the system described by column (2) is less than that of the system described by column (4). In particular, the total cost of the system described by column (2) is $916,068$, while the model described by column (4) has a total cost of $3,189,260$. This follows from the fact that when no preferences are considered, allocation is done based on distance, which by definition has the minimum costs.

Facility disruption is common in some settings. Many facilities may become unavailable due to natural disasters, terrorist attacks, or labor strikes. Qi et al. (2010) mention examples of disruptions caused by hurricanes Katrina and Rita in 2005. In healthcare context, Berman et al. (2007) describe a situation where some hospitals reach their capacity limit

Table 2.1: Effect of Customer Preferences on Failure Costs. Column (3) Shows the failure costs when no preference is considered, while column (6) shows the failure costs when preference is considered in the original allocation and the re-allocation.

(1)	(2)	(3)	(4)	(5)	(6)
	Cui et al (2010)		Cui et al (2010) + Preferences		Cost increase
	Demand Covered (%)	Failure Cost	Demand Covered (%)	Failure Cost	
Sacramento, CA	19	838,308	13	1,812,526	116%
Austin, TX	9	594,411	14	1,459,756	146%
Harrisburg, PA	29	714,066	6	1,735,023	143%
Lansing, MI	12	537,818	14	1,914,027	256%
Montgomery, AL	17	634,892	11	1,562,375	146%
Des Moines, IA	15	547,005	3	1,647,376	201%
Dummy	0	-	40	-	-

in emergency rooms, and they notify the ambulance dispatch, which redirects ambulances to the next closest open facility.

Therefore, this chapter incorporates the possibility of disruption for facilities, together with customer preferences into a facility location model. The aim is to minimize the costs of assigning and reassigning of customers, as well as the fixed cost of opening facilities at candidate locations. The model balances a trade-off between opening too many facilities and the excessive travel costs resulting from opening too few. Applications of this model can be found in locating service centers, warehouses, hospitals, etc.

2.1.3 Contributions

The contributions of this chapter are mainly in two areas. Firstly, the reliable location model with customer preferences is introduced in two forms: preliminary and modified. This model is shown to be able to save cost significantly. Secondly, branch-and-bound

and branch-and-cut algorithms are developed to solve this model. Using a combination of techniques, the proposed algorithms are able to solve instances of different sizes, including extremely large datasets.

2.2 Preliminary Model

The RUFLO model is described and formulated here. Section 2.3 present a solution mechanism for this model. Another formulation that changes the way preferences are enforced and with smaller number of variables is shown in Section 2.5.

2.2.1 Model Description

A central authority is responsible for opening and operating facilities among a set of candidate locations. Each customer has a preference ordering over the candidate facility locations. The preferences of customers are strictly followed whenever the assignment decision is made. In other words, if a customer prefers facility m over n (and both are open), then the customer is initially assigned to facility m regardless of their proximity to the customer.

Preferences of each customer are exogenous inputs to the model. Details about the underlying utility functions that may produce these preferences are beyond the scope of this work. However, different preferences sets are generated based on some criteria (including measures of distance, size, and quality of service), and are used in Section 2.4 to study the behavior of the model.

Each candidate facility has a distinct fixed probability of failure. The events of facility disruptions are assumed to be independent. When an open facility fails, the clients assigned to this location have to be reassigned to another open facility. Because of customer preferences, the customers are assigned to the most preferred facility that is still available. In our model, it is assumed that failures happen, if any, before customers make trips to any facilities. In other words, by the time the demand requires fulfillment, it will be known

with certainty which facilities are available and which are not. Available facilities from that point on are expected to remain available throughout the fulfillment period.

Each customer is assigned (and reassigned) to up to $|\mathcal{R}|$ facilities, where $|\mathcal{R}|$ is the number of backup levels in the model. We assume that customers have complete information about failures. Unlike models that assume incomplete information (e.g., Albareda-Sambola et al., 2015; Berman et al., 2009), customers in our model travel from their location to the intended available facility directly (i.e., without re-routing or backtracking).

We introduce a penalty cost ϕ_i of not serving customer i. To model this, a dummy facility, indexed by $j = J$, is introduced. This dummy facility has fixed cost $f_J = 0$, failure probability $q_J = 0$, and transportation cost $d_{iJ} = \phi_i$ for all customers $i \in \mathcal{I}$. In the current model, ϕ_i can be incurred even if some facilities are open, if ϕ_i is less than the cost of serving customer i through any of these open facilities. This means the central planner would intervene and 'override' customers' preference lists if needed. This assumption is changed in the modified model of Section 2.5, where customer preferences are enforced regardless of the values of service costs.

The aim of the model is to decide how many facilities to open, where to open them, and how customers are assigned and reassigned to available facilities. The objective is to to minimize the total cost consisting of the fixed cost for opening and operating facilities, and the expected transportation cost across all levels.

2.2.2 Notation

Sets

$\mathcal{I}$: Set of customers, $i \in \mathcal{I} = \{1, \ldots, I\}$,

$\mathcal{J}$: Set of candidate facilities, $j \in \mathcal{J} = \{1, \ldots, J\}$, (facility J is the Dummy facility)

$\mathcal{R}$: Set of backup levels, $r \in \mathcal{R} = \{1, \ldots, R\}$,

Parameters

f_j : Fixed cost for opening and operating location $j \in \mathcal{J}$,

q_j : Probability of failure of location $j \in \mathcal{J}$, where $0 \leq q_j \leq 1$,

η_i : Demand for customer $i \in \mathcal{I}$,

d_{ij} : Cost of serving customer $i \in \mathcal{I}$ from location $j \in \mathcal{J}$,

Decision Variables

$$X_j = \begin{cases} 1 & \text{if facility } j \in \mathcal{J} \text{ is open;} \\ 0 & \text{otherwise,} \end{cases}$$

$$Y_{ijr} = \begin{cases} 1 & \text{if customer } i \in \mathcal{I} \text{ is assigned to facility } j \in \mathcal{J} \text{ at level } r \in \mathcal{R}; \\ 0 & \text{otherwise,} \end{cases}$$

P_{ijr} : Probability that customer $i \in \mathcal{I}$ is served by facility $j \in \mathcal{J}$ at level $r \in \mathcal{R}$.

Note that we could have defined $\hat{d}_{ij} = \eta_i d_{ij}$ and used it in the objective function. However, we still need to know d_{ij} to compare with ϕ_i, as explained above. Also, the datasets used in testing report d_{ij} and η_i separately. Therefore, the formulation will keep these parameters separated.

2.2.3 Model Formulation

The mixed integer program (MIP) formulation of the problem is as follows.

$$\min \sum_{j\in\mathcal{J}} f_j X_j + \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\sum_{r\in\mathcal{R}} \eta_i d_{ij} P_{ijr} Y_{ijr} \tag{2.1a}$$

$$\text{s.t.} \sum_{j\in\mathcal{J}\setminus\{J\}} Y_{ijr} + \sum_{s\in\mathcal{R}:s\le r} Y_{iJs} = 1 \qquad i\in\mathcal{I}, r\in\mathcal{R} \tag{2.1b}$$

$$\sum_{r\in\mathcal{R}} Y_{ijr} \le X_j \qquad i\in\mathcal{I}, j\in\mathcal{J} \tag{2.1c}$$

$$\sum_{r\in\mathcal{R}} Y_{iJr} = 1 \qquad i\in\mathcal{I} \tag{2.1d}$$

$$1 - X_m + \sum_{s\in\mathcal{R}:s\le r} Y_{imk} \ge Y_{inr}, \qquad i\in\mathcal{I}, r\in\mathcal{R},\ m,n\in\mathcal{J} \tag{2.1e}$$

$$P_{ij1} = 1 - q_j \qquad i\in\mathcal{I}, j\in\mathcal{J} \tag{2.1f}$$

$$P_{ijr} = (1-q_j)\sum_{k\in\mathcal{J}} \frac{q_k}{1-q_k} P_{i,k,r-1} Y_{i,k,r-1} \qquad i\in\mathcal{I}, j\in\mathcal{J}, r\in\mathcal{R}\setminus\{1\} \tag{2.1g}$$

$$X_j, Y_{ijr} \in \{0,1\},\ P_{ijr} \ge 0 \qquad i\in\mathcal{I}, j\in\mathcal{J}, r\in\mathcal{R} \tag{2.1h}$$

The objective function (2.1a) minimizes the sum of the opening and operating the centers, and the expected transportation costs. Constraints (2.1b) enforce each customer i either to be assigned to a regular facility at level r, or assigned to the dummy facility J at certain level $s \leq r$. Assignment to a closed facility is prohibited by constraints (2.1c). Constraints (2.1d) guarantee that each customer is assigned to the dummy facility at exactly one assignment level. Constraints (2.1e) are the preference enforcing constraints. They state that if both m and n are open, and m is preferred over n by customer i, then m must be given a lower backup rank than n. This, however, depends on if they are both to be assigned. Recall that not all open facilities are assigned to a particular customer i; a maximum of $|\mathcal{R}|$ facilities can be assigned. Moreover, these constraints allow that if n is open and m is closed, then n can be given any rank. Constraints (2.1f) and (2.1g) are the assignment (and reassignment) probability equations. In the first level $r = 0$ (primary assignment), Constraints (2.1f) state that P_{ijr} is the probability that j remains available. These probability equations are similar to those in Cui et al. (2010). Finally, the binary requirements on X_j and Y_{ijr} are enforced by constraints (2.1h).

The nonlinear term in the objective function and in constraint (2.1g) , $P_{ijr}Y_{ijr}, i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}$, is a product of a continuous variable and a binary variable. Thus, we replace each $P_{ijr}Y_{ijr}$ by W_{ijr}, and enforce $W_{ijr} = P_{ijr}Y_{ijr}$ using the following set of new constraints for all $i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}$: $W_{ijr} \leq P_{ijr}$, $W_{ijr} \leq Y_{ijr}$, $W_{ijr} \geq P_{ijr} + Y_{ijr} - 1$, $W_{ijr} \geq 0$.

The number of variables in this models is $|\mathcal{J}| + 3 \times |\mathcal{I}| \times |\mathcal{J}| \times |\mathcal{R}|$.

2.3 Solution Techniques: Preliminary Model

The model in Section 2.2 is hard to solve using commercial software. As such, a dedicated algorithm is developed here to solve the model efficiently. The methodology is mainly a Lagrangian relaxation embedded into a branch-and-bound structure. The details and development of the algorithm are presented below.

We start by defining a variable Z_{ij} for each customer $i \in \mathcal{I}$ as the copy of the variable

X_j, and adding constraint (2.2) to model (2.1).

$$X_j - Z_{ij} = 0 \quad j \in \mathcal{J},\ i \in \mathcal{I}. \tag{2.2}$$

For each customer $i \in \mathcal{I}$, let $\mathbf{Z}_i := \{Z_{ij},\ j \in \mathcal{J}\}$, $\mathbf{W}_i := \{W_{ijr},\ j \in \mathcal{J}, r \in \mathcal{R}\}$, $\mathbf{Y}_i := \{Y_{ijr},\ j \in \mathcal{J}, r \in \mathcal{R}\}$, $\mathbf{P}_i := \{P_{ijr},\ j \in \mathcal{J}, r \in \mathcal{R}\}$, and define the solution set $S_i := \{(\mathbf{Z}_i, \mathbf{Y}_i, \mathbf{W}_i, \mathbf{P}_i) : (2.3a) - (2.3h)\}$.

$$\sum_{j \in \mathcal{J}\setminus\{J\}} Y_{ijr} + \sum_{s \in \mathcal{R}: s \le r} Y_{iJs} = 1 \qquad r \in \mathcal{R} \tag{2.3a}$$

$$\sum_{r \in \mathcal{R}} Y_{ijr} \le Z_{ij} \qquad j \in \mathcal{J} \tag{2.3b}$$

$$\sum_{r \in \mathcal{R}} Y_{iJr} = 1 \tag{2.3c}$$

$$1 - Z_{im} + \sum_{s \in \mathcal{R}: s \le r} Y_{ims} \ge Y_{inr} \qquad m, n \in \mathcal{J}, r \in \mathcal{R} \tag{2.3d}$$

$$P_{ij1} = 1 - q_j \qquad j \in \mathcal{J} \tag{2.3e}$$

$$P_{ijr} = (1 - q_j) \sum_{k \in \mathcal{J}} \frac{q_k}{1 - q_k} W_{i,k,r-1} \qquad j \in \mathcal{J}, r \in \mathcal{R} \setminus \{1\} \tag{2.3f}$$

$$W_{ijr} \le P_{ijr},\ W_{ijr} \le Y_{ijr},\ W_{ijr} \ge P_{ijr} + Y_{ijr} - 1, \qquad j \in \mathcal{J}, r \in \mathcal{R} \tag{2.3g}$$

$$Z_{ij}, Y_{ijr} \in \{0, 1\}, W_{ijr}, P_{ijr} \ge 0 \qquad j \in \mathcal{J}, r \in \mathcal{R}. \tag{2.3h}$$

Then, the *split-variable formulation* of model (2.1) is given by:

$$\begin{aligned} \Psi^* = \min\ & \sum_{i \in \mathcal{I}} \left(\sum_{j \in \mathcal{J}} \frac{1}{|\mathcal{I}|} f_j Z_{ij} + \sum_{j \in \mathcal{J}} \sum_{r \in \mathcal{R}} \eta_i d_{ij} W_{ijr} \right) \\ & \text{s.t. } (\mathbf{Z}_i, \mathbf{Y}_i, \mathbf{W}_i, \mathbf{P}_i) \in S_i && i \in \mathcal{I} \\ & X_j - Z_{ij} = 0 && j \in \mathcal{J}, i \in \mathcal{I} \end{aligned}$$

Let $\mathcal{S}_i$ denote the index set of S_i, that is, $S_i = \{(\hat{\mathbf{Z}}_i^s, \hat{\mathbf{Y}}_i^s, \hat{\mathbf{W}}_i^s, \hat{\mathbf{P}}_i^s)\ : s \in \mathcal{S}_i\}$. We can express any solution in S_i by:

$$(\mathbf{Z}_i, \mathbf{Y}_i, \mathbf{W}_i, \mathbf{P}_i) = \sum_{s \in \mathcal{S}_i} \lambda_i^s (\hat{\mathbf{Z}}_i^s, \hat{\mathbf{Y}}_i^s, \hat{\mathbf{W}}_i^s, \hat{\mathbf{P}}_i^s),\ \sum_{s \in \mathcal{S}_i} \lambda_i^s = 1,\ \lambda_i^s \in \{0, 1\}. \tag{2.4}$$

Using (2.4), the linear relaxation of the Dantzig-Wolfe reformulation of the split variable formulation is given by:

$$
\begin{aligned}
\text{[DW]}\ \min\ & \sum_{i\in\mathcal{I}}\sum_{s\in\mathcal{S}_i}\left(\sum_{j\in\mathcal{J}}\frac{1}{|\mathcal{I}|}f_j\hat{Z}^s_{ij}+\sum_{j\in\mathcal{J}}\sum_{r\in\mathcal{R}}\eta_i d_{ij}\hat{W}^s_{ijr}\right)\lambda^s_i & & \\
\text{s.t. } & X_j-\sum_{s\in\mathcal{S}_i}\hat{Z}^s_{ij}\lambda^s_i=0 & & j\in\mathcal{J}, i\in\mathcal{I},\ (\mu_{ij})\\
& \sum_{s\in\mathcal{S}_i}\lambda^s_i=1 & & i\in\mathcal{I},\ (\theta_i)\\
& \lambda^s_i\geq 0 & & i\in\mathcal{I}, s\in\mathcal{S}_i.
\end{aligned}
$$

The DW can be solved efficiently using column generation. In this chapter, we use a subgradient method to solve the dual of the DW, which is stated as:

$$
\begin{aligned}
\text{[DW-Dual]}\ \max\ & \sum_{i\in\mathcal{I}}\theta_i & & \\
\text{s.t. } & \sum_{i\in\mathcal{I}}\mu_{ij}=0 & & j\in\mathcal{J},\\
& \theta_i-\sum_{j\in\mathcal{J}}\hat{Z}^s_{ij}\mu_{ij}\ \leq\sum_{j\in\mathcal{J}}\frac{1}{|\mathcal{I}|}f_j\hat{Z}^s_{ij}+\sum_{j\in\mathcal{J}}\sum_{r\in\mathcal{R}}\eta_i d_{ij}\hat{W}^s_{ijr} & & i\in\mathcal{I}, s\in\mathcal{S}_i.
\end{aligned}
$$

Let $L_i(\boldsymbol{\mu}_i)=\sum_{j\in\mathcal{J}}\frac{1}{|\mathcal{I}|}f_jZ_{ij}+\sum_{j\in\mathcal{J}}\sum_{r\in\mathcal{R}}\eta_i d_{ij}W_{ijr}+\sum_{j\in\mathcal{J}}\mu_{ij}Z_{ij}$, and define

$$
D_i(\boldsymbol{\mu}_i)=\min_{\mathbf{Z}_i,\mathbf{W}_i,\mathbf{Y}_i}\{L_i(\boldsymbol{\mu}_i):(\mathbf{Z}_i,\mathbf{W}_i,\mathbf{Y}_i)\in S_i\}\qquad i\in\mathcal{I}. \tag{2.5}
$$

Then, DW-Dual, which is also known as the Lagrangian problem, can be simplified as:

$$
\begin{aligned}
\Psi^*_{LD}=\max\ & \sum_{i\in\mathcal{I}}\theta_i & & & \text{(2.6a)}\\
\text{s.t. } & \sum_{i\in\mathcal{I}}\mu_{ij}=0 & & j\in\mathcal{J}, & \text{(2.6b)}\\
& \theta_i\leq D_i(\boldsymbol{\mu}_i) & & i\in\mathcal{I}. & \text{(2.6c)}
\end{aligned}
$$

Note that $D_i(\boldsymbol{\mu}_i)$ is concave in $(\boldsymbol{\mu}_i)$, and its subgradient at $(\boldsymbol{\mu}_i^k)$ is $\left(\mathbf{Z}_i^k\right)$, where $(\mathbf{Z}_i^k, \mathbf{W}_i^k, \mathbf{Y}_i^k)$ is an optimal solution to (2.5). It follows from the subgradient inequality that

$$\theta_i \leq D_i(\boldsymbol{\mu}_i) \leq D_i(\boldsymbol{\mu}_i^k) + \sum_{j \in \mathcal{J}} Z_{ij}^k (\mu_{ij} - \mu_{ij}^k).$$

We can solve the DW-Dual using a cutting plane method that replaces each $D_i(\boldsymbol{\mu}_i)$ with a relaxation based on a set of subgradients $\mathcal{K}$, and solves the linear program (2.7) at iteration k:

$$\max \sum_{i \in \mathcal{I}} \theta_i \tag{2.7a}$$

$$\text{s.t.} \sum_{i \in \mathcal{I}} \mu_{ij} = 0 \qquad j \in \mathcal{J}, \tag{2.7b}$$

$$\theta_i \leq D_i(\boldsymbol{\mu}_i^k) + \sum_{j \in \mathcal{J}} Z_{ij}^k (\mu_{ij} - \mu_{ij}^k) \qquad i \in \mathcal{I}, k \in \mathcal{K}. \tag{2.7c}$$

However, that cutting plane method is unstable and converges slowly for practical instances. To improve convergence, we use a proximal bundle method proposed in Lubin et al. (2013) that subtracts a quadratic penalty term from the objective (2.7a) weighted by $\tau \geq 0$:

$$\max \sum_{i \in \mathcal{I}} \theta_i - \frac{1}{2} \tau \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} (\mu_{ij} - \mu_{ij}^+)^2,$$

where $(\boldsymbol{\mu}_i^+)$ is the current proximity center. Table 2.2 presents the steps of the method.

Because of the nonconvexity caused by discrete variable sets X and Y, Ψ_{LD}^* provides a lower bound on Ψ^*, that is, the Lagrangian bound. To close the gap, we use Ψ_{LD}^* in a branch-and-bound algorithm, where branching is based on the disagreements in the copy variables $\mathbf{Z}_i$, $i \in \mathcal{I}$. This approach is first proposed by Carøe and Schultz (1999) for two-stage stochastic integer programs. Table 2.3 presents each step of the branch-and-bound algorithm. In Table 2.3, $\mathcal{P}$ denotes the list of current problems with associated lower bounds $\Psi_{LD}^*(P)$. This overall procedure will be referred to the Lagrangian Branch-and-Bound (LBB) algorithm.

Table 2.2: Proximal Bundle Method Used to Solve the DW-Dual (Equation 2.6)

Initialize:	Choose a relative convergence tolerance ϵ. Set $k \leftarrow 1$, $\tau \leftarrow 1$, $m \leftarrow 0.1$, $\boldsymbol{\mu}_i^+ \leftarrow 0$, $i \in \mathcal{I}$. Solve (2.5) with $(\boldsymbol{\mu}_i^k) = (\boldsymbol{\mu}_i^+)$ for each $i \in \mathcal{I}$, $curObj \leftarrow \sum_i D_i(\boldsymbol{\mu}_i^+)$
Step 1	Solve (2.7) to obtain $(\boldsymbol{\mu}_i^k)$, and let $v = \sum_i \theta_i^k - curObj$. If $v/(1 + \lvert curObj \rvert) < \epsilon$, terminate. Else $k \leftarrow k + 1$.
Step 2	Solve (2.5) with $(\boldsymbol{\mu}_i^k)$ $\forall i \in \mathcal{I}$, $newObj \leftarrow \sum_i D_i(\boldsymbol{\mu}_i^k)$.
Step 3	Update $\tau \leftarrow \min(\max(u, \tau/10, 10^{-4}), 10\tau)$, where $u = 2\tau(1 - (newObj - curObj)/v)$.
Step 4	If $(newObj - curObj > m.v)$, update $(\boldsymbol{\mu}_i^+) \leftarrow (\boldsymbol{\mu}_i^k)$, $curObj \leftarrow newObj$. Go to Step 1.

Table 2.3: Branch-and-bound Algorithm Used to Close the Gap Between Ψ^*_{LD} and Ψ^*.

Initialize:	Set $\bar{\Psi} = \infty$ and let $\mathcal{P}$ consist of problem (2.1)
Step 1 (*Termination*)	If $\mathcal{P} = \phi$, then the solution corresponding to $\bar{\Psi}$ is optimal.
Step 2 (*Node Selection*)	Select and delete a problem P from $\mathcal{P}$, solve the Lagrangian problem (2.6) to obtain the lower bound $\Psi^*_{LD}(P)$. If P is infeasible ($\Psi^*_{LD}(P) = -\infty$), go to Step 1.
Step 3a (*Bounding*)	If $\bar{\Psi} \leq \Psi^*_{LD}(P)$, go to Step 1 (this step is executed as soon as the $\Psi_{LD}(P)$ exceeds $\bar{\Psi}$).
Step 3b (*Feasible Solution*)	Else, if the customer solutions constitute a feasible solution to problem (2.1) with an objective function value $\hat{\Psi}$, then $\bar{\Psi} := \min\left\{\bar{\Psi}, \hat{\Psi}\right\}$, and delete from $\mathcal{P}$ all problems P' with $\Psi^*_{LD}(P') \geq \bar{\Psi}$. Go to Step 1.
Step 3c (*Heuristic Solution*)	Else compute the average $\bar{X}_j = \frac{1}{\lvert\mathcal{I}\rvert}\sum_{i\in\mathcal{I}} \hat{Z}_{ij}$ and round it to the closest integer to obtain $\bar{X}^R_j$ for $j \in \mathcal{J}$. If $\bar{\mathbf{X}}^R$ is feasible to problem (2.1) with an objective function value Ψ^R, then $\bar{\Psi} := \min\left\{\bar{\Psi}, \Psi^R\right\}$, and delete from $\mathcal{P}$ all problems P' with $z_{LD}(P') \geq \bar{z}$. Go to Step 4.
Step 4 (*Branching*)	Select a component j of $\bar{\mathbf{X}}$ such that $\bar{X}_j$ is fractional. Add two new problems to $\mathcal{P}$ obtained from P by adding the constraints $X_j = 0$ and $X_j = 1$. Go to Step 1.

2.4 Computational Results and Analyses: Preliminary Model

In this section, the model that is described in Section 2.2 is verified. Also, the algorithm which is explained in Section 2.3 is tested for convergence and performance. First, details regarding the computational environment and datasets used in these tests are presented. Then, preliminary testing is based on one implementation of the LBB algorithm is presented. After that, two more implementations of the LBB algorithm are discussed and extensively tested. Insights are drawn from each experimentation procedure. These insights are essential to the development of the procedures in Section 2.6.

The machine used for testing is Intel Xeon CPU E-2680 with 2 processors and 24 threads each. The machine is running Windows Server 2012. The optimization software is ILOG CPLEX 12.6 at 64-bit architecture. Unless otherwise mentioned, this machine is used at full power for any implementation.

In different parts of this section, some parameters are fine-tuned to ensure desired performance, while other parameters are varied to conduct sensitivity analysis on. The implementation mechanism itself is varied to test the performance of each. In particular, the type of the node queue is varied and the resulting performance is studied. In Section 2.4.2, the variation in preference will be studied through different values of b (defined below). Changing the value of b may affect performance of the solution algorithm by moving the optimal solution closer or further from the optimal solution had preference constraints been neglected. Moreover, the effect of the number of assignment levels $|\mathcal{R}|$, dataset size, and ϕ will be studied. The number of levels affects directly the size of the instance, while the choice of ϕ can have an effect on the time needed for the solution algorithm to converge. In Section 2.4.3, different values of k (defined below) will be tested to fine-tune the algorithm. Higher values of k indicates more iterations done, which means more time spent at each node. A trade-off is potentially needed between the quality generated by higher values of k and the time needed to achieve that. Also, some model features will be examined such as the equality vs. inequality of probability constraints, and the usage of reduced numerical precision and lazy constraints in the solution procedure.

These features can have direct effect on the time required for the algorithm to converge. Finally, Section 2.4.4 will run instances of different sizes and with different values of k.

In the tables below, the running time refers to the time to achieve optimality. Unless otherwise mentioned, the algorithm (LBB or default CPLEX) is allowed to run until it concludes with optimality. If running time exceeds some value (varies for each dataset), the experiment is terminated, and the time is indicated with a 'larger than' symbol (>).

The gap reported in the tables represents the relative gap between the best lower bound (found by relaxation), and the best integer feasible solution achieved up to that point. Specifically,

$$\text{Gap} = \frac{\text{best integer - best LB}}{\text{best integer}}$$

This formula is adjusted to account for values of zero in the denominator, as well as the sign of each value, whenever necessary.

2.4.1 Data Sources

The datasets used in this section are as follows.

- Randomly Generated: these are datasets that were created based on the parameters shown in Table 2.4. **Unless mentioned otherwise, the randomly generated instances are used in the numerical experiments**.
- `USmap49` and `USmap88`: The 'real' map of the US with one node for each state. The demand is proportional to the population of each city. The data set is from Larry Snyder's data (`http://coral.ie.lehigh.edu/larry/research/data-sets-for-stochastic-p-robust-location-problems/`). The failure probability of each location is proportional to the distance between that location and New Orleans, LA. The parameter values are taken similar to those in Aboolian et al. (2013). These datasets are based on 1990 census data, with each node representing one of the 49 capitol cities in the United States. The demand η_i of city i is set to

the city's population divided by 10^4, and the fixed cost f_j is set to the median home value in the city. The transportation cost d_{ij} is calculated as the great circle distance between node i and j. In these datasets, the set of candidate locations $\mathcal{J}$ is equal to the set of customers $\mathcal{I}$, which means that each demand node is a candidate location for a facility. Penalty cost ϕ_i is set to 10,000 for all customers i. Failure probabilities q_j are calculated using $q_j = \beta + 0.1\alpha e^{-d_j/400}$, where $\beta = 0.01$ and d_j is the great circle distance (in miles) between point j and New Orleans. This formula is similar to the one used in Aboolian et al. (2013) and uses the assumption that cities close to New Orleans should have higher values of q_j. This assumption is based on hurricane Katrina disaster in 2005 which was centered around New Orleans.

- `random50` and `random100:` these data sets were also drawn from Larry Snyder's datasets, and represent nodes in a unit square. Failure probabilities are randomly generated. The value of ϕ used is $1,000$.

Table 2.4: Parameters for the Randomly Generated Instances

Parameter	**Range**
Facility fixed cost	$[1,000, 11,000]$
Service Cost	$[100, 500]$
Failure probability	$[0.01, 0.11]$
Demand	$[10, 110]$
dummy cost	$1,000$

Preferences of each customer can be random, based on distance, based on facility quality, or a combination of the distance and quality. Facility quality is measured based on the population (demand) at any given facility; the more populated the node, the higher quality is the facility. Random preferences are used for the randomly generated instances. For other data sets, a utility function for generating preferences is created. The *preferenceScore* formula provides an easy way to combine the effect of distance and the effect of quality (in measures of demand) into a single parameter. The function considers the distance and population of each facility, and, depending on a parameter b, combine the

two measures and gives the preference ordering for each customer. In particular, customer i gives the following the preference score *preferenceScore* for facility j:

$$preferenceScore_{ij} = b(distanceScore_{ij}) + (1 - b)(demandScore_j)$$

where $b \in [0, 1]$ is a parameter used to control the preference scheme, and

$$distanceScore_{ij} = 100 - 100\frac{d_{ij} - d^i_{min}}{d^i_{max} - d^i_{min}}$$

and

$$demandScore_j = 100\frac{\eta_j - \eta_{min}}{\eta_{max} - \eta_{min}}$$

where d^i_{min} is the distance from i to its nearest facility, d^i_{max} is the distance from i to its furthest facility, η_{min} is the least demand among all facilities, and η_{max} is the largest demand among all facilities.

Obviously, the closest facility would get the highest $distanceScore_{ij}$, and the most populated node would get the highest *demandScore*. Based on the values of $preferenceScore_{ij}$ for a particular i, facilities can be ordered accordingly. The value of b would be used in the experiments to control the emphasis of the utility function.

Using the above formulas, for each customer i, the facility j with the highest score $preferenceScore_{ij}$ is the most preferred, followed by facility k with $preferenceScore_{ik}$ such that $preferenceScore_{ik} < preferenceScore_{ij}$, and so on.

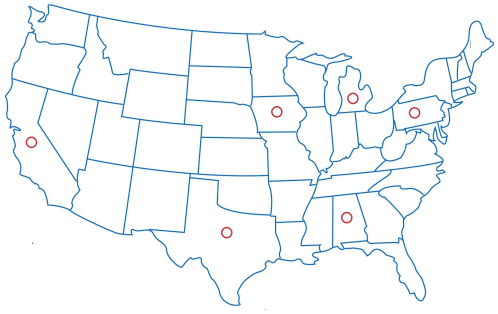
2.4.2 Basic Analysis: CB-LBB

This section describes the preliminary analysis of running our algorithm, and compares that with a default commercial software. The algorithm in this subsection is coded in C++ within callback functions of CPLEX, and therefore denoted by Callbacks Implementation of Lagrangian Branch-and-Bound (CB-LBB). Using this structure, our implementation is using the main skeleton of CPLEX structure, but amending some functions such that our methodology is applied. This approach has the advantages of (1) using the well-organized CPLEX structure, which means fewer code components to track, and (2) benefiting from

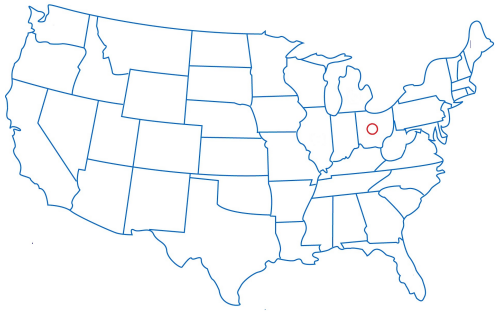
the pre-processing and heuristics that CPLEX applies (by default) before starting the branching process, which would greatly reduce the complexity and size of the problem, making it easier and faster to solve.

Model Validation: Effect of Preference

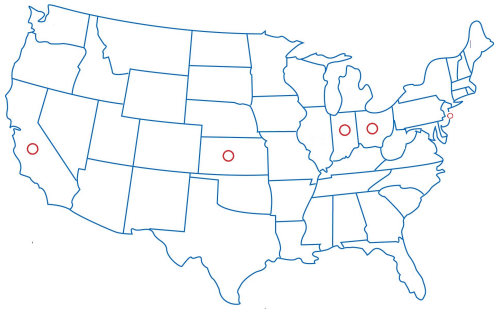
Figure 2.1 shows the difference in location decisions when different preference schemes are used for the `USmap49` dataset with two levels of backup. In 2.1a, the preferences of all customers are based on distance (the closer the facility, the more preferred it is). In 2.1b, the preferences of all customers are identical since it is based on quality (a particular facility is the most preferred for all customers). As can be seen in this case, only one facility is opened. This can be understood by noting that if more than one facility is open, all customers would be assigned to the most preferred among them. Hence, there is no reason to open another facility (in this case of $|\mathcal{R}| = 2$). By increasing $|\mathcal{R}|$, it may be feasible to open more facilities accordingly. In 2.1c, random preferences are generated, and different set of facilities are opened accordingly.



(a) Distance-based preferences



(b) Quality-based preferences



(c) Random preferences

Figure 2.1: Locations Open with Different Preference Orderings with Dataset `USmap49` and $|\mathcal{R}| = 2$

Effect of $|\mathcal{R}|$

Recall that at each node, the LBB algorithm iterates between solving (2.7) and (2.5). Solving these two problems once each is considered one iteration. The procedure of Table 2.2 suggests going through iterations until a stopping criterion is met. Specifically, when the procedure does not improve the solution anymore, it stops. There is no limit on the number of iterations to run at any node. In the tables below, the maximum number of iterations needed at any node is recorded and listed under '*Max* k needed'. This value provides a sense of the maximum time spent at any node in solving the original problem.

It is clear from the tables below that the behavior of the LBB algorithm and default CPLEX is consistent, and predictable with respect to the number of levels $|\mathcal{R}|$. Hence, the numerical experiments below are limited to four levels of backup or less. Experimenting larger values or $|\mathcal{R}|$ would not provide additional significant insights.

Table 2.5 shows the experiments conducted on the `random50` dataset. *Max* k represents the maximum number of iterations needed in any node. The running time for our algorithm is compared to the running time of the default CPLEX MIP solver. The gap after the root node, and *Max* k needed are also provided. As can be seen, the LBB algorithm outperforms default CPLEX in all combinations of $|\mathcal{R}|$ and b presented. Also, for five out of the six combinations, the LBB algorithm is able to arrive to optimality by only solving the root node. Within these instances, the time needed by the default CPLEX is between 44 and 231 folds the time needed by the LBB algorithm. The CPLEX algorithm does little to reduce the gap at the root node. Moreover, the LBB algorithm does not need to iterate more than 13 in the worst case for any node. In fact, for half of the instances (3 out of 6), the overall algorithm terminates at the root node, which in turn needed only one iteration. This shows the efficiency of the procedure, which contributes in the short processing time of the overall LBB algorithm. The instance which does not find the optimal solution at the root node converges after few branches. It is not expected that all instances would converge at the root node. Randomness in generating the dataset allows for such differences.

Table 2.6 shows the running times of data set `USmap49`, as well as the gap after the root node. As can be seen, the processing times are higher for most instances using default CPLEX. Also, the gap after the root node is much lower for the root node when the LBB

Table 2.5: Effect of $|\mathcal{R}|$ on CB-LBB Performance with Dataset `random50`

$\|\mathcal{R}\|$	b	**CB-LBB**			**CPLEX Default**	
		Time (sec)	Gap After Root Node	Max k Needed	Time (Sec)	Gap After Root Node
3	0	27	0.00%	1	1,200	99.97%
4	0	212	0.00%	13	48,985	99.84%
3	0.5	40	0.00%	1	1,941	99.96%
4	0.5	167	67.40%	4	12,980	99.53%
3	1	29	0.00%	1	2,607	99.91%
4	1	134	0.00%	7	15,726	99.78%

Table 2.6: Effect of $|\mathcal{R}|$ on CB-LBB Performance with Dataset `USmap49`

$\|\mathcal{R}\|$	b	**CB-LBB**		**CPLEX Default**	
		Time (Sec)	Gap After Root Node	Time (sec)	Gap After Root Node
3	0	5,759	75.34%	577	76.55%
4	0	74,482	77.04%	2,273	76.30%
3	0.5	2,323	40.43%	15,289	66.94%
4	0.5	7,659	44.57%	7,102	69.28%
3	1	5,271	32.91%	>32,400	66.76%
4	1	1,995	22.28%	14,480	69.06%

algorithm is in use.

Table 2.7 shows the results for dataset `random100`. As can be seen, the default CPLEX is unable to arrive at optimality in any instance before 5 hours, whereas the LBB algorithm concludes optimal for four instances (with a maximum processing time of 70 minutes), and reduces the gap for the other two instances into less than 0.5% within 5 hours or less.

Table 2.7: Effect of $|\mathcal{R}|$ on CB-LBB Performance with Dataset `random100`

$\|\mathcal{R}\|$	b	**CB-LBB** Time to Optimal (Sec)	**CPLEX Default** Gap After 5 Hours
3	0	413	99.9%
4	0	4,176	98.2%
3	0.5	190	99.9%
4	0.5	5 hours at 0.31% gap	99.9%
3	1	1,708	99.98%
4	1	4 hours at 0.1% gap	99.9%

Table 2.8 shows the results for the randomly generated instances. The preferences were randomly generated. *Max* k represents the maximum number of iterations needed in any node. These instances are combinations of either 20 or 40 facilities with 48 or 96 customers. Each combination is repeated three times with different random number generator's seed. The running times are limited to 3 hours (10,800 seconds). As can be seen, the LBB algorithm arrives optimality much faster than the default CPLEX. Also, the LBB algorithm is able to terminate immediately after the root node. The CPLEX algorithm closes less than 1% gap for all instances.

Table 2.8: Performance of CB-LBB with Randomly Generated Datasets

$\lvert\mathcal{I}\rvert \times \lvert\mathcal{J}\rvert \times \lvert\mathcal{R}\rvert$	**CB-LBB**			**CPLEX Default**	
	Time (sec)	Gap after root node	Max k needed	Time (sec)	Gap after root node
$48 \times 20 \times 4$	23	0.00%	1	>10,800	99.39%
$48 \times 20 \times 4$	64	0.00%	12	4,290	99.51%
$48 \times 20 \times 4$	120	54.14%	17	>10,800	98.02%
$48 \times 40 \times 4$	4,947	47.63%	59	>10,800	99.03%
$48 \times 40 \times 4$	1,675	48.99%	42	>10,800	99.61%
$48 \times 40 \times 4$	1,328	62.23%	26	3,405	99.62%
$96 \times 20 \times 4$	76	0.00%	23	1,632	99.48%
$96 \times 20 \times 4$	44	0.00%	7	3,880	99.51%
$96 \times 20 \times 4$	140	60.26%	9	2,859	99.71%
$96 \times 40 \times 4$	1,431	43.86%	23	>10,800	99.55%
$96 \times 40 \times 4$	4,444	92.06%	36	>10,800	99.61%
$96 \times 40 \times 4$	925	48.3%	24	>10,800	99.64%

Effect of ϕ

The value of $\phi = \phi_i,\ \forall i \in \mathcal{I}$, is an important parameter of the model. Higher values of ϕ means higher penalties for not fulfilling demand. Accordingly, the assignment strategy would be to avoid having this penalty at lower levels of backup, and push it to an advanced level of backup (where the probability of enduring these penalties is small). Interestingly, the LBB algorithm can exploit higher values of ϕ and arrive at optimal solutions faster, as Table 2.9 shows, whereas default CPLEX would take a longer time as ϕ increases.

Table 2.9: Effect of Changing the Value of ϕ on the Performance of CB-LBB with Dataset `random50` and $|\mathcal{R}| = 3$.

	$\phi = 80$		$\phi = 800$		$\phi = 8,000$		$\phi = 80,000$	
b	CB-LBB	CPLEX	CB-LBB	CPLEX	CB-LBB	CPLEX	CB-LBB	CPLEX
0	529	102	58	477	40	395	27	1,200
0.5	121	541	59	647	75	579	40	1,941
1	128	>1,000	47	>1,000	37	>1,000	29	2,607

Table 2.10 shows the effect of the value of the dummy cost on the location and allocation decisions. The dataset used in `USmap49` with two levels $|\mathcal{R}| = 2$. For smaller values of ϕ, higher portions of the demand is lost (assigned to the dummy facility). However, by increasing the dummy cost, more facilities are opened, and less demand is lost.

Table 2.10: Effect of ϕ on the Location and Allocation Decisions with Dataset `USmap49` and $|\mathcal{R}| = 2$

ϕ	% Assigned to Dummy at first level	Number of open locations (excluding Dummy)
100	100%	0
500	58%	1
800	19%	5
1,000	13%	5
1,200	6%	5
1,500	5%	5
1,800	4%	5
2,500	0%	5

Insights From Preliminary Analysis and CB-LBB Testing

From the results shown above, we can conclude that the preference of customers indeed has an effect on the decisions of how many and where to locate facilities. This effect can be significant if the preferences differ greatly from the distance-based preferences. Moreover, as can be expected, increasing the value of $|\mathcal{R}|$ would result in increasing the size of the problem, which in turn results in longer time needed to attain optimality. The effect of ϕ is model-dependent. It seems that our algorithm is able to respond to higher values of ϕ faster by pushing the assignment of the dummy facility to the last level R. CPLEX, however, needs longer time as ϕ increase, which may be due to computational difficulties.

As explained at the beginning of this section, our algorithm was implemented by amending particular functions of CPLEX tree structure using callbacks. The goal was to have fewer code components to write and to use the additional features that come with CPLEX such as preprocessing and heuristics.

While the results above are promising, further experiments showed that the performance of our algorithm is not consistent; for some instances of different sizes, CPLEX is significantly better in terms of solution time and quality. Also, there was no trend

or explanation of why some instances are easier to solve by our algorithm than others. This inconsistency and inability to explain behavior, encouraged us to look for further implementation technique.

Furthermore, careful observation of the implementation showed that our implementation does not go through all the pre-processing and heuristics that default CPLEX does. After investigation, it was revealed that using callbacks causes CPLEX to switch off the use of some of these methods, since the data handling is no longer safe (i.e., when a default function is replaced a user-defined callback, the output of that function is not guaranteed to be in the same form and structure that the next function expects. Therefore, CPLEX only keeps the main tree structure without going through the additional features). Knowing this, the use of CPLEX structure has proven to be less appealing. Consequently, a new approach is required, which will be discussed in the following section.

2.4.3 Stack Queue Tree: Stack-LBB

In the previous section, it was argued that using callbacks is not ideal and does not provide consistent results. In this section, the main branch-and-bound tree structure is built using C++ without using the CPLEX callbacks. Thus, this implementation is independent of the CPLEX skeleton. However, CPLEX is used to solve the DW-Dual and customer subproblems as discussed in Section 2.3. These problems are sent to CPLEX and solved to optimality with no other dependence on CPLEX.

A major feature in any branch-and-bound tree implementation is the order of nodes to be explored and solved. In this particular implementation, the newly created nodes are ordered in a stack (LIFO) queue. Therefore, this implementation will be referred to as Stack Queue Implementation of Lagrangian Branch-and-Bound (Stack-LBB). Such ordering would process depth-first and explore one path of nodes until an integer solution is found or it is proven to be unpromising. In doing so, and while going down the tree, the upper bound is improved, which makes the exploration of the next branches faster by fathoming all nodes with lower bound that exceeds the best integer solution found. However, the disadvantage of such an approach is spending too much time within a particular branch because of the continued creation of nodes within that branch. This branch

may eventually be fathomed without contributing significantly in improving the lower and upper bounds. Unless the upper bound is significantly improved, the tree would keep on growing and becomes harder to solve. Also, memory usage and tractability would become serious issues that hinder the successful termination of the algorithm.

Stack-LBB Algorithm Performance

Table results below show the numerical experiments for different randomly generated instances. Recall that k represents the maximum number of iterations (described in Section 2.3) allowed at each node. Once the k iterations are performed, the bounds are fixed and the node branches accordingly. The procedure is compared with the default CPLEX (with all preprocessing and heuristics features switched on) and with CPLEX when the preprocessing methods are switched off. The former represents the best CPLEX can do. The latter represents the embedded branch and bound structure in CPLEX without the reductions done at the beginning before starting the tree, which makes it equivalent to our method since we do not do preprocessing or reductions before starting the tree structure.

As can be shown in Table 2.11, the LBB algorithm needs less time than default CPLEX to arrive to optimality in 12 out of 17 instances. In the other five instances, default CPLEX outperforms the LBB algorithm. The LBB algorithm always outperforms the no-preprocessing CPLEX. Moreover, the time needed for the LBB algorithm to converge decreases as the value of k decreases. This suggests that the iteration process is more computationally expensive than the branching and the creation of new nodes. Accordingly, this also suggests that an improvement in the iteration process would significantly enhance the performance of the algorithm.

Table 2.12 shows the number of nodes needed to arrive at optimality. It is immediately apparent that the LBB algorithm requires far less nodes than the default CPLEX and the no-preprocessing CPLEX. Another observation is that, generally, the number of nodes needed decreases as k decreases. This is counterintuitive since one would expect that with higher k, the node bounds would improve, which would lead to the fathoming of more nodes and branches which, in turn, would result in smaller number of nodes needed. A possible explanation for this pattern is that more nodes of good quality are created, which

would require solving larger number of nodes. If the majority of nodes have poor bounds, once the upper bound of the tree surpass a particular level, a large portion of these nodes will automatically be fathomed.

Table 2.11: Processing Time (in Seconds) Using Stack-LBB Algorithm

$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{R}\|$	$k = 10$	$k = 7$	$k = 5$	$k = 3$	$k = 1$	$k = 0$	CPLEX Default	CPLEX (no preprocessing)
48	20	4	508	490	472	210	228	219	744	>3,600
48	20	20	882	891	845	427	447	450	5,890	>9,976
48	20	4	598	574	532	295	289	242	1,008	>3,600
48	20	20	34,560	1,055	961	552	514	514	>24 hrs	>9,239
48	20	4	647	671	603	253	256	370	2,118	>3,600
48	20	20	1,102	1,019	507	489	489	>6 hrs	>24 hrs	>24 hrs
48	21	4	1,338	1,292	1,372	403	430	427	144	3,620
48	21	21	2,237	2,178	2,269	810	842	841	1,521	12,779
48	22	4	3,121	3,080	3,412	646	590	590	220	9,923
48	22	22	4,704	4,712	5,187	1,235	1,245	1,246	1,249	>24hours
48	23	4	6,106	6,077	5,945	1,267	1,251	1,257	211	8,219
48	23	23	9,398	9,755	9,556	2,413	2,355	2,349	1,722	48,564
48	24	4	7,915	9,510	6,865	665	676	676	619	13,660
48	24	24	11,784	13,666	10,771	1,406	1,459	1,458	4,621	>84,755
48	25	4	8,569	9,513	7,155	1,265	688	686	178	6,672
48	25	25	11,868	14,935	11,058	1,265	1,731	1,503	1,357	220,609
48	30	4	>50hrs	>50hrs	120,907	9,076	11,192	11,016	1,134	43,367

Table 2.12: Number of Nodes Explored Using Stack-LBB Algorithm

$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{R}\|$	$k = 10$	$k = 7$	$k = 5$	$k = 3$	$k = 1$	$k = 0$	CPLEX Default	CPLEX (no preprocessing)
48	20	4	715	697	719	535	559	559	1,771,376	>2,000,000
48	20	20	725	743	739	551	579	579	2,452,833	>243,693
48	20	4	849	801	793	665	671	671	2,735,558	>3,000,000
48	20	20	855	851	817	625	601	601	>1,000,000	>1,830,439
48	20	4	897	925	883	633	651	651	4,661,872	>5,000,000
48	20	20	897	897	863	633	629	629	>7,000,000	>504,369
48	20	4	1,419	1,445	1,515	829	871	860	187,274	3,814,008
48	21	21	1,415	1,439	1,519	873	873	873	441,039	1,872,257
48	21	4	2,465	2,501	2,783	1,077	1,009	1,009	314,785	5,285,216
48	22	22	2,323	2,435	2,703	1,073	1,091	1,091	276,388	>1,811,366
48	22	4	3,665	3,665	3,919	1,813	1,775	1,775	200,163	2,304,875
48	23	23	3,629	3,809	3,921	1,779	1,783	1,783	174,416	2,219,227
48	20	4	3,647	4,205	3,579	899	923	923	707,185	12,319,311
48	24	24	3,635	4,083	3,607	927	957	957	761,758	>11,844,992
48	24	4	3,661	4,071	3,601	945	963	963	150,631	1,927,315
48	20	25	3,617	4,361	3,663	945	1,143	1,023	61,916	1,401,060
48	20	4	>18,000	>18,000	15,401	3,861	4,411	4,411	1,040,130	12,250,381

Evaluating Features

In pursuit of improving the performance of the algorithm, three main features of the Stack-LBB implementation in C++ are analyzed.

Equality vs. Inequality in Probability Constraints
The constraints (2.1g) (and their equivalent in later reformulations) can have very small variable coefficients. Due to various mathematical operations, numerical rounding would result in computational difficulties. This is especially important since they are equality constraints. Brief testing shows that the optimal point does not change by changing the sense of constraints (2.1g) from $=$ to $\geq$. However, the time needed and nodes explored change. Since an equality constraint is more restrictive, it results in less time and fewer nodes. Results are shown in Table 2.13.

Table 2.13: Effect of the Sense of Constraints (2.1g) on the Performance of Stack-LBB

				Time (Seconds)		Nodes Explored	
$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{J}\rvert$	$\lvert\mathcal{R}\rvert$	Precision	$\geq$	$=$	$\geq$	$=$
48	20	4	Default	12,937	1,909	16,720,240	2,868,844
48	20	4	Reduced	2,597	2,548	4,167,587	2,915,280
48	20	4	Default	2,271	839	3,760,331	1,492,474
48	20	4	Reduced	1,708	838	2,450,213	1,624,500

Reduced Precision
Another way to control the numerical errors resulting from rounding is by fixing the precision of the numbers by controlling the number of decimal digits. Table 2.13 shows that by limiting the number of decimal digits, performance is improved.

Regular vs. Lazy DW-Dual Constraints
The algorithm in Section 2.3 describes how a new realization of constraint (2.7c) is added to the DW-Dual problem (2.7) after each iteration. In building the DW-Dual problem object in CPLEX, the constraints (2.7c) can be added either as regular constraints or as

lazy constraints. In general, constraints are added to a problem as lazy constraints if there is a belief that some of them might be redundant and/or to reduce the computational footprint of the problem. The mathematical meaning and implications of lazy constraints are beyond the scope of the current discussion. Table 2.14 shows that although there is a slight advantage of regular constraints, the decision of adding (2.7c) as regular or lazy constraint does not have a significant effect on the performance of the algorithm.

Table 2.14: Effect of Regular and Lazy Constraints on the Performance of Stack-LBB

			Time (Seconds)		Nodes Explored	
$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{R}\|$	Regular	Lazy	Regular	Lazy
48	20	4	5,741	5,561	913	913
48	20	4	4,397	5,532	583	563
48	20	4	5,059	>7,000	563	>1,100
48	20	4	5,179	8,987	659	649

Insights from Stack Queue Tree

The results of experiments done on the branch-and-bound tree with stack queue show that the algorithm is efficient for some instances, especially small ones. However, for larger instances, the algorithm performance deteriorates significantly. Moreover, while there is a clearer pattern of behavior for this implementation compared to the implementation using CPLEX callbacks structure, there are still unexplained points in this implementation as well. Furthermore, the increase in the time needed to reach optimality with increasing k is counterintuitive, problematic, and may be a sign of a bigger problem that we are unaware of. Therefore, it is discouraged to move on with this implementation.

Ordering the nodes in a stack (LIFO) queue ignores the quality of the nodes in the queue. The next implementation uses the properties of the created nodes to order them. This has the potential of improving the performance of the whole algorithm.

2.4.4 Priority Queue Tree: PQ-LBB

Each node created in the branch and bound tree comes with inherited and acquired characteristics. These characteristics include the variable bounds and the objective value bounds. Specifically, the lower bound (in a minimization problem) of the objective function value of the node is a very important characteristic of the node. Let A and B be two nodes that are waiting to be processed in a branch-and-bound tree. Whenever the upper bound (feasible solution) for node A (equivalently, for a branch starting from node A) is found to be less than the lower bound of node B, node B should be fathomed. Fathoming nodes would reduce the size of the queue and potentially reduce the time needed to arrive to the optimal solution. Therefore, it is justified to design a branch-and-bound tree such that the queue of nodes prioritizes the nodes with the lowest lower bounds. Processing these nodes first would potentially eliminate the need to process nodes of less quality.

This is how the tree is designed here; by ordering the created nodes in a queue such that the ones with the lowest lower bound on top and processed first. Therefore, this implementation will be referred to as Priority Queue Implementation of Lagrangian Branch-and-Bound (PQ-LBB)

PQ-LBB Algorithm Performance

Table 2.15 shows the time required to arrive to optimality for randomly generated instances using the priority queue tree implementation. It is apparent from the table that these times are significantly less than those of the stack queue tree implementation above. This shows that the algorithm indeed benefits from prioritizing nodes with lower lower bounds. Moreover, the algorithm is generally better than default CPLEX, and significantly surpasses the no-preprocessing CPLEX. Table 2.16 shows the number of nodes explored. A similar pattern is observed here too.

Table 2.15: Processing Time (in Seconds) Using PQ-LBB

$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{R}\|$	$k=9$	$k=7$	$k=5$	$k=3$	$k=1$	$k=0$	CPLEX Default	CPLEX (no preprocessing)
48	20	4	66	80	71	71	67	71	744	3,253
48	20	20	542	568	498	472	481	459	5,890	>9,976
48	21	4	122	118	130	109	114	107	144	3,620
48	21	21	531	470	567	444	473	447	1,521	12,779
48	22	4	118	119	120	105	98	102	220	9,923
48	22	22	528	571	569	495	471	491	1,249	>24hours
48	23	4	186	212	203	183	156	175	211	8,219
48	23	23	985	975	1025	965	833	897	1,722	48,564
48	24	4	118	115	116	113	109	108	619	13,660
48	24	24	650	620	636	643	639	612	4,621	>84,755
48	25	4	106	105	111	104	99	104	178	6,672
48	25	25	629	656	657	633	586	619	1,357	220,609

Table 2.16: Number of Nodes Explored Using PQ-LBB

$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{R}\|$	$k=9$	$k=7$	$k=5$	$k=3$	$k=1$	$k=0$	CPLEX Default	CPLEX (no preprocessing)
48	20	4	2,325	2,847	2,603	2,547	2,773	2,995	1,771,376	3,542,148
48	20	20	3,825	4,361	4,477	4,763	4,454	4,583	2,452,833	>2,574,439
48	21	4	4,825	4,863	5,387	4,409	4,933	4,717	187,274	3,814,008
48	21	21	4,937	4,351	5,511	4,191	4,997	4,615	441,039	1,872,257
48	21	4	4,275	4,389	4,481	3,897	3,909	4,115	314,785	5,285,216
48	22	22	4,059	4,417	4,519	3,903	3,933	4,315	276,388	>1,811,366
48	21	4	7,433	8,507	8,405	7,441	6,683	7,729	200,163	2,304,875
48	23	23	7,501	7,559	8,089	7,477	6,689	7,507	174,416	2,219,227
48	21	4	3,251	3,215	3,379	3,395	3,663	3,487	707,185	12,319,311
48	24	24	3,383	3,207	3,367	3,403	3,669	4,395	761,758	>11,844,992
48	25	4	2,999	3,011	3,395	3,155	3,247	3,445	150,631	1,927,315
48	25	25	3,159	3,225	3,415	3,291	3,255	3,459	61,916	1,401,060

PQ-LBB Algorithm Performance: Larger Instances

Since this implementation shows promising results on small instances, larger datasets are now tested to confirm the suitability of the procedure. Tables 2.17-2.20 show results for randomly generated datasets. The LBB algorithm is applied only at the root node. After the root node, the procedures performs branching without going into Lagrangian iterations.

As can be seen in Tables 2.17 and 2.18, the LBB requires far less time than CPLEX and still achieve comparable gaps after the root node. The gap percentage after one hour of running is significant for both methods. However, the LBB tends to close the gap more than CPLEX most of the time (Table 2.19). Since the Lagrangian iterations are performed only at the root node, the LBB algorithm is able to go through a large number of nodes, as appears in Table 2.20.

Table 2.17: Processing Time (in Seconds) of the Root Node Using PQ-LBB with Large Datasets

$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{R}\|$	$k=40$	$k=30$	$k=20$	$k=10$	$k=9$	$k=7$	$k=5$	$k=3$	$k=1$	$k=0$	CPLEX
96	30	4	35	26	16	9	8	6	5	1	1	1	536
96	40	4	65	46	30	15	14	12	10	8	3	2	245
96	50	4	109	72	46	23	21	17	15	10	6	2	487
96	50	4	117	80	49	25	21	18	15	11	5	3	635

Table 2.18: Gap (%) After Processing the Root Node Using PQ-LBB with Large Datasets

$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{R}\|$	$k=40$	$k=30$	$k=20$	$k=10$	$k=9$	$k=7$	$k=5$	$k=3$	$k=1$	$k=0$	CPLEX
96	30	4	42.11	44.56	45.72	46.98	47.46	48.92	52.31	55.32	55.32	55.32	46.11
96	40	4	42.63	43.29	44.07	45.24	45.79	46.92	50.51	52.84	52.84	52.84	44.10
96	50	4	45.11	45.65	46.30	47.39	47.80	49.30	52.49	53.74	53.74	53.74	84.66
96	50	4	43.79	44.34	45.05	46.19	46.58	48.36	51.15	53.22	53.22	53.22	41.69

Next, the US cities datasets are tested. Tables 2.21-2.24 show results of testing the `USmap88` dataset with $|\mathcal{R}| = 4$ and preferences are randomly generated. Recall that in this dataset, failure probabilities q_j are calculated using $q_j = \beta + 0.1\alpha e^{-d_j/400}$, where

Table 2.19: Gap (%) After 60 Minutes Using PQ-LBB with Large Datasets

$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{R}\|$	$k = 40$	$k = 30$	$k = 20$	$k = 10$	$k = 9$	$k = 7$	$k = 5$	$k = 3$	$k = 1$	$k = 0$	CPLEX
96	30	4	7.79	8.22	8.19	8.29	8.15	7.58	8.11	8.32	7.6	7.55	8.51
96	40	4	20.09	20.18	19.74	20.17	19.37	19.44	19.94	19.1	19.22	19.11	32.13
96	50	4	24.46	23.96	24.28	24.09	24.44	24.16	23.96	23.95	23.78	24.44	37.28
96	50	4	23.06	23.39	23.52	23.59	23.86	23.4	23.11	23.03	23	23.19	38.72

Table 2.20: Number of Nodes Explored After 60 Minutes Using PQ-LBB with Large Datasets

$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{R}\|$	$k = 40$	$k = 30$	$k = 20$	$k = 10$	$k = 9$	$k = 7$	$k = 5$	$k = 3$	$k = 1$	$k = 0$	CPLEX
96	30	4	131,645	118,317	123,655	120,427	123,019	124,075	124,433	120,907	130,115	130,779	227,345
96	30	4	55,441	58,225	55,287	56,237	56,581	59,171	56,131	57,627	60,281	59,661	50,954
96	40	4	38,337	39,249	39,541	40,183	39,019	39,809	39,883	41,127	42,233	41,527	24,262
96	50	4	36,057	36,213	36,129	36,733	36,495	36,985	37,883	37,965	39,627	39,547	24,716

$\beta = 0.01$ and d_j is the great cycle distance (in miles) between point j and New Orleans, LA. Therefore, different values of α are used in the experiments. Since these runs took a long time before convergence, all runs were stopped after 60 minutes and the important statistics were collected. Also, the LBB algorithm is applied only on the root node. After the root node, the procedure continues with branching without going through the Lagrangian iterations.

The LBB algorithm requires far less time at the root node than the default CPLEX. This is true across different values of k and α, as seen in Table 2.21. However, the two methods are fairly comparative based on the efficiency at the root node (measured by the gap after fathoming the root node), as it is represented in Table 2.22. This shows that the extra time needed by CPLEX for some instances may be justified.

It is clear from Table 2.23 that the gap percentage is still significant after running both methods for 60 minutes, except for one instance for which CPLEX found an optimal within this time. Apart from this instance, the PQ-LBB algorithm appears to be reducing the gap better than CPLEX. The number of nodes explored after running for 60 minutes is comparable for both methods, as demonstrated in Table 2.24. Recall that in these runs,

Table 2.21: Processing Time (in Seconds) of the Root Node Using PQ-LBB with `USmap88` and $|\mathcal{R}| = 4$

α	$k = 9$	$k = 7$	$k = 5$	$k = 3$	$k = 1$	$k = 0$	CPLEX
1	30	28	23	19	13	6	352
1.05	31	27	23	19	12	5	444
1.1	33	28	24	20	12	5	368
1.15	34	28	23	19	12	5	524
1.2	34	31	25	21	13	6	260
1.25	34	29	24	21	13	6	377
1.3	35	31	26	23	14	6	330
1.35	35	29	26	21	13	5	343
1.4	35	30	26	21	14	5	235
1.45	35	32	26	21	13	6	380

Table 2.22: Gap (%) After Processing the Root Node Using PQ-LBB with `USmap88` and $|\mathcal{R}| = 4$

α	$k = 9$	$k = 7$	$k = 5$	$k = 3$	$k = 1$	$k = 0$	CPLEX
1	69.68	71.63	74.88	80.35	81.58	81.58	85.37
1.05	69.69	71.65	75.43	80.52	81.48	81.48	64.29
1.1	69.68	71.64	75.40	80.50	81.42	81.42	49.89
1.15	69.54	71.53	75.29	80.46	81.31	81.31	72.47
1.2	69.48	71.52	75.32	80.44	81.24	81.24	85.88
1.25	69.43	71.38	75.02	80.40	81.09	81.09	73.39
1.3	69.36	71.33	75.23	80.36	81.03	81.03	90.68
1.35	69.36	71.32	74.98	80.16	80.91	80.91	90.66
1.4	69.35	71.23	75.03	80.08	80.81	80.81	90.77
1.45	69.34	71.30	75.23	80.16	80.69	80.69	90.71

PQ-LBB is applying the Lagrangian iterations at the root node only, which explains the large number of nodes exposed within one hour.

Table 2.23: Gap (%) After 60 Minutes Using PQ-LBB with `USmap88` and $|\mathcal{R}| = 4$

α	$k = 9$	$k = 7$	$k = 5$	$k = 3$	$k = 1$	$k = 0$	CPLEX
1	24.56	22.08	23.17	22.25	22.81	22.70	28.66
1.05	22.49	21.67	24.18	22.34	23.19	22.61	0
1.1	23.10	22.28	24.57	22.60	22.94	22.89	32.58
1.15	23.56	23.45	22.74	22.85	23.00	22.49	28.37
1.2	24.81	24.82	23.03	23.07	23.25	23.03	27.13
1.25	24.23	23.10	24.83	23.28	23.48	23.17	31.22
1.3	23.39	23.68	22.15	23.31	23.53	23.27	30.62
1.35	23.37	23.26	23.61	24.02	23.48	23.72	34.77
1.4	22.83	23.25	23.26	23.80	23.32	23.69	26.96
1.45	22.83	24.62	23.22	22.86	23.12	23.57	34.45

Table 2.24: Number of Nodes Explored After 60 Minutes Using PQ-LBB with `USmap88` and $|\mathcal{R}| = 4$

α	$k = 9$	$k = 7$	$k = 5$	$k = 3$	$k = 1$	$k = 0$	CPLEX
1	16,587	16,019	15,965	16,361	17,429	17,577	11,167
1.05	16,215	16,311	16,421	16,347	18,109	17,597	6,719
1.1	16,093	15,969	16,763	16,229	17,407	17,501	12,127
1.15	15,535	15,795	15,665	16,081	17,277	17,371	11,850
1.2	15,735	16,485	15,599	15,871	17,049	17,427	15,600
1.25	1,5473	15,635	16,087	15,813	16,959	17,295	21,123
1.3	15,551	15,675	15,515	15,807	16,831	17,109	11,567
1.35	15,321	16,075	15,341	16,225	16,647	16,907	11,053
1.4	15,141	15,411	15,073	15,743	16,321	16,833	15,958
1.45	15,483	14,931	14,779	15,485	16,425	16,697	14,983

Stack-LBB vs. PQ-LBB Comparison

The two implementations, Stack-LBB and PQ-LBB, are compared in Table 2.25 in terms of running time in seconds. The values inside the table are averages over k for all instances of the same size, if available. As shown, the PQ-LBB dominates Stack-LBB for all datasets tested.

Table 2.25: Processing Time (in Seconds) Using Stack-LBB and PQ-LBB: A Comparison

$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{R}\|$	Stack-LBB	PQ-LBB
48	20	4	414	67
48	20	20	2,688	292
48	21	4	877	117
48	21	21	1,530	489
48	22	4	1,907	110
48	22	22	3,055	521
48	23	4	3,651	186
48	23	23	5,971	947
48	24	4	4,385	113
48	24	24	6,757	633
48	25	4	4,646	105
48	25	25	7,060	630

Shortcomings of PQ-LBB Algorithm

While the previous results show that PQ-LBB is superior to Stack-LBB, the behavior of PQ-LBB is not consistent and not always explainable. Table 2.26 shows results of some of the datasets for which PQ-LBB performed poorly. This poor performance can partly be explained by the increased size of the problem, especially higher $|\mathcal{J}|$ values. However, other instances are identical to the ones where PQ-LBB has performed very well. This demonstrates the shortcomings of this implementation and the need to arrive at a better implementation technique.

Table 2.26: Processing Time (in Seconds) for Datasets with Poor Performance of PQ-LBB

Dataset	LBB	CPLEX
Random Gen. $40 \times 40 \times 4$	1,967	672
Random Gen. $40 \times 40 \times 4$	2,471	679
Random Gen. $48 \times 30 \times 4$	>3,600	431
`USmap49`	1,525	114
`USmap49`	210	104
`USmap49`	1,699	619

Insights from PQ-LBB Algorithm

From the extensive analysis conducted on the priority queue implementation of the LBB algorithm, it is obvious that the PQ-LBB is superior to the Stack-LBB in terms of the time required and nodes explored until optimality. The PQ-LBB behavior is more consistent and shows patterns.

The results also show that the PQ-LBB algorithm does not perform well with larger datasets. When tested with still larger datasets, the PQ-LBB algorithm's performance deteriorates sharply, leading to excessive processing times and large optimality gaps. A major consideration in the procedure is deciding on which nodes to apply the Lagrangian iterations on. Applying the Lagrangian iterations on a high number of nodes means slower progress at these nodes. This can be fruitful only if there was a significant improvement in the bounds after processing these nodes. Experiments have shown that this is not guaranteed with the current version of PQ-LBB.

Therefore, the procedure needs to be reviewed thoroughly. It is essential to find an iteration procedure that requires less computational resources and needs less time to complete. A suitable algorithm would start from a new formulation of the model itself that reduces the size of the problem, and builds on that to tighten the relaxation in order to have an efficient solution algorithm. This will be shown in Sections 2.5-2.7.

2.5 Modified Model

The model developed in this section builds on the model in Section 2.2. The major difference between the two models relies on the way allocation of customers is done to open facilities when the cost of service is higher than the penalty cost. A detailed discussion is given after the model formulation. Moreover, the modified model has less number of variables than the preliminary. This will help in improving the performance of the solution algorithms.

In this section, we reformulate the uncapacitated fixed-charge location problem which considers the preferences of customers and the reliability of facilities. A central planner selects facility locations from a set of candidate sites to minimize the total cost of opening facilities and providing service. Each customer has a strict preference order over a subset of the candidate sites, and uses her most preferred available facility. If that facility fails due to a disruptive event, the customer attends her next preferred available facility. This model bridges the gap between the location models that consider the preferences of customers and the ones that consider the reliability of facilities.

The main contribution of this section is providing a formulation of the problem that (1) gives customers full power to decide on the allocation based on their preference order, and (2) reduces the number of variables.

Let $\mathcal{I}$ be the set of customer types and $\mathcal{J}$ be the set of candidate facility sites. Each customer type, referred to as "customer" hereafter, is characterized by a preference list over the set of candidate facility sites, a service cost vector and a demand value.

We denote the demand of customer $i \in \mathcal{I}$ by η_i and the fixed cost of opening facility $j \in \mathcal{J}$ by f_j. Let d_{ij} be the cost of serving customer i at location j. Each customer i has a strict preference order over a subset of the candidate sites. Let $h(i, j)$ be the order of facility $j \in \mathcal{J}$ in the preference list of customer i. If facility $j \in \mathcal{J}$ is less preferred than facility $k \in \mathcal{J}$ for customer i, that is, $h(i, j) > h(i, k)$, we denote it by $j <_i k$.

The facility at site j fails with probability $q_j \in [0, 1]$ independent of other facilities. Each customer may patronize up to $|\mathcal{R}| \geq 1$ backup facilities in the order of preference, and attends the backup facility at level $r \leq |\mathcal{R}|$ if and only if the facilities at levels

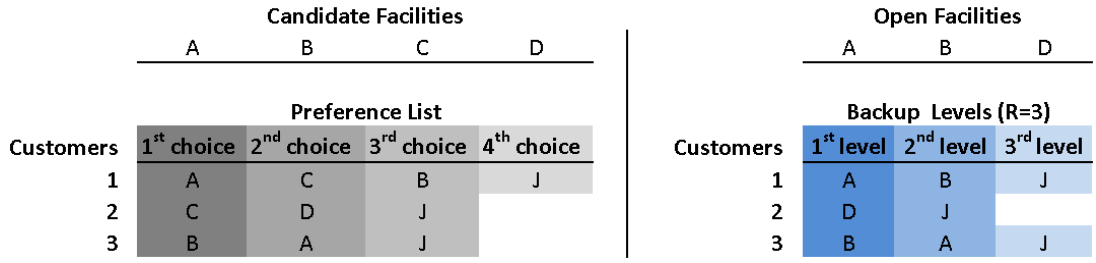


Figure 2.2: Demonstration of Assignment Based on $|\mathcal{R}|$ and Preference List: There are 4 facilities and 3 customers in this small example. The number of backup levels $|\mathcal{R}| = 3$ and J is the Dummy facility. The preference list of each customer is given on the left. Customer 2 does not prefer to use facility A and B. Similarly, customer 3 does not prefer to use facilities C and D. A feasible solution is given on the right. Facilities A, B and D are open. Customers use their most preferred open facility in the first backup level. They are served by the Dummy facility J in the last backup level.

$1, \ldots, r-1$ fail. Note that the preference list of a customer may include less than $|\mathcal{R}|$ facilities (see Figure 2.2), in which case the number of backup facilities for that customer will be limited by the size of her preference list. The first backup level always includes the most preferred available facility that is used by the customer in the absence of any failure. In other words, the facility at the first backup level is not actually a "backup" facility, but rather it is used by the customer under normal operating conditions when there is no disruption. Furthermore, the last backup level always includes a dummy facility J, which is assumed to serve the customer when all facilities in the earlier backup levels fail (see Figure 2.2). Customer i incurs a disutility cost of not accessing service, ϕ_i when she has to use the dummy facility. For the dummy facility, it is assumed that the fixed cost $f_J = 0$, the failure probability $q_J = 0$ and the service cost $d_{iJ} = \phi_i$.

In our model, it is assumed that failures happen, if any, before customers make trips to any facilities. In other words, by the time the demand requires fulfillment, it will be known with certainty which facilities are available and which are not. Available facilities from that point on are expected to remain available throughout the fulfillment period, and no failed facility would become available again.

Let binary variable X_j be one if facility j is open, and zero otherwise. If $j \in \mathcal{J}$ is the s^{th} most preferred facility for customer i, then she may have facility j only at a backup level that is less than or equal to s. This is because, if customer i has facility j at backup level $r > s$, then all facilities in the first $(r-1) \geq s$ backup levels should be more preferred than facility j, but this contradicts the fact that j is the s^{th} most preferred facility. For a given backup level r and customer i, we denote the set of facilities that satisfy $h(i,j) \geq r$ by $\mathcal{J}_{ir} \subseteq \mathcal{J}$. The Dummy facility J is included in $\mathcal{J}_{ir}$. We denote the set of backup levels by $\mathcal{R} = \{1, \ldots, R\}$. For a given facility j and customer i, the set of backup levels that satisfy $r \leq h(i,j)$ is given by $\mathcal{R}_{ij} \subseteq \mathcal{R}$. We define binary variable Y_{ijr} to be one if customer i chooses facility j at backup level r, and zero otherwise. For customer i, binary variable Y_{ijr} is defined only for (j,r) pairs such that $j \in \mathcal{J}_{ir}$ and $r \in \mathcal{R}_{ij}$.

Finally, let variable P_{ir} denote the probability that customer i seeks service at backup level r.

The reliable fixed-charge facility location problem with order (RUFLO) is formulated as:

$$\text{(RUFLO)}\ \Psi^* = \min \sum_{j\in\mathcal{J}} f_j X_j + \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\sum_{r\in\mathcal{R}_{ij}} \eta_i d_{ij}(1-q_j)P_{ir}Y_{ijr} \tag{2.8a}$$

$$\text{s.t.} \sum_{j\in\mathcal{J}_{ir}} Y_{ijr} \le 1 \qquad i\in\mathcal{I}, r\in\mathcal{R}, \tag{2.8b}$$

$$Y_{ijr} \le \sum_{k\in\mathcal{J}_{i,r-1}, j<_i k} Y_{i,k,r-1} \qquad i\in\mathcal{I}, r\in\mathcal{R}\setminus\{1\}, j\in\mathcal{J}_{ir}, \tag{2.8c}$$

$$\sum_{r\in\mathcal{R}_{ij}} Y_{ijr} \le X_j \qquad i\in\mathcal{I}, j\in\mathcal{J}, \tag{2.8d}$$

$$\sum_{r\in\mathcal{R}_{iJ}} Y_{iJr} = 1 \qquad i\in\mathcal{I}, \tag{2.8e}$$

$$1 - X_m + \sum_{s<r} Y_{ims} \ge \sum_{j<_i m} Y_{ijr} \qquad i\in\mathcal{I}, r\in\mathcal{R}, m\in\mathcal{J}_{ir}, \tag{2.8f}$$

$$P_{i1} = 1 \qquad i\in\mathcal{I}, \tag{2.8g}$$

$$P_{ir} \ge P_{i,r-1} \sum_{k\in\mathcal{J}_{i,r-1}} q_k Y_{i,k,r-1} \qquad i\in\mathcal{I}, r\in\mathcal{R}\setminus\{1\}, \tag{2.8h}$$

$$X_j, Y_{ijr} \in \{0,1\} \qquad i\in\mathcal{I}, j\in\mathcal{J}, r\in\mathcal{R}_{ij}. \tag{2.8i}$$

The objective function (2.8a) minimizes the fixed cost of opening facilities plus the expected service cost. Constraints (2.8b) ensure that customers do not use more than one facility at each backup level. Constraints (2.8c) state that if customer i chooses facility j at backup level r then she should have chosen a more preferred facility at level $r-1$. Constraints (2.8d) prohibit using a closed facility and ensures that an open facility is used at most in one backup level for each customer. Constraints (2.8e) guarantee that each customer chooses the dummy facility at some backup level.

Constraints (2.8f) assert that if customer i chooses facility j at backup level r, she should have chosen the more preferred facility m at an earlier level $s<r$ given that m is open. If facility m is closed (i.e., $X_m = 0$), constraints (2.8f) are not active. On the one hand, constraints (2.8f) do not guarantee that if customer i uses a facility at level r, then she should be using other facilities at all levels before r. On the other hand, constraints (2.8c) do not ensure that if a facility preferred by customer i is open, then it must be used before

any other less preferred facility is used. Thus, we need constraints (2.8c) and (2.8f) both.

Constraints (2.8g) and (2.8h) model the recursive probability equations. We set $P_{i1} = 1$ because customers always seek service at the first backup level. For $r > 1$, P_{ir} equals the probability that customer seeks service at backup level $r-1$, and the facility at that backup level fails. Finally, constraints (2.8i) enforce binary restrictions.

Note that we could have defined $\hat{d}_{ij} = \eta_i d_{ij}$ and used in the objective function. However, the inclusion of capacity constraints (see Section 2.9) requires the knowledge of the demand parameter separately. Also, the datasets used in testing report d_{ij} and η_i separately. Therefore, the formulation will keep these parameters separated.

We now discuss the modifications of this model over the model of Section 2.2. Consider customer $i \in \mathcal{I}$ that has facility $j \in \mathcal{J}$ in its preference list, and j is open. Facility j is currently the most preferred available facility to customer i. If the service cost d_{ij} is higher than the penalty cost ϕ_i, the model in Section 2.2 states that customer is assigned to the Dummy facility J at that particular level. However, the modified model of this section would respect the preferences of customers regardless of the service cost. In other words, the central authority has no power in amending the preference list of customers. This is a major change in the models, and can have significant impact on the resulting location and allocation decisions. The other main difference between the preliminary and modified models is the number of variables. Note that this formulation is almost similar in function, but different in the number of variables compared to the model in Section 2.2. Specifically, because P_{ijr} is replaced by P_{ir}, there are $|\mathcal{I}| \times (|\mathcal{J}|-1) \times |\mathcal{R}|$ less variables. This significant difference will prove to be useful in faster convergence, as is shown in Section 2.7.

The nonlinear term $P_{ir}Y_{ijr}$ in (2.8a) and (2.8h) is a product of a continuous variable and a binary variable. Thus, we can replace $P_{ir}Y_{ijr}$ with an auxiliary variable W_{ijr}, and enforce $W_{ijr} = P_{ir}Y_{ijr}$ by $W_{ijr} \leq P_{ir}$, $W_{ijr} \leq Y_{ijr}$, $W_{ijr} \geq P_{ir} + Y_{ijr} - 1$, $W_{ijr} \geq 0$.

2.6 Solution Techniques: Modified Model

This section aims at developing solution techniques to the modified model presented in Section 2.5. The main contribution of this section are: (1) proposing a constraint which

significantly tightens the LP relaxation of the formulation, and (2) developing a Lagrangian branch-and-bound approach and a branch-and-cut approach based on a relaxed formulation. As a result of the modifications of the model itself which are presented in Section 2.5, and the solution techniques shown here, these contributions result in a significant improvement in the performance of the algorithm. More results and analysis are discussed in Section 2.7.

2.6.1 Tighter LP

Our computational experiments reveal that formulation (2.8) has weak linear programming (LP) relaxation, and therefore off-the-shelf solvers such as CPLEX may exhibit poor performance for large problem instances. We propose a set of constraints to tighten the LP relaxation of RUFLO based on the following observation that is valid for every integer feasible solution. Customer $i \in \mathcal{I}$ either selects a facility $j \in \mathcal{J}_{ir} \setminus \{J\}$ at backup level $r \in \mathcal{R}$, or she selects the Dummy facility J at an earlier backup level $s < r$. This relation can be enforced by the constraint set offered in Remark 1.

Remark 1 *There exist an optimal solution to RUFLO in which*

$$\sum_{j \in \mathcal{J}_{ir}} Y_{ijr} + \sum_{s<r} Y_{iJs} = 1 \quad i \in \mathcal{I}, r \in \mathcal{R}. \tag{2.9}$$

Constraints (2.8b) are dominated by (2.9). We show that constraints (2.8c) are also implied and not necessary after adding (2.9). In particular, constraints (2.8c) enforce that $Y_{ijr} \leq u$ when $\sum_{k \in \mathcal{J}_{i,r-1}, j <_i k} Y_{i,k,r-1} = u \in [0,1]$. There can be three different cases when $\sum_{k \in \mathcal{J}_{i,r-1}, j <_i k} Y_{i,k,r-1} = u$.

Case 1: $Y_{i,j,r-1} = 1 - u$. In this case, $Y_{ijr} \leq u$ from (2.8d).

Case 2: $\sum_{k \in \mathcal{J}_{i,r-1}, k <_{ij}} Y_{i,k,r-1} = 1 - u$. In this case, $1 - X_j + \sum_{s<r-1} Y_{ijs} \geq 1 - u$ from (2.8f). Therefore, $\sum_{s \geq r-1} Y_{ijs} \leq u$, and so $Y_{ijr} \leq u$ from (2.8d).

Case 3: $\sum_{s<r-1} Y_{iJs} = 1 - u$. In this case, $\sum_{s<r} Y_{iJs} \geq 1 - u$, and thus $Y_{ijr} \leq u$ from (2.9).

As a result, $Y_{ijr} \leq u$ in all three cases when $\sum_{k \in \mathcal{J}_{i,r-1}, j<_i k} Y_{i,k,r-1} = u$. The Reformulation of RUFLO Using the Proposed Valid Inequality (RUFLO-R) is given by:

$$\text{(RUFLO-R)} \quad \min \sum_{j \in \mathcal{J}} f_j X_j + \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \sum_{r \in \mathcal{R}_{ij}} \eta_i d_{ij} (1 - q_j) W_{ijr} \tag{2.10a}$$

$$\text{s.t.} \sum_{j \in \mathcal{J}_{ir}} Y_{ijr} + \sum_{s<r} Y_{iJs} = 1 \qquad i \in \mathcal{I}, r \in \mathcal{R}, \tag{2.10b}$$

$$(2.8d) - (2.8f)$$

$$P_{i1} = 1 \qquad i \in \mathcal{I}, \tag{2.10c}$$

$$P_{ir} \geq \sum_{k \in \mathcal{J}_{i,r-1}} q_k W_{i,k,r-1} \qquad i \in \mathcal{I}, r \in \mathcal{R} \setminus \{1\}, \tag{2.10d}$$

$$W_{ijr} \leq P_{ir},\ W_{ijr} \leq Y_{ijr},\ W_{ijr} \geq P_{ir} + Y_{ijr} - 1, \quad i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}_{ij}, \tag{2.10e}$$

$$X_j, Y_{ijr} \in \{0,1\}, W_{ijr} \geq 0, \qquad i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}_{ij}. \tag{2.10f}$$

Corollary 1 *The LP relaxation of RUFLO-R is tighter than the LP relaxation of RUFLO.*

Corollary 1 follows since constraints (2.8b) are dominated by (2.9). Next, we present two solution algorithms: a Lagrangian branch-and-bound algorithm and a branch-and-cut algorithm. We also propose a neighborhood search method to generate upper bounds.

2.6.2 Lagrangian Branch-and-Bound Algorithm

Let $\mathcal{L}$ be a partition of customers such that $\bigcup_{\ell \in \mathcal{L}} \mathcal{I}_\ell = \mathcal{I}$. We define binary variable $Z_{\ell j}$ to be a copy of the X_j variable for each customer subset $\ell \in \mathcal{L}$. In any feasible solution to RUFLO-R, the value of $Z_{\ell j}$ must be the same in all customer groups for each facility site j. We later relax this requirement to decompose the RUFLO-R into customer group subproblems.

Let $\mathbf{Z}_\ell := \{Z_{\ell j},\ j \in \mathcal{J}\}$, $\mathbf{W}_\ell := \{W_{ijr},\ i \in \mathcal{I}_\ell, j \in \mathcal{J}, r \in \mathcal{R}_{ij}\}$,
$\mathbf{Y}_\ell := \{Y_{ijr},\ i \in \mathcal{I}_\ell, j \in \mathcal{J}, r \in \mathcal{R}_{ij}\}$, $\mathbf{P}_\ell := \{P_{ir},\ i \in \mathcal{I}_\ell, r \in \mathcal{R}\}$, and define the solution set

$S_\ell := \{(\mathbf{Z}_\ell, \mathbf{Y}_\ell, \mathbf{W}_\ell, \mathbf{P}_\ell) : (2.11a) - (2.11h)\}$, where

$$\sum_{j \in \mathcal{J}_{ir}} Y_{ijr} + \sum_{s<r} Y_{iJs} = 1 \qquad i \in \mathcal{I}_\ell, r \in \mathcal{R}, \tag{2.11a}$$

$$\sum_{r \in \mathcal{R}_{ij}} Y_{ijr} \leq Z_{\ell j} \qquad i \in \mathcal{I}_\ell, j \in \mathcal{J}, \tag{2.11b}$$

$$\sum_{r \in \mathcal{R}_{iJ}} Y_{iJr} = 1 \qquad i \in \mathcal{I}_\ell, \tag{2.11c}$$

$$1 - Z_{\ell m} + \sum_{s<r} Y_{ims} \geq \sum_{j <_i m} Y_{ijr} \qquad i \in \mathcal{I}_\ell, r \in \mathcal{R}, m \in \mathcal{J}_{ir}, \tag{2.11d}$$

$$P_{i1} = 1 \qquad i \in \mathcal{I}_\ell, \tag{2.11e}$$

$$P_{ir} \geq \sum_{k \in \mathcal{J}_{i,r-1}} q_k W_{i,k,r-1} \qquad i \in \mathcal{I}_\ell, r \in \mathcal{R} \setminus \{1\}, \tag{2.11f}$$

$$W_{ijr} \leq P_{ir},\ W_{ijr} \leq Y_{ijr},\ W_{ijr} \geq P_{ir} + Y_{ijr} - 1 \qquad i \in \mathcal{I}_\ell, j \in \mathcal{J}, r \in \mathcal{R}_{ij}, \tag{2.11g}$$

$$Z_{ij}, Y_{ijr} \in \{0,1\}, W_{ijr} \geq 0 \qquad i \in \mathcal{I}_\ell, j \in \mathcal{J}, r \in \mathcal{R}_{ij}. \tag{2.11h}$$

The split-variable reformulation of RUFLO-R is then given by:

$$\Psi^* = \min \sum_{\ell \in \mathcal{L}} \left(\sum_{j \in \mathcal{J}} \frac{1}{|\mathcal{L}|} f_j Z_{\ell j} + \sum_{i \in \mathcal{I}_\ell} \sum_{j \in \mathcal{J}} \sum_{r \in \mathcal{R}_{ij}} \eta_i d_{ij} (1 - q_j) W_{ijr} \right) \tag{2.12a}$$

$$\text{s.t. } (\mathbf{Z}_\ell, \mathbf{Y}_\ell, \mathbf{W}_\ell, \mathbf{P}_\ell) \in S_\ell \qquad \ell \in \mathcal{L}, \tag{2.12b}$$

$$X_j - Z_{\ell j} = 0 \qquad \ell \in \mathcal{L}, j \in \mathcal{J}. \tag{2.12c}$$

Note that constraints (2.12c) ensure that the objective function (2.12a) is exactly equal to the original objective function (2.8a). We relax the split-variable formulation by replacing constraints (2.12b) with

$$(\mathbf{Z}_\ell, \mathbf{Y}_\ell, \mathbf{W}_\ell, \mathbf{P}_\ell) \in \text{conv}(S_\ell),\ \ell \in \mathcal{L}, \tag{2.13}$$

where $\text{conv}(S_\ell)$ denotes the convex hull of S_ℓ. Let $\mathcal{S}_\ell$ be the index set of solutions in S_ℓ, that is, $S_\ell = \{(\hat{\mathbf{Z}}_\ell^s, \hat{\mathbf{Y}}_\ell^s, \hat{\mathbf{W}}_\ell^s, \hat{\mathbf{P}}_\ell^s) : s \in \mathcal{S}_\ell\}$. Then, the convex hull of S_ℓ can be expressed as:

$$\text{conv}(S_\ell) = \left\{ \sum_{s \in \mathcal{S}_\ell} \lambda_\ell^s (\hat{\mathbf{Z}}_\ell^s, \hat{\mathbf{Y}}_\ell^s, \hat{\mathbf{W}}_\ell^s, \hat{\mathbf{P}}_\ell^s),\ \sum_{s \in \mathcal{S}_\ell} \lambda_\ell^s = 1,\ \lambda_\ell^s \geq 0 \right\}. \tag{2.14}$$

Based on (2.14), we obtain the relaxation of the split-variable formulation as:

$$\min \sum_{\ell\in\mathcal{L}}\sum_{s\in\mathcal{S}_\ell}\left(\sum_{j\in\mathcal{J}}\frac{1}{|\mathcal{L}|}f_j\hat{Z}^s_{\ell j}+\sum_{i\in\mathcal{I}_\ell}\sum_{j\in\mathcal{J}}\sum_{r\in\mathcal{R}_{ij}}\eta_i d_{ij}(1-q_j)\hat{W}^s_{ijr}\right)\lambda^s_\ell \tag{2.15a}$$

$$\text{s.t. } X_j-\sum_{s\in\mathcal{S}_\ell}\hat{Z}^s_{\ell j}\lambda^s_\ell=0 \qquad j\in\mathcal{J},\ell\in\mathcal{L},\ (\mu_{\ell j}) \tag{2.15b}$$

$$\sum_{s\in\mathcal{S}_\ell}\lambda^s_\ell=1 \qquad \ell\in\mathcal{L},\ (\theta_\ell) \tag{2.15c}$$

$$\lambda^s_\ell\geq 0 \qquad \ell\in\mathcal{L},s\in\mathcal{S}_\ell. \tag{2.15d}$$

The number of λ^s_ℓ variables in formulation (2.15) equals $\sum_{\ell\in\mathcal{L}}|\mathcal{S}_\ell|$, which can be enormous even for moderate-size instances. Therefore, we use a subgradient-based cutting plane method to solve the dual of problem (2.15) that is given by:

$$\max \sum_{\ell\in\mathcal{L}}\theta_\ell \tag{2.16a}$$

$$\text{s.t. } \sum_{\ell\in\mathcal{L}}\mu_{\ell j}=0, \qquad j\in\mathcal{J}, \tag{2.16b}$$

$$\theta_\ell-\sum_{j\in\mathcal{J}}\hat{Z}^s_{\ell j}\mu_{\ell j}\ \leq\sum_{j\in\mathcal{J}}\frac{1}{|\mathcal{L}|}f_j\hat{Z}^s_{\ell j}+\sum_{i\in\mathcal{I}_\ell}\sum_{j\in\mathcal{J}}\sum_{r\in\mathcal{R}_{ij}}\eta_i d_{ij}(1-q_j)\hat{W}^s_{ijr}, \quad \ell\in\mathcal{L},s\in\mathcal{S}_\ell. \tag{2.16c}$$

Let $F_\ell(\boldsymbol{\mu}_\ell)=\sum_{j\in\mathcal{J}}\left(\frac{1}{|\mathcal{L}|}f_j+\mu_{\ell j}\right)Z_{\ell j}+\sum_{i\in\mathcal{I}_\ell}\sum_{j\in\mathcal{J}}\sum_{r\in\mathcal{R}_{ij}}\eta_i d_{ij}(1-q_j)W_{ijr}$ and define the *customer group subproblem* $\ell\in\mathcal{L}$ as:

$$D_\ell(\boldsymbol{\mu}_\ell)=\min\{F_\ell(\boldsymbol{\mu}_\ell):(\mathbf{Z}_\ell,\mathbf{Y}_\ell,\mathbf{W}_\ell,\mathbf{P}_\ell)\in S_\ell\}, \quad \ell\in\mathcal{L}. \tag{2.17}$$

Then, formulation (2.16), also known as the Lagrangian dual problem, can be written as:

$$\Psi^*_{LD}=\max \sum_{\ell\in\mathcal{L}}\theta_\ell \tag{2.18a}$$

$$\text{s.t. } \sum_{\ell\in\mathcal{L}}\mu_{\ell j}=0 \qquad j\in\mathcal{J}, \tag{2.18b}$$

$$\theta_\ell\leq D_\ell(\boldsymbol{\mu}_\ell) \qquad \ell\in\mathcal{L}. \tag{2.18c}$$

Note that $D_\ell(\boldsymbol{\mu}_\ell)$ is concave in $\boldsymbol{\mu}_\ell$, and its subgradient at $\boldsymbol{\mu}_\ell^k$ is $\mathbf{Z}_\ell^k$, where $(\mathbf{Z}_\ell^k, \mathbf{W}_\ell^k, \mathbf{Y}_\ell^k, \mathbf{P}_\ell^k)$ is an optimal solution to (2.17). It follows from the subgradient inequality that

$$\theta_\ell \leq D_\ell(\boldsymbol{\mu}_\ell) \leq D_\ell(\boldsymbol{\mu}_\ell^k) + \sum_{j \in \mathcal{J}} Z_{\ell j}^k (\mu_{\ell j} - \mu_{\ell j}^k), \quad \ell \in \mathcal{L}. \tag{2.19}$$

The Lagrangian dual problem (2.18) can be solved optimally by a cutting plane method that enforces constraints (2.18c) with subgradient inequalities (2.19) as follows:

$$\max \sum_{\ell \in \mathcal{L}} \theta_\ell \tag{2.20a}$$

$$\text{s.t.} \sum_{\ell \in \mathcal{I}} \mu_{\ell j} = 0 \qquad j \in \mathcal{J}, \tag{2.20b}$$

$$\theta_\ell \leq D_\ell(\boldsymbol{\mu}_\ell^k) + \sum_{j \in \mathcal{J}} Z_{\ell j}^k (\mu_{\ell j} - \mu_{\ell j}^k) \qquad \ell \in \mathcal{L}, k \in \mathcal{K}_\ell, \tag{2.20c}$$

where $\mathcal{K}_\ell$ is the set of subgradients for $\ell \in \mathcal{L}$. The customer group subproblem (2.17) must be solved to generate the subgradient inequality (2.19) for each $\ell \in \mathcal{L}$. This can be computationally expensive if the size of (2.17) is large. We can alleviate this difficulty by not optimally solving subproblem (2.17). In particular, let $U_\ell(\boldsymbol{\mu}_\ell)$ be the objective function of a feasible solution $(\tilde{\mathbf{Z}}_\ell^k, \tilde{\mathbf{W}}_\ell^k, \tilde{\mathbf{Y}}_\ell^k, \tilde{\mathbf{P}}_\ell^k)$ to the customer group subproblem $\ell \in \mathcal{L}$. Then, the subgradient inequality(2.19) can be formulated as:

$$\theta_\ell \leq U_\ell(\boldsymbol{\mu}_\ell) \leq U_\ell(\boldsymbol{\mu}_\ell^k) + \sum_{j \in \mathcal{J}} \tilde{Z}_{\ell j}^k (\mu_{\ell j} - \mu_{\ell j}^k), \quad \ell \in \mathcal{L}. \tag{2.21}$$

The feasible solution $(\tilde{\mathbf{Z}}_\ell^k, \tilde{\mathbf{W}}_\ell^k, \tilde{\mathbf{Y}}_\ell^k, \tilde{\mathbf{P}}_\ell^k)$ can be obtained using a heuristic, or it can be set as the incumbent solution of a branch-and-bound algorithm after a certain running time.

The convergence of the proposed Lagrangian decomposition is usually slow. Especially during the initial iterations, the algorithm can move from one $\boldsymbol{\mu}^k$ to another one without making significant progress. Therefore, we utilize a proximal bundle method that optimizes the objective function (2.22) obtained by subtracting a weighted penalty term from the objective function (2.20a).

$$\max_{\boldsymbol{\theta}, \boldsymbol{\mu}} \sum_{\ell \in \mathcal{L}} \theta_\ell - \frac{1}{2} \tau \sum_{\ell \in \mathcal{L}} \sum_{j \in \mathcal{J}} (\mu_{\ell j} - \mu_{\ell j}^+)^2, \tag{2.22}$$

Table 2.27: Proximal Bundle Method to Solve the Lagrangian Dual problem (2.18)

Initialization	Set $\epsilon \leftarrow 10^{-5}$, $k \leftarrow 1$, $\tau \leftarrow 10^{-5}$, $\boldsymbol{\mu}_\ell^+ \leftarrow 0$, $\ell \in \mathcal{L}$. Solve problem (2.17) with $\boldsymbol{\mu}_\ell^k = \boldsymbol{\mu}_\ell^+$ for each $\ell \in \mathcal{L}$, $curObj \leftarrow \sum_\ell D_\ell(\boldsymbol{\mu}_\ell^+)$
Step 1	Solve problem (2.20) with objective function (2.22) to obtain $\boldsymbol{\mu}^k$, and let $v = \sum_\ell \theta_\ell^k - curObj$. If $v/(1 + \lvert curObj \rvert) < \epsilon$, **terminate**. Else $k \leftarrow k + 1$.
Step 2	Solve problem (2.17) with $\boldsymbol{\mu}_\ell^k$, for all $\ell \in \mathcal{L}$, $newObj \leftarrow \sum_\ell D_\ell(\boldsymbol{\mu}_\ell^k)$.
Step 3	Update $\tau \leftarrow \min(\max(u, \tau/10, 10^{-7}), 10\tau)$, where $u = 1.5\tau(1 - (newObj - curObj)/v)$.
Step 4	If $newObj - curObj \geq 10^{-5}v$, update $\boldsymbol{\mu}^+ \leftarrow \boldsymbol{\mu}^k$, $curObj \leftarrow newObj$. Goto Step 1.

where $\boldsymbol{\mu}^+$ is the current proximity center and $\tau \geq 0$ is the weight of the quadratic penalty term. We slightly modify the updating rules proposed by Lubin et al. (2013) based on preliminary computations. Table 2.27 summarizes the steps of our implementation.

Due to the nonconvexities caused by binary variables $\mathbf{Z}_\ell$ and $\mathbf{Y}_\ell$ in constraints (2.12b), Ψ_{LD}^* will provide a lower bound on Ψ^*. We use the branch-and-bound algorithm presented in Table 2.28 to reduce the gap between Ψ_{LD}^* and Ψ^*. In Step 2 of this algorithm, problem P, the parent of which has the lowest Lagrangian bound among all processed problems, is solved. We fathom P if it is infeasible or if its lower bound is greater than the incumbent objective value. In Step 3, a heuristic neighborhood search is performed to improve the upper bound. In Step 4, branching is performed on the copy variables $\mathbf{Z}_\ell$, $\ell \in \mathcal{L}$. We select facility j with the highest expected service cost that is not agreed by all customer group problems for branching. This variable selection rule is used to make a significant impact on the customer group subproblem solutions.

Table 2.28: Lagrangian Branch-and-Bound Algorithm.

Step 0(*Initialization*)	- Set the upper bound $\bar{\Psi} = \infty$. - Let the set of unsolved problems $\mathcal{P}$ include problem (2.10).
Step 1(*Stopping*)	- If $\mathcal{P} = \emptyset$, the solution corresponding to $\bar{\Psi}$ is optimal.
Step 2 (*Processing*)	- Select and delete a problem P from $\mathcal{P}$. - Solve problem (2.18) to get the Lagrangian bound $\Psi^*_{LD}(P)$. - If P is infeasible, or if $\Psi^*_{LD}(P) \geq \bar{\Psi}$, go to Step 1.
Step 3 (*Heuristic Solution*)	- For each customer group $\ell \in \mathcal{L}$, let $\Psi^N_\ell(P)$ be the objective value of the best solution in the *neighborhood* of $\mathbf{Z}^P_\ell$, update $\bar{\Psi} := \min\left\{\bar{\Psi}, \Psi^N_\ell(P)\right\}$ (see Eq. (2.24) for the neighborhood definition). - Delete from $\mathcal{P}$ all problems P' with $\Psi_{LD}(P') \geq \bar{\Psi}$.
Step 4 (*Branching*)	- If customer group subproblem solutions $\mathbf{Z}^P_\ell$ are the same for all $\ell \in \mathcal{L}$, that is, $\mathbf{Z}^P_\ell = \mathbf{Z}^P \ \forall \ell \in \mathcal{L}$, then go to Step 1. - Select a facility j such that $\mathbf{Z}^P_{\ell j}$ is not identical for all $\ell \in \mathcal{L}$. - Add two new problems P_1 and P_2 to $\mathcal{P}$ obtained from P by adding the constraints $Z_{\ell j} = 0$ and $Z_{\ell j} = 1 \ \forall \ell \in \mathcal{L}$. Go to Step 1.

2.6.3 Branch-and-Cut Algorithm

We propose a primal relaxation of the RUFLO-R by replacing $(1-q_j)P_{ir}$ term in the objective function (2.8a) with fixed but *optimistic* failure probability estimates Q_{ijr}. Aboolian et al. (2013) proposed a similar relaxation idea for the reliable facility location problem. In this model, unlike Aboolian et al. (2013), we incorporate customer preferences. Moreover, we tighten the failure probability estimates progressively when solving the relaxed problem. Consider assigning customer $i \in \mathcal{I}$ to facility $j \in \mathcal{J}_{ir}$ at level $r \in \mathcal{R}$. Let $q_{[1]} \leq q_{[2]} \leq \ldots \leq q_{[r-1]}$ be an ordering of failure probabilities of the $r-1$ most reliable sites in $\{k \in \mathcal{J}_{ir} \mid j <_i k\}$. We initialize $Q_{ijr} = (1-q_j)\prod_{t=1}^{r-1} q_{[t]}$.

Lemma 1 *If $Y_{ijr} = 1$, then $Q_{ijr} \leq (1 - q_j)P_{ir}$ for all $i \in \mathcal{I}, r \in \mathcal{R}, j \in \mathcal{J}_{ir}$ in any feasible solution to the RUFLO-R.*

Proof: From constraints (2.8g) and (2.8h), if $Y_{ijr} = 1$, then $(1 - q_j)P_{ir}$ is equal to the probability that facilities serving customer i at backup levels $1, \ldots, r-1$ all fail independently, and facility j does not fail. Only those facilities that customer i prefers to j, that is, from set $\{k \in \mathcal{J}_{ir} \mid j <_i k\}$, can serve customer i at levels $1, \ldots, r-1$. By definition, Q_{ijr} assumes that the most reliable $r-1$ facilities in $\{k \in \mathcal{J}_{ir} \mid j <_i k\}$ serves customer i at levels $1, \ldots, r-1$. □

Replacing $(1 - q_j)P_{ir}$ with fixed failure probabilities Q_{ijr} in the formulation of the RUFLO and using an auxiliary variable Z that represents the total service cost we obtain the following mixed-integer program:

$$\Psi' = \min \sum_{j \in \mathcal{J}} f_j X_j + Z \tag{2.23a}$$

$$\text{s.t. } (2.8d) - (2.8i),\ (2.10c)$$

$$Z \geq \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \sum_{r \in \mathcal{R}_{ij}} \eta_i d_{ij} Q_{ijr} Y_{ijr}, \tag{2.23b}$$

$$X_j, Y_{ijr} \in \{0, 1\} \qquad i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}_{ij}.$$

It follows from Lemma (1) that $\Psi' \leq \Psi^*$. Note that constraint (2.23b) and auxiliary variable Z are not necessary to formulate the relaxed problem, however they will be useful in our implementation.

Any location vector $\hat{\mathbf{X}} \in \{0, 1\}^{|\mathcal{J}|}$ obtained by solving the relaxed problem (2.23) generates a feasible solution to RUFLO-R. This is simply achieved by assigning customers to open facilities in $\hat{\mathbf{X}}$ with respect to their preferences. The value of this feasible solution provides an upper bound to RUFLO-R. To improve this upper bound, we perform a neighborhood search. In particular, distance-h neighborhood of a given facility location vector $\hat{\mathbf{X}}$ is defined as:

$$N_h(\hat{\mathbf{X}}) = \{\mathbf{X}' \in \{0, 1\}^{|\mathcal{J}|} : | \sum_{j \in \mathcal{J}} |\hat{X}_j - X'_j| \leq h\}. \tag{2.24}$$

The neighborhood search procedure calculates the objective value of all facility location vectors in $N_h(\hat{\mathbf{X}})$ and returns the best one. If the objective value corresponding to a location vector in $N_h(\hat{\mathbf{X}})$ is better than the objective value associated with $\hat{\mathbf{X}}$, the search restarts from this new location vector. This procedure is repeated until no further improvement is achieved. All location vectors in $N_h(\hat{\mathbf{X}})$ can be removed from the feasible region of problem (2.23) since their objective values are already examined. Let $S_{\hat{\mathbf{X}}}$ denote the set of open facility locations in $\hat{\mathbf{X}}$, that is, $S_{\hat{\mathbf{X}}} = \{j \in \mathcal{J} : \hat{X}_j = 1\}$. We use the following supervalid inequality to remove all location vectors in $N_h(\hat{\mathbf{X}})$ from the feasible region of problem (2.23):

$$\sum_{j \in S_{\hat{\mathbf{X}}}} X_j - \sum_{j \in \mathcal{J} \setminus S_{\hat{\mathbf{X}}}} X_j \leq |S_{\hat{\mathbf{X}}}| - h - 1. \tag{2.25}$$

Note that an improved lower bound and another location vector can be obtained by resolving the relaxed problem (2.23) after adding the cut (2.25). This overall process of obtaining upper bounds from the neighborhood search and generating lower bounds by solving the relaxed problem (2.23) with additional cuts of type (2.25) can be repeated a given number of times or until the gap between the lower and upper bounds is sufficiently small. As also noted by Aboolian et al. (2013), this search-and-cut procedure is very similar to a branch-and-bound algorithm because it must exhaust all of the possible solutions to find a global optimal solution.

When solving the relaxed problem (2.23), we use the `lazyconstraint` callback function of CPLEX in which the neighborhood search is performed for each integer solution found in the branch-and-bound tree. We then add the cutting plane (2.25) globally and update the upper bound as necessary. Furthermore, we tighten the failure probability estimates Q_{ijr} locally based on the local upper bound $U_j \in \{0, 1\}$ of each X_j variable at the incumbent branch-and-bound tree node. In particular, only those facilities that customer i prefers to j whose upper bound is one, that is, from set $\{k \in \mathcal{J}_{ir} \mid j <_i k \text{ and } U_k = 1\}$, can serve customer i at levels $1, \ldots, r-1$. Let $q'_{[1]} \leq q'_{[2]} \leq \ldots \leq q'_{[r-1]}$ be an ordering of failure probabilities of the first $r-1$ most reliable sites in $\{k \in \mathcal{J}_{ir} \mid j <_i k \text{ and } U_k = 1\}$. We set the updated failure probability estimate $Q'_{ijr} = (1 - q_j) \prod_{t=1}^{r-1} q'_{[t]}$. Note that if there is less than $r-1$ facilities in $\{k \in \mathcal{J}_{ir} \mid j <_i k \text{ and } U_k = 1\}$, then $Q'_{ijr} = 0$. After updating the

failure probability estimates, we add the following local cut to the node subproblem:

$$Z \geq \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \sum_{r \in \mathcal{R}_{ij}} \eta_i d_{ij} Q'_{ijr} Y_{ijr} \tag{2.26}$$

This implementation turned out to be consistently faster than re-solving the relaxed problem each time a cutting plane is added to the model.

2.7 Computational Results and Analyses: Modified Model

We test the computational performance of the proposed solution methods. Similar to Cánovas et al. (2007), the datasets used in our experiments are partly taken from Beasley (1990). Specifically, the facility opening cost f_j, service cost d_{ij} and demand η_i values are taken from the data file *capa* for the uncapacitated warehouse location problem in Beasley (1990). This data file has 1,000 customers and 100 facilities. The number of facilities is less than 100 in all of our test instances, so we sample facilities from the *capa* data file without replacement. To generate an instance with less than 1,000 customers, we sample customers from the *capa* data file without replacement. To generate an instance with more than 1,000 customers, we first include all customers in the *capa* file, and generate additional customers by taking the average of two randomly sampled customers.

If there are n customers, then for each facility $j \in \mathcal{J}$, we multiply the opening cost f_j by $n \times 10^{-3}$ to adjust for the fact that facilities can serve up to 1,000 customers in the *capa* data file. The failure probability of each facility except the dummy facility is calculated using the formula $q_j = 0.01 + 1.5e^{-\max\{1, d_{1j}/6000\}}$. According to this formula, failure probability of facility j decreases as the cost of serving customer 1 at j (or the distance between customer 1 and facility j) increases. Aboolian et al. (2013) used a similar formula to generate facility failure probabilities. Recall that $q_J = 0$ for the dummy facility. We vary the number of backup levels $|\mathcal{R}| \in \{4, 5, 6\}$, and set the disutility cost $\phi_i = 5 \times 10^5$ for all customers. Finally, we generate the preferences of the customers randomly using a method proposed by Cánovas et al. (2007) (see Appendix A).

The preliminary experiments showed that the Lagrangian branch-and-bound algorithm (LB&B) and the branch-and-cut algorithm (B&C) may not be competitive for relatively small instances which can be solved by CPLEX within 3 hours using default settings. The results in Section 2.4 show that the previous implementations of the algorithm can handle smaller instances. Therefore, the problem sizes in Table 2.29 are chosen such that most instances cannot be solved by CPLEX within 3 hours.

The LP Ratio column in Table 2.29 shows the ratio of the LP relaxation of the RUFLO-R to the LP relaxation of the RUFLO. The optimality gaps of the LB&B, B&C and CPLEX are reported after 3-hour run time. For the LB&B algorithm, we partition the set of customers into 12 groups, that is, $|\mathcal{L}| = 12$, and fixed the distance parameter $h = 3$ in the neighborhood search algorithm. We solve the group subproblems in parallel with 12 cores using CPLEX 12.7 in single thread mode on each core. Maximum 100 iterations are allowed in the proximal bundle method after which we branch as described in Section 2.6.2.

The neighborhood search distance parameter h has more significant impact on the performance of the B&C algorithm compared to the LB&B algorithm. Therefore, we run the B&C algorithm three times with $h \in \{2, 3, 4\}$ for each instance, and report the smallest optimality gap along with the corresponding h. We run the B&C algorithm on a single core as its implementation uses callback functions of CPLEX, and therefore do not allow for parallelization easily. To ensure fair comparisons, we also run CPLEX in the single thread mode.

As can be seen in Table 2.29, the LP relaxation of the RUFLO-R is at least 2.5 times tighter than the LP relaxation of the RUFLO. Therefore, we consider the RUFLO-R in the rest of our experiments. CPLEX 12.7 returns the smallest optimality gap for six instances with 96 customers and 50 facilities after 3-hour run time. The B&C returns smaller gaps for all other instances. The LB&B returns larger optimality gap than the B&C for all instances in Table 2.29. The LB&B algorithm, however, is still practically valuable, because it can be applied to instances with much larger number of customers than the ones reported in Table 2.29. This is due to the fact that the LB&B algorithm can decompose the set of customers into smaller groups, whereas the B&C does not allow for such decomposition.

We consider instances with extensively larger number of customers in Table 2.30. Each

customer (or customer type) in our model has a preference list. In practice, solving problems with large number of customers might be required to increase granularity when considering diverse preferences of several different customer types. For the LB&B algorithm, we partition the set of customers into 48 groups, that is, $|\mathcal{L}| = 48$, and fix the distance parameter $h = 3$ in the neighborhood search algorithm. We solve the group subproblems in parallel with 48 cores using CPLEX 12.7 on each core. The group subproblems cannot be solved optimally within a reasonable time due to their gigantic size. Therefore, we use the lower bound obtained by CPLEX after one hour in the Lagrangian bound calculations. We perform only one iteration in the proximal bundle method. We also attempt to solve each instance using CPLEX 12.7 without applying any decomposition. We run CPLEX for one hour with its default settings on a single compute node with 32 GB memory and 8 cores, which compose the global limits set by our computing environment.

The Lag/CPLEX column in Table 2.30 reports the ratio of the Lagrangian bound to the lower bound of CPLEX after one hour. As can be seen, the Lagrangian lower bound is at least ten times stronger across all instances. The optimality gap of the LB&B algorithm is significantly smaller compared to CPLEX 12.7. Furthermore, the optimality gap of the LB&B decreases as the number of facilities decreases, although the number of customers increases. In summary, our results suggest that the LB&B algorithm outperforms CPLEX for large-scale instances, while the B&C algorithm should be used for smaller problems. Also, the results show that the gap generated after a fixed time does not grow with larger instances. This is an important feature showing the stability of the algorithms.

Table 2.29: Problem Sizes of Large Test Instances and Optimality Gaps After 3-Hour Run Time

$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{R}\|$	LP Ratio	LB&B (%)	B&C (%)	h	CPLEX (%)
96	50	3	8.5	18.5	11.0	3	10.2
96	50	3	4.0	14.9	9.7	3	(3,026.6 s)†
96	50	3	10.9	15.7	8.0	3	3.9
96	50	4	25.6	24.3	9.3	3	8.8
96	50	4	10.5	12.8	6.4	3	(8,547.9 s)†
96	50	4	36.8	21.3	6.6	3	9.6
96	50	5	84.0	23.9	9.9	2	16.2
96	50	5	21.5	17.0	5.5	3	2.2
96	50	5	117.0	22.8	6.8	3	14.0
192	40	3	5.0	19.3	12.5	2	26.3
192	40	3	4.0	14.3	6.2	3	12.8
192	40	3	5.4	16.9	10.2	2	22.1
192	40	4	17.0	21.2	15.3	2	26.3
192	40	4	13.4	16.2	5.2	2	11.7
192	40	4	20.6	16.6	8.2	2	22.5
192	40	5	57.6	21.5	17.5	4	29.7
192	40	5	55.2	16.4	6.6	3	18.9
192	40	5	67.5	17.0	10.9	3	26.4
240	30	3	3.6	11.1	8.7	4	10.9
240	30	3	2.5	11.8	5.7	4	10.7
240	30	3	3.7	13.4	10.7	2	13.9
240	30	4	12.3	12.7	6.6	4	12.7
240	30	4	5.0	13.3	3.9	4	6.0
240	30	4	12.3	16.1	10.7	4	16.0
240	30	5	43.5	12.8	8.9	4	22.0
240	30	5	7.7	14.3	5.7	4	15.4
240	30	5	43.0	16.7	12.6	4	21.1

† CPLEX found the optimal solution in less than 3 hours.

Table 2.30: Problem Sizes of Extremely Large Test Instances and Optimality Gaps After 1-Hour Run Time

$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{R}\|$	Lag/CPLEX	LB&B (%)	CPLEX (%)
7,200	50	3	21.9	31.6	99.8
7,200	50	3	23.9	31.0	99.8
7,200	50	3	18.7	32.0	99.8
7,200	50	4	21.3	32.4	99.8
7,200	50	4	23.8	30.4	99.8
7,200	50	4	18.5	31.3	99.8
7,200	50	5	20.8	33.9	99.8
7,200	50	5	23.8	30.4	99.8
7,200	50	5	18.5	31.2	99.8
8,382	40	3	14.0	28.2	95.4
8,382	40	3	16.5	27.3	96.1
8,382	40	3	23.5	28.2	97.2
8,382	40	4	17.2	27.7	99.8
8,382	40	4	19.4	26.6	99.8
8,382	40	4	26.3	27.1	99.8
8,382	40	5	17.1	27.9	99.8
8,382	40	5	19.4	26.6	99.8
8,382	40	5	26.1	27.5	99.8
9,600	30	3	12.3	21.5	94.4
9,600	30	3	10.4	21.0	93.3
9,600	30	3	19.3	21.1	99.8
9,600	30	4	14.7	21.6	99.7
9,600	30	4	13.4	20.6	99.6
9,600	30	4	19.1	21.3	99.8
9,600	30	5	14.4	23.1	99.7
9,600	30	5	13.4	20.7	99.7
9,600	30	5	19.1	21.1	99.7

2.8 Applications on Healthcare

The models discussed here can be applied in various settings. Some applications are discussed here with the aim of showing importance and applicability of the model.

2.8.1 Cancer Screening

The proposed methodology can be applied in locating preventative healthcare facilities such as those providing breast and colorectal cancer screening. Mammography and colonoscopy screening reduce cancer risk and improve health-outcomes (Ayer et al., 2012; Erenay et al., 2014). Therefore, several models are proposed for locating mammography and endoscopy centers (Akhundov, 2015; Haase and Müller, 2015; Uzunlar et al., 2012; Verter and Lapierre, 2002; Verter and Zhang, 2015; Vidyarthi and Kuzgunkaya, 2015; Zhang et al., 2009, 2010, 2012c). These applications are important because accessibility of screening services is a key factor for the compliance of individuals at risk to the screening programs (Zhang et al., 2009).

However, the allocation of patients to the preventative healthcare facilities should be based on user choice (Verter and Lapierre, 2002; Zhang et al., 2009). This is because patients consider both ease of access and quality of care when choosing the facility they attend. That is, a significant portion of patients bypass the closest endoscopy facility for having their colonoscopy screening in a better clinic both in rural and urban areas (Charlton et al., 2015). Furthermore, patients may not receive their service form a preferred preventative medicine facility due to stochastic factors such as unfavourable road conditions, scheduling issues, or congestion in the waiting list. In such a case, patients may need to visit the next facility in their preference list. Therefore, it is desirable to consider patient preferences and facility availability when determining the optimal locations of preventative healthcare facilities.

Verter and Lapierre (2002) and Zhang et al. (2012c) assumed that patients attend to the mammography center with shortest travel time. However, the probability of requesting breast cancer screening linearly decrease with the distance to the closest facility. In the models of Zhang et al. (2009) and Zhang et al. (2010), patients choose attending

the mammography center that serve them within the shortest expected service time and participation rate to breast cancer screening linearly decreases as service time increases. Zhang et al. (2012c) also proposed a second model assuming that patients may attend each open mammography center with particular probabilities. These probabilities are modeled as multinomial logit functions and they are proportional to patients' utility of receiving service from a facility which primarily depends on distance.

Particular facility location models allow customers to patronize one of the open facilities based on various preference mechanisms. For instance, some studies used accessibility-based proxy preference measures such as distance, travel time, and service time (Verter and Lapierre, 2002; Zhang et al., 2010), while some other studies assumed stochastic facility choice with probabilities proportional to utility of using the open facilities (Müller et al., 2009; Haase and Müller, 2013). Most models for locating preventative healthcare facilities used one of these preference mechanisms.

Cancer screening tests can improve survival and decrease mortality by detecting cancer at an early stage when treatment is more effective. Moreover, regular use of cervical and colorectal cancer screening tests can prevent the development of cancer through identification and removal or treatment of premalignant abnormalities (American Cancer Society, 2017).

Most CRC cases originate from benign growths on the inner surface of the colon and rectum (called adenomatous polyps), which may progress to CRC (Loeve et al., 2004). This natural progression of the CRC makes it possible for testing procedures to discover these lesions and adenomatous polyps early on. CRC screening is one of the preventive healthcare operations that are recommended at certain time window, but not urgent. Hence, patients are usually given option to choose the facility at which a test is conducted.

This system also can be modeled using assumptions discussed above. Patients have well known preferences of clinics and health centers based on quality, distance, and familiarity with physicians and staff. They would consider the most preferred available option, but they also might decide to go to the next most preferred available option, depending on other factors.

2.8.2 Senior Centers

Senior centers have become one of the most widely used services among America's older adults; one million senior citizens are served every day through 11,000 senior centers (National Council on Aging, 2015). Senior centers provide services like nutrition programs, health and wellness programs, employment assistance, and social and recreational activities, among others. The National Council on Aging reports that 75% participants visit their center 1 to 3 times a week (National Council on Aging, 2015). Also, it has been reported that participants do not necessarily go to the nearest center to them (CMU Center for Economic Development, 2007). Instead, they might choose the one with better services or more friends.

This system can be solved by our model. We assume seniors have preferences, based on quality, distance, and other factors. Since visiting a senior center is voluntary and not urgent, seniors with known preference might decide not to go to their most preferred center, even if it is open. Instead, they would consider the next one on their list, and so on.

Considering these two factors (availability of facilities and preferences of patients) is important when locating senior centers which provide recreational and social activities for elderly (Hickerson et al., 2008). Existing location models for senior centers mainly consider accessibility and distance-based service demand (Drobne and Bogataj, 2015; Johnson et al., 2005). However, utilization of these centers depends on many other factors including alternative activities, availability of friends, affiliation with the center (Demko, 1980). Therefore, a senior may prefer to travel longer to visit a center that is closer to friends/relatives or provides more relevant activities rather than attending to the closest facility. In addition, a center may temporarily fail to provide services due to stochastic factors such as unplanned maintenances, accessibility issues, and health risks.

2.8.3 Emergency Response

The proposed models can also capture dynamics of locating emergency response facilities such as emergency operation centers, medical aid stations, evacuation points, etc. Given the

limited government funding, it is critical to locate such facilities efficiently. Citizens should access to or should be accessed from these facilities as quickly as possible to provide timely relief (Chen and Yu, 2016). However, the opened emergency response facilities may become inaccessible due to the effect of the disaster/emergency, for example, unsafe/blocked roads, or damaged/non-operational facilities (Akgün et al., 2015; An et al., 2013; Verma and Gaukler, 2015). In addition, it may not be reasonable to expect the citizens to travel to assigned emergency facilities as they may visit another one (possibly traveling more) due to having close-by relatives, seeking better quality service, or safety and welfare concerns (Teng et al., 2014). For example, in case of civil conflicts like that in Syria, refugees do not always travel to the camps in the closest neighboring regions or countries, but travel more distances (even under serious safety risks) to more developed countries for welfare concerns (Pecanha and Wallace, 2015).

2.9 Conclusions

The classical facility location models assume that a central planner makes both the location and allocation decisions. For example, this is the case when a firm ships products to its customers from different distribution centers. However, if the customers travel to the facilities to obtain service, they would attend the facility of their choice. That is, once the facilities are open, customers may not comply with the minimum cost allocation of the central planner any more. Furthermore, different customer types may have different preferences over the set of available facilities based on several factors such as social class, habits, work, age, to name a few.

We introduced the reliable facility location problem with customer preferences. This model opens facilities and allocates customers to a number of facilities in the order of their preferences. The goal of the model is to minimize the total cost of opening facilities plus the expected service cost. Less preferred facilities work as a backup if more preferred facilities fail. The proposed model bridges the gap between the location models that consider the preferences of customers and the ones that consider the reliability of facilities.

The demand, and other parameters in the model, are considered known inputs and

do not change. However, the demand values in this model are taken as proportion with respect to other demand values in the system. Therefore, even if the demand grows in absolute terms, the input to the model may still be valid, provided that the proportion of each demand to other demands is preserved. In other words, the percentages of demand are what matters in the system, not the actual values. Nevertheless, when designing for the long term, and proportions of demand may change. In this case, the model can still be used. Values of demand that are obtained from time series models or other forecast techniques will be the inputs for the model. This would ensure that the model will still be valid after the demand growth/change.

It was shown through experimentation that the preferences of customers do affect the location and allocation decisions. This shows the value of incorporating such characteristics in the model. If the decision maker does not include the preferences in the model, and customers take actions based on their preferences, the realized cost of the system would be significantly more than the expected cost.

The proposed model is more realistic, but at the same time it is more difficult to solve. A Lagrangian-based branch-and-bound procedure was developed to solve the model. Computationally, the LBB algorithm was presented and tested using three different implementations. Using callbacks within CPLEX did not prove useful due to the limitations imposed by CPLEX to guarantee safe data handling. The PQ-LBB was shown to be superior to the Stack-LBB. The three implementations were unable to perform well with larger instances. The PQ-LBB algorithm required excessive amount of time to process each node, and the improvement after each step was limited. Moreover, there was no clear pattern of the the performance. In practice, solving problems with large number of customers might be required to increase granularity when considering diverse preferences of several different customer types.

The modified model was then introduced. Unlike the preliminary model, the modified model gives the customers complete control over the allocation decisions The modified model also reduces the number of variables defined. We developed a Lagrangian branch-and-bound approach and a branch-and-cut approach based on a relaxed formulation. We also proposed a constraint which significantly tightens LP relaxation of the formulation. Our numerical experiments showed that the proposed solution algorithms can be applied to

problems with extremely large number of customers. In real life location problems, decision makers may need to consider a large number of customer types in terms of their preferences over the set of candidate sites. For instance, when choosing preventative healthcare facilities, patients might decide based on proximity and service quality (Verter and Zhang, 2015). There could be several preference types because patients might weigh proximity and service quality differently.

Some extensions of this work remain for future research. One direction would be improving the solution algorithms. Solving the Lagrangian dual can be time consuming due to the need to solve many mixed-integer subproblems. We may alleviate this difficulty using a Benders decomposition within an LP based branch-and-bound method. A pure Benders decomposition approach, however, may yield weak relaxations, leading to a large branch-and-bound tree. Therefore, we will try to use integrality constraints to obtain improved LP relaxations within the Benders decomposition framework (Bodur et al., 2017).

Another future direction would be adding a budget constraint or a limit on the number of facilities to open. This will be similar to the p-median problem. Also, we will explore other applications of the reliability models with customer preferences, especially in preventative healthcare.

Furthermore, if the demand is stochastic, and the variability is found to be significant and cannot be ignored, this deterministic model can still be used. To account for the variability of demand, a sampling is done on the values of each demand, and these 'realizations' are used as model parameters. In particular, best/worst case analysis can be conducted to examine the departure of the resulting solution from the deterministic case. The resulting solutions found by using different realizations are then compared to find an estimate of the sensitivity of the model to the variations in demand. In practice, if the variations are significant, the estimates of demand are continuously updated to ensure the lowest deviation of actual values from estimated ones. Alternatively, a set of possible scenarios, Ω, would be created, and the parameter in question would be indexed by $\omega \in \Omega$. Then, a set of constraints is added for each constraint in the original formulation representing different scenarios. The size of the set Ω is decided on by the modeler. A bigger Ω would account for more scenarios, but will also result in a bigger program, which will be computationally more expensive to solve.

Finally, it is possible to introduce capacity of facilities into the model. Allocating customers to capacitated facilities based on preference would be a non-trivial extension of our model, because in this case the model must determine which customers are denied service if there is not enough capacity at a highly preferred facility. It is also possible to model capacity levels for each facility as decision variables.

Capacity constraints can be incorporated in the model by adding the following set of constraints.

$$\sum_{i\in\mathcal{I}}\sum_{r\in\mathcal{R}}\eta_i X_j Y_{ijr} \leq L_j^{max} \qquad \forall\ j \in \mathcal{J} \tag{2.27}$$

where L_j^{max} is the maximum capacity of location $j \in \mathcal{J}$.

Constraints (2.27) can be added to the model (2.10) to form the *Capacitated RUFLO-R* or *RUFLO-RC*.

The resulting *RUFLO-RC* can be solved using two approaches. The first approach is to modify Step 3 of the LB&B algorithm in Table 2.28 to also include a check on the integer solution found to guarantee it respects the capacity constraints (the model is initially solved by ignoring the capacity constraints, then this check is performed). If the integer solution is not feasible, a cut is added to remove this solution from all nodes. The procedure would continue as usual afterwards.

The other approach is to add constraints (2.27) to (2.12) and relax it in a similar manner as (2.12c) using Lagrangian relaxation. Given that there will be two sets of constraints relaxed in this case, the lower bound obtained is expected to be worse than the lower bound obtained by relaxing only one set of constraints.

The efficiency of these two approaches, and other practical issues associated with implementing them are venues for future research.

Chapter 3

Resource Allocation in Colorectal Cancer Screening

Nearly 15.5 million Americans with a history of cancer were alive on January 1, 2016. About 1.7 million new cancer cases are expected to be diagnosed, and approximately 600,920 Americans are expected to die of cancer in 2017 (about 1,620 people per day). Nearly 1 of every 4 deaths in the US is caused by cancer, making cancer the second most common cause of death in the US, exceeded only by heart disease. The five-year Relative Survival Rate (RSR) for all cancers diagnosed in the US was 68% in 2006-2012, up from 49% in 1975-1977 (American Cancer Society, 2017).

About 810,045 Canadians (or 2.4% of all Canadians) had been diagnosed with cancer in the decade leading up to 2009. It is estimated that 206,200 Canadians will develop cancer and 80,800 will die of cancer in 2017. Cancer is the leading cause of death in Canada, responsible for nearly 30% of all deaths, followed by cardiovascular diseases and chronic lower respiratory diseases. The five-year RSR in Canada is 60% (Canadian Cancer Society, 2017).

Colorectal Cancer (CRC) is the third most common cancer in both men and women in the US, with an estimated of 135,430 new cases expected to be diagnosed, and an estimated 50,260 deaths expected to occur from it in 2017. In Canada, colorectal cancer is the second most common cancer in males, and the third most common in females. Approximately,

26,000 new cases of colorectal cancer are expected to be diagnosed and 10,000 deaths are expected to occur in 2017 (American Cancer Society, 2017; Canadian Cancer Society, 2017).

Declining incident rates and improvements in early detection and treatments have led to decline in the overall death rate. From 2007 to 2011, the overall colorectal cancer death rate declined by 2.5% per year in the US, and by 2.5% per year since 2004 for Canadian males and by 1.8% per year since 2001 for Canadian females (American Cancer Society, 2017; Canadian Cancer Society, 2017). However, the American Cancer Society estimates that the annual number of cancer deaths are growing. Between 2010 and 2015, the number of cancer deaths grew by 3.5% (American Cancer Society, 2017).

This chapter is devoted to discuss analytical frameworks aimed at finding an optimal screening policy for CRC for a representative population with limited screening resources. Before discussing the mathematical models, Section 3.1 provides an overview and history of CRC screening benefits and guidelines, as well as a review of related literature. Then, Section 3.2 presents a Markov decision process model. Since this model is hard to solve due the extremely large probability matrix, Section 3.3 discusses a mixed integer programming model that can be solved in reasonable time. In addition, the latter model accounts for factors not accounted for in the Markov decision process model, such as the age groups of the population, gender, and personal history of colonoscopy. Since the accurate estimation of different parameters is essential to obtain reasonable results in this model, a description of the data sources is presented in Section 3.4. Numerical results and analysis are shown in Section 3.5, followed by final remarks in Section 3.6.

3.1 Introduction

In this introduction, a description of the fundamentals in CRC screening is given, followed by a brief history of the development of the CRC screening guidelines (mainly in the US). Then, a review of the related literature is given. Finally, the model of CRC disease progression is discussed. The goal is to extend this one-patient model to consider all individuals in the target population.

3.1.1 Colorectal Cancer Screening

Cancer screening tests can improve survival and decrease mortality by detecting cancer at an early stage, when treatment is more effective. Moreover, regular use of cervical and colorectal cancer screening tests can prevent the development of cancer through identification and removal or treatment of premalignant abnormalities (American Cancer Society, 2017).

Most CRC cases originate from benign growths on the inner surface of the colon and rectum (called adenomatous polyps), which may progress to CRC (Loeve et al., 2004). This natural progression of the CRC makes it possible for testing procedures to discover these lesions and adenomatous polyps early on.

CRC screening can be accomplished using various methods. These include colonoscopy, sigmoidoscopy, Computed Tomography (CT) colonography (virtual colonoscopy), double-contrast barium enema, DNA stool test, and Fecal Occult Blood Test (FOBT) (Pignone et al., 2002). The screening methods can be roughly categorized into two distinct groups: tests that primarily detect cancer, and structural exams that detect both cancer and pre-cancerous polyps. The methods for structural examinations, which detect both cancer and advanced lesions, include flexible sigmoidoscopy, colonoscopy, CT colonography, and double-contrast barium enema (McFarland et al., 2008). Methods in the cancer detection group are mainly stool tests, which include occult blood or exfoliated DNA (Levin et al., 2008).

The details of each testing procedure is beyond the scope of the current discussion. However, it is worth mentioning that these tests vary in their accuracy and disutility (see Table 3.1). Accuracy of a test is the chance of correctly detecting colorectal lesions (CRC and polyps). The disutility arises from pain, uneasiness, and anxiety associated with the screening procedures, preparations, complications, and time delay before obtaining pathology results. The FOBT test, for example, requires drug and dietary restrictions before the test, and may not detect a tumor that is not bleeding (National Cancer Institute, 2016). In general, invasive screening methods, such as colonoscopy, have higher disutility but also higher accuracy. Other screening methods have lower accuracy as well as disutility.

There have been calls to state a preference for colonoscopy above all other options (Allison and Lawson, 2006). Colonoscopy is the most accurate and commonly recommended

Table 3.1: CRC Screening Methods

Screening Method	Test Type	Accuracy	Disutility
Colonoscopy	Invasive test	Very high	Very high
Sigmoidoscopy	Invasive test	High	High
CT Colonography	X-ray test	High	Low
Barium Enema	X-ray test	Low	Low
DNA Stool Test	Stool test	High	Lower
Fecal Occult Blood Test	Stool test	Lower	Lower

Source: Erenay et al. (2014); National Cancer Institute (2016)

screening test in the US (Krist et al., 2007). Moreover, colonoscopy is the standard screening test for the CRC follow-up and surveillance (Winawer, 2007).

Randomized trials and observational studies have demonstrated mortality reductions associated with early detection of invasive disease, as well as removal of adenomatous polyps (Hardcastle et al., 1996; Kronborg et al., 1996; Mandel et al., 2000; Selby et al., 1992). Moreover, there is both direct and indirect clinical evidence that CRC screening methods are effective for CRC prevention (Pignone et al., 2002).

Screening is also beneficial after detecting and removing a polyp (polypectomy) because the lifetime risk of CRC is not completely eliminated. After a polypectomy, a missed (synchronous) or new (metachronous) adenomatous polyp may progress to CRC (Yang et al., 1998). Thus, the risk of CRC remains even after CRC treatment because patients may suffer recurrence of their disease (Kjeldsen et al., 1997; Scholefield and Steele, 2002). In addition, patients who are successfully treated for CRC may develop new adenomatous polyps and these new polyps may also progress to CRC (metachronous CRC) (Fajobi et al., 1998; Park et al., 2006).

As such, screening guidelines were developed to help patients and physicians in preventing and early detecting of CRC occurrence or re-occurrence. In the past decade, there has been progress in reducing CRC incidence and death rates. These declines can be attributed to improved utilization of CRC screening on early detection and prevention

through polypectomy, risk-factor reduction (e.g., declining tobacco use), and improved treatments (Edwards et al., 2010).

The prevalence of CRC screening has stabilized in more recent years and still lags behind breast and cervical screening prevalence (American Cancer Society, 2015), although CRC claims more lives. In addition, Klabunde et al. (2009) report that 43% of clinicians recommend more frequent colonoscopy screening than the guidelines for low-risk patients. Therefore, it is important to initiate CRC screening procedures for a larger portion of the population in order to both prevent cancer and detect it early.

Howlader et al. (2015) note that the relative five-year survival rate is 90% for CRC patients diagnosed at an early, localized stage, while only 40% of cases are diagnosed with this stage. Table 3.2 shows the percentage of American adults that had undergone CRC screening in 2013. Endoscopy is the general term used for medical procedures in which an instrument, called an endoscope, is put into the body to look inside. In colon and rectal regions, this procedure is referred to as colonoscopy and sigmoidoscopy. As can be seen, 58.6% were up-to-date with screening (either an FOBT within the past year or a sigmoidoscopy within the past five years or a colonoscopy within the past 10 years). Compared to 46.8% in 2005 (American Cancer Society, 2009). This represents an increase in cancer screening compliance in the US. The US Preventive Services Task Force (USPSTF) recommends only routine screening for CRC up to age 75 (US Preventive Services Task Force, 2008). For this population (ages 50-75 years), 57.2% were up-to-date with USPSTF screening recommendations.

Most screening guidelines recommend initiating CRC screening at a later age. This can be justified by looking at Table 3.3, which shows the probability of developing invasive cancer during selected age intervals for American adults in 2009-2011. These figures are for those who are free of cancer at the beginning of each age interval. "All sites" excludes basal cell and squamous cell skin cancers and in situ cancers except urinary bladder. As depicted in the table, there is a higher chance of developing cancer for older individuals.

The overall incident rate of CRC in Canada has decreased slightly since 2000 (Canadian Cancer Society, 2016). This decline is prominent among older adults, as rates are increasing among young adults (under the age of 50 years) in Canada (BC Cancer Agency, 2013; CCO,

Table 3.2: Colorectal Cancer Screening Rate (%) in the US 2013

	Fecal Occult Blood Test		**Endoscopy**		**Combined FOBT/Endoscopy**	
	50 to 75 years	50 years and older	50 to 75 years	50 years and older	50 to 75 years	50 years and older
Male	8.0	7.8	53.6	56.1	56.3	58.8
Female	7.7	7.7	55.2	55.8	58.1	58.6
Overall	7.8	7.8	54.4	55.9	57.2	58.6

Source: American Cancer Society (2015)

Table 3.3: Probability (%) of Developing Invasive Cancer during Selected Age Intervals by Sex, US, 2009-2011.

		Birth to 49	**50 to 59**	**60 to 69**	**70 and Older**	**Birth to Death**
All sites	Male	3.4 (1 in 29)	6.7 (1 in 15)	15.1 (1 in 7)	36.0 (1 in 3)	43.3 (1 in 2)
	Female	5.4 (1 in 19)	6.0 (1 in 17)	10.0 (1 in 10)	26.4 (1 in 4)	37.8 (1 in 3)
Colon & Rectum	Male	0.3 (1 in 300)	0.7 (1 in 148)	1.3 (1 in 80)	3.9 (1 in 26)	4.8 (1 in 21)
	Female	0.3 (1 in 326)	0.5 (1 in 193)	0.9 (1 in 112)	3.5 (1 in 28)	4.5 (1 in 22)

Source: American Cancer Society (2017)

2016; Patel and De, 2016) and the United States (Austin et al., 2014).

CRC screening can identify and remove precancerous polyps and reduce cancer incidence. Starting from 2007, CRC screening programs began in some provinces in Canada for people aged 50 and older who are at average risk of the disease. As of 2016, all 10 provinces had implemented or considered implementing organized colorectal cancer screening programs (CPAC, 2017).

The discussion above shows the benefit of CRC screening in the prevention and early detection of polyps and cancerous lesions. With improved awareness and various national efforts to increase the compliance rate of CRC screening, the demand on medical screening resources is expected to increase. Also, the aging population of western countries imposes higher demand on healthcare resources including cancer screening procedures. These factors, coupled with the scarcity of cancer screening resources, make it vital for healthcare systems to plan ahead for the best usage of resources. Mathematical and analytical models are powerful tools to understand the current challenges, and to provide solutions and recommendations to the policy makers using statistical, mathematical, and computational procedures.

3.1.2 Development of Colorectal Cancer Screening Guidelines

During the 1990's, the US Agency for Health Care Policy and Research assembled an expert panel to prepare clinical practice guidelines for colorectal cancer screening, and an accompanying rationale based on the best available evidence. The Panel published a report (Winawer et al., 1997) highlighting a substantial body of research evidence favoring colorectal cancer screening. Afterwards, guidelines for CRC screening were published by the American Cancer Society (Smith et al., 2001), the USPSTF (US Preventive Services Task Force, 2002), the American College of Gastroenterology (Rex et al., 2000), and the American Society of Colon and Rectal Surgeons (Simmang et al., 1999). This showed a national consensus favoring colorectal cancer screening.

These guidelines are subject to change and update. Smith et al. (2015) list the major updates these guidelines have been through, as shown in Table 3.4. The most recent Society

guidelines, which were in collaboration with the American College of Radiology and the US Multi-Society Task Force on Colorectal Cancer (a consortium representing the American College of Gastroenterology, the American Society of Gastrointestinal Endoscopy, and the American Gastroenterological Association), were released in 2016. The official statement can be found in the US Preventive Services Task Force (2016).

Table 3.4: History of Recent Updates to American Cancer Society Cancer Early Detection Guidelines for Colorectal Cancer

Year	**Update**
2001:	Complete update
2003:	Technology update
2006:	Update for postpolypectomy and postcolorectal cancer resection surveillance
2008:	Complete update
2016:	Update for asymptotic patients †

Source: Smith et al. (2015). † Update was not announced at the time of Smith et al. (2015)

The 2016 update reaffirms the 2008 guidelines on the benefits of screening adults, 50-75 years of age. The new guidelines recommend screening of adults ages 76-85 on an individualized basis, depending on the patient's health and previous screening history. This is different than the 2008 recommendation against subjecting individuals of this age group to routine screening. In its 2008 recommendation, the task force discussed screening with flexible sigmoidoscopy every five years, combined with either Fecal Immunochemical Test (FIT) or gFOBT every three years. The current recommendation statement specifically discusses screening with flexible sigmoidoscopy every 10 years, combined with an annual FIT. Note that all these updates are for asymptomatic patients at low risk. Choi et al. (2017) provide a comparison of the most recent recommendations of different organizations.

Table 3.5 shows the Society's CRC screening guidelines for average-risk asymptomatic people (American Cancer Society, 2017). The American Cancer Society and other organiza-

tions recommend more intensive surveillance for individuals at higher risk [1] of CRC (Smith et al., 2001; Winawer et al., 2003). The model developed here considers both low-risk and high-risk patients. Moreover, post-CRC patients are clustered into a separate risk level to more accurately describe the real-life dynamics.

Most Canadian provinces have organized colorectal cancer screening programs, each with specific guidelines that may differ in each province and territory. Nevertheless, the Canadian Cancer Society recommends that men and women of age 50 and over have a stool test at least every 2 years, with appropriate follow up (Canadian Cancer Society, 2015). The latest guidelines by the Canadian Task Force on Preventive Health Care can be found in the Canadian Task Force on Preventive Health Care (2016).

While the knowledge of guidelines is considered high (Rex et al., 2015), the evidence shows that the actual practice tend to both overuse the surveillance examination in low-risk patients and underuse it in high-risk patients (Schoen et al., 2010). Therefore, the impact of recommended policies as well as the actual practice need to be quantified. Models discussed in this thesis help in this regard.

[1] These include: individuals with a history of adenomatous polyps, individuals with a personal history of curative-intent resection of CRC, individuals with a family history of either CRC or colorectal adenomas diagnosed in a first-degree relative, individuals at significantly higher risk because of a history of inflammatory bowel disease of significant duration, or individuals at significantly higher risk because of the known or suspected presence of a hereditary syndrome, such as Lynch syndrome, or familial adenomatous polyposis (Smith et al., 2015)

Table 3.5: American Cancer Society's CRC Screening Guidelines 2016 for Men and Women Ages 50+

Test or Procedure	**Frequency**
gFOBT with at least 50% test sensitivity for cancer, or FIT with at least 50% test sensitivity for cancer, **or**	Annual testing of spontaneously passed stool specimens. Single stool testing during a clinician office visit is not recommended, nor are throw in the toilet bowl tests. In comparison with guaiac-based tests for the detection of occult blood, immunochemical tests are more patient-friendly and are likely to be equal or better in sensitivity and specificity. There is no justification for repeating FOBT in response to an initial positive finding.
Stool DNA test, **or**	Every 3 years
Flexible sigmoidoscopy (FSIG), **or**	Every 5 years alone, or consideration can be given to combining FSIG performed every 5 years with a highly sensitive gFOBT or FIT performed annually.
Double-Contrast Barium Enema (DCBE), **or**	Every 5 years
Colonoscopy	Every 10 years
CT Colonography	Every 5 years

Source: American Cancer Society (2017)

3.1.3 Review of Related Literature

There is a number of studies discussing different aspects of CRC screening. These includes: the utility of screening for older patients (Schoen, 2006), the timing of CRC screening termination (Maheshwari et al., 2008), and the impact of new screening modalities such as CT colonography on CRC prevention (Regueiro, 2005). This chapter discusses models that efficiently allocate limited cancer screening resources among a population of different risk levels, ages, and personal cancer history.

Multiple risk factors need to be considered when a cancer screening policy is developed. For example, an analysis based on a microsimulation model (Ramsey et al., 2010) suggests that early screening colonoscopy in subjects with a family history of CRC may be cost-effective. Ladabaum et al. (2010) suggest that persons with a family history of CRC could benefit the most from screening, and screening for them could be most cost effective. Pfister et al. (2004) discuss the best screening schedule for patients after curative treatment of CRC. In the models discussed in this chapter, the family history is implicitly accounted for by classifying a proportion of the population to be at high risk. Patients who undergo CRC treatment are also considered in the model.

The mathematical models that are built to study various aspects of CRC, including the one discussed in this chapter, depend on the understanding the natural history of the CRC, such as the rate of progression from adenomatous polyp to CRC. Several studies on CRC aim to estimate unobservable CRC progression parameters using publicly available databases, such as Surveillance, Epidemiology, and End Results Program (SEER) database as benchmark statistics. Erenay et al. (2011) estimate a set of parameters revealing some of the characteristics of metachronous CRC. Moreover, Roberts et al. (2007) build a more detailed discrete-time simulation model that mimics the progression of CRC. They also use the simulation model to measure the performances of different CRC screening policies. These studies, among others, will be used to specify the inputs of the models in this chapter.

Partially Observable Markov Decision Process (POMDP) models are important tools used to solve stochastic systems. POMDP models are used in other cancer screening problems. Ayer et al. (2012) provide a POMDP model to determine patient-specific mammography screening times. Maillart et al. (2008) use a partially observable Markov chain formu-

lation to examine the value of dynamic screening policies in which the length of screening interval can be a function of patient age. Zhang et al. (2012b) propose a POMDP model for screening for prostate cancer, while Zhang et al. (2012a) develop another POMDP application to determine the optimal timing of biopsy, based on annual prostate-specific antigen test results. The model of Section 3.2 is a POMDP model that is specific to CRC screening and aims at finding the optimal screening policy in a limited resource environment.

Leshno et al. (2003) develop a hidden Markov chain with two states for different polyp sized and three states for CRC stages. They evaluate the performances of six screening policies. Erenay et al. (2014) propose a similar model but with a dynamic programming mechanism and solve it optimally. They also account for personal history of CRC. These models of a single patient are referred to in building the models in this chapter.

Yaesoubi and Cohen (2011) propose a simplified Markov chain model for infectious disease spread. Their framework and some of their notation are used here as a building block for our model. Ayvaci et al. (2012) develop a Markov Decision Process (MDP) model to capture diagnosis decision after mammography under restricted resources. Their objective differs from the objective of the models here, which is allocating screening resources among a representative population.

The demand for cancer screening is projected to increase as a result of increased compliance. The American Cancer Society joined the National Colorectal Cancer Roundtable in its '80% by 2018' initiative in 2013 (American Cancer Society, 2015). The goal of this campaign is to increase the rate of regular colorectal cancer screening among adults 50 and older to 80% by 2018, with an emphasis on economically disadvantaged individuals, who are least likely to be tested. Higher rates of compliance would bring more challenges to the screening programs regarding capacity and resources available. Some of these challenges are discussed in Güneş et al. (2015), where they provide an analysis indicating what would happen if clinicians use more frequent screening schedule, or if compliance rate increases. They show that the benefits of screening programs can be realized only if the available service capacity matches the increasing demand. The objective of their model is to minimize the incidence rate or the mortality rate. This is different than our objective of maximizing the total expected quality adjusted life years.

The MDP model developed in this chapter is hard to solve. Therefore, a deterministic mixed integer program is presented. The use of mathematical programs in healthcare is well-established (see Chapter 1 for discussion). There are also applications of integer or mixed integer programs in the context of cancer screening. For example, Kim et al. (2006) develop binary integer programming model to identify an optimal package of health services to be provided during a single visit for a particular target population. Demarteau et al. (2012) present a linear program that determine the combination of the different prevention options to minimize cervical cancer screening coverage and vaccination coverage constraints.

Some of the other modeling and solution methodologies used in the cancer screening applications are as follows. Güneş et al. (2015) develop a compartmental model for the allocation of colonoscopy resource among preventive and diagnosis activities. Simulation models have been used to study the effectiveness of cancer screening strategies. The Microsimulation SCreening ANalysis (MISCAN) has been used to compare the effectiveness of different colorectal screening strategies (Loeve et al., 1999). MISCAN has been applied to other types of cancer as well (Fone et al., 2003). The models developed in this chapter are not bound by a set of policies. Instead, they find the optimal policy based on the input parameters and constraints.

3.1.4 Colorectal Cancer Natural Progression for a Single Patient

The current health state of a single CRC patient is represented using three cancer stages: without lesion, (having adenomatous) polyp, and (having) CRC. Also, patients are categorized based on risk level into: low-risk, high-risk, and post-CRC. This is based on the American Gastroenterology Association classification of patients based on personal history. Low-risk patients are those asymptomatic without personal or family history of CRC. High-risk patients are those with a history of adenomatous polyp, while post-CRC are patients with history of CRC (Winawer et al., 2003). High-risk patients may also include individuals with a family history of either CRC or colorectal adenomas diagnosed in a first degree relative, individuals with known or suspected presence of a hereditary syndrome, such as Lynch syndrome, or familial adenomatous polyposis.

These risk levels are completely observable, and the patient moves from one risk level to another after a completely observable event occurs (Erenay et al., 2014). Having different risk levels is important since current CRC screening guidelines provide different recommendations for patients who already had polypectomy and CRC (see Table 3.5 for screening guidelines for low-risk patients).

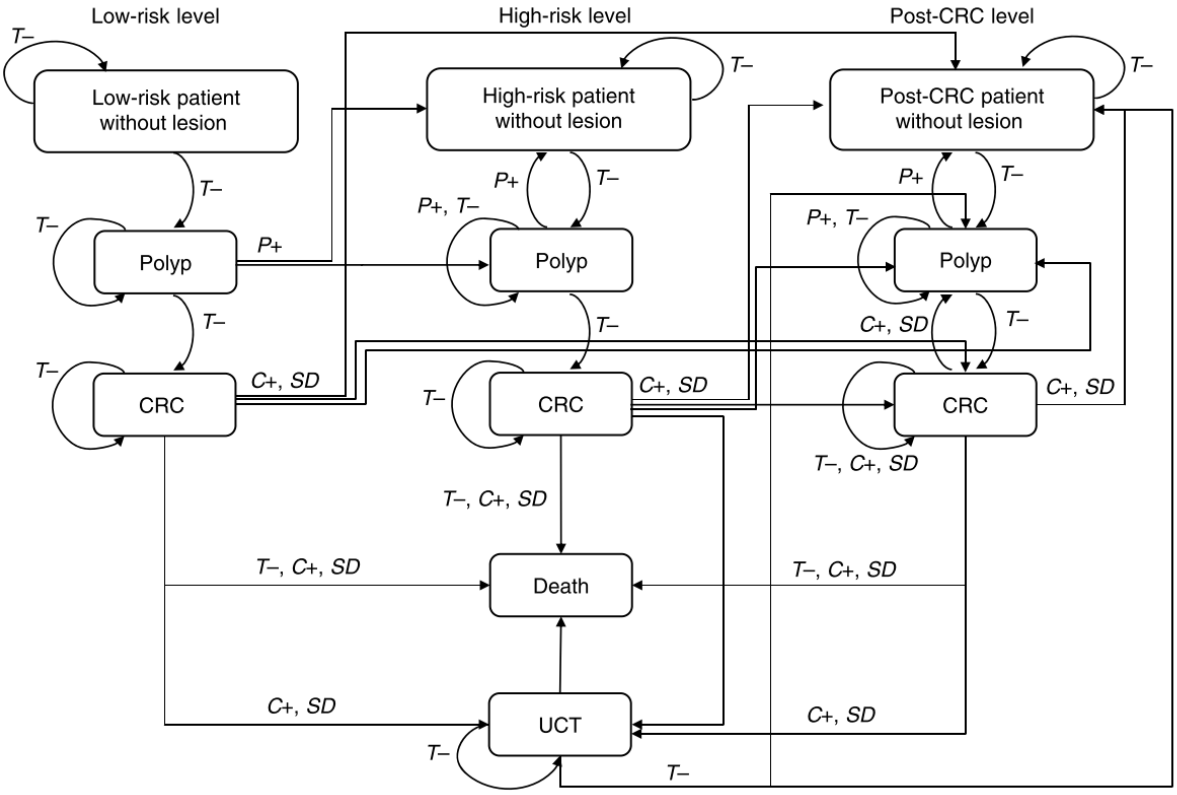


Figure 3.1: Core Health State Transitions for an Individual Patient According to the Screening Results. *Source: Erenay et al. (2014)*

A colonoscopy may detect an adenomatous polyps, a CRC lesion, or nothing suspicious in the colon or rectum. Furthermore, a patient may experience sever CRC symptoms and undergo a diagnostic colonoscopy screening, which is termed self-diagnosis (SD).

Figure 3.1 shows how core health states for a single patient randomly change based on colonoscopy results. Screening results T-, $P+$, $C+$, and SD refer to test negative, detection

of adenomatous polyp via colonoscopy, detection of CRC via colonoscopy, and self-diagnosis of CRC, respectively. There are 11 core health states. Arrows represent possible core state transitions, each based on some screening result. When the colonoscopy test is positive, the transitions $P+$ and $C+$ occur, whereas, T- and SD result from the natural progression of the disease and can occur at any time period. Transition to mortality (D) can occur from any core health state. However, the graph omits them to improve readability.

A complete description of Figure 3.1 can be found in Erenay et al. (2014). To understand the process, it is important to introduce some concepts regarding the CRC screening mechanism. The probability of accurately identifying the patients with no colorectal lesions is equal to 1 (Frazier et al., 2000). Moreover, the sensitivity of colonoscopy is the probability of accurately detecting CRC lesions. A similar definition applies for sensitivity of polyps.

A brief description of transition dynamics is now given for low-risk level. High-risk and post-CRC levels follow a similar logic. A low-risk patient will always have a screening result T-, as long as the patient has no lesions. A patient may develop an adenomatous polyp and move to the polyp state within the year. Otherwise, the patient stays in the same health state. If the patient has an adenomatous polyp at the beginning of the year, and the test missed the polyp (T-), this polyp either stays as an adenomatous polyp or turns into a CRC within the year. This can occur with probability equal to 1-sensitivity of colonoscopy. If the test detects and removes the polyp ($P+$) at the beginning of the current year, the patient either develops a new adenomatous polyp and moves to the polyp state in the high-risk level or moves to the high-risk patient without lesion core health state. If the cancer treatment is not successful, the patient either dies (D) or becomes under cancer treatment (UCT) during that year.

Thus, the core health (or disease progression) states of an individual patient are {*LR0*, *LR1*, *LR2*, *HR0*, *HR1*, *HR2*, *PC0*, *PC1*, *PC2*, *UCT*, *D*}.

3.1.5 Contributions

In this chapter, two models are formulated for the problem of allocating CRC screening resources among a representative population of individuals. The first is an MDP model

that keeps track of the stochastic nature of the transitions in the system. The second model is MIP that adds extra dimensions and population dynamics to the MDP model. However, the MIP uses deterministic transitions to model the behavior of large group of individuals in the system. The results from the MIP model show that the current guidelines are not always optimal, and the system favors females and younger individuals, as will be discussed later on.

3.2 Discrete-Time Markov Decision Process Model

A full description of a discrete-time MDP model is given here. It will be explained later that this model is hard to solve. Therefore, a mixed integer program is introduced in Section 3.3, which will be solved and analyzed.

This section starts with a description of the model and an explanation of the mathematical notation. Then, the model is formally introduced. This is followed by a discussion of state aggregation and approximate dynamic programming; two main techniques that may help is solving the model. Finally, the last section explains the difficulties in solving this model, and the need for alternative modeling concept.

The following is a verbal description of the model. The aim of this model is to develop a screening policy for a representative population such that the capacity and available resources are taken into consideration. The objective of the decision maker is to maximize a social welfare measure, which is Quality Adjusted Life Years (QALYs) for the population. Patients are categorized according to the risk level into: low-risk, high-risk, and post-CRC. These levels are completely observable by the decision maker. Within each risk level, the patients are divided into three unobservable clusters depending on cancer progression. These clusters are: no lesions, (having) polyp, and (having) CRC. Given it is a standard and commonly recommended procedure, colonoscopy is considered as the screening method. The policy maker decides on the number (or percentage) of patients to undergo colonoscopy with each level.

As such, the MDP model developed here defines a state as the number of individuals in each core health state. As explained above, there are nine core health states (three risk

levels, with three cancer progression levels in each of them), in addition to under cancer treatment (UCT) and death (D) core health states. This brings the total of core health states to 11. Actions are defined as the percentage of patients in each risk level to undergo colonoscopy at a given year. It is assumed in this model that the population starts at age 50, and colonoscopy is no longer performed after age 75.

3.2.1 Model Description

In this section, the MDP model is formally introduced. The model developed aims at finding optimal screening policies for a representative population whose individuals are subject the disease progression pattern described in the previous section. The major consideration for the policy maker is the limited CRC screening resources available. The decision maker must allocate scarce screening resource such that the society welfare (e.g., total QALYs) is maximized.

One screening method is considered in this model, which is colonoscopy, since it is the most commonly recommended screening procedure (Krist et al., 2007). A patient is faced with a decision to undergo screening (colonoscopy) in each year or not. This decision is made by the policy maker, and it is assumed that patients accurately follow the suggested policy, that is, 100% of patients who are recommended to undergo a colonoscopy in a given year will indeed perform it. This perfect compliance will be relaxed in Section 3.3.

Yaesoubi and Cohen (2011) use a discrete-time Markov model to formulate the spread of infectious disease among a particular fixed size population. We adopt some of their notations and definitions in this model. However, due to the significant difference in the two areas of application, more/new notation is introduced to accurately model the CRC screening system.

An individual patient can be in any one of the core health states of the system at time 0. The core health states are denoted by s_i, where $i \in \{0, 1, \ldots, M\}$. Therefore, there are $M + 1$ core health states in the system. In particular, $(M + 1) = 11$. This is a discrete-time model, which starts at time 0, and terminates at time t^{max}. At a given time $t \in \mathcal{T} = \{0, 1, \ldots, t^{max}\}$, the number of individuals in core health state s_i is $X_i, i \in$

$\{0, 1, \ldots, M\}$. In this model, it is assumed that the population is fixed and equal to N. Thus, the following holds.

$$\sum_{i=0}^{M} X_i(t) = N \tag{3.1}$$

Equation (3.1) also means that the system state is fully identified by a sub-vector of only M variables from $X(t) = \{X_0(t), X_1(t), \ldots, X_M(t)\}$. The variable remaining can be determined by the values of the others.

The states of the system are defined as the number of individuals in each health state. There are $M+1$ core health states, and a fixed population N. In general, without any more restrictions, there are $\binom{N+(M+1)-1}{(M+1)-1} = \binom{N+M}{M}$ states of the system. Hence, a system state $X(t)$ at time $t \in \mathcal{T} = \{0, 1, \ldots, t^{max}\}$ can be defined as $X(t) = \{X_0(t), X_1(t), \ldots, X_{10}(t)\} \in \mathcal{X}$, such that (3.1) holds.

Each core health state $s_i, i \in \{0, \ldots, M\}$ is accessible from a set of core health states. In other words, individuals in a particular core health state can transition to one of the health states that are accessible from where they originally are. This can be thought of as a set of inflows (denoted by $\overline{s_i}$), or a set of outflows (denoted by $\underline{s_i}$) of each core health state, respectively. Indeed, these sets can be empty (denoted by ϕ). Table 3.6 shows these sets for all health states in our model.

The driving event $u_{ij}(t), s_j \in \underline{s_i}, i \in \{0, \ldots, M\}$ is a non-negative discrete random variable representing the number of transitions (number of individuals transitioning) from health state s_i to s_j during the interval $[t, t+\Delta t], t \in \{0, \Delta t, 2\Delta t, \ldots, t^{max}\}$. The assumption in this model that $\Delta t = 1$. Let $P_{u_{ij}(t)}(\cdot)$ denote the probability mass function for the random variable $u_{ij}(t)$; that is

$$P_{u_{ij}(t)}(c) = \Pr\{u_{ij}(t) = c\}, s_j \in \underline{s_i},$$

for some integer c.

Furthermore, the stochastic flow coming *out* of a health state $s_i, i \in \{0, \ldots, M\}$ at time $t \in \mathcal{T} = \{0, 1, \ldots, t^{max}\}$ is $\underline{X_i(t)}$, and the flow coming *into* a health state s_i at time

Table 3.6: Sets of States to and from All Health States s_i

s_i	$\overline{s_i}$	$\underline{s_i}$
s_0 (LR0)	ϕ	$\{s_1, s_{10}\}$
s_1 (LR1)	$\{s_0\}$	$\{s_2, s_3, s_4, s_{10}\}$
s_2 (LR2)	$\{s_1\}$	$\{s_6, s_7, s_8, s_9, s_{10}\}$
s_3 (HR0)	$\{s_1, s_4\}$	$\{s_4, s_{10}\}$
s_4 (HR1)	$\{s_1, s_3\}$	$\{s_3, s_5, s_{10}\}$
s_5 (HR2)	$\{s_4\}$	$\{s_6, s_7, s_8, s_9, s_{10}\}$
s_6 (PC0)	$\{s_2, s_5, s_7, s_8\}$	$\{s_7, s_{10}\}$
s_7 (PC1)	$\{s_2, s_5, s_6, s_8\}$	$\{s_6, s_9, s_{10}\}$
s_8 (PC2)	$\{s_2, s_5, s_7\}$	$\{s_6, s_7, s_9, s_{10}\}$
s_9 (UCT)	$\{s_2, s_5, s_7\}$	$\{s_6, s_7, s_{10}\}$
s_{10} (D)	$\{s_1, s_2, \ldots, s_9\}$	ϕ

t is $\overline{X_i(t)}$. This flow can be decomposed based on which health state it is coming from or going to. Specifically, if an individual can transfer from health states s_j and s_k to health state s_i, that is $\overline{s_i} = \{s_j, s_k\}$, then $u_{j,i}(t) + u_{k,i}(t) = \overline{X_i(t)}$. Similarly, if individuals are transferred from health state s_i to health states s_y and s_z, that is $\underline{s_i} = \{s_y, s_z\}$, then $\underline{X_i(t)} = u_{i,y}(t) + u_{i,z}(t)$. Table 3.7 shows the variables used in the MDP model and their descriptions.

It is important to note that $u_{ij}(t), s_j \in \underline{s_i}, i \in \{0, \ldots, M\}$, and consequently, $\overline{X_i(t)}$ and $\underline{X_i(t)}$ for $i \in \{0, \ldots, M\}$ are action-dependent. This means that the values of these variables depend on the action taken. The random variables $u_{i,j}(t)$ are assumed to be independently distributed for all health states s_i and $s_j \in \underline{s_i}, i \in \{0, \ldots, M\}$. Also, these random variables are only determined by the state of the system at time t, which is $X(t) = \{X_0(t), X_1(t), \ldots, X_M(t)\}$.

The set of dynamic driving constraints summarizes the relationships among the driving events during interval $[t, t + \Delta t]$ and the state of the system at time t and $t + \Delta t$.

$$X_0(t+\Delta t) = X_0(t) - \underline{X_0(t)} \tag{3.2a}$$

$$X_i(t+\Delta t) = X_i(t) + \overline{X_i(t)} - \underline{X_i(t)} \qquad \text{for } i \in \{1, 2, \ldots, M-1\} \tag{3.2b}$$

$$X_M(t+\Delta t) = X_M(t) + \overline{X_M(t)} \tag{3.2c}$$

This can be translated into:

$\underline{X_0(t)} = u_{0,1}(t) = \overline{X_1(t)}$, $\underline{X_1(t)} = u_{1,2}(t) + u_{1,3}(t) + u_{1,4}(t)$ and so on.

The actions of the MDP model represent the proportion of individuals in each core health state to undergo a colonoscopy at time t. However, disease progression is unobservable within each risk level. There are five risk levels, $\mathcal{R} = \{LR, HR, PC, UCT, D\}$. For example, low-risk patients are indistinguishable to the policy maker, and are stochastically distributed into without lesion, polyp, and CRC levels within the same risk-level. Therefore, the decision maker performs action $a(t) = \{a^{LR}(t), a^{HR}(t), a^{PC}(t)\} \in \mathcal{A}$ at time $t \in \mathcal{T}$, where $a^{LR}(t), a^{HR}(t)$, and $a^{PC}(t)$ are, respectively, the proportion of patients in LR, HR, and PC level to undergo colonoscopy. Alternatively, action $a(t) \in \mathcal{A}$ at time $t \in \mathcal{T}$ can be expressed as: $a(t) = \{a_0(t), a_1(t), \ldots, a_8(t)\} \in \mathcal{A}$, such that:

$$\sum_{i=0}^{8} a_i(t) X_i(t) = L^{max} \tag{3.3a}$$

$$a_i(t) = a^{LR}(t) \qquad \text{for } i \in \{0, 1, 2\} \tag{3.3b}$$

$$a_i(t) = a^{HR}(t) \qquad \text{for } i \in \{3, 4, 5\} \tag{3.3c}$$

$$a_i(t) = a^{PC}(t) \qquad \text{for } i \in \{6, 7, 8\} \tag{3.3d}$$

where L^{max} represents the capacity limit for the available colonoscopy resources. Equation (3.3a) enforces the number of people to undergo colonoscopy cannot exceed the capacity limit. Equations (3.3b)-(3.3d) state that within each risk level, the actions must be identical because of the unobservable disease progression levels.

Time periods are in years; $t \in \mathcal{T} = \{0, 1, \ldots, t^{max}\}$, which represents the number of years after age 50. This assumption is made since almost all guidelines suggest initiating

CRC screenings at or after age 50 (Levin et al., 2008; Winawer et al., 2003). The maximum age t^{max} is the age after which colonoscopy offer little to no value. It is assumed in this model that $t^{max} = 25$, meaning that no patient will undergo colonoscopy after age 75. Studies on the termination of CRC screening suggest to stop screening at ages 75-85 (Maheshwari et al., 2008; Zauber et al., 2008), and The USPSTF only recommends routine screening for CRC up to age 75 (US Preventive Services Task Force, 2008).

The assumption of initiating screening at age 50 means the model is considering a particular age group only. While this assumption would help computationally, future extensions to this model would add more age groups, as well as gender-specific classification. It is also worth noting that with larger population N and, consequently, large number of individuals in each core health state, it is possible to approximate the binomial distribution by the normal distribution. This is particularly useful since different normal distributions with different variances can be added and characterized. This would be one direction of future work in this model.

Table 3.7: Variable Description for MDP Model

Variable	Description
$\mathcal{I}$	The set of disease progression states, indexed by i,
$\mathcal{R}$	The set of risk levels, indexed by R, $\mathcal{R} := \{LR, HR, PC, UCT, D\}$,
$\mathcal{O}$	The set of observations, indexed by o, $\mathcal{O} := \{T\text{-}, P+, C+, SD\}$,
$\mathcal{A}$	The set of action vectors, indexed by $a(t)$,
$\mathcal{T}$	The set of time periods, indexed by t,
$\mathcal{S}$	The set of health states, indexed by s,
$\mathcal{X}$	The set of system states, indexed by $X(t)$,
s	The health state vector, $s := \{s_0, s_1, \ldots, s_M\}$,
s_i	The core health state, where $i \in \{0, 1, \ldots, M\}$,
$X_i(t)$	Number of individuals in state $s_i, i \in \{0, \ldots, M\}$ at time $t \in \mathcal{T}$,
$X(t)$	A vector of $X_i(t)$ for all $i \in \{0, \ldots, M\}$,
$\overline{s_i}$	The set of health states that leads to health state s_i,
$\underline{s_i}$	The set of health states that health state s_i leads to,

$\overline{X_i(t)}$ Random variable that represents the number of individuals transitioning into health state s_i during the interval $[t, t + \Delta t]$,

$\underline{X_i(t)}$ Random variable that represents the number of individuals transitioning from health state s_i during the interval $[t, t + \Delta t]$,

$u_{ij}(t)$ Random variable that represents the number of individuals transitioning from health state s_i to health state s_j during the interval $[t, t + \Delta t]$,

$\mathbf{u}(t)$ A vector of $u_{ij}(t)$ for all $i \in \{0, \ldots, M\}$,

t^{max} The duration (in years) that the model is run for,

$a_i(t)$ Action taken at time t, which is the proportion of individuals in core health state s_i to undergo colonoscopy,

$a^R(t)$ Action taken at time t, which is the proportion of R risk level individuals to undergo colonoscopy, where Equations (3.3b)-(3.3d) apply,

$a(t)$ Action vector at time $t \in \mathcal{T}$, or $a(t) = \{a^{LR}(t), a^{HR}(t), a^{PC}(t)\} \in \mathcal{A}$, and $t \in \mathcal{T}$,

$\hat{a}$ Treatment given to an individual patient; either undergo colonoscopy, or do nothing. $\hat{a} \in \{dn, cl\}$,

L^{max} The CRC screening capacity limit,

$p_t(s_j|s_i, \hat{a}, o)$ The probability that an individual patient will be in core health state s_j in year $t + 1$ given that the patient is in core health state s_i, treatment $\hat{a} \in \{dn, cl\}$ is selected, and screening result $o \in \mathcal{O}$ is observed in year t, where $s_j \in \underline{s_i}$,

$q(s_i, \hat{a}, o, s_j)$ The expected reward (in QALYs) of individual patient for going from core health state s_i at time t to core health state $s_j, s_j \in \underline{s_i}, i \in \{0, \ldots, M\}$ at time $t + 1$ when treatment $\hat{a} \in \{dn, cl\}$ is taken and observation $o \in \mathcal{O}$ is seen,

$g_t(s_j|s_i, \hat{a})$ The probability that a patient will be in core health state $s_j \in \underline{s_i}$ in year $t + 1$ given that the patient is in core health state s_i and treatment $\hat{a} \in \{dn, cl\}$ is selected in year t,

$P_{u_{ij}(t)}(c\|a_i(t))$	The probability that the c individuals would move from core health state s_i at time t to core health state $s_j \in \underline{s_i}$ at time $t+1$ when action $a_i(t)$ is performed,
$r_t(c\|s_i, a_i(t))$	The immediate reward of transitioning c patients from core health state s_i to core health state s_j after action $a_i(t)$ is taken,
$P_t(X', X, a(t))$	The probability of going from system state X at time $t \in \mathcal{T}$ to system state s' at time $t+1$ when action $a(t) \in \mathcal{A}$ is taken,
$\hat{u}_{ij}(t)$	A realizations of $u_{ij}(t), s_j \in \underline{s_i}$,
$\hat{\mathbf{u}}(X', X)$	The vector of $\hat{u}_{ij}$ such that the transition from system state X at time t to health state X' at time $t+1$ is feasible,
$\hat{\mathbf{U}}(X', X)$	The set of all $\hat{\mathbf{u}}(X', X)$ vectors,
$V_t^*(X)$	The maximum expected TQALYs from for a system at state X in year t to year t^{max},
$r_{t^{max}}(X)$	Terminal reward,
λ	Discount factor.

3.2.2 Model Formulation

The model is formulated here by introducing the governing formulas of the MDP process.

The transition probability of the system is an important characteristic of the MDP model. The following is a discussion regarding expressing this probability in terms of known and estimated parameters. Two formulations are presented and discussed.

Define $p_t(s_j|s_i, \hat{a}, o)$ as the probability that a patient will be in health state s_j in year $t+1$ given that the patient is in health state s_i, treatment $\hat{a} \in \{dn, cl\}$ is selected, and screening result $o \in \{T-, P+, C+, SD\}$ is observed in year t, where $s_j \in \underline{s_i}$. Thus, $P_{u_{ij}(t)}(c|s_i, a_i(t))$ is defined as the probability that the c individuals would move from core health state s_i at time t to core health state $s_j, s_j \in \underline{s_i}$ at time $t+1$ when action $a_i(t)$ is performed, which is equivalent to:

$$
\begin{aligned}
&P_{u_{ij}(t)}(c|a_i(t)) = \Pr(u_{ij}(t) = c|a_i(t)) \\
&= \sum_{\hat{c}=0}^{c} \Bigg\{ \sum_{\{m_k\}\in\Omega_m} \left[\frac{a_i(t)X_i(t)!}{m_1!\dots m_K!} \left(\prod_{1\le k\le K} f_t(o_k|s_i, cl)^{m_k} \right) \sum_{\{c_k\}\in\Omega_c} \sum_{1\le k\le K} \binom{m_k}{c_k} p_t(s_j|s_i, cl, o_k)^{c_k} (1 - p_t(s_j|s_i, cl, o_k))^{m_k - c_k} \right] \\
&+ \sum_{\{m'_k\}\in\Omega'_m} \left[\frac{(1-a_i(t))X_i(t)!}{m'_1!\dots m'_K!} \left(\prod_{1\le k\le K} f_t(o_k|s_i, dn)^{m'_k} \right) \sum_{\{c'_k\}\in\Omega'_c} \sum_{1\le k\le K} \binom{m'_k}{c'_k} p_t(s_j|s_i, dn, o_k)^{c'_k} (1 - p_t(s_j|s_i, dn, o_k))^{m'_k - c'_k} \right] \Bigg\}
\end{aligned}
\tag{3.4}
$$

such that $\Omega_m := \big\{\{m_1, \dots, m_K\} : \sum_{1\le k\le K} m_k = a_i(t)X_i(t)\big\}$,
$\Omega'_m := \big\{\{m'_1, \dots, m'_K\} : \sum_{1\le k\le K} m'_k = (1 - a_i(t))X_i(t)\big\}$,
$\Omega_c := \big\{\{c_1, \dots, c_K\} : \sum_{1\le k\le K} c_k = \hat{c}\big\}$,
$\Omega'_c := \big\{\{c'_1, \dots, c'_K\} : \sum_{1\le k\le K} c'_k = c - \hat{c}\big\}$.

Equation (3.4) is mainly composed of two parts. The first line represents the probability that $\hat{c}$ individuals move from core health state s_i to core health state s_j as a result of undergoing colonoscopy ($\hat{a} = cl$). The number of individuals that are subjected to this treatment is $a_i(t)X_i(t)$. These individuals are divided into K groups such that m_k individuals would get observation o_k with probability $f_t(o_k|s_i, cl)$, where $1 \le k \le K$. Specifically, they are divided into $\{m_1, \dots, m_K\}$ such that $\sum_{1\le k\le K} m_k = a_i(t)X_i(t)$. The set Ω_m contains all feasible values of the vector $\{m_1, \dots, m_K\}$. This division process has a multinomial distribution. Now, of each m_k individuals, c_k individuals end up in core health state s_j with probability $p_t(s_j|s_i, cl, o_k)$, which has a binomial distribution. The values of $\{c_1, \dots, c_K\}$ must add up to $\hat{c}$. The set Ω_c contains all such vectors. The second line of the equation has a similar structure, but represents the probability that $c - \hat{c}$ individuals move from s_i to s_j as a result of $\hat{a} = dn$ (not doing a colonoscopy). Finally, the whole equation is summed over the possible values that $\hat{c}$ can take, which is $0 \le \hat{c} \le c$.

To simplify notation, let:

$$
\mathcal{L}_t(c_k, s_i, cl, o_k) = \binom{m_k}{c_k} p_t(s_j|s_i, cl, o_k)^{c_k} (1 - p_t(s_j|s_i, cl, o_k))^{m_k - c_k} \tag{3.5a}
$$

$$
\mathcal{L}'_t(c'_k, s_i, dn, o_k) = \binom{m'_k}{c'_k} p_t(s_j|s_i, dn, o_k)^{c'_k} (1 - p_t(s_j|s_i, dn, o_k))^{m'_k - c'_k} \tag{3.5b}
$$

As such, Equations (3.4) can be written as:

$$
\begin{aligned}
&P_{u_{ij}(t)}(c|a_i(t)) = \Pr(u_{ij}(t) = c|a_i(t)) \\
&= \sum_{\hat{c}=0}^{c} \Bigg\{ \sum_{\{m_k\}\in\Omega_m} \left[\frac{a_i(t)X_i(t)!}{m_1!\dots m_K!} \left(\prod_{1\le k\le K} f_t(o_k|s_i, cl)^{m_k} \right) \sum_{\{c_k\}\in\Omega_c} \sum_{1\le k\le K} \mathcal{L}_t(c_k, s_i, cl, o_k) \right] \\
&+ \sum_{\{m'_k\}\in\Omega'_m} \left[\frac{(1-a_i(t))X_i(t)!}{m'_1!\dots m'_K!} \left(\prod_{1\le k\le K} f_t(o_k|s_i, dn)^{m'_k} \right) \sum_{\{c'_k\}\in\Omega'_c} \sum_{1\le k\le K} \mathcal{L}'_t(c'_k, s_i, dn, o_k) \right] \Bigg\}
\end{aligned} \tag{3.6}
$$

such that $\Omega_m := \big\{\{m_1, \dots, m_K\} : \sum_{1\le k\le K} m_k = a_i(t)X_i(t)\big\}$,
$\Omega'_m := \big\{\{m'_1, \dots, m'_K\} : \sum_{1\le k\le K} m'_k = (1-a_i(t))X_i(t)\big\}$,
$\Omega_c := \big\{\{c_1, \dots, c_K\} : \sum_{1\le k\le K} c_k = \hat{c}\big\}$,
$\Omega'_c := \big\{\{c'_1, \dots, c'_K\} : \sum_{1\le k\le K} c'_k = c - \hat{c}\big\}$.

The reward of the system in each time period, is the sum of individual rewards of patients in that time period. The immediate reward is expressed in QALYs, which is defined as the difference between the total lifetime and total disutility of having undetected CRC, undergoing CRC screening, and undergoing CRC treatment. Define $q(s_i, \hat{a}, o, s_j)$ as the expected reward (in QALYs) of individual patient for going from core health state s_i at time t to core health state $s_j, s_j \in \underline{s_i}, i \in \{0, \dots, M\}$ at time $t+1$ when treatment $\hat{a} \in \{dn, cl\}$ is taken and observation $o \in \mathcal{O}$ is seen. The reward $r_t(c|s_i, a_i(t))$ is defined as the immediate reward of transitioning c patients from core health state s_i to core health state s_j after action $a_i(t)$ is taken. This is equivalent to:

$$
\begin{aligned}
&r_t(c|s_i, a_i(t)) = r(u_{ij}(t) = c|s_i, a_i(t)) \\
&= \sum_{\hat{c}=0}^{c} \Bigg\{ \sum_{\{m_k\}\in\Omega_m} \left[\frac{a_i(t)X_i(t)!}{m_1!\dots m_K!} \left(\prod_{1\le k\le K} f_t(o_k|s_i, cl)^{m_k} \right) \sum_{\{c_k\}\in\Omega_c} \sum_{1\le k\le K} \mathcal{L}_t(c_k, s_i, cl, o_k) q(s_i, cl, o_k, s_j) c_k \right] \\
&+ \sum_{\{m'_k\}\in\Omega'_m} \left[\frac{(1-a_i(t))X_i(t)!}{m'_1!\dots m'_K!} \left(\prod_{1\le k\le K} f_t(o_k|s_i, dn)^{m'_k} \right) \sum_{\{c'_k\}\in\Omega'_c} \sum_{1\le k\le K} \mathcal{L}'_t(c'_k, s_i, dn, o_k) q(s_i, dn, o_k, s_j) c'_k \right] \Bigg\}
\end{aligned} \tag{3.7}
$$

such that $\Omega_m := \big\{\{m_1, \dots, m_K\} : \sum_{1\le k\le K} m_k = a_i(t)X_i(t)\big\}$,
$\Omega'_m := \big\{\{m'_1, \dots, m'_K\} : \sum_{1\le k\le K} m'_k = (1-a_i(t))X_i(t)\big\}$,
$\Omega_c := \big\{\{c_1, \dots, c_K\} : \sum_{1\le k\le K} c_k = \hat{c}\big\}$,
$\Omega'_c := \big\{\{c'_1, \dots, c'_K\} : \sum_{1\le k\le K} c'_k = c - \hat{c}\big\}$.

Equation (3.7) has the same structure as Equation (3.6). The expected QALYs for an individual patient $q(s_i, \hat{a}, o_k, s_j)$ is multiplied by the number of patients c_k having the action $\hat{a}$, and seeing the same observation, o_k.

For the system to transition from system state X at time t to system state X' at time $t+1$, the values of $X_i(t)$ and $X_i(t+1), i \in \{0, \ldots, M\}$ must be known. Define a vector $\hat{\mathbf{u}}(X', X)$ as a vector of $\hat{u}_{ij}(t)$ which makes the transition from X at time t to X' at time $t+1$ feasible for any $t \in \mathcal{T}$. Recall that $\hat{u}_{ij}(t)$ is a realization of $u_{ij}(t)$. The set that contains all $\hat{\mathbf{u}}(X', X)$ vectors is $\hat{\mathbf{U}}(X', X)$.

Now, define $P_t(X', X, a(t))$ as the probability of going from state X at time $t \in \mathcal{T}$ to state X' at time $t+1$ when action $a(t) \in \mathcal{A}$ is taken. We have:

$$P_t(X', X, a(t)) = \sum_{\hat{\mathbf{u}}(X',X) \in \hat{\mathbf{U}}(X',X)} \left[\prod_{c \in \hat{\mathbf{u}}(X',X)} P_{u_{ij}(t)}(c|a_i(t)) \right] \tag{3.8}$$

The inner brackets of Equation (3.8) contain the multiplication of one realization array $\hat{\mathbf{u}}(X', X)$ that makes the transition from X to X' feasible. Since there are many possible feasible paths to go from X to X', the products are summed over all members of the set $\hat{\mathbf{U}}(X', X)$.

If it is impossible to transition from state X to state X' (i.e. there are no combination values of $u_{ij}(t), s_j \in \underline{s_i}, i \in \{0, \ldots, M\}$ that can make the transition valid), then the set $\hat{\mathbf{U}}(X', X)$ is assumed to be empty, and the value of $P_t(X', X, a(t))$ is given a value of zero.

The total reward of the system which is at state X at time t after performing action $a(t) \in \mathcal{A}$, denoted by $r_t(X, a(t))$, can be found by considering Equations (3.6) and (3.7), as follows.

$$r_t(X, a(t)) = \sum_{\hat{\mathbf{u}}(X',X) \in \hat{\mathbf{U}}(X',X)} \left[\prod_{c \in \hat{\mathbf{u}}(X',X)} P_{u_{ij}(t)}(c|a_i(t)) r_t(c|s_i, a_i(t)) \right] \tag{3.9}$$

The previous equations are complex and can be troublesome, especially Equations (3.6) and (3.7). Although this form is required to analyze some aspects of the model (e.g., when

allocation decision depends on observation like SD), a less complicated form can also be beneficial, which will be the focus for the remaining of this section.

Define $g_t(s_j|s_i, \hat{a})$ as the probability that a patient will be in core health state $s_j \in \underline{s_i}$ in year $t+1$ given that the patient is in core health state s_i and treatment $\hat{a} \in \{dn, cl\}$ is selected in year t. That is,

$$g_t(s_j|s_i, \hat{a}) = \sum_{o \in \mathcal{O}} p_t(s_j|s_i, \hat{a}, o) f_t(o|s_i, \hat{a}) \quad \forall s_i, i \in \{0, \ldots, M\}, s_j \in \underline{s_i}, \hat{a} \in \{dn, cl\}, \text{ and } t < t^{max} \tag{3.10}$$

Equation (3.10) is basically a weighted average of probabilities of going from core health state s_i to core health state s_j given treatment $\hat{a}$, weighted to the probability of seeing observation $o \in \mathcal{O}$. Consequently, the following holds.

$$\begin{aligned} &P_{u_{ij}(t)}(c|s_i, a_i(t)) = \Pr(u_{ij}(t) = c|s_i, a_i(t)) \\ &= \sum_{\hat{c}=0}^{c} \left\{ \binom{a_i(t)X_i}{\hat{c}} g_t(s_j|s_i, cl)^{\hat{c}} (1 - g_t(s_j|s_i, cl)^{a_i(t)X_i - \hat{c}} \right. \\ &\left. + \binom{(1-a_i(t))X_i}{c-\hat{c}} g_t(s_j|s_i, dn)^{c-\hat{c}} (1 - g_t(s_j|s_i, dn))^{(1-a_i(t))X_i - c + \hat{c}} \right\} \end{aligned} \tag{3.11}$$

Equation (3.11) is composed of two main parts. The first represents the probability of having $\hat{c}$ individuals move from core health state s_i to core health state s_j as a result of undergoing colonoscopy, $\hat{a} = cl$. This has a binomial distribution with probability of success equal to $g_t(s_j|s_i, cl)$. The second part is explained similarly, but $\hat{a} = dn$, and the number of individuals to transfer in this case is $c - \hat{c}$. The summation at the beginning of the equation is to account for the fact that $\hat{c}$ can have any integer values between 0 and c.

The reward of the system state X at time t and action $a(t) \in \mathcal{A}$ is:

$$r_t(X, a(t)) = \sum_{i \in \{0,\ldots,M\}} [a_i(t) X_i q_t(s_i, cl) + (1 - a_i(t)) X_i q_t(s_i, dn)] \tag{3.12}$$

We end this section by giving a formula for the objective function. The objective is to maximize the expected TQALYs, which is the sum of the expected immediate rewards. Define $V_t^*(X)$ as the maximum expected TQALYs from year t to year t^{max}. Also, define $r_{t^{max}}(X)$ as the terminal reward, which is the QALYs after screening program is terminated, t^{max}. We have:

$$V_t^*(X) = \max_{a(t)\in\mathcal{A};(3.3a)} r_t(X, a(t)) + \lambda \sum_{X'\in\mathcal{X}} P_t(X', X, a(t)) V_{t+1}^*(X) \tag{3.13}$$

where λ is a discount factor. Equation 3.13 is a sum of immediate reward, and the discounted expected value of future rewards.

3.2.3 State Aggregation

For a population of size N, the transition probability matrix of the Markov process $\{X_0(t), \ldots, X_M(t) : t \in \mathcal{T}\}$ can be of size $\binom{N+M}{M} \times \binom{N+M}{M}$, which grows substantially with larger values of M and N. Therefore, it would be hard to solve this model for moderately large populations on personal computers.

One way to address this issue is by state aggregation, which would reduce the size of the state space. This can be done through aggregating a number of states into one. This can reduce the state space by a factor depending on the size of each state group.

Yaesoubi and Cohen (2011) propose an approach in which the state space $\{X_0(t), \ldots, X_M(t) : t \in \mathcal{T}\}$ can be represented by $\{\Theta_0(t), \ldots, \Theta_M(t) : t \in \mathcal{T}\}$, where $\Theta_i(t), i \in \{0, \ldots, M\}$ is the proportion of the population in core health state s_i at time t, and can only take a limited number of values from the set $\{\theta_i^1, \theta_i^2, \ldots, \theta_i^{d_i}\}$, where d_i represent the number of distinct possible values that $\Theta_i(t)$ can take.

To determine the set $\{\theta_i^1, \theta_i^2, \ldots, \theta_i^{d_i}\}$, define the points $\{b_i^1, b_i^2, \ldots, b_i^{d_i}\}$, such that $b_i^1 = 0, b_i^{d_i} = 1$, and $b_i^1 < b_i^2 < \ldots < b_i^{d_i}$. These points divide the interval $[0, 1]$ into d_i regions. Consequently, a possible value of $\Theta_i(t)$, say item j in the set $\{\theta_i^1, \theta_i^2, \ldots, \theta_i^{d_i}\}$, can be determined by the formula. $\theta_i^j = \frac{b_i^{j-1} + b_i^j}{2}$, for $j \in \{1, \ldots, d_i\}$. Then, the transition probability for $\{\Theta_0(t), \ldots, \Theta_M(t) : t \in \mathcal{T}\}$ can be calculated for $\{X_0(t), \ldots, X_M(t)\} = \{\lfloor N\Theta_0 \rfloor, \ldots, \lfloor N\Theta_M \rfloor\}$.

Another method of aggregating states is the fixed-wight aggregation technique (Heyman and Sobel, 1984). This method is used by Higginson and Bookbinder (1995) to simplify their MDP model for shipment consolidation. This method sets a batch size b, then groups the states of the original model to create a new, smaller state space. In this technique, the new larger state is formed by grouping and giving each original state (commonly equal) weights within that group. In particular, let α and β represent sets of original unaggregated states, and k and m represent states in the new aggregated model. Also, define a normalizing constant $\omega_\alpha \geq 0$, such that $\sum_{\alpha \in k} \omega_\alpha = 1$. Then,

$$g'_t(k, m, \hat{a}) = \sum_{\alpha \in k} \left[\omega_\alpha \sum_{\beta \in m} g_t(\alpha, \beta, \hat{a}) \right]$$

Then the system probabilities and reward can be calculated accordingly. It is common to have equal weights for the states within the larger states, that is, $\omega_\alpha = \frac{1}{b}$. This assumes that all original states are equally likely to happen within each new aggregated state. This assumption may not always be justified, which represents a disadvantage of this method.

A more advanced method of reducing the size of a state space is the grid-based approximation technique. Lovejoy (1991) uses a uniform grid consisting of a finite subset of points from the state space that do not change throughout an iterative procedure. This method provides an efficient interpolation, although the grid exponentially grows as the number of core health states increases. Moreover, Hauskrecht (1997) uses a nonuniform grid method that starts from an arbitrary set of points and enhances the grid at each iteration by including more points according to various heuristics. This method has the advantage of efficient use of computation power, whereas its disadvantage is its use of heuristic rules throughout the process. Sandikci et al. (2013) exploit the structure of their model by building a finite subset of plausible (feasible) states, and assign a belief distribution over the subset. Also, they select states such that all of the nonzero values in a selected state vector are positive integer multiples of $\frac{1}{q}$, where q is a positive integer representing the grid resolution.

The structure of the MDP model can be exploited to eliminate the system states that are unlikely to occur. For example, a patient who leaves the low-risk level cannot go back. The same applies for high-risk level. Moreover, a patient in low-risk level with polyp cannot

transfer to low-risk without lesion, and so on. Given these special characteristics of the model, the state space reduces significantly.

Furthermore, by limiting computations on the states with higher chances to occur, the method of Sandikci et al. (2013) provides a significant computational advantage. In their case, the state space is reduced by a factor as high as one million.

3.2.4 Approximate Dynamic Programming

Given the intractability of the MDP model, the Approximate Dynamic Programming (ADP) approach can be used to solve the model. The idea of this method is to approximate the $V_t^*(X)$ function using linear functions and continuously update the approximate function based on the new information generated by the X vector at each iterations. Interested readers are referred to Powell and Topaloglu (2006) for general structure of the ADP.

The use of ADP is possible in theory to solve our MDP model. However, since the state space is prohibitively large, the performance and quality of the ADP approach might suffer. Further research can be done on this area to investigate the effectiveness of such an approach.

3.2.5 Complexity of the Model

As mentioned above, for a population of size N, the transition probability matrix of the Markov process $\{X_0(t), \ldots, X_M(t) : t \in \mathbf{T}\}$ is of size $\binom{N+M}{M} \times \binom{N+M}{M}$, which grows substantially with larger values of M and N. Table 3.8 shows sample values of N and the resulting size of the transition matrix. As can be seen, the size of the probability transition matrix is prohibitively large for even moderately large population size N. Therefore, it would be hard to solve this model on personal computers.

An alternative approach to model and solve this problem is to use the fact that with large populations, the stochasticity of the system can be approximated by a deterministic model. When considering a collection of stochastic events that share the main features, the detailed differences play insignificant role compared to the overall trend of the system.

Table 3.8: The Size of Probability Transition Matrix for Different Population Sizes N

N	M	$\binom{N+M}{M}$
10	10	184,756
50	10	75,394,027,566
100	10	46,897,636,623,981
1,000	10	291,098,519,807,782,000,000,000

Specifically, for patients of the same gender, age, and risk level, the trend of their behavior can be approximated by a single flow. Although it is known that their individual behavior can be different, the main features of the system are preserved, and therefore, it is possible to make observations and recommendations on the overall system.

This is the idea of the next section. The deterministic optimization model developed in Section 3.3 will be an approximation of the MDP model discussed in this section. The loss of information and insights due to the usage of average values is justified by the less complexity and the possibility of solving and analyzing. This approach is used in many application in healthcare, including liver transplant (Akan et al., 2012).

Given that only very small values of N ($N \leq 10$) can be solved in the current MDP model form, solution generated would have insignificant insights. Also, as will be later discussed, the next section adds more factors into consideration. Therefore, no numerical results are discussed for the MDP model.

3.3 Mixed Integer Program Model

This section aims at building a mixed integer program to model the population dynamics and resource allocation for CRC screening policies. Similar to Section 3.2, the objective of the decision maker is to maximize a social welfare measure, which is QALYs for the population. Patients are categorized according to the risk level into: low-risk, high-risk, and post-CRC, which are completely observable by the decision maker. Within each risk level, the patients are divided into three unobservable clusters depending on cancer progression:

no lesions, (having) polyp, and (having) CRC. Once again, colonoscopy is considered as the screening method in this method as well. The policy maker decides on the number (or percentage) of patients to undergo colonoscopy with each level. Alternatively, the decision maker would decides which patient group (classification of patients will be discussed later) to undergo colonoscopy and which will not.

Adding the UCT and D, the total number of core health states is 11. Actions are defined as whether or not patients in each risk level undergo colonoscopy at a given year. An extension to the model would be by considering actions as percentages of individuals in each group to undergo colonoscopy. This can have the action variables as either integer or continuous, depending on the level of granularity required. A special case is to have binary actions. In other words, the same action will be performed for all members of a given group of patients. It is assumed in this model that the population starts at age 50, and colonoscopy is no longer performed after age 75.

3.3.1 Model Description

In this section, the deterministic model is formally introduced. The model aims at finding optimal screening policies for a representative population whose individuals are subject the disease progression pattern described in Section 3.1.4. The major consideration for the policy maker is the limited CRC screening resources available. The decision maker must allocate scarce screening resources such that the total QALYs (or any other measure of social welfare) is maximized.

While this model shares a lot of features with the one developed in Section 3.2, many variable definitions and formulas are distinctively different. A full description is given below. Please note that some variable definitions might overlap with the ones in Section 3.2, while others might be different from them.

In this model, the disease progression for an individual patient follows the pattern described in Section 3.1.4 and depicted in Figure 3.1. Specifically, $i \in \mathcal{I}$ represent the disease progression state. This model has 11 disease progression states, as explained in Section 3.1.4. There are five risk levels, $R \in \mathcal{R}$, in this model. Table 3.9 shows the risk levels and the disease progression states associated with each.

Table 3.9: Risk Levels and Associated Disease Progression States

Risk Level	Associated Disease Progression State
0	$\{0, 1, 2\}$
1	$\{3, 4, 5\}$
2	$\{6, 7, 8\}$
3	$\{9\}$
4	$\{10\}$

Let j represent an age group, where $j \in \mathcal{J}$. Since it is assumed in this model that screening starts at age 50 for both genders, the age groups start at 50 and end at age 75. Age group's length can range from one year up to several years. Shorter group's length would mean more accurate representation of the population because it is more granular. However, adopting short group length poses computational challenges to the model. Moreover, age groups of length of more than 10 years is not recommended due to the possible loss of accuracy. Almost all guidelines suggest initiating CRC screenings at or after age 50 (Levin et al., 2008; Winawer et al., 2003). The maximum age (the year at the end of the last age group $j \in \mathcal{J}$ is the age at which the consideration for colonoscopy stops. Normally, this is the age after which colonoscopy offer little to no value. Studies on the termination of CRC screening suggest to stop screening at ages 75-85 (Maheshwari et al., 2008; Zauber et al., 2008), and the USPSTF only recommends routine screening for CRC up to age 75 (US Preventive Services Task Force, 2008).

The population is also divided based on gender, $k \in \mathcal{K}$. The fourth factor in classifying individuals in the population is the personal history. Screening tests with negative results (i.e., no abnormal lesions found) do not change the disease progression state of a low risk patient. An index for personal history, $h \in \mathcal{H}$, is introduced to capture the dynamics that might happen to patients who go through this experience. A discussion on how the value of h changes will be presented later. Having an index for personal history allows, among others, for tracking the disutility of repetitive screening tests, and personalize compliance rates. It can be argued that a patient is less motivated to undergo, say, a third screening test if the first two resulted in finding no abnormalities (the author is yet to found evidence

of such behavior). Therefore, let $s^i_{j,k,h}$ represent the state of the system having disease progression i, age group j, gender k, and history h. The number of individuals in this state at time t is denoted by $X^i_{j,k,h}(t)$.

In this model, it is assumed that the initial population equals N_0. Thus, the following holds.

$$\sum_{i,j,k,h} X^i_{j,k,h}(0) = N_0 \tag{3.14}$$

Time periods are in years; $t \in \mathcal{T} = \{0, 1, \ldots, t^{max}\}$, which represents time epochs for which the model is run. The higher t^{max}, the more aging dynamics is allowed, which would capture more details in the population transformations.

It will be assumed that $h \in \{0, 1, 2\}$. Recall that h is the index of colonoscopy history indicating how many test negatives ($o = T$-) the patient (or group of patients) has received. If a polyp or cancer has been detected and removed, h is reset to zero. If the patient tests negative twice successively, h becomes 2, but further negative tests would not change the value of h.

The actions of the this model represent whether or not the individuals at each core health state undergo a colonoscopy at time t. However, disease progression is unobservable within each risk level. For example, low-risk patients are indistinguishable to the policy maker, and are stochastically distributed into: without lesion, polyp, and CRC levels within the same risk-level. Therefore, the actions are denoted by $a^R_{j,k,h}(t)$ omitting the index for the disease progression within a certain risk level. Thus, the vector a_t is the vector of all actions at time t, or $a_t = \left\{a^R_{j,k,h}(t)\right\}\ \forall\ R, j, k$, and h. The vector a_t is a member of the set $\mathcal{A}$, which represents the set of all action vectors (i.e., all possible actions).

Requiring a patient, or group of patients, to undergo screening does not guarantee that they will comply. Only a proportion will comply, and actually undergo colonoscopy. This is called compliance rate (denoted by $\theta^R_{j,k,h}$). Here again, the index i is omitted. The subgroup of individuals that actually undergo colonoscopy are said to be subjected to treatment $\hat{a} = \{cl\}$, while those individuals who do not undergo colonoscopy (either not scheduled or not compliant with regulations) are subjected to treatment $\hat{a} = \{dn\}$.

As a result of treatment $\hat{a}$, the individual faces an observation. There are four observations available in this model, $o \in \{T\text{-}, P+, C+, SD\}$. These observations are: test negative T-, polyp found $P+$, cancer found $C+$, and self-diagnosed SD. Depending on the action and disease progression (see Section 3.1.4), some of these observations may not be possible.

A list of the variables used in this model are shown in Table 3.10.

Table 3.10: Variable Description for MIP Model

Variable	Description
$\mathcal{I}$	The set of disease progression states, indexed by i,
$\mathcal{J}$	The set of age groups, indexed by j,
$\mathcal{K}$	The set of genders, indexed by k,
$\mathcal{R}$	The set of risk levels, indexed by R, $\mathcal{R} := \{LR, HR, PC, UCT, D\}$,
$\mathcal{O}$	The set of observations, indexed by o, $\mathcal{O} := \{T\text{-}, P+, C+, SD\}$,
$\mathcal{A}$	The set of all action vectors a_t,
$\mathcal{H}$	The set of all history states, indexed by h,
$\mathcal{T}$	The set of time periods, indexed by t, $\mathcal{T} = \{0, 1, \ldots, t^{max}\}$, where t^{max} represents the last time epoch for which the model is run,
$\mathcal{L}$	The set of possible policies,
$s^i_{j,k,h}$	The health state defined by disease progression stage $i \in \mathcal{I}$, age group $j \in \mathcal{J}$, gender $k \in \mathcal{K}$, and disease history $h \in \mathcal{H}$,
$X^i_{j,k,h}(t)$	The number of people in the health state defined at time t by disease progression stage $i \in \mathcal{I}$, age group $j \in \mathcal{J}$, gender $k \in \mathcal{K}$, and disease history $h \in \mathcal{H}$,
$\widetilde{X}^i_{j,k,h}(t)$	The adjusted $X^i_{j,k,h}(t)$ after aging,
$a^R_{j,k,h}(t)$	The action of individuals in state $s^i_{j,k,h}$ to either undergo colonoscopy or not at time t. Since disease progression is unobservable within each risk level, the index i does not appear here,
a_t	The vector of all actions $a^R_{j,k,h}(t)$ at time t, $a_t = \left\{a^R_{j,k,h}(t)\right\}\ \forall\ R, j, k,$ and h. Thus, $a_t \in \mathcal{A}$,
$\hat{a}$	The type of treatment a subgroup is subjected to, $\hat{a} \in \{cl, dn\}$,

$f\left(o\|s^i_{j,k,h},\hat{a}\right)$	The rate of observing observation $o \in \mathcal{O}$ at time t when action $\hat{a} \in \{cl, dn\}$ is taken on state $s^i_{j,k,h}$,
$p\left(s^{i'}_{j',k',h'}\|s^i_{j,k,h},\hat{a},o\right)$	The rate at which individuals will be in state $s^{i'}_{i',j',k'}$ given that they are in state $s^i_{j,k,h}$, action $\hat{a} \in \{cl, dn\}$ is taken, and screening result o is observed,
$q\left(s^{i'}_{j',k',h'}, s^i_{j,k,h},\hat{a},o\right)$	Immediate rewards (in expected QALYs) for all individuals going from state $s^i_{j,k,h}$ to state $s^{i'}_{j',k',h'}$ given action $\hat{a}$ and screening result o is observed,
$q_{t^{max}}(s^i_{j,k,h})$	Terminal reward for individuals in state $s^i_{j,k,h}$ at the last time period, t^{max},
L^{max}	The capacity limit for colonoscopy resource available.
υ_j	Rate of aging to age group j from an immediate predecessor age group $\underline{j}$,
$\theta^R_{j,k,h}$	Compliance rate for age individuals in states $s^i_{j,k,h}$ $\forall i$,
τ_P	Sensitivity of colonoscopy to polyps,
τ_C	Sensitivity of colonoscopy to cancer,
$\omega_{\hat{a}}$	Probability of CRC self-detection given action $\hat{a}$,
$\rho^{i,i'}_j$	Lesion progression rate from states i to state i',
$\delta^{i,j}_{\hat{a},o}$	Rate of mortality in state i given treatment $\hat{a}$ and observation o.
$\gamma_{i,j}$	Rate of completion within year t of treatment initiated at state i,
$\underline{h}$	An immediate predecessor of h (e.g., if $h = 1$ then $\underline{h} = 0$),
$\underline{j}$	An immediate predecessor of j (e.g., if $j = 1$ then $\underline{j} = 0$),
h_0	1 if $h = 0$, 0 otherwise,
λ_t	Discount factor in year t,
$q^i_{J,h,k}$	Terminal reward (QALYs after age J) for state $s^i_{j,k,h}$,
d_C, d_{CT}, d_{UCT}	Disutility of undetected CRC, CRC treatment, and being in the UCT state,
$d_{poly}(cl)$	Disutility of undergoing colonoscopy with polypectomy,
$d_{\neg poly}(cl)$	Disutility of undergoing colonoscopy without polypectomy,
$\kappa^{i,j}_{\hat{a},o}$	Probability of immediate mortality from screening complications,
κ^j_{UCT}	Probability of immediate mortality from treatment at age j.

m_ℓ Binary variable set to one if policy $\ell \in \mathcal{L}$ is selected, zero otherwise.

3.3.2 Model Formulation

The system state transition rate function, $g\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h}, a^{R}_{j,k,h}(t)\right)$, is defined as

$$\begin{aligned} g\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h}, a^{R}_{j,k,h}(t)\right) =& \theta^{R}_{j,k,h} a^{R}_{j,k,h}(t) \sum_{o\in\mathcal{O}} p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h}, cl, o\right) f\left(o|s^{i}_{j,k,h}, cl\right) \\ &+ \left[1 - \theta^{R}_{j,k,h} a^{R}_{j,k,h}(t)\right] \sum_{o\in\mathcal{O}} p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h}, dn, o\right) f\left(o|s^{i}_{j,k,h}, dn\right) \end{aligned} \tag{3.15}$$

The first term of equation (3.15) represents the rate of transition from state $s^{i}_{j,k,h}$ to state $s^{i'}_{j',k',h'}$ as a result of conducting colonoscopy. This rate is only applied to the proportion of individuals in the state who are compliant with the guidelines. The second term represents those who are not supposed to undergo colonoscopy, as well as non-compliant individuals who are supposed to undergo colonoscopy but choose not to.

Detailed formulations of $p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h}, \hat{a}, o\right)$, $f\left(o|s^{i}_{j,k,h}, \hat{a}\right)$, and $q\left(s^{i'}_{j',k',h'}, s^{i}_{j,k,h}, \hat{a}\right)$ (see below) can be found in Appendix C.

Equation (3.15) can be written as

$$\begin{aligned} g\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h}, a^{R}_{j,k,h}(t)\right) =& a^{R}_{j,k,h}(t)\theta^{R}_{j,k,h} \left[\sum_{o\in\mathcal{O}} p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h}, cl, o\right) f\left(o|s^{i}_{j,k,h}, cl\right) - \sum_{o\in\mathcal{O}} p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h}, dn, o\right) f\left(o|s^{i}_{j,k,h}, dn\right)\right] \\ &+ \sum_{o\in\mathcal{O}} p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h}, dn, o\right) f\left(o|s^{i}_{j,k,h}, dn\right) \end{aligned} \tag{3.16}$$

Let $\hat{g}(s'|s,\hat{a}) = \hat{g}\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h}, \hat{a}\right) = \sum_{o\in\mathcal{O}} p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h}, \hat{a}, o\right) f\left(o|s^{i}_{j,k,h}, \hat{a}\right)$. Therefore, we can write equations (3.15) as

$$g\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h}, a^{R}_{j,k,h}\right) = a^{R}_{j,k,h}\theta^{R}_{j,k,h}\left[\hat{g}(s'|s, cl) - \hat{g}(s'|s, dn)\right] + \hat{g}(s'|s, dn) \tag{3.17}$$

Formula (3.17) will be used for its brevity.

Aging Constraints

It is assumed that aging happens before other transitions in the population. Consequently, actions will be applied to the *adjusted* number of individuals in each system state $\widetilde{X}^i_{j,k,h}(t)$, which is defined as

$$\widetilde{X}^i_{j,k,h}(t) = (1 - \upsilon_j)X^i_{j,k,h}(t) + (\upsilon_{\underline{j}})X^i_{\underline{j},k,h}(t) \qquad , j > 0 \tag{3.18}$$

where $\underline{j}$ represents the age group immediately preceding age group j. Equation (3.18) states that the number of individuals in a particular state after the event of aging is the sum of two components. The first is the proportion of individuals who did not proceed to the next age group. In other words, their new age is still within the range of the current age group. The other component comes from the individuals who left their (younger) age group to join the current, if applicable.

Also, it is assumed that the number of individuals in the youngest age group ($j = 0$) remains constant over time. This stems from the assumption that the number of individuals leaving this youngest age group is equal to the number of people joining (i.e., becoming part of the target population). This assumption is expressed mathematically as follows. For all i, k, h, and t, the following set of constraints hold.

$$\widetilde{X}^i_{0,k,h}(t) = X^i_{0,k,h}(t) \tag{3.19}$$

Transition Function

It is now possible to write the main transition function in this model. The following represents the formula governing the transition of individuals from any system state at time t to the state $X^i_{j,k,h}(t + 1)$ at time $t + 1$. For a given action set $a_t = \{a^R_{j,k,h}\}$, the transition equations have the form

$$X^i_{j,k,h}(t+1) = \left(1 - \sum_{i',j',k',R',h'} g\left(s^{i'}_{j',k',h'} | s^i_{j,k,h}, a^R_{j,k,h}(t)\right)\right) \widetilde{X}^i_{j,k,h}(t) + \sum_{i'',j'',k'',R'',h''} \left[g\left(s^i_{j,k,h} | s^{i''}_{j'',k'',h''}, a^R_{j,k,h}(t)\right) \widetilde{X}^{i''}_{j'',k'',h''}(t)\right] \tag{3.20}$$

where $s^{i'}_{j',k',h'}$ are the states that $s^i_{j,k,h}$ leads to, and $s^{i''}_{j'',k'',h''}$ are the states that lead to $s^i_{j,k,h}$.

The first line of the right hand side of equation (3.20) represents the number individuals who were in state $X^i_{j,k,h}$ at time t and did not leave to another state at time $t+1$. The second line represent all individuals who transferred to state $X^i_{j,k,h}(t+1)$ at time $t+1$ from any other state at time t.

To provide visual clarity, whenever possible, the following replacements will be made. s will be used instead of $s^i_{j,k,h}$, s' will be used instead of $s^{i'}_{j',k',h'}$, and s'' will be used instead of $s^{i''}_{j'',k'',h''}$.

Equation (3.20) can be simplified as follows.

$$X(t+1) = \left(1 - \sum_{s'} g\left(s' | s, a(t)\right)\right) \widetilde{X}(t) + \sum_{s''} \left[g\left(s | s'', a(t)\right) \widetilde{X}''(t)\right] \tag{3.21}$$

Substituting the value of g function of formula (3.17) into (3.21) gives

$$X(t+1) = \left(1 - \sum_{s'} \hat{g}\left(s' | s, dn\right)\right) \widetilde{X}(t) - \sum_{s'} \left[\hat{g}\left(s' | s, cl\right) - \hat{g}\left(s' | s, dn\right)\right] \theta a(t) \widetilde{X}(t) + \sum_{s''} \hat{g}(s | s'', dn) \widetilde{X}''(t) + \sum_{s''} \left[\hat{g}\left(s | s'', cl\right) - \hat{g}\left(s | s'', dn\right)\right] \theta'' a''(t) \widetilde{X}''(t) \tag{3.22}$$

In the previous equation, the state s can be excluded from all summations since they would add up to zero anyway. This gives

$$X(t+1) = \left(1 - \sum_{s' \setminus s} \hat{g}\left(s'|s, dn\right)\right) \widetilde{X}(t) - \sum_{s' \setminus s} \left[\hat{g}\left(s'|s, cl\right) - \hat{g}\left(s'|s, dn\right)\right] \theta a(t) \widetilde{X}(t) + \sum_{s'' \setminus s} \hat{g}(s|s'', dn) \widetilde{X}''(t) + \sum_{s'' \setminus s} \left[\hat{g}\left(s|s'', cl\right) - \hat{g}\left(s|s'', dn\right)\right] \theta'' a''(t) \widetilde{X}''(t) \tag{3.23}$$

The explicit formulations of the transition equations are shown in Appendix B.

Capacity Constraints

The capacity constraints can be expressed as

$$\sum_{R,i,j,k,h} a^R_{j,k,h}(t) X^i_{j,k,h}(t) \leq L^{max}, \quad \forall\, t \tag{3.24}$$

where L^{max} is the maximum capacity of screening resources for the system.

Note that the capacity is calculated based on the scheduled colonoscopies, not the actual.

Reward

The reward of the system at each time period, is the sum of individual rewards of patients at that time period. The immediate reward is expressed in quality-adjusted life years (QALYs).

The reward of the system is defined in the following manner. Define $q\left(s^{i'}_{j',k',h'}, s^i_{j,k,h}, \hat{a}\right)$ as the reward of the system resulted from going from state $s^i_{j,k,h}$ at time t to state $s^{i'}_{j',k',h'}$ at time $t+1$ when treatment $\hat{a}$ is applied. $q\left(s^{i'}_{j',k',h'}, s^i_{j,k,h}, \hat{a}\right)$ is defined as follows.

$$q\left(s^{i'}_{j',k',h'}, s^i_{j,k,h}, \hat{a}\right) = \sum_{o \in \mathcal{O}} p\left(s^{i'}_{j',k',h'} | s^i_{j,k,h}, \hat{a}, o\right) f\left(o|s^i_{j,k,h}, \hat{a}\right) q\left(s^{i'}_{j',k',h'}, s^i_{j,k,h}, \hat{a}, o\right) \tag{3.25}$$

Similarly, define $q\left(s^i_{j,k,h}, \hat{a}\right)$ as the total reward achieved from going to any state at time $t+1$ given the (group of) patients are in state $s^i_{j,k,h}$ at time t and treatment $\hat{a}$ is applied.

$$q\left(s^i_{j,k,h}, \hat{a}\right) = \sum_{j',k',h',i'} \sum_{o \in \mathcal{O}} p\left(s^{i'}_{j',k',h'} | s^i_{j,k,h}, \hat{a}, o\right) f\left(o | s^i_{j,k,h}, \hat{a}\right) q\left(s^{i'}_{j',k',h'}, s^i_{j,k,h}, \hat{a}, o\right) \tag{3.26}$$

It is now possible to define $r\left(s^i_{j,k,h}, a^R_{j,k,h}(t)\right)$, which is the total reward for all individuals who are in state $s^i_{j,k,h}$ at time t when action $a^R_{j,k,h}(t)$ is applied.

$$\begin{aligned} r\left(s^i_{j,k,h}, a^R_{j,k,h}\right) =& \theta^R_{j,k,h} a^R_{j,k,h} \sum_{j',k',h',i'} \sum_{o \in \mathcal{O}} p\left(s^{i'}_{j',k',h'} | s^i_{j,k,h}, cl, o\right) f\left(o | s^i_{j,k,h}, cl\right) q\left(s^{i'}_{j',k',h'}, s^i_{j,k,h}, cl, o\right) \\ &+ \left[1 - \theta^R_{j,k,h} a^R_{j,k,h}\right] \sum_{j',k',h',i'} \sum_{o \in \mathcal{O}} p\left(s^{i'}_{j',k',h'} | s^i_{j,k,h}, dn, o\right) f\left(o | s^i_{j,k,h}, dn\right) q\left(s^{i'}_{j',k',h'}, s^i_{j,k,h}, dn, o\right) \end{aligned} \tag{3.27}$$

The first line of (3.27) represents the reward achieved from individual compliant with the policy. Since the compliance rate $\theta^R_{j,k,h}$ can be less than one (not all patients who are instructed to undergo screening will actually do it), the second line of the formula accounts for non-compliant individuals.

Let $\hat{r}(s^i_{j,k,h}, \hat{a}) = \sum_{j',k',h',i'} \sum_{o \in \mathcal{O}} p\left(s^{i'}_{j',k',h'} | s^i_{j,k,h}, \hat{a}, o\right) f\left(o | s^i_{j,k,h}, \hat{a}\right) q\left(s^{i'}_{j',k',h'}, s^i_{j,k,h}, \hat{a}, o\right)$

Then, equation (3.27) can be simplified as follows.

$$r\left(s^i_{j,k,h}, a^R_{j,k,h}(t)\right) = \theta^R_{j,k,h} a^R_{j,k,h}(t) \hat{r}(s^i_{j,k,h}, cl) + \left[1 - \theta^R_{j,k,h} a^R_{j,k,h}(t)\right] \hat{r}(s^i_{j,k,h}, dn) \tag{3.28}$$

which can be re-written as

$$r\left(s^i_{j,k,h}, a^R_{j,k,h}(t)\right) = \left[\hat{r}(s^i_{j,k,h}, cl) - \hat{r}(s^i_{j,k,h}, dn)\right] \theta^R_{j,k,h} a^R_{j,k,h}(t) + \hat{r}(s^i_{j,k,h}, dn) \tag{3.29}$$

Objective function

The model aims at maximizing the QALYs of the whole population. The objective function is expressed as follows.

$$\max_{a} \sum_{i,j,k,h,t} r\left(s^i_{j,k,h}, a^R_{j,k,h}(t)\right) X^i_{j,k,h}(t) + \sum_{i,j,k,h} q_{t^{max}}(s^i_{j,k,h}) X^i_{j,k,h}(t^{max}) \tag{3.30}$$

The first term represents the reward achieved as a result of actions across all time periods for all states. The second term represent the terminal rewards, which are the estimated remaining QALYs from this point onwards. This is applied at the final time period t^{max}.

Rewriting the formula above gives

$$\begin{aligned}\max_{a} \sum_{i,j,k,h,t} & \left(\left[\hat{r}(s^i_{j,k,h}, cl) - \hat{r}(s^i_{j,k,h}, dn)\right] \theta^R_{j,k,h} a^R_{j,k,h}(t) X^i_{j,k,h}(t) + \hat{r}(s^i_{j,k,h}, dn) X^i_{j,k,h}(t)\right) \\ & + \sum_{i,j,k,h} q_{t^{max}}(s^i_{j,k,h}) X^i_{j,k,h}(t^{max})\end{aligned} \tag{3.31}$$

3.3.3 Full Problem

The full program is shown here. The variables are $X^i_{j,k,h}(t)$ and $\widetilde{X}^i_{j,k,h}(t), \forall i, j, k, h, t$ are continuous, while variables $a^R_{j,k,h}(t), \forall R, j, k, h, t$ are binary. The following formulation omits the indices of the first constraints to make it visually clearer.

$$\max \sum_{i,j,k,h,t} \left(\left[\hat{r}(s, cl) - \hat{r}(s^i_{j,k,h}, dn) \right] \theta a(t) X(t) + \hat{r}(s, dn) X(t) \right) + \sum_{i,j,k,h} q_{t^{max}}(s) X(t^{max}) \tag{3.32a}$$

$$\text{s.t. } X(t+1) = \left(1 - \sum_{s' \setminus s} \hat{g}\left(s'|s, dn\right) \right) \widetilde{X}^i_{j,k,h}(t) - \sum_{s' \setminus s} \left[\hat{g}\left(s'|s, cl\right) - \hat{g}\left(s'|s, dn\right) \right] \theta a(t) \widetilde{X}(t) + \sum_{s'' \setminus s} \hat{g}(s|s'', dn) \widetilde{X}''(t) + \sum_{s'' \setminus s} \left[\hat{g}\left(s|s'', cl\right) - \hat{g}\left(s|s'', dn\right) \right] \theta'' a''(t) \widetilde{X}''(t) , \forall\ i, j, k, h, t \tag{3.32b}$$

$$\widetilde{X}^i_{j,k,h}(t) = (1 - \upsilon_j) X^i_{j,k,h}(t) + (\upsilon_{\underline{j}}) X^i_{\underline{j},k,h}(t), \quad \forall\ i, j > 0, k, h, t \tag{3.32c}$$

$$\widetilde{X}^i_{0,k,h}(t) = X^i_{0,k,h}(t), \quad \forall\ i, k, h, t \tag{3.32d}$$

$$\sum_{i,j,k,h} X^i_{j,k,h}(0) = N_0 \tag{3.32e}$$

$$\sum_{R,i,j,k,h} a^R_{j,k,h}(t) X^i_{j,k,h}(t) \leq L^{max}, \quad \forall\ t < t^{max} \tag{3.32f}$$

$$a^R_{j,k,h}(t) \in \{0, 1\}, \quad \forall\ R, j, k, h, t \tag{3.32g}$$

The model (3.32) is nonlinear. The product $a^R_{j,k,h}(t) X^i_{j,k,h}(t)$ appears in the objective function and the first constraint. However, this product of a binary variable and a continuous variable can be replaced by $W^i_{j,k,h}(t)$ by enforcing the following

$$W^i_{j,k,h}(t) \leq N_0 a^R_{j,k,h}(t), \quad \forall\ i, j, k, h, t \tag{3.33a}$$

$$W^i_{j,k,h}(t) \leq X^i_{j,k,h}(t), \quad \forall\ i, j, k, h, t \tag{3.33b}$$

$$W^i_{j,k,h}(t) \geq X^i_{j,k,h}(t) - (1 - a^R_{j,k,h}(t)) N_0, \quad \forall\ i, j, k, h, t \tag{3.33c}$$

$$W^i_{j,k,h}(t) \geq 0, \quad \forall\ i, j, k, h, t \tag{3.33d}$$

Similarly, the product $a^R_{j,k,h}(t) \widetilde{X}^i_{j,k,h}(t)$ appears in the objective function and the first

constraint. However, this product of a binary variable and a continuous variable can be replaced by $\widetilde{W}^i_{j,k,h}(t)$ by enforcing the following

$$\widetilde{W}^i_{j,k,h}(t) \leq N_0 a^R_{j,k,h}(t), \quad \forall\ i,j,k,h,t \tag{3.34a}$$

$$\widetilde{W}^i_{j,k,h}(t) \leq X^i_{j,k,h}(t), \quad \forall\ i,j,k,h,t \tag{3.34b}$$

$$\widetilde{W}^i_{j,k,h}(t) \geq X^i_{j,k,h}(t) - (1 - a^R_{j,k,h}(t))N_0, \quad \forall\ i,j,k,h,t \tag{3.34c}$$

$$\widetilde{W}^i_{j,k,h}(t) \geq 0, \quad \forall\ i,j,k,h,t \tag{3.34d}$$

The full linear program is shown below.

$$\max \sum_{i,j,k,h,t} \left(\left[\hat{r}(s^i_{j,k,h}, cl) - \hat{r}(s^i_{j,k,h}, dn) \right] \theta^R_{j,k,h} W^i_{j,k,h}(t) + \hat{r}(s^i_{j,k,h}, dn) X^i_{j,k,h}(t) \right) + \sum_{i,j,k,h} q_{t^{max}}(s^i_{j,k,h}) X^i_{j,k,h}(t^{max}) \tag{3.35a}$$

$$\text{s.t. } X(t+1) = \left(1 - \sum_{s' \backslash s} \hat{g}\left(s'|s, dn\right) \right) \widetilde{X}(t) - \sum_{s' \backslash s} \left[\hat{g}\left(s'|s, cl\right) - \hat{g}\left(s'|s, dn\right) \right] \theta \widetilde{W}(t) + \sum_{s'' \backslash s} \hat{g}(s|s'', dn) \widetilde{X}''(t) + \sum_{s'' \backslash s} \left[\hat{g}\left(s|s'', cl\right) - \hat{g}\left(s|s'', dn\right) \right] \theta'' \widetilde{W}''(t) \tag{3.35b}$$

$$\widetilde{X}^i_{j,k,h}(t) = (1 - \upsilon_j) X^i_{j,k,h}(t) + (\upsilon_{\underline{j}}) X^i_{\underline{j},k,h}(t) \tag{3.35c}$$

$$\sum_{i,j,k,h} X^i_{j,k,h}(0) = N_0 \tag{3.35d}$$

$$\sum_{R,i,j,k,h} W^i_{j,k,h}(t) \leq L^{max}, \quad \forall\ t < t^{max} \tag{3.35e}$$

$$\sum_{R,i,j,k,h} W^i_{j,k,h}(t) C^i_j \leq C^{max}, \quad \forall\ t < t^{max} \tag{3.35f}$$

$$W^i_{j,k,h}(t) \leq N_0 a^R_{j,k,h}(t), \quad \forall\ i,j,k,h,t \tag{3.35g}$$

$$W^i_{j,k,h}(t) \leq X^i_{j,k,h}(t), \quad \forall\ i,j,k,h,t \tag{3.35h}$$

$$W^i_{j,k,h}(t) \geq X^i_{j,k,h}(t) - (1 - a^R_{j,k,h}) N_0, \quad \forall\ i,j,k,h,t \tag{3.35i}$$

$$\widetilde{W}^i_{j,k,h}(t) \leq N_0 a^R_{j,k,h}(t), \quad \forall\ i,j,k,h,t \tag{3.35j}$$

$$\widetilde{W}^i_{j,k,h}(t) \leq \widetilde{X}^i_{j,k,h}(t), \quad \forall\ i,j,k,h,t \tag{3.35k}$$

$$\widetilde{W}^i_{j,k,h}(t) \geq \widetilde{X}^i_{j,k,h}(t) - (1 - a^R_{j,k,h}) N_0, \quad \forall\ i,j,k,h,t \tag{3.35l}$$

$$\tag{3.35m}$$

3.3.4 Set of Policies

The model above has ultimate freedom to choose a policy that is optimal. However, this optimal policy might not be practical (e.g., too complicated for physicians and patients,

too demanding on specific group of patients, etc.). Instead, the model can be adjusted such that it chooses an optimal policy from a set of policies. Ideally, this set of policies would be comprehensive and not restricting. Mathematically, the model can be amended to account for this setup.

Let $\ell \in \mathcal{L}$ be a specific policy. Therefore, the number of candidate policies is $|\mathcal{L}|$. Also, let m_ℓ be a binary variable that takes a value of one if policy ℓ is selected, and zero otherwise. The action for individual system state, $a^R_{j,k,h}$, can be either part of policy ℓ (i.e., having a value of one) or not (i.e., having a value of zero), but not both. For clarity, let a_ℓ represent any action that is part of policy ℓ, and $\bar{a}_\ell$ represents any action that is not part of policy ℓ.

Therefore, amending the model with the following constraints makes it possible to select from a set of policies.

$$\sum_{\ell \in \mathcal{L}} m_\ell = 1 \tag{3.36a}$$

$$a_\ell \geq m_\ell \qquad \forall \ell \in \mathcal{L} \tag{3.36b}$$

$$\bar{a}_\ell \leq (1 - m_\ell) \qquad \forall \ell \in \mathcal{L} \tag{3.36c}$$

Constraints (3.36a) guarantees that only one policy is selected. Constraints (3.36b) state that if policy ℓ is selected, all actions belong to that policy should be equal to one, and if policy ℓ is not selected, the constraints are non-restricting. Similarly, constraints (3.36c) state that if policy ℓ is selected, all actions that do not belong to that policy must equal to zero, and if policy ℓ is not selected, these constraints are non-restricting.

3.4 Data Sources

The accuracy of the our models depends largely on the quality of the input parameters. Therefore, a special care is given to the choice and calibration of the data that is used

as input to the models. The following is a brief description of the data sources that were referred to to extract the necessary parameters.

The lesion progression rates, $\rho_j^{i,i'}$, from disease progression state i to disease progression state i' for age group j are derived from Erenay et al. (2011) and Loeve et al. (2004). These include the polyp onset probabilities, polyp-to-CRC progression probabilities, and lesion progression probabilities after CRC treatment and for post-CRC individuals.

The probabilities of mortality, $\delta_{i,o}^t(\hat{a})$, at disease progression state i when treatment $\hat{a}$ is taken at time t with observation o are extracted from the US life tables (Arias, 2015; Erenay et al., 2011) and checked against Xu et al. (2016), where applicable. These include the probabilities of mortality for CRC-free patients, mortality from undetected CRC, mortality for post-CRC patients, mortality in the UCT disease progression state, mortality after colonoscopy, and mortality after CRC treatment. The same for $\kappa_{\hat{a},o}^{i,j}$, the probability of immediate mortality from screening complications, and κ_{UCT}^j, the probability of immediate mortality from treatment at age j.

The probabilities that the CRC treatment will be completed within one year, $\gamma_{i,j}$, are withdrawn from SEER data (Erenay et al., 2014; Longo et al., 2000; Ohlsson and Pålsson, 2003; Yun et al., 2008). The disutility values $d_C, d_{CT}, d_{UCT}, d_{poly}(cl)$, and $d_{\neg poly}(cl)$ are based on values from Erenay et al. (2011, 2014), and Howlader et al. (2017). This is also the case for the terminal rewards $q_{t^{max}}(s_{j,k,h}^i)$. Moreover, the sensitivity of colonoscopy and CRC is based on Frazier et al. (2000) and Vijan et al. (2007), while the probability of CRC self detection $\omega_{\hat{a}}$ is extracted from Erenay et al. (2014) and Howlader et al. (2017).

Capacity parameters are estimated using methods used in Butterly et al. (2007) and Güneş et al. (2015). Population demographic parameters are estimated using Ramsey et al. (2010)'s simulation study. Initial number of individuals in each core health states are estimated based on Loeve et al. (2004) for people with polypectomy, Wilschut et al. (2011) for people with family history, and Erenay et al. (2011) for post-CRC patients.

Other sources of data consulted and checked include Arora et al. (2009); Brenner et al. (2011); Butterly et al. (2007); Gatto et al. (2003); Ladabaum and Song (2005); and Seeff et al. (2004).

3.5 Computational Results and Analyses of MIP Model

In this section, we present preliminary results from the MIP model that would guide further testing and analysis. The analysis will focus on the effects of changing particular parameters on the overall optimal policy. For this preliminary analysis, the parameters of interest are mainly the prevalence and capacity level. A sensitivity analysis will be conducted on these parameters to see the significance of each. Other parameters (mainly disease-related) are assumed to be fixed for now. Future work will include comprehensive sensitivity analysis for all major parameters.

Butterly et al. (2007) estimate that 35% of the population (>50 years) are at increased risk, while 65% of the population are at average risk. Matching figures appear in Ladabaum and Song (2005). The 35% of the increased risk population include those with personal history of polyps.

Regarding compliance, the base values for low risk, high risk, and post-CRC populations will be assumed 0.6, 0.8, and 1.0, respectively. This in part is based on figures from Butterly et al. (2007); Frazier et al. (2000); and Güneş et al. (2015).

Vijan et al. (2004) estimate the screening colonoscopy capacity at around two million per year. This estimation is based on a database of 400 gastrointestinal endoscopists in the USA. The sample is extrapolated to all gastroenterologists, and then extended to account for screenings done by providers other than gastroenterologists. Seeff et al. (2004) estimate the current capacity of colonoscopy (diagnosis and screening) at 14.2, and the potential capacity (that could be available if there is demand for it) at 22.4 million. This estimate is based on telephone surveys of medical facilities known to have purchased or leased colonoscopy (or sigmoidoscopy) equipment within a certain time frame. Butterly et al. (2007) and Güneş et al. (2015) estimate the percentage of colonoscopy capacity allocated to screening at 55% (range 50%-60%). Therefore, the base case for capacity will be 5 million annually, with range (2-7.5) million.

The test instances shown in this section share some common parameters. The number of age groups $|\mathcal{J}| = 5$. These age groups are 50-54, 55-59, 60-64, 65-69, and 70-74. Increasing the number of age groups would significantly increase the size of the model,

as well as generate more accurate results. Also, $t^{max} = 8$. This will keep the model at reasonable size. Future work includes increasing this value to better capture the aging of the population.

The number of variables is $|\mathcal{R}||\mathcal{J}||\mathcal{K}||\mathcal{H}||\mathcal{T}|+4\times|\mathcal{I}||\mathcal{J}||\mathcal{K}||\mathcal{H}||\mathcal{T}|$. Of those, $|\mathcal{R}||\mathcal{J}||\mathcal{K}||\mathcal{H}||\mathcal{T}|$ are binary.

3.5.1 Effects of Prevalence

To test the effects of having accurate estimates of the CRC prevalence in the target population, we start with some hypothetical cases that would help in understanding the dynamics of the model and the resulting policies.

The prevalence values are given to a total of 90 system states; given for both males and females: $|\mathcal{J}| = 5$, and number of disease progression 9 (states UCT and D are excluded), with $t = 0$ and $h = 0$.

Equally Distributed Population

In this instance, we assume that the initial population is equally distributed among all population groups at 1.11% for each group. Obviously, any capacity less than $(1.11 \times 3 =)$ 3.33 % of the target population would mean that colonoscopy would never be applied (at least at time $t = 0$). This is proven when experimenting with the lower limit of the colonoscopy capacity. As mentioned above, the lower limit is 2 million annually, which corresponds to 2.34% of the target population. The optimal policy in such a case calls for screening of the youngest PC males and females at $h = 2$ almost annually. This is accompanied by occasional screening for lower h values.

Increasing the capacity limit to 5 million annually corresponds to 5.84% of the population. This gives more freedom to screen more population groups. The youngest PC females with $h = 0$ and $h = 2$ are screened almost annually, while those with $h = 1$ are screened biannually. The same trend appears for the second youngest age group. This shows the emphasis on screening the PC females in general. For the LR population, screening is

concentrated on males with $j = 3, h = 2$ almost annually after $t = 6$. It is necessary for this to be feasible that there are some capacity allocated to the same population group at $h = 0$ and $h = 1$ in order to be have patients at $h = 2$.

Finally, increasing the capacity limit to 7.5 million annually corresponds to 8.76% of the target population. The optimal policy in this case is to screen the second youngest LR males and females almost every 3 years, with T- results. The screening continues with aging. Also, the males and females at PC risk level are screened annually covering more age groups and h levels. This shows that when capacity is available, more resources are directed toward LR. This can be explained by the tendency of the model to remove more lesions, and have lesion free population.

The analysis here shows the significance of the capacity limit on the optimal policy. As detailed above, the optimal policy can change dramatically based on the number of screening tests available. This is very important since the current guidelines are insensitive to the available capacity, and therefore, might be suboptimal as well.

Low-Risk-Dominated Population

In this scenario, the target population is of low-risk only. Out of this, 2.5% have polyps and another 1% have cancer. Since the number of individuals in each initial population group is either zero or a value higher than the lower bound of the capacity (2.34%), the lower bound of capacity is of no use and offers no help in screening the population. When using the base case capacity of 5.84%, the same happens. The capacity is so small that it is not enough to meaningfully screen any group entirely. At the upper limit on capacity of 8.76%, the optimal policy focus on screening the oldest LR males at least once every 5 years. Unlike previous optimal policies, this policy's focus is on males, rather than females. This is an interesting observation and will be studies further in future research. However, one explanation for this might be that the model is tending to keep the bulk of the population at the lowest risk possible. By screening the LR individuals at disease progression state $i = 0$, it is more likely to keep them there with higher compliance. Hence, it would be a high percentage of T- observations.

Representative Population

For the representative population, we will assume the low risk population to be 65%, while the high risk and post-CRC populations are, respectively, 30% and 5%. This is based on figures given in Butterly et al. (2007) and Ladabaum and Song (2005). Within the low-risk groups, it will be assumed that 96.5% have no polyps, 2.5% have polyps, and 1% have cancers. Similarly, within the high-risk and post-CRC groups, it will be assumed that the proportion of polyps and cancers are double those for low-risk patients. In particular, the percentages for no polyps, polyps, and cancers are, respectively, 93%, 5%, and 2%.

When the capacity is scarce (2.34% of the target population), the optimal policy tends to focus on the oldest *HR* males and females, with screening programs at least once every 7-8 years. A slightly more frequent screening program is given to *PC* patients, especially the youngest age group. This is understandable given the limited capacity to prioritize those patients with higher chances of survival and where impact is maximized. With a moderate estimate of capacity at 5.84%, *LR* patients are screened almost every 5 years up to age 60. *HR* and *PC* patients are screened annually from age 50-64. Afterwards, they are screened biannually. With the upper bound on capacity (8.76% of the target population), *LR* patients are screened every 3 years, while *HR* and *PC* are screened annually.

Table 3.11 shows a graphical representation of the results discussed earlier. As can be seen, there is a clear trend regarding the effect of capacity on the optimal policies. A sample policy is given in Table 3.12. Typically, the policies would differ for different age groups. When designing an easy-to-follow policy, it is important to aggregate and combine policies of different age groups and genders in order to arrive at a policy with minimal number of variables. Patients and physicians are more likely to follow such policies.

3.5.2 Effects of Capacity

To study the effects of changing the available colonoscopy capacity, we will use the example of representative population from the previous sections. In particular, we will assume we have a population of 65% low risk, 30% high risk, and 5% post-CRC distribution. Within

Table 3.11: Comparison between Optimal Policies and Current Guidelines

Prevalence	Capacity		
	Low	Base	High
LR-Dominated	*LR*,*HR*,*PC*	*LR*,*HR*,*PC*	$\overline{LR}$,*HR*,*PC*
Representative	*LR*,*HR*,*PC*	*LR*,$\overline{HR}$,$\overline{PC}$	$\overline{LR}$,$\overline{HR}$,$\overline{PC}$

Legend:
underlined (and red): Less frequent than guidelines
normal (and black): As frequent as guidelines
$\overline{\textit{upper bar}}$ *(and green)*: More frequent than guidelines

Table 3.12: Sample Screening Policy: Optimal Policy for Representative Population with Base Case Capacity

Age	*LR*	*HR*	*PC*
50-54	5	1	
55-59			
60-64	10		
65-69		2	
70-74			

the low risk groups, 96.5% have no polyps, 2.5% have polyps, and another 1% have cancers. For high risk and post-CRC groups, the distribution is 93%, 5%, and 2%, respectively.

From previous experiments, we know that with shortage in available capacity, the optimal policy tends to focus on HR and PC groups, especially young ones. Now, we consider the potential capacity of 11 million screening colonoscopies annually. Recall that this figure comes from an estimated potential capacity of 22.4 million annually from Seeff et al. (2004), and a proportion of about 50% given to screening (Butterly et al., 2007). With this scenario, capacity is around 12.85% of the target population. With this capacity, HR and PC are screened almost annually, while LR are screened every 5 years.

It is apparent from analyzing different scenarios of available capacity that it is a very important parameter to consider when designing the screening policy for a population. The current guidelines are not necessary optimal for all capacity levels. Therefore, failure to address capacity would result in suboptimal screening programs, which would translate into poorer health outcomes.

3.6 Conclusions

This chapter presented two modeling procedures for the problem of allocating colonoscopy screening resources among a population of potential CRC patients. The MDP procedure is a stochastic technique that has the potential to be very accurate. However, due to the extremely large number of system states, solving this model is a challenge. The MIP procedure benefits from the large number of individuals in each population subgroup and provides an approximation to the stochastic process. Consequently, the MIP model can account for population dynamics that are not considered in the MDP model, such as different age groups, gender, and personal history of screening tests.

The analysis of this chapter has focused on the significance and importance of incorporating capacity availability and requirements into the design of CRC screening guidelines. It was found that the current guidelines are not always optimal. In particular, if colonoscopy resources are abundant, optimal screening programs recommend higher screening rates for low-risk patients than the current guidelines, and almost similar screening rate for

post-CRC patients. When screening capacity is scarce, the low risk patients are screened less frequently than the guidelines. This shows the significance of incorporating screening capacity into the decision of optimal screening policies.

It was found that the population composition can have an influence on the optimal policy. In particular, it was shown through experimentation that in situations where capacity is scarce, the optimal policy tends to favor females and post-CRC patients. This is explained by the higher expected age and lower mortality rate for females, as well as the major health benefits of removing potential polyps and cancerous lesions from the post-CRC patients.

Extensions to this model include increasing the number age groups. Ideally, age groups would be of length one year. This will cause $|\mathcal{J}|$ to increase, but also it would eliminate the need for $\widetilde{X}$. The resulting effect needs to be studied. Also, the changes in the population distribution and dynamics, especially with large values of t^{max} are interesting research questions to pursue.

Moreover, other considerations can be added to make the model more realistic. For example, the cost of the screening represents a major factor in designing optimal policy that would be sustainable and acceptable by the society and medical community. As such, incorporating the budget constraint would be vital to successful implementation of the guidelines. In particular, the following represents the cost limitation that considers the available budget.

$$\sum_{R,i,j,k,h} a^R_{j,k,h}(t) X^i_{j,k,h}(t) C^i_j \leq C^{max}, \quad \forall\, t < t^{max} \tag{3.37}$$

where C^i_j is the cost of conducting colonoscopy for one patient in risk level R and disease progression stage i and age group j, and C^{max} is the cost limit (budget) for colonoscopy in one year.

Furthermore, to obtain the society's support of preventive care policies, it is important to quantify and limit the mortality rate that is likely to occur as a result of applying such policy. The following constraints aim to do that.

$$\sum_{R,i,j,k,h,\hat{a}} a^{R}_{j,k,h}(t) X^{i}_{j,k,h}(t) \delta_{R,i,o}(\hat{a}) f\left(o|s^{i}_{j,k,h}, \hat{a}\right) \leq Mortality^{max}, \quad \forall\ t < t^{max} \tag{3.38}$$

where $Mortality^{max}$ is the limit on annual mortality that is set in priori. The discussion on ethical issues associated specifying a mortality limit is worth having, but it is beyond the scope of this work.

Chapter 4

Conclusions and Future Work

This chapter gives a summary of the models and results discussed in this thesis. Also, a glimpse into the future research work directions is presented, with a brief discussion on the challenges and potential of each.

4.1 Summary

This thesis discussed resource allocation in healthcare applications. In Chapter 1, a review of major resource allocation models in healthcare was presented. A brief introduction to the target problems and research questions were then listed.

Chapter 2 discussed a facility location model where facilities are subject to failure, and customers have preferences. The aim was to find the best number and location of facilities to open, and the best assignment strategy for customer demand. It was shown that this integration is important and would significantly save costs. Then, a solution methodology was presented. The Lagrangian based procedure was then implemented in three different ways. It was shown that each implementation has its own features. Among them, PQ-LBB emerged as the most efficient due to its special characteristic of prioritizing nodes that have the potential to be optimal. Later on, a reformulation with fewer number of variables was presented, and solution methodologies based on a Lagrangian relaxation

embedded within a branch-and-bound and a branch-and-cut structures. Also, important results and characteristics of the methodology were proven. Moreover, the chapter proposed a constraint which significantly tightens LP relaxation of the formulation. The numerical experiments showed that the proposed solution algorithms can be applied to problems with extremely large number of customers.

Chapter 3 discussed allocating CRC screening resources among a representative population. Screening is effective in the prevention and early detection of CRC. However, the full benefit of screening is attained by periodic testing. This can be a challenge given the limited screening resources available. An analytical framework was advised based on a MDP. The challenges and advantages of using such model were presented. The model built on a published model on the individual patient CRC screening process. A formal representation of the MDP components was given, including the probability transition matrix, reward, and objective function. The state space for this model would be huge as the population size grows. As a result, some methodologies for state aggregation were discussed. Later on, a MIP was developed for the same problem. This model was solved and significant insights were drawn. It was found that current guidelines are not always optimal. In particular, if colonoscopy resources are abundant, optimal screening programs recommend higher screening rates for low-risk patients than the current guidelines, and almost similar screening rate for post-CRC patients. When screening capacity is scarce, the low risk patients are screened less frequently than the guidelines. This shows the significance of incorporating screening capacity into the decision of optimal screening policies.

4.2 Future Research Directions

There are many directions for future research. The following are some examples. These suggestions are organized into two main problems: the reliable facility location model and the CRC cancer screening allocation model.

4.2.1 Reliable Facility Location Model with Customer Preferences

In preventive healthcare, patients (customers) have more chance of selecting when and where to receive a service. As opposed to urgent healthcare interventions, customers' preferences in preventive healthcare play a major part in service assignments. Examples of preventive healthcare include screenings for cancers, diabetes, and cardiovascular diseases, regular monitoring of weight and cholesterol levels, and general advice regarding tobacco and alcohol use, to name a few. The decisions of how many preventive healthcare facilities, where to locate them, and which customers to assign to them can be drawn upon building on classical location models, as well as literature related to customer preferences.

We were able to solve extremely large instances. In real life location problems, decision makers may need to consider a large number of customer types in terms of their preferences over the set of candidate sites. For instance, when choosing preventive healthcare facilities, patients might decide based on proximity and service quality (Verter and Zhang, 2015). There could be several preference types because patients might weigh proximity and service quality differently.

One possible extension for this work would be improving the solution algorithms. Solving the Lagrangian dual can be time consuming due to the need to solve many mixed-integer subproblems. We may alleviate this difficulty using a Benders decomposition within an LP based branch-and-bound method. A pure Benders decomposition approach, however, may yield weak relaxations, leading to a large branch-and-bound tree. Therefore, we will try to use integrality constraints to obtain improved LP relaxations within the Benders decomposition framework (Bodur et al., 2017)

Another future direction would be adding a budget constraint or a limit on the number of facilities to open. This will be similar to the p-median problem. It is also possible to introduce facility capacities into the model. Allocating customers to capacitated facilities based on preference would be a non-trivial extension of our model, because in this case the model must determine which customers are denied service if there is not enough capacity at a highly preferred facility. It is also possible to model capacity levels for each facility as decision variables. Furthermore, we will explore other applications of the reliability models

with customer preferences, especially in preventative healthcare.

Moreover, facility location models can be split into two interrelated problems. As explained by Hanjoul and Peeters (1987), the two problems are solved under the assumption that the other is optimized accordingly. The first subproblem is the locating (how many and where facilities are to be opened), while the other is the allocation (which customer is assigned to which facility). In this context and with assumption about the strategy and optimality criteria of both 'players', the solution of this location-allocation problem can be viewed as Stackelberg equilibrium. By introducing the preferences of customers into this problem, it would be interesting to study the dynamics of the game if both players have conflicting goals. For example, not revealing the true preferences of customers to promote closer assignment for certain customers. Camacho-Vallejo et al. (2014) present an algorithm to solve the bilevel UFLPUP based on a Stackelberg equilibrium scheme with an evolutionary algorithm.

4.2.2 Resource Allocation in Colorectal Cancer Screening

The aim of the models described in Chapter 3 was to develop a framework to find an optimal CRC screening policy for a representative population. For this purpose, an MDP model was developed. Given the complexity of the problem, certain assumptions were made to make it tractable. These assumptions are to be relaxed in future extensions of the model, including: adding age- and gender-based groups to the population, and allowing different or multiple initiation and termination times for CRC screening.

Moreover, it is assumed that the process is completely observable, and that actions applied to the individual patient are partially observable. A more realistic representation would be to assume that every individual in the population has a probability of developing a polyp, and a probability of developing cancer. In this case, the system state would consider this probability mix. This, however, can dramatically affect the state space as well as solution efficiency.

Although the MDP model is capable of explicitly capturing the randomness, making it more realistic, the difficulty of finding a solution calls for different modeling techniques.

For example, in the compartmental model or the fluid model, transitions are deterministic, and the system can be defined by a set of differential equations. This would allow to handle a bigger population and, therefore, build more efficient solution.

Extensions for the MIP model include increasing the number age groups. Ideally, age groups would be of length one year. This will cause $|\mathcal{J}|$ to increase, but also it would eliminate the need for $\widetilde{X}$. The resulting effect needs to be studied. Also, the changes in the population distribution and dynamics, especially with large values of t^{max} are interesting research questions to pursue.

Moreover, other considerations can be added to make the model more realistic. For example, the cost of the screening represents a major factor in designing an optimal policy that would be sustainable and acceptable by the society and medical community. As such, incorporating the budget constraint would be vital to successful implementation of the guidelines. Another modification in the model would be removing the index of time from the action variables. This would allow for more intuitive policies that do not depend on time.

Furthermore, to obtain they society's support of preventive care policies, it is important to quantify and limit the mortality rate that is likely to occur as a result of applying such policy. The discussion on ethical issues associated specifying a mortality limit is worth having, but it is beyond the scope of this work.

Finally, for average risk population, all organized CRC screening programs in Canada target individuals in the 50-74 age group (CPAC, 2017). It is worth noting that approximately 75% of all CRCs occur among persons at average risk, at any age (Winawer et al., 1991). Therefore, the current guidelines have not helped in increasing identification and prevention of CRC among younger adults at average risk. In fact, around 60-75% of young-onset of CRC cases are attributed to reasons other than family history of the disease and genetic predisposition (Patel and De, 2016). This high, and trending upwards, percentage of undetected cancers would result in excessive costs and strain the healthcare system due to costly and long projected treatments in the near and far future. There are significant health and cost benefits associated with analyzing the average risk population and eventually designing targeted screening programs that would help in identifying early onset of

CRC among average risk individuals. This is a direction of research that the author is currently investigating.

References

Aboolian, R., Cui, T., and Shen, Z.-J. M. (2013). An efficient approach for solving reliable facility location models. *INFORMS Journal on Computing*, 25(4):720–729.

Ahmadi-Javid, A. and Seddighi, A. H. (2013). A location-routing problem with disruption risk. *Transportation Research Part E: Logistics and Transportation Review*, 53(1):63–82.

Akan, M., Alagoz, O., Ata, B., Erenay, F. S., and Said, A. (2012). A broader view of designing the liver allocation system. *Operations Research*, 60(4):757–770.

Akgün, ., Gümübua, F., and Tansel, B. (2015). Risk based facility location by using fault tree analysis in disaster management. *Omega*, 52:168–179.

Akhundov, N. (2015). Optimal location, patient routing, and capacity decisions for endoscopy clinical network in western ontario: A Simulation-based optimization approach. Master's thesis, University of Waterloo.

Akkerman, R. and Knip, M. (2004). Reallocation of beds to reduce waiting time for cardiac surgery. *Health Care Management Science*, 7(2):119–126.

Aksen, D., Aras, N., and Piyade, N. (2013). A bilevel p-median model for the planning and protection of critical facilities. *Journal of Heuristics*, 19(2):373–398.

Alagoz, O., Ayer, T., and Erenay, F. S. (2011). Operations research models for cancer screening. *Wiley Encyclopedia of Operations Research and Management Science*.

Alagoz, O., Maillart, L. M., Schaefer, A. J., and Roberts, M. S. (2004). The optimal timing of living-donor liver transplantation. *Management Science*, 50(10):1420–1430.

Alagoz, O., Maillart, L. M., Schaefer, A. J., and Roberts, M. S. (2007). Determining the acceptance of cadaveric livers using an implicit model of the waiting list. *Operations Research*, 55(1):24–36.

Albareda-Sambola, M., Hinojosa, Y., and Puerto, J. (2015). The reliable p-median problem with at-facility service. *European Journal of Operational Research*, 245(3):656–666.

Alessandra, A. J. and Grazman, T. E. (1978). Using simulation in hospital planning. *Simulation*, 30(2):62–67.

Allison, J. E. and Lawson, M. (2006). Screening tests for colorectal cancer: A Menu of options remains relevant. *Current Oncology Reports*, 8(6):492–498.

American Cancer Society (2009). Cancer prevention and early detection facts and figures. Technical report, American Cancer Society. Atlanta, GA.

American Cancer Society (2015). Cancer prevention and early detection. Technical report, American Cancer Society. Atlanta, GA.

American Cancer Society (2017). Cancer facts and figures 2017. Technical report, American Cancer Society. Atlanta, GA.

An, S., Cui, N., Li, X., and Ouyang, Y. (2013). Location planning for transit-based evacuation under the risk of service disruptions. *Transportation Research Part B: Methodological*, 54:1–16.

Arias, E. (2015). United states life tables 2011. *National Vital Statistics Reports-National Center for Health Statistics*, 64(11):1–63.

Arora, G., Mannalithara, A., Singh, G., Gerson, L. B., and Triadafilopoulos, G. (2009). Risk of perforation from a colonoscopy in adults: A Large population-based study. *Gastrointestinal Endoscopy*, 69(3 SUPPL.):654–664.

Austin, H., Jane Henley, S., King, J., Richardson, L. C., and Eheman, C. (2014). Changes in colorectal cancer incidence rates in young and older adults in the united states: What does it tell us about screening. *Cancer Causes & Control*, 25(2):191–201.

Ayer, T., Alagoz, O., and Stout, N. K. (2012). OR Forum - A POMDP approach to personalize mammography screening decisions. *Operations Research*, 60(5):1019–1034.

Ayvaci, M. U., Alagoz, O., and Burnside, E. S. (2012). The effect of budgetary restrictions on breast cancer diagnostic decisions. *Manufacturing & Service Operations Management*, 14(4):600–617.

Baldwin, L.-M., Cai, Y., Larson, E. H., Dobie, S. A., Wright, G. E., Goodman, D. C., Matthews, B., and Hart, L. G. (2008). Access to cancer services for rural colorectal cancer patients. *The Journal of Rural Health*, 24(4):390–399.

Baltacioglu, T., Ada, E., Kaplan, M. D., Yurt And, O., and Cem Kaplan, Y. (2007). A new framework for service supply chains. *The Service Industries Journal*, 27(2):105–124.

Baron, O., Milner, J., and Naseraldin, H. (2011). Facility location: A Robust optimization approach. *Production and Operations Management*, 20(5):772–785.

BC Cancer Agency (2013). Colorectal Cancer Incidence in BC 1974 to 2009. http://www.bccancer.bc.ca/statistics-and-reports-site/Documents/IncidenceColorectal.pdf. [Online; accessed 09-Apr-2017].

Beasley, J. E. (1990). OR-Library: Distributing test problems by electronic mail. *Journal of the Operational Research Society*, 41(11):1069–1072.

Berman, O., Krass, D., and Menezes, M. B. (2007). Facility reliability issues in network p-median problems: Strategic centralization and co-location effects. *Operations Research*, 55(2):332–350.

Berman, O., Krass, D., and Menezes, M. B. (2009). Locating facilities in the presence of disruptions and incomplete information. *Decision Sciences*, 40(4):845–868.

Bertsimas, D., Farias, V. F., and Trichakis, N. (2013). Fairness, efficiency, and flexibility in organ allocation for kidney transplantation. *Operations Research*, 61(1):73–87.

Blount, S., Galambosi, A., and Yakowitz, S. (1997). Nonlinear and dynamic programming for epidemic intervention. *Applied Mathematics and Computation*, 86(2):123–136.

Bodur, M., Dash, S., and Günlük, O. (2017). Cutting planes from extended LP formulations. *Mathematical Programming*, 161(1-2):159–192.

Brailsford, S., Harper, P., Patel, B., and Pitt, M. (2009). An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation*, 3(3):130–140.

Brailsford, S. and Vissers, J. (2011). OR in healthcare: A European perspective. *European Journal of Operational Research*, 212(2):223–234.

Brandeau, M. L., Zaric, G. S., and Richter, A. (2003). Resource allocation for control of infectious diseases in multiple independent populations: Beyond cost-effectiveness analysis. *Journal of Health Economics*, 22(4):575–598.

Brenner, H., Jenny, C.-C., Seiler, C., Rickert, A., and Hoffmeister, M. (2011). Protection from colorectal cancer after colonoscopy. *Annals of Internal Medicine*, 154(1):22–30.

Butler, T. W., Leong, G. K., and Everett, L. N. (1996). The operations management role in hospital strategic planning. *Journal of Operations Management*, 14(2):137–156.

Butterly, L., Olenec, C., Goodrich, M., Carney, P., and Dietrich, A. (2007). Colonoscopy demand and capacity in new hampshire. *American Journal of Preventive Medicine*, 32(1):25–31.

Camacho-Vallejo, J. F., Cordero-Franco, A. E., and González-Ramírez, R. G. (2014). Solving the bilevel facility location problem under preferences by a stackelberg-evolutionary algorithm. *Mathematical Problems in Engineering*, 2014.

Camm, J. D., Chorman, T. E., Dill, F. A., Evans, J. R., Dennis, J., and Wegryn, G. W. (1997). Blending OR/MS , judgment , and GIS : Restructuring P&G's supply chain. *Interfaces*, 27(1):128–142.

Canadian Cancer Society (2015). Screening for colorectal cancer. http://www.cancer.ca/en/cancer-information/cancer-type/colorectal/screening/?region=on. [Online; accessed 13-Oct-2015].

Canadian Cancer Society (2016). Canadian cancer statistics 2016. Technical report, Toronto, Ontario.

Canadian Cancer Society (2017). Canadian cancer statistics 2017. Technical report, Toronto, Ontario.

Canadian Task Force on Preventive Health Care (2016). Recommendations on screening for colorectal cancer in primary care. *Canadian Medical Association Journal*, 188(5):340–348.

Cánovas, L., García, S., Labbé, M., and Marín, A. (2007). A strengthened formulation for the simple plant location problem with order. *Operations Research Letters*, 35(2):141–150.

Carøe, C. C. and Schultz, R. (1999). Dual decomposition in stochastic integer programming. *Operations Research Letters*, 24(1):37–45.

Carter, M. (2002). Diagnosis: Mismanagement of resources. *OR MS Today*, 29(2):26–33.

Catlin, A., Cowan, C., Hartman, M., Heffler, S., Team, N. H. E. A., et al. (2008). National health spending in 2006: a year of change for prescription drugs. *Health Affairs*, 27(1):14–29.

Cayirli, T. and Veral, E. (2003). Outpatient scheduling in health care: A Review of literature. *Production and Operations Management*, 12(4):519.

CCO (2016). Colorectal cancer increasing in younger adults. https://www.cancercare.on.ca/ocs/csurv/ont-cancer-facts. Cancer Care Ontario [Online; accessed 09-Apr-2017].

CDC (2014). Influenza activity,united states, 2013,14 season and composition of the 2014,15 influenza vaccines. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6322a2.htm. Centers for Disease Control and Prevention [Online; accessed 08-Sep-2015].

Chan, S.-Y. E., Ohlmann, J., Dunbar, S., Dunbar, C., Ryan, S., and Savory, P. (2002). Operations research methods applied to workflow in a medical records department. *Health Care Management Science*, 5(3):191–199.

Chao, D. L., Halloran, M. E., Obenchain, V. J., and Longini Jr, I. M. (2010). FluTE, A Publicly available stochastic influenza epidemic simulation model. *PLoS Computational Biology*, 6(1):e1000656.

Charlton, M. E., Matthews, K. A., Gaglioti, A., Bay, C., McDowell, B. D., Ward, M. M., and Levy, B. T. (2015). Is travel time to colonoscopy associated with late-stage colorectal cancer among medicare beneficiaries in Iowa? *The Journal of Rural Health*, 32(4):363–373.

Chen, A. Y. and Yu, T. Y. (2016). Network based temporary facility location for the emergency medical services considering the disaster induced demand and the transportation infrastructure in disaster response. *Transportation Research Part B: Methodological*, 91:408–423.

Chen, L., Olhager, J., and Tang, O. (2014). Manufacturing facility location and sustainability: A Literature review and research agenda. *International Journal of Production Economics*, 149:154–163.

Chen, Q., Li, X., and Ouyang, Y. (2011). Joint inventory-location problem under the risk of probabilistic facility disruptions. *Transportation Research Part B: Methodological*, 45(7):991–1003.

Choi, Y., Sateia, H. F., Peairs, K. S., and Stewart, R. W. (2017). Screening for colorectal cancer. *Seminars in Oncology*, 44(1):34–44.

CIHI (2015a). Canadian organ replacement register annual report: Treatment of end-stage organ failure in canada, 2004 to 2013. Technical report, Canadian Institute for Health Information, Ottawa, Ontario.

CIHI (2015b). National health expenditure trends, 1975 to 2014. Technical report, Canadian Institute for Health Information, Ottawa, Ontario.

CMU Center for Economic Development (2007). AAA Senior Center Sourcebook. Technical report, Pittsburgh, PA.

Cohen, J. T., Neumann, P. J., and Weinstein, M. C. (2008). Does preventive care save money? Health economics and the presidential candidates. *New England Journal of Medicine*, 358(7):661–663.

CPAC (2017). Colorectal cancer screening in canada: Monitoring & evaluation of quality indica-

tors - results report, January 2013 - December 2014. Technical report, Canadian Partnership Against Cancer, Toronto, Ontario.

Cui, T., Ouyang, Y., and Shen, Z.-J. M. (2010). Reliable facility location design under the risk of disruptions. *Operations Research*, 58(4-Part-1):998–1011.

Current, J. and Weber, C. (1994). Application of facility location modeling constructs to vendor selection problems. *European Journal of Operational Research*, 76(3):387–392.

Daskin, M. and Dean, L. (2005). Location of healthcare facilities. *Operations Research and Health Care (Brandeau, M and Sainfort, F and Pierskalla, W, Eds)*, 70:43–76.

Daskin, M. S. (1995). *Network and Discrete Location: Models, Algorithms, and Applications*. John Wiley & Sons, New York.

De Vries, G., Bertrand, J., and Vissers, J. (1999). Design requirements for health care production control systems. *Production Planning & Control*, 10(6):559–569.

Demarteau, M. N., Breuer, T., and Standaert, B. (2012). Selecting a mix of prevention strategies against cervical cancer for maximum efficiency with an optimization program. *Pharmacoeconomics*, 30(4):337–353.

Demirtas, E. A. and Üstün, Ö. (2008). An integrated multiobjective decision making process for supplier selection and order allocation. *Omega*, 36(1):76–90.

Demko, D. J. (1980). Utilization, attrition, and the senior center. *Journal of Gerontological Social Work*, 2(2):87–93.

Dobrzykowski, D., Deilami, V. S., Hong, P., and Kim, S.-C. (2014). A structured analysis of operations and supply chain management research in healthcare (1982–2011). *International Journal of Production Economics*, 147:514–530.

Dobrzykowski, D. D. (2012). Examining heterogeneous patterns of electronic health records use: A Contingency perspective and assessment. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, 7(2):1–16.

Drobne, S. and Bogataj, M. (2015). Optimal allocation of public service centres in the central places of functional regions. *IFAC-PapersOnLine*, 28(3):2362–2367.

Earnshaw, S. R. and Dennett, S. L. (2003). Integer/linear mathematical programming models. *Pharmacoeconomics*, 21(12):839–851.

Earnshaw, S. R., Richter, A., Sorensen, S. W., Hoerger, T. J., Hicks, K. A., Engelgau, M., Thompson, T., Narayan, K. V., Williamson, D. F., Gregg, E., et al. (2002). Optimal

allocation of resources across four interventions for type 2 diabetes. *Medical Decision Making*, 22(suppl 1):s80–s91.

Edwards, B. K., Ward, E., Kohler, B. A., Eheman, C., Zauber, A. G., Anderson, R. N., Jemal, A., Schymura, M. J., Lansdorp-Vogelaar, I., Seeff, L. C., et al. (2010). Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. *Cancer*, 116(3):544–573.

Epstein, D. M., Chalabi, Z., Claxton, K., and Sculpher, M. (2007). Efficiency, equity, and budgetary policies informing decisions using mathematical programming. *Medical Decision Making*, 27(2):128–137.

Erenay, F. S., Alagoz, O., Banerjee, R., and Cima, R. R. (2011). Estimating the unknown parameters of the natural history of metachronous colorectal cancer using discrete-event simulation. *Medical Decision Making*, 31(4):611–624.

Erenay, F. S., Alagoz, O., and Said, A. (2014). Optimizing colonoscopy screening for colorectal cancer prevention and surveillance. *Manufacturing & Service Operations Management*, 16(3):381–400.

Esserman, L., Shieh, Y., and Thompson, I. (2009). Rethinking screening for breast cancer and prostate cancer. *Jama*, 302(15):1685–1692.

Fajobi, O., Yiu, C., Sen-Gupta, S., and Boulos, P. (1998). Metachronous colorectal cancers. *British Journal of Surgery*, 85(7):897–901.

Fakhimi, M. and Propert, J. (2013). Operations research within uk healthcare: A Review. *Journal of Enterprise Information Management*, 26(1):21–49.

Falkenheimer, S. A. (2004). The adequacy of preventive health care: Does the health care provider matter? https://cbhd.org/content/adequacy-preventive-health-care-does-health-care-provider-matter. [Online; accessed 13-Oct-2015].

Fone, D., Hollinghurst, S., Temple, M., Round, A., Lester, N., Weightman, A., Roberts, K., Coyle, E., Bevan, G., and Palmer, S. (2003). Systematic review of the use and value of computer simulation modelling in population health and health care delivery. *Journal of Public Health*, 25(4):325–335.

Frazier, A. L., Colditz, G. A., Fuchs, C. S., and Kuntz, K. M. (2000). Cost-effectiveness of screening for colorectal cancer in the general population. *Jama*, 284(15):1954–1961.

Garcia-Herreros, P., Wassick, J. M., and Grossmann, I. E. (2014). Design of resilient supply chains with risk of facility disruptions. *Industrial & Engineering Chemistry Research*.

Gatto, N. M., Frucht, H., Sundararajan, V., Jacobson, J. S., Grann, V. R., and Neugut, A. I. (2003). Risk of perforation after colonoscopy and sigmoidoscopy: A Population-based study. *Journal of the National Cancer Institute*, 95(3):230–236.

Griffin, P. M., Scherrer, C. R., and Swann, J. L. (2008). Optimization of community health center locations and service offerings with statistical need estimation. *IIE Transactions*, 40(9):880–892.

Gu, W., Wang, X., and McGregor, S. E. (2010). Optimization of preventive health care facility locations. *International Journal of Health Geographics*, 9(1):17.

Günal, M. M. and Pidd, M. (2010). Discrete event simulation for performance modelling in health care: A Review of the literature. *Journal of Simulation*, 4(1):42–51.

Güneş, E. D., Örmeci, E. L., and Kunduzcu, D. (2015). Preventing and diagnosing colorectal cancer with a limited colonoscopy resource. *Production and Operations Management*, 24(1):1–20.

Gupta, D. and Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40(9):800–819.

Haase, K. and Müller, S. (2013). Management of school locations allowing for free school choice. *Omega*, 41(5):847–855.

Haase, K. and Müller, S. (2015). Insights into clients choice in preventive health care facility location planning. *OR Spectrum*, 37(1):273–291.

Hanjoul, P. and Peeters, D. (1987). A facility location problem with clients' preference orderings. *Regional Science and Urban Economics*, 17(3):451–473.

Hans, E. W., van Houdenhouven, M., and Hulshof, P. J. H. (2012). A framework for health care planning and control. *International Series in Operations Research & Management Science (Hall RW, Ed)*, 168:303–320.

Hansen, P., Kochetov, Y., and Mladenovi, N. (2004). *Lower bounds for the uncapacitated facility location problem with user preferences*. GERAD, HEC Montréal.

Hardcastle, J. D., Chamberlain, J. O., Robinson, M. H., Moss, S. M., Amar, S. S., Balfour, T. W., James, P. D., and Mangham, C. M. (1996). Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. *The Lancet*, 348(9040):1472–1477.

Harrison, G. W., Shafer, A., and Mackay, M. (2005). Modelling variability in hospital bed occupancy. *Health Care Management Science*, 8(4):325–334.

Hauskrecht, M. (1997). Incremental methods for computing bounds in partially observable markov decision processes. In *AAAI/IAAI*, pages 734–739. Citeseer.

Heidenberger, K. (1996). Strategic investment in preventive health care: Quantitative modelling for programme selection and resource allocation. *Operations-Research-Spektrum*, 18(1):1–14.

Herrera, R., Kalcsics, J., and Nickel, S. (2008). *Reliability Models for the Uncapacitated Facility Location Problem with User Preferences*. Springer.

Heyman, D. P. and Sobel, M. (1984). *Stochastic Models in Operations Research, Volume II: Stochastic Optimization*. McGraw Hill.

Hickerson, B., Moore, A., Oakleaf, L., Edwards, M., James, P. a., Swanson, J., and Henderson, K. a. (2008). The role of a senior center in promoting physical activity for older adults. *Journal of Park & Recreation Administration*, 26(1):22–39.

Higginson, J. K. and Bookbinder, J. H. (1995). Markovian decision processes in shipment consolidation. *Transportation Science*, 29(3):242–255.

Howlader, N., Noone, A., Krapcho, M., Garshell, J., Miller, D., Altekruse, S., Kosary, C., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D., Chen, H., Feuer, E., and Cronin, K. (2015). SEER cancer statistics review, 1975-2012. Technical report, National Cancer Institute. Bethesda, MD.

Howlader, N., Noone, A., Krapcho, M., Miller, D., Bishop, K., Kosary, C., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D., Chen, H., Feuer, E., and Cronin, K. (2017). SEER cancer statistics review, 1975-2014. Technical report, National Cancer Institute. Bethesda, MD. https://seer.cancer.gov/csr/1975_2014/.

Hulshof, P. J., Kortbeek, N., Boucherie, R. J., Hans, E. W., and Bakker, P. J. (2012). 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.

Ishii, H., Lee, Y. L., and Yeh, K. Y. (2007). Fuzzy facility location problem with preference of candidate sites. *Fuzzy Sets and Systems*, 158(17):1922–1930.

Jarrett, G. P. (1998). Logistics in the health care industry. *International Journal of Physical Distribution & Logistics Management*, 28(9/10):741–772.

Jayaraman, V., Srivastava, R., and Benton, W. C. (1999). Supplier selection and order quantity

allocation: A Comprehensive model. *The Journal of Supply Chain Management*, 35(2):50–58.

Johnson, M. P., Gorr, W. L., and Roehrig, S. (2005). Location of service facilities for the elderly. *Annals of Operations Research*, 136(1):329–349.

Jun, J., Jacobson, S., and Swisher, J. (1999). Application of discrete-event simulation in health care clinics: A Survey. *Journal of the Operational Research Society*, pages 109–123.

Kim, J. J., Salomon, J. A., Weinstein, M. C., and Goldie, S. J. (2006). Packaging health services when resources are limited: The example of a cervical cancer screening visit. *PLoS Medicine*, 3(11):e434.

Kjeldsen, B., Kronborg, O., Fenger, C., and Jørgensen, O. (1997). A prospective randomized study of follow-up after radical surgery for colorectal cancer. *British Journal of Surgery*, 84(5):666–669.

Klabunde, C. N., Lanier, D., Nadel, M. R., McLeod, C., Yuan, G., and Vernon, S. W. (2009). Colorectal cancer screening by primary care physicians: Recommendations and practices, 2006–2007. *American Journal of Preventive Medicine*, 37(1):8–16.

Klafehn, K. A., Owens, D. L., Felter, R. A., Vonneman, N., and McKinnon, C. J. (1989). Evaluating the linkage between emergency medical services and the provision of scarce resources through simulation. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 335. American Medical Informatics Association.

Klose, A. and Drexl, A. (2005). Facility location models for distribution system design. *European Journal of Operational Research*, 162(1):4–29.

Kong, N., Schaefer, A. J., Hunsaker, B., and Roberts, M. S. (2010). Maximizing the efficiency of the us liver allocation system through region design. *Management Science*, 56(12):2111–2122.

Krist, A. H., Jones, R. M., Woolf, S. H., Woessner, S. E., Merenstein, D., Kerns, J. W., Foliaco, W., and Jackson, P. (2007). Timing of repeat colonoscopy: Disparity between guidelines and endoscopists recommendation. *American Journal of Preventive Medicine*, 33(6):471–478.

Kronborg, O., Fenger, C., Olsen, J., Jørgensen, O. D., and Søndergaard, O. (1996). Randomised study of screening for colorectal cancer with faecal-occult-blood test. *The Lancet*, 348(9040):1467–1471.

Ladabaum, U., Ferrandez, A., and Lanas, A. (2010). Cost-effectiveness of colorectal cancer

screening in high-risk spanish patients: Use of a validated model to inform public policy. *Cancer Epidemiology Biomarkers & Prevention*, 19(11):2765–2776.

Ladabaum, U. and Song, K. (2005). Projected national impact of colorectal cancer screening on clinical and economic outcomes and health services demand. *Gastroenterology*, 129(4):1151–1162.

Lane, D. C., Monefeldt, C., and Rosenhead, J. (2000). Looking in the wrong place for healthcare improvements: A System dynamics study of an accident and emergency department. *Journal of the Operational Research Society*, pages 518–531.

Lee, J. M. and Lee, Y. H. (2012). Facility location and scale decision problem with customer preference. *Computers & Industrial Engineering*, 63(1):184–191.

Leshno, M., Halpern, Z., and Arber, N. (2003). Cost-effectiveness of colorectal cancer screening in the average risk population. *Health Care Management Science*, 6(3):165–174.

Levin, B., Lieberman, D. A., McFarland, B., Smith, R. A., Brooks, D., Andrews, K. S., Dash, C., Giardiello, F. M., Glick, S., Levin, T. R., et al. (2008). Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A Joint guideline from the american cancer society, the us multi-society task force on colorectal cancer, and the american college of radiology. *CA: A Cancer Journal for Clinicians*, 58(3):130–160.

Li, L. X., Benton, W., and Leong, G. K. (2002). The impact of strategic operations management decisions on community hospital performance. *Journal of Operations Management*, 20(4):389–408.

Li, Q., Zeng, B., and Savachkin, A. (2013). Reliable facility location design under disruptions. *Computers & Operations Research*, 40(4):901–909.

Li, X. and Ouyang, Y. (2010). A continuum approximation approach to reliable facility location design under correlated probabilistic disruptions. *Transportation Research Part B: Methodological*, 44(4):535–548.

Lim, M., Daskin, M. S., Bassamboo, A., and Chopra, S. (2010). A facility reliability problem : Formulation , properties , and algorithm. *Naval Research Logistics*, 57(1):58–70.

Lim, M. K., Bassamboo, A., Chopra, S., and Daskin, M. S. (2013). Facility location decisions with random disruptions and imperfect estimation. *Manufacturing & Service Operations Management*, 15(2):239–249.

Loeve, F., Boer, R., van Oortmarssen, G. J., van Ballegooijen, M., and Habbema, J. D. F. (1999).

The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. *Computers and Biomedical Research*, 32(1):13–33.

Loeve, F., Boer, R., Zauber, A. G., van Ballegooijen, M., van Oortmarssen, G. J., Winawer, S. J., and Habbema, J. D. F. (2004). National polyp study data: evidence for regression of adenomas. *International journal of cancer*, 111(4):633–639.

Longo, W. E., Virgo, K. S., Johnson, F. E., Oprian, C. A., Vernava, A. M., Wade, T. P., Phelan, M. A., Henderson, W. G., Daley, J., and Khuri, S. F. (2000). Risk factors for morbidity and mortality after colectomy for colon cancer. *Diseases of the Colon & Rectum*, 43(1):83–91.

Lovejoy, W. S. (1991). Computationally feasible bounds for partially observed markov decision processes. *Operations Research*, 39(1):162–175.

Lowery, J. C. and Martin, J. B. (1991). Design and validation of a critical care simulation model. *Journal of the Society for Health Systems*, 3(3):15–36.

Lu, M., Ran, L., and Shen, Z.-j. M. (2015). Reliable Facility Location Design Under Uncertain Correlated Disruptions. *Manufacturing & Service Operations Management*, 17(4):445–455.

Lubin, M., Martin, K., Petra, C. G., and Sandıkçı, B. (2013). On parallelizing dual decomposition in stochastic integer programming. *Operations Research Letters*, 41(3):252–258.

Mahar, S., Bretthauer, K. M., and Salzarulo, P. A. (2011). Locating specialized service capacity in a multi-hospital network. *European Journal of Operational Research*, 212(3):596–605.

Maheshwari, S., Patel, T., and Patel, P. (2008). Screening for colorectal cancer in elderly persons who should we screen and when can we stop? *Journal of Aging and Health*, 20(1):126–139.

Maillart, L. M., Ivy, J. S., Ransom, S., and Diehl, K. (2008). Assessing dynamic breast cancer screening policies. *Operations Research*, 56(6):1411–1427.

Mandel, J. S., Church, T. R., Bond, J. H., Ederer, F., Geisser, M. S., Mongin, S. J., Snover, D. C., and Schuman, L. M. (2000). The effect of fecal occult-blood screening on the incidence of colorectal cancer. *New England Journal of Medicine*, 343(22):1603–1607.

Marić, M., Stanimirović, Z., and Milenković, N. (2012). Metaheuristic methods for solving the bilevel uncapacitated facility location problem with clients preferences. *Electronic Notes in Discrete Mathematics*, 39:43–50.

May, R. M. and Anderson, R. M. (1984). Spatial heterogeneity and the design of immunization programs. *Mathematical Biosciences*, 72(1):83–111.

McFarland, E. G., Levin, B., Lieberman, D. A., Pickhardt, P. J., Johnson, C. D., Glick, S. N., Brooks, D., and Smith, R. A. (2008). Revised colorectal screening guidelines: Joint effort of

the american cancer society, u.s. multisociety task force on colorectal cancer, and american college of radiology. *Radiology*, 248(3):717–720.

Medlock, J. and Galvani, A. P. (2009). Optimizing influenza vaccine distribution. *Science*, 325(5948):1705–1708.

Melo, M. T., Nickel, S., and Saldanha-da Gama, F. (2009). Facility location and supply chain management–A Review. *European Journal of Operational Research*, 196(2):401–412.

Mestre, A. M., Oliveira, M. D., and Barbosa-Póvoa, A. P. (2015). Location-allocation approaches for hospital network planning under uncertainty. *European Journal of Operational Research*, 240(3):791–806.

Molinari, N.-A. M., Ortega-Sanchez, I. R., Messonnier, M. L., Thompson, W. W., Wortley, P. M., Weintraub, E., and Bridges, C. B. (2007). The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine*, 25(27):5086–5096.

Müller, J. (1998). Optimal vaccination patterns in age-structured populations. *SIAM Journal on Applied Mathematics*, 59(1):222–241.

Müller, S., Haase, K., and Kless, S. (2009). A multiperiod school location planning approach with free school choice. *Environment and Planning A*, 41(12):2929–2945.

National Cancer Institute (2016). Tests to Detect Colorectal Cancer and Polyps. http://www.cancer.gov/types/colorectal/screening-fact-sheet. [Online; accessed 13-May-2017].

National Council on Aging (2015). Senior centers fact sheet. https://www.ncoa.org/news/resources-for-reporters/get-the-facts/senior-center-facts/. [Online; accessed 20-Jun-2016].

Ness, R. M., Holmes, A. M., Klein, R., and Dittus, R. (2000). Cost-utility of one-time colonoscopic screening for colorectal cancer at various ages. *The American Journal of Gastroenterology*, 95(7):1800–1811.

Ohlsson, B. and Pålsson, B. (2003). Follow-up after colorectal cancer surgery. *Acta Oncologica*, 42(8):816–826.

OPTN (2014). OPTN/SRTR 2012 annual data report. Technical report, Organ Procurement and Transplantation Network. Rockville, MD.

Owen, S. H. and Daskin, M. S. (1998). Strategic facility location: A Review. *European Journal of Operational Research*, 111(3):423–447.

Park, I., Yu, C., Kim, H., Jung, Y., Han, K., and Kim, J. (2006). Metachronous colorectal cancer. *Colorectal Disease*, 8(4):323–327.

Patel, P. and De, P. (2016). Trends in colorectal cancer incidence and related lifestyle risk factors in 15-49-year-olds in canada, 1969-2010. *Cancer Epidemiology*, 42:90–100.

Pecanha, S. and Wallace, T. (2015). The flight of refugees around the globe. *The New York Times*.

Peng, P., Snyder, L. V., Lim, A., and Liu, Z. (2011). Reliable logistics networks design with facility disruptions. *Transportation Research Part B: Methodological*, 45(8):1190–1211.

Pfister, D. G., Benson III, A. B., and Somerfield, M. R. (2004). Surveillance strategies after curative treatment of colorectal cancer. *New England Journal of Medicine*, 350(23):2375–2382.

Pierskalla, W. P. and Brailer, D. J. (1994). Applications of operations research in health care delivery. *Handbooks in Operations Research and Management Science*, 6:469–505.

Pignone, M., Rich, M., Teutsch, S. M., Berg, A. O., and Lohr, K. N. (2002). Screening for colorectal cancer in adults at average risk: A Summary of the evidence for the us preventive services task force. *Annals of internal medicine*, 137(2):132–141.

Poulin, E. (2003). Benchmarking the hospital logistics process a potential cure for the ailing health care sector. *CMA Management*, 77(1):20–23.

Powell, W. B. and Topaloglu, H. (2006). Approximate dynamic programming for large-scale resource allocation problems. In *Models, Methods, and Applications for Innovative Decision Making*, pages 123–147. INFORMS.

Qi, L., Shen, Z.-J. M., and Snyder, L. V. (2010). The effect of supply disruptions on supply chain design decisions. *Transportation Science*, 44(2):274–289.

Rais, A. and Viana, A. (2011). Operations research in healthcare: A Survey. *International Transactions in Operational Research*, 18(1):1–31.

Ramsey, S. D., Wilschut, J., Boer, R., and van Ballegooijen, M. (2010). A decision-analytic evaluation of the cost-effectiveness of family history–based colorectal cancer screening programs. *The American Journal of Gastroenterology*, 105(8):1861–1869.

Regueiro, C. R. (2005). Aga future trends committee report: Colorectal cancer: A Qualitative review of emerging screening and diagnostic technologies. *Gastroenterology*, 129(3):1083–1103.

ReVelle, C. S. and Eiselt, H. A. (2005). Location analysis: A Synthesis and survey. *European Journal of Operational Research*, 165(1):1–19.

Revelle, C. S., Eiselt, H. A., and Daskin, M. S. (2008). A bibliography for some fundamental problem categories in discrete location science. *European Journal of Operational Research*, 184(3):817–848.

Rex, D. K., Johnson, D. A., Lieberman, D. A., Burt, R. W., and Sonnenberg, A. (2000). Colorectal cancer prevention 2000: Screening recommendations of the american college of gastroenterology. *The American Journal of Gastroenterology*, 95(4):868–877.

Rex, D. K., Schoenfeld, P. S., Cohen, J., Pike, I. M., Adler, D. G., Fennerty, M. B., Lieb, J. G., Park, W. G., Rizk, M. K., Sawhney, M. S., et al. (2015). Quality indicators for colonoscopy. *The American journal of gastroenterology*, 110(1):72.

Roberts, S., Wang, L., Klein, R., Ness, R., and Dittus, R. (2007). Development of a simulation model of colorectal cancer. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 18(1):4.

Royston, G. (2009). One hundred years of operational research in health. *Journal of the Operational Research Society, pages S169–S179*, page 17.

Sandikci, B., Maillart, L. M., Schaefer, A. J., Alagoz, O., and Roberts, M. S. (2008). Estimating the patient's price of privacy in liver transplantation. *Operations Research*, 56(6):1393–1410.

Sandikci, B., Maillart, L. M., Schaefer, A. J., and Roberts, M. S. (2013). Alleviating the patient's price of privacy through a partially observable waiting list. *Management Science*, 59(8):1836–1854.

Scaparra, M. P. and Church, R. L. (2008). A bilevel mixed-integer program for critical infrastructure protection planning. *Computers & Operations Research*, 35(6):1905–1923.

Schoen, R. E. (2006). Debate: Should screening colonoscopy be performed on an 88-yr-old healthy patient&quest. *The American Journal of Gastroenterology*, 101(8):1713–1715.

Schoen, R. E., Pinsky, P. F., Weissfeld, J. L., Yokochi, L. A., Reding, D. J., Hayes, R. B., Church, T., Yurgalevich, S., Doria-Rose, V. P., Hickey, T., et al. (2010). Utilization of surveillance colonoscopy in community practice. *Gastroenterology*, 138(1):73–81.

Scholefield, J. and Steele, R. (2002). Guidelines for follow up after resection of colorectal cancer. *Gut*, 51(suppl 5):v3–v5.

Seeff, L. C., Richards, T. B., Shapiro, J. A., Nadel, M. R., Manninen, D. L., Given, L. S., Dong, F. B., Winges, L. D., and McKenna, M. T. (2004). How many endoscopies are performed for colorectal cancer screening? Results from CDCs survey of endoscopic capacity. *Gastroenterology*, 127(6):1670–1677.

Selby, J. V., Friedman, G. D., Quesenberry Jr, C. P., and Weiss, N. S. (1992). A case–control study of screening sigmoidoscopy and mortality from colorectal cancer. *New England Journal of Medicine*, 326(10):653–657.

Shechter, S. M., Bryce, C. L., Alagoz, O., Kreke, J. E., Stahl, J. E., Schaefer, A. J., Angus, D. C., and Roberts, M. S. (2005). A clinically based discrete-event simulation of end-stage liver disease and the organ allocation process. *Medical Decision Making*, 25(2):199–209.

Shen, Z.-J. M., Zhan, R. L., and Zhang, J. (2011). The reliable facility location problem: Formulations, heuristics, and approximation algorithms. *INFORMS Journal on Computing*, 23(3):470–482.

Simmang, C. L., Senatore, P., Lowry, A., Hicks, T., Burnstein, M., Dentsman, F., Fazio, V., Glennon, E., Hyman, N., Kerner, B., et al. (1999). Practice parameters for detection of colorectal neoplasms: The standards committee, the american society of colon and rectal surgeons. *Diseases of the Colon & Rectum*, 42(9):1123–1129.

Smith, R. A., Manassaram-Baptiste, D., Brooks, D., Doroshenk, M., Fedewa, S., Saslow, D., Brawley, O. W., and Wender, R. (2015). Cancer screening in the united states, 2015: A Review of current american cancer society guidelines and current issues in cancer screening. *CA: A Cancer Journal for Clinicians*, 65(1):30–54.

Smith, R. A., von Eschenbach, A. C., Wender, R., Levin, B., Byers, T., Rothenberger, D., Brooks, D., Creasman, W., Cohen, C., Runowicz, C., et al. (2001). American cancer society guidelines for the early detection of cancer: Update of early detection guidelines for prostate, colorectal, and endometrial cancers. *CA: A Cancer Journal for Clinicians*, 51(1):38–75.

Snyder, L. and Daskin, M. S. (2006). Stochastic p-robust location problems. *IIE Transactions*, 38(11):971–985.

Snyder, L. V., Atan, Z., Peng, P., Rong, Y., Schmitt, A. J., and Sinsoysal, B. (2016). Or/ms models for supply chain disruptions: A Review. *IIE Transactions*, 48(2):89–109.

Snyder, L. V. and Daskin, M. S. (2005). Reliability models for facility location: The expected failure cost case. *Transportation Science*, 39(3):400–416.

Stevenson, C. (1995). Statistical models for cancer screening. *Statistical Methods in Medical Research*, 4(1):18–32.

Stinnett, A. A. and Paltiel, A. D. (1996). Mathematical programming for the efficient allocation of health care resources. *Journal of Health Economics*, 15(5):641–653.

Syam, S. S. and Côté, M. J. (2010). A location-allocation model for service providers with application to not-for-profit health care organizations. *Omega*, 38(3-4):157–166.

Talbot, H. K., Zhu, Y., Chen, Q., Williams, J. V., Thompson, M. G., and Griffin, M. R. (2013). Effectiveness of influenza vaccine for preventing laboratory-confirmed influenza hospitalizations in adults, 2011-2012 influenza season. *Clinical Infectious Diseases*.

Teng, J., Zhang, B., Bai, X., Yang, Z., and Xuan, D. (2014). Incentive-driven and privacy-preserving message dissemination in large-scale mobile networks. *IEEE Transactions on Parallel and Distributed Systems*, 25(11):2909–2919.

Teo, C.-P. and Shu, J. (2004). Warehouse-retailer network design problem. *Operations Research*, 52(3):396–408.

The Economist (2004). A survery of health care finances. *The Economist*, 8384:3–14.

Uribe-Sánchez, A., Savachkin, A., Santana, A., Prieto-Santa, D., and Das, T. K. (2011). A predictive decision-aid methodology for dynamic mitigation of influenza pandemics. *OR spectrum*, 33(3):751–786.

US Preventive Services Task Force (2002). Screening for colorectal cancer: Recommendation and rationale. *Annals of Internal Medicine*, 137(2):129.

US Preventive Services Task Force (2008). Screening for colorectal cancer: US preventive services task force recommendation statement. *Annals of Internal Medicine*, 149(9).

US Preventive Services Task Force (2016). Screening for colorectal cancer: US preventive services task force recommendation statement. *JAMA*, 315(23):2564–2575.

Uzunlar, O., Ceyhan, M. E., Benneyan, J. C., Watts, B. V., and Shiner, B. (2012). Optimal longitudinal relocation of colonoscopy services within the veterans health administration. In *IIE Annual Conference Proceedings*, page 1. Institute of Industrial Engineers-Publisher.

Van Zon, A. and Kommer, G. (1999). Patient flows and optimal health-care resource allocation at the macro-level: A Dynamic linear programming approach. *Health Care Management Science*, 2(2):87–96.

Vasilyev, I. and Klimentova, K. (2010). The branch and cut method for the facility location problem with clients preferences. *Journal of Applied and Industrial Mathematics*, 4(3):441–454.

Vasilev, I., Klimentova, K., and Kochetov, Y. A. (2009). New lower bounds for the facility location problem with clients preferences. *Computational Mathematics and Mathematical Physics*, 49(6):1010–1020.

Verma, A. and Gaukler, G. M. (2015). Pre-positioning disaster response facilities at safe locations: An evaluation of deterministic and stochastic modeling approaches. *Computers & Operations Research*, 62:197–209.

Verter, V. and Lapierre, S. D. (2002). Location of preventive health care facilities. *Annals of Operations Research*, 110(1-4):123–132.

Verter, V. and Zhang, Y. (2015). Location models for preventive care. In *Applications of Location Analysis*, pages 223–241. Springer.

Vidyarthi, N. and Kuzgunkaya, O. (2015). The impact of directed choice on the design of preventive healthcare facility network under congestion. *Health Care Management Science*, 18(4):459–474.

Vijan, S., Hwang, I., Inadomi, J., Wong, R. K., Choi, J. R., Napierkowski, J., Koff, J. M., and Pickhardt, P. J. (2007). The cost-effectiveness of ct colonography in screening for colorectal neoplasia. *The American Journal of Gastroenterology*, 102(2):380–390.

Vijan, S., Inadomi, J., Hayward, R., Hofer, T., and Fendrick, A. (2004). Projections of demand and capacity for colonoscopy related to increasing rates of colorectal cancer screening in the united states. *Alimentary Pharmacology & Therapeutics*, 20(5):507–515.

Wagner, S. L., Shubair, M. M., and Michalos, A. C. (2010). Surveying older adults opinions on housing: Recommendations for policy. *Social Indicators Research*, 99(3):405–412.

Wilschut, J. A., Steyerberg, E. W., van Leerdam, M. E., Lansdorp-Vogelaar, I., Habbema, J. D. F., and van Ballegooijen, M. (2011). How much colonoscopy screening should be recommended to individuals with various degrees of family history of colorectal cancer? *Cancer*, 117(18):4166–4174.

Winawer, S., Fletcher, R., Rex, D., Bond, J., Burt, R., Ferrucci, J., Ganiats, T., Levin, T., Woolf, S., Johnson, D., et al. (2003). Colorectal cancer screening and surveillance: Clinical guidelines and rationaleupdate based on new evidence. *Gastroenterology*, 124(2):544–560.

Winawer, S. J. (2007). Colorectal cancer screening. *Best Practice & Research Clinical Gastroenterology*, 21(6):1031–1048.

Winawer, S. J., Fletcher, R. H., Miller, L., Godlee, F., Stolar, M., Mulrow, C., Woolf, S., Glick, S., Ganiats, T., Bond, J., et al. (1997). Colorectal cancer screening: Clinical guidelines and rationale. *Gastroenterology*, 112(2):594–642.

Winawer, S. J., Schottenfeld, D., and Flehinger, B. J. (1991). Colorectal cancer screening. *Journal of the National Cancer Institute*, 83(4):243–53.

Xu, J., Murphy, S. L., Kochanek, K. D., and Bastian, B. (2016). US deaths : Final data for 2013. *National Vital Statistics Reports-National Center for Health Statistics*, 64(2).

Yaesoubi, R. and Cohen, T. (2011). Generalized markov models of infectious disease spread: A Novel framework for developing dynamic health policies. *European Journal of Operational Research*, 215(3):679–687.

Yang, G., Yu, H., Zheng, S., Zheng, W., Shu, X.-O., Sun, Q.-R., Li, W.-D., Shen, G.-F., Shen, Y.-Z., and Potter, J. D. (1998). Pathologic features of initial adenomas as predictors for metachronous adenomas of the rectum. *Journal of the National Cancer Institute*, 90(21):1661–1665.

Yun, H., Lee, L., Park, J., Cho, Y., Cho, Y., Lee, W., Kim, H., Chun, H., and Yun, S. (2008). Local recurrence after curative resection in patients with colon and rectal cancers. *International Journal of Colorectal Disease*, 23(11):1081–1087.

Zaric, G. S. and Brandeau, M. L. (2001). Optimal investment in a portfolio of HIV prevention programs. *Medical Decision Making*, 21(5):391–408.

Zaric, G. S. and Brandeau, M. L. (2002). Dynamic resource allocation for epidemic control in multiple populations. *Mathematical Medicine and Biology*, 19(4):235–255.

Zauber, A. G., Lansdorp-Vogelaar, I., Knudsen, A. B., Wilschut, J., van Ballegooijen, M., and Kuntz, K. M. (2008). Evaluating test strategies for colorectal cancer screening: A Decision analysis for the us preventive services task force. *Annals of Internal Medicine*, 149(9):659–669.

Zhang, J., Denton, B. T., Balasubramanian, H., Shah, N. D., and Inman, B. A. (2012a). Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management*, 14(4):529–547.

Zhang, J., Denton, B. T., Balasubramanian, H., Shah, N. D., and Inman, B. A. (2012b). Optimization of PSA screening policies a comparison of the patient and societal perspectives. *Medical Decision Making*, 32(2):337–349.

Zhang, Y., Berman, O., Marcotte, P., and Verter, V. (2010). A bilevel model for preventive healthcare facility network design with congestion. *IIE Transactions*, 42(12):865–880.

Zhang, Y., Berman, O., and Verter, V. (2009). Incorporating congestion in preventive healthcare facility network design. *European Journal of Operational Research*, 198(3):922–935.

Zhang, Y., Berman, O., and Verter, V. (2012c). The impact of client choice on preventive healthcare facility network design. *OR Spectrum*, 34(2):349–370.

APPENDICES

Appendix A

Generating Customer Preferences in Section 2.7

This appendix describes how the preferences of customers are generated for the problem instances considered in computational studies in Section 2.7. We use a method from Cánovas et al. (2007) to generate preferences randomly based on service costs d_{ij} with some rationality. For example, if d_{i1} and d_{i2} are the two lowest costs facilities for customer i and there is a significant gap up to the third lowest cost, it is likely that one of these two facilities will be the most preferred by this customer. Moreover, in this method the bigger (smaller) the differences among costs are, the easier (more difficult) for the customer to decide which facility is more attractive. We describe the outline of the procedure.

- Generate fake costs $\tilde{d}_{ij}$ for each pair (i, j) from triangular probability distribution. Let $m_i = \min_{j \in \mathcal{J}}\{d_{ij}\}$ and $M_i = \max_{j \in \mathcal{J}}\{d_{ij}\}$. The triangular distribution is defined over $[m_i, M_i]$ and d_{ij} is the peak point.
- Order this fake cost $\{\tilde{d}_{ij}\}_j$ for each customer i. Then, facility j_1 with the lowest value $\tilde{d}_{ij_1}$ will be the most preferred for customer i and so on until the highest fake cost for the least preferred facility.

Appendix B

Detailed Transition Equations of the MIP Model

The explicit formulation of equation (3.20) are given here.

$$X^0_{j,k,h}(t+1) = \left(\theta^L_{j,k,\underline{h}} a^L_{j,k,\underline{h}} \left(1-\delta^{0,cl}_{j,T\text{-}}\right) + \left[1-\theta^L_{j,k,\underline{h}} a^L_{j,k,\underline{h}}\right]\left(1-\delta^{0,dn}_{j,T\text{-}}\right)\right) \rho^{0,0}_j \widetilde{X}^0_{j,k,\underline{h}}(t)$$

$$X^1_{j,k,h}(t+1) = (1-\tau_P)\theta^L_{j,k,\underline{h}} a^L_{j,k,\underline{h}} \sum_{i\in\{0,1\}} \left(\rho^{i,1}_j \left(1-\delta^{i,cl}_{j,T\text{-}}\right) \widetilde{X}^i_{j,k,\underline{h}}(t)\right) + \left[1-\theta^L_{j,k,\underline{h}} a^L_{j,k,\underline{h}}\right] \sum_{i\in\{0,1\}} \rho^{i,1}_j \left(1-\delta^{i,dn}_{j,T\text{-}}\right) \widetilde{X}^i_{j,k,\underline{h}}(t)$$

$$\begin{aligned} X^2_{j,k,h}(t+1) = {} & (1-\tau^{cl}_C)(1-\omega_{cl})\theta^L_{j,k,\underline{h}} a^L_{j,k,\underline{h}} \left(\rho^{1,2}_j \left(1-\delta^{1,cl}_{j,T\text{-}}\right) \widetilde{X}^1_{j,k,\underline{h}}(t) + \left(1-\delta^{2,cl}_{j,T\text{-}}\right) \widetilde{X}^2_{j,k,\underline{h}}(t)\right) \\ & +(1-\omega_{dn}) \left[1-\theta^L_{j,k,\underline{h}} a^L_{j,k,\underline{h}}\right] \left(\rho^{1,2}_j \left(1-\delta^{1,dn}_{j,T\text{-}}\right) \widetilde{X}^1_{j,k,\underline{h}}(t) + \left(1-\delta^{2,dn}_{j,T\text{-}}\right) \widetilde{X}^2_{j,k,\underline{h}}(t)\right) \end{aligned}$$

$$\begin{aligned} X^3_{j,k,h}(t+1) = {} & \left(\theta^H_{j,k,\underline{h}} a^H_{j,k,\underline{h}} \left(1-\delta^{3,cl}_{j,T\text{-}}\right) + \left[1-\theta^H_{j,k,\underline{h}} a^H_{j,k,\underline{h}}\right]\left(1-\delta^{3,dn}_{j,T\text{-}}\right)\right) \rho^{3,3}_j \widetilde{X}^3_{j,k,\underline{h}}(t) \\ & +h_0\tau_P\rho^{3,3}_j \sum_{(i,R)\in\{(1,L),(4,H)\}} \left(1-\delta^{i,cl}_{j,P+}\right) \sum_{\hat{h}} \theta^R_{j,k,\hat{h}} a^R_{j,k,\hat{h}} \widetilde{X}^i_{j,k,\hat{h}}(t) \end{aligned}$$

$$\begin{aligned} X^4_{j,k,h}(t+1) = {} & (1-\tau_P)\theta^H_{j,k,\underline{h}} a^H_{j,k,\underline{h}} \sum_{i\in\{3,4\}} \left(\rho^{i,4}_j \left(1-\delta^{i,cl}_{j,T\text{-}}\right) \widetilde{X}^i_{j,k,h-}(t)\right) + \left[1-\theta^H_{j,k,\underline{h}} a^H_{j,k,\underline{h}}\right] \sum_{i\in\{3,4\}} \left(\rho^{i,4}_j \left(1-\delta^{i,dn}_{j,T\text{-}}\right) \widetilde{X}^i_{j,k,\underline{h}}(t)\right) \\ & +h_0\tau_P\rho^{3,4}_j \sum_{(i,R)\in\{(1,L),(4,H)\}} \left(1-\delta^{i,cl}_{j,P+}\right) \sum_{\hat{h}} \theta^R_{j,k,\hat{h}} a^R_{j,k,\hat{h}} \widetilde{X}^i_{j,k,\hat{h}}(t) \end{aligned}$$

$$
\begin{aligned}
X^{5}_{j,k,h}(t+1) = \quad & (1-\tau^{cl}_{C})(1-\omega_{cl})\theta^{H}_{j,k,\underline{h}}a^{H}_{j,k,\underline{h}}\left(\rho^{4,5}_{j}\left(1-\delta^{4,cl}_{j,T\text{-}}\right)\widetilde{X}^{4}_{j,k,\underline{h}}(t)+\left(1-\delta^{cl}_{5,T\text{-}}\right)\widetilde{X}^{5}_{j,k,\underline{h}}(t)\right)\\
& +(1-\omega_{dn})\left[1-\theta^{H}_{j,k,\underline{h}}a^{H}_{j,k,\underline{h}}\right]\left(\rho^{4,5}_{j}\left(1-\delta^{4,dn}_{j,T\text{-}}\right)\widetilde{X}^{4}_{j,k,\underline{h}}(t)+\left(1-\delta^{dn}_{5,T\text{-}}\right)\widetilde{X}^{5}_{j,k,\underline{h}}(t)\right)
\end{aligned}
$$

$$
\begin{aligned}
X^{6}_{j,k,h}(t+1) = \quad & \left(\theta^{P}_{j,k,\underline{h}}a^{P}_{j,k,\underline{h}}\left(1-\delta^{6,cl}_{j,T\text{-}}\right)+\left[1-\theta^{P}_{j,k,\underline{h}}a^{P}_{j,k,\underline{h}}\right]\left(1-\delta^{6,dn}_{j,T\text{-}}\right)\right)\rho^{6,6}_{j}\widetilde{X}^{6}_{j,k,\underline{h}}(t)\\
& +\left(\theta^{U}_{j,k,\underline{h}}a^{U}_{j,k,\underline{h}}\gamma_{9,j}\left(1-\delta^{9,cl}_{j,T\text{-}}\right)+\left[1-\theta^{U}_{j,k,\underline{h}}a^{U}_{j,k,\underline{h}}\right]\left(1-\delta^{9,dn}_{j,T\text{-}}\right)\right)\rho^{6,6}_{j}\widetilde{X}^{9}_{j,k,\underline{h}}(t)+h_{0}\tau_{P}\rho^{6,7}_{j}\left(1-\delta^{7,cl}_{j,P+}\right)\sum_{\hat{h}}\theta^{P}_{j,k,\hat{h}}a^{P}_{j,k,\hat{h}}\widetilde{X}^{7}_{j,k,\hat{h}}(t)\\
& +h_{0}\sum_{\hat{a}\in\{cl,dn\}}\sum_{o\in\{C+,SD\}}\tau^{\hat{a}}_{o}\sum_{(i,R)\in\{(2,L),(5,H),(8,P)\}}\rho^{i,6}_{j}\gamma_{i,j}\left(1-\delta^{i,\hat{a}}_{j,o}\right)\sum_{\hat{h}}\theta^{R}_{j,k,\hat{h}}a^{R}_{j,k,\hat{h}}\widetilde{X}^{i}_{j,k,\hat{h}}(t)
\end{aligned}
$$

$$
\begin{aligned}
X^{7}_{j,k,h}(t+1) = \quad & (1-\tau_{P})\quad\left(\sum_{i\in\{6,7\}}\theta^{P}_{j,k,\underline{h}}a^{P}_{j,k,\underline{h}}\rho^{i,7}_{j}\left(1-\delta^{i,cl}_{j,T\text{-}}\right)\widetilde{X}^{i}_{j,k,\underline{h}}(t)+\theta^{U}_{j,k,\underline{h}}a^{U}_{j,k,\underline{h}}\rho^{6,7}_{j}\gamma_{9,j}\left(1-\delta^{9,cl}_{j,T\text{-}}\right)\widetilde{X}^{9}_{j,k,\underline{h}}(t)\right)\\
& +\left[1-\theta^{P}_{j,k,\underline{h}}a^{P}_{j,k,\underline{h}}\right]\sum_{i\in\{6,7\}}\rho^{i,7}_{j}\left(1-\delta^{i,dn}_{j,T\text{-}}\right)\widetilde{X}^{i}_{j,k,\underline{h}}(t)+\left[1-\theta^{U}_{j,k,\underline{h}}a^{U}_{j,k,\underline{h}}\right]\rho^{6,7}_{j}\gamma_{9,j}\left(1-\delta^{9,dn}_{j,T\text{-}}\right)\widetilde{X}^{9}_{j,k,\underline{h}}(t)\\
& +h_{0}\tau_{P}\theta^{P}_{j,k,h}a^{P}_{j,k,h}\rho^{6,7}_{j}\left(1-\delta^{7,cl}_{j,P+}\right)\widetilde{X}^{7}_{j,k,h}(t)\\
& +h_{0}\sum_{\hat{a}\in\{cl,dn\}}\sum_{o\in\{C+,SD\}}\tau^{\hat{a}}_{o}\sum_{(i,R)\in\{(2,L),(5,H),(8,P)\}}\rho^{i,7}_{j}\gamma_{i,j}\left(1-\delta^{i,\hat{a}}_{j,o}\right)\sum_{\hat{h}}\theta^{R}_{j,k,\hat{h}}a^{R}_{j,k,\hat{h}}\widetilde{X}^{i}_{j,k,\hat{h}}(t)
\end{aligned}
$$

$$
\begin{aligned}
X^{8}_{j,k,h}(t+1) = \quad & (1-\tau_{P})\quad\theta^{P}_{j,k,\underline{h}}a^{P}_{j,k,\underline{h}}\left(\rho^{7,8}_{j}\left(1-\delta^{7,cl}_{j,T\text{-}}\right)\widetilde{X}^{7}_{j,k,\underline{h}}(t)+\left(1-\delta^{8,cl}_{j,T\text{-}}\right)\widetilde{X}^{8}_{j,k,\underline{h}}(t)\right)\\
& +\left[1-\theta^{P}_{j,k,\underline{h}}a^{P}_{j,k,\underline{h}}\right]\left(\rho^{7,8}_{j}\left(1-\delta^{dn}_{7,T\text{-}}\right)\widetilde{X}^{7}_{j,k,\underline{h}}(t)+\left(1-\delta^{8,dn}_{j,T\text{-}}\right)\widetilde{X}^{8}_{j,k,\underline{h}}(t)\right)\\
& +h_{0}\sum_{\hat{a}\in\{cl,dn\}}\sum_{o\in\{C+,SD\}}\tau^{\hat{a}}_{o}\sum_{(i,R)\in\{(2,L),(5,H),(8,P)\}}\rho^{i,8}_{j}\gamma_{i,j}\left(1-\delta^{i,\hat{a}}_{j,o}\right)\sum_{\hat{h}}\theta^{R}_{j,k,\hat{h}}a^{R}_{j,k,\hat{h}}\widetilde{X}^{i}_{j,k,\hat{h}}(t)
\end{aligned}
$$

$$
\begin{aligned}
X^{9}_{j,k,h}(t+1) = \quad & \left(\theta^{U}_{j,k,\underline{h}}a^{U}_{j,k,\underline{h}}\left(1-\delta^{9,cl}_{j,T\text{-}}\right)+\left[1-\theta^{U}_{j,k,\underline{h}}a^{U}_{j,k,\underline{h}}\right]\left(1-\delta^{9,dn}_{j,T\text{-}}\right)\right)(1-\gamma_{9,j})\widetilde{X}^{9}_{j,k,\underline{h}}(t)\\
& +h_{0}\sum_{\hat{a}\in\{cl,dn\}}\sum_{o\in\{C+,SD\}}\tau^{\hat{a}}_{o}\sum_{(i,R)\in\{(2,L),(5,H),(8,P)\}}(1-\gamma_{i,j})\left(1-\delta^{i,\hat{a}}_{j,o}\right)\sum_{\hat{h}}\theta^{R}_{j,k,\hat{h}}a^{R}_{j,k,\hat{h}}\widetilde{X}^{i}_{j,k,\hat{h}}(t)
\end{aligned}
$$

$$
X^{10}_{j,k,h}(t+1) = \quad \sum_{\hat{a}}\sum_{o}\sum_{i}\delta^{i,\hat{a}}_{j,o}
$$

Appendix C

Parameter Tables

C.1 $f\left(o|s^i_{j,k,h},\hat{a}\right)$

The values of the observation rates are derived from (Erenay et al., 2014).

Table C.1: $f\left(o|s^i_{j,k,h},\hat{a}\right)$ Values

	T-	*P+*	*C+*	*SD*
$s^0_{j,k,h},\ \forall j,k,h$	1	0	0	0
$s^1_{j,k,h},\ \forall j,k,h$	$1-\tau^{\hat{a}}_P$	$\tau^{\hat{a}}_P$	0	0
$s^2_{j,k,h},\ \forall j,k,h$	$(1-\tau^{\hat{a}}_C)(1-\omega_{\hat{a}})$	0	$\tau^{\hat{a}}_C$	$(1-\tau^{\hat{a}}_C)\omega_{\hat{a}}$
$s^3_{j,k,h},\ \forall j,k,h$	1	0	0	0
$s^4_{j,k,h},\ \forall j,k,h$	$1-\tau^{\hat{a}}_P$	$\tau^{\hat{a}}_P$	0	0
$s^5_{j,k,h},\ \forall j,k,h$	$(1-\tau^{\hat{a}}_C)(1-\omega_{\hat{a}})$	0	$\tau^{\hat{a}}_C$	$(1-\tau^{\hat{a}}_C)\omega_{\hat{a}}$
$s^6_{j,k,h},\ \forall j,k,h$	1	0	0	0
$s^7_{j,k,h},\ \forall j,k,h$	$1-\tau^{\hat{a}}_P$	$\tau^{\hat{a}}_P$	0	0
$s^8_{j,k,h},\ \forall j,k,h$	$(1-\tau^{\hat{a}}_C)(1-\omega_{\hat{a}})$	0	$\tau^{\hat{a}}_C$	$(1-\tau^{\hat{a}}_C)\omega_{\hat{a}}$
$s^U_{j,k,h},\ \forall j,k,h$	1	0	0	0
$s^D_{j,k,h},\ \forall j,k,h$	1	0	0	0

Table C.2: $\tau_o^{\hat{a}}$ Values

$\hat{a}$	$C+$	SD
cl	τ_C	$(1-\tau_C)\omega_{cl}$
dn	0	ω_{dn}

C.2 $p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h},\hat{a},o\right)$

The following represent the values of $p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h},\hat{a},o\right)$. To ease demonstration, the following notation will be used. When an observation at state $s^{i}_{j,k,h}$ results in $o = T$- and the no polyp is formed, the system state will transfer to state $s^{i}_{j,k,h+}$ for given R, i, j, k and h. In this context, $h+$ means the immediate succession of the value of h. If h assumes the maximum value (in this case, $h = 2$), then $h+$ would simply mean that h remains the same $h = 2$. This would allow to have only two 'categories' of $s^{i}_{j,k,h}$ for particular R, i, j, k; namely $s^{i}_{j,k,0}$ and $s^{i}_{j,k,h+}$.

Tables (C.3 - C.6) show the values of $p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h},\hat{a},T-\right)$. The values missing from these tables are $p\left(s^{D}_{j',k',h'}|s^{D}_{j,k,h},\hat{a},T-\right) = 1,\ \forall j,k,h.$

Table C.3: $p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h},\hat{a},o\right)$ Values for $R = LR$ and $o = T-$

	$s^{0}_{j,k,0}$	$s^{0}_{j,k,h+}$	$s^{1}_{j,k,0}$	$s^{1}_{j,k,h+}$	$s^{2}_{j,k,0}$	$s^{2}_{j,k,h+}$	$s^{10}_{j,k,h+}$
$s^{0}_{j,k,h}$	0	$\rho^{0,0}_{j}\left[1-\delta^{0,j}_{\hat{a},T\text{-}}\right]$	0	$\rho^{1,0}_{j}\left[1-\delta^{0,j}_{\hat{a},T\text{-}}\right]$	0	0	$\delta^{0,j}_{\hat{a},T\text{-}}$
$s^{1}_{j,k,h}$	0	0	0	$\rho^{1,1}_{j}\left[1-\delta^{1,j}_{\hat{a},T\text{-}}\right]$	0	$\rho^{2,1}_{j}\left[1-\delta^{1,j}_{\hat{a},T\text{-}}\right]$	$\delta^{1,j}_{\hat{a},T\text{-}}$
$s^{2}_{j,k,h}$	0	0	0	0	0	$1-\delta^{2,j}_{\hat{a},T\text{-}}$	$\delta^{2,j}_{\hat{a},T\text{-}}$

Table C.4: $p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h},\hat{a},o\right)$ Values for $R = HR$ and $o = T-$

	$s^3_{j,k,0}$	$s^3_{j,k,h+}$	$s^4_{j,k,0}$	$s^4_{j,k,h+}$	$s^5_{j,k,0}$	$s^5_{j,k,h+}$	$s^{10}_{j,k,h+}$
$s^3_{j,k,h}$	0	$\rho^{3,3}_j\left[1-\delta^{3,j}_{\hat{a},T-}\right]$	0	$\rho^{4,3}_j\left[1-\delta^{3,j}_{\hat{a},T-}\right]$	0	0	$\delta^{3,j}_{\hat{a},T-}$
$s^4_{j,k,h}$	0	0	0	$\rho^{4,4}_j\left[1-\delta^{4,j}_{\hat{a},T-}\right]$	0	$\rho^{5,4}_j\left[1-\delta^{4,j}_{\hat{a},T-}\right]$	$\delta^{4,j}_{\hat{a},T-}$
$s^5_{j,k,h}$	0	0	0	0	0	$1-\delta^{5,j}_{\hat{a},T-}$	$\delta^{5,j}_{\hat{a},T-}$

Table C.5: $p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h},\hat{a},o\right)$ Values for $R = PC$ and $o = T-$

	$s^6_{j,k,0}$	$s^6_{j,k,h+}$	$s^7_{j,k,0}$	$s^7_{j,k,h+}$	$s^8_{j,k,0}$	$s^8_{j,k,h+}$	$s^{10}_{j,k,h+}$
$s^6_{j,k,h}$	0	$\rho^{6,6}_j\left[1-\delta^{6,j}_{\hat{a},T-}\right]$	0	$\rho^{7,6}_j\left[1-\delta^{6,j}_{\hat{a},T-}\right]$	0	0	$\delta^{6,j}_{\hat{a},T-}$
$s^7_{j,k,h}$	0	0	0	$\rho^{7,7}_j\left[1-\delta^{7,j}_{\hat{a},T-}\right]$	0	$\rho^{8,7}_j\left[1-\delta^{7,j}_{\hat{a},T-}\right]$	$\delta^{7,j}_{\hat{a},T-}$
$s^8_{j,k,h}$	0	0	0	0	0	$1-\delta^{8,j}_{\hat{a},T-}$	$\delta^{8,j}_{\hat{a},T-}$

Table C.6: $p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h},\hat{a},o\right)$ Values for $R = UCT$ and $o = T-$

$s^6_{j,k,0}$	$s^6_{j,k,h+}$	$s^7_{j,k,0}$	$s^7_{j,k,h+}$	$s^8_{j,k,0}$	$s^8_{j,k,h+}$	$s^9_{j,k,0}$	$s^{10}_{j,k,0}$
$\rho^{6,6}_j\gamma_{9,j}\left[1-\delta^{9,j}_{\hat{a},T-}\right]$	0	$\rho^{7,6}_j\gamma_{9,j}\left[1-\delta^{9,j}_{\hat{a},T-}\right]$	0	0	0	$(1-\gamma_{9,j})\left[1-\delta^{9,j}_{\hat{a},T-}\right]$	$\delta^{9,j}_{\hat{a},T-}$

Table C.7: $p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h},\hat{a},o\right)$ Values for $\hat{a} = Co$, and $o = P+$

	$s^3_{j,k,0}$	$s^4_{j,k,0}$	$s^5_{j,k,0}$	$s^{PC,0}_{j,k,0}$	$s^{PC,1}_{j,k,0}$	$s^{PC,2}_{j,k,0}$	$s^{10}_{j,k,h}$
$s^1_{j,k,h}$	$\rho^{3,3}_j\left[1-\delta^{1,j}_{cl,P+}\right]$	$\rho^{4,3}_j\left[1-\delta^{1,j}_{cl,P+}\right]$	0	0	0	0	$\delta^{1,j}_{cl,P+}$
$s^4_{j,k,h}$	$\rho^{3,3}_j\left[1-\delta^{4,j}_{cl,P+}\right]$	$\rho^{4,3}_j\left[1-\delta^{4,j}_{cl,P+}\right]$	0	0	0	0	$\delta^{4,j}_{cl,P+}$
$s^7_{j,k,h}$	0	0	0	$\rho^{6,6}_j\left[1-\delta^{7,j}_{cl,P+}\right]$	$\rho^{7,6}_j\left[1-\delta^{7,j}_{cl,P+}\right]$	0	$\delta^{7,j}_{cl,P+}$

Table C.8: $p\left(s^{i'}_{j',k',h'}|s^{i}_{j,k,h},\hat{a},o\right)$ Values for All $\hat{a}$ when $o \in \{C+, SD\}$

	$s^{6}_{j,k,0}$	$s^{7}_{j,k,0}$	$s^{8}_{j,k,0}$	$s^{9}_{j,k,0}$	$s^{10}_{j,k,0}$
$s^{2}_{j,k,h}$	$\rho^{6,2}_{j}\gamma_{2,j}\left[1-\delta^{2,j}_{\hat{a},o}\right]$	$\rho^{7,2}_{j}\gamma_{2,j}\left[1-\delta^{2,j}_{\hat{a},o}\right]$	$\rho^{8,2}_{j}\gamma_{2,j}\left[1-\delta^{2,j}_{\hat{a},o}\right]$	$(1-\gamma_{2,j})\left[1-\delta^{2,j}_{\hat{a},o}\right]$	$\delta^{2,j}_{\hat{a},o}$
$s^{5}_{j,k,h}$	$\rho^{6,5}_{j}\gamma_{5,j}\left[1-\delta^{5,j}_{\hat{a},o}\right]$	$\rho^{7,5}_{j}\gamma_{5,j}\left[1-\delta^{5,j}_{\hat{a},o}\right]$	$\rho^{8,5}_{j}\gamma_{5,j}\left[1-\delta^{5,j}_{\hat{a},o}\right]$	$(1-\gamma_{5,j})\left[1-\delta^{5,j}_{\hat{a},o}\right]$	$\delta^{5,j}_{\hat{a},o}$
$s^{8}_{j,k,h}$	$\rho^{6,8}_{j}\gamma_{8,j}\left[1-\delta^{8,j}_{\hat{a},o}\right]$	$\rho^{7,8}_{j}\gamma_{8,j}\left[1-\delta^{8,j}_{\hat{a},o}\right]$	$\rho^{8,8}_{j}\gamma_{8,j}\left[1-\delta^{8,j}_{\hat{a},o}\right]$	$(1-\gamma_{8,j})\left[1-\delta^{8,j}_{\hat{a},o}\right]$	$\delta^{8,j}_{\hat{a},o}$

C.3 $q\left(s^{i'}_{j',k',h'}, s^{i}_{j,k,h}, \hat{a}, o\right)$

It is assumed here that the value of $q\left(s^{i'}_{j',k',h'}, s^{i}_{j,k,h}, \hat{a}, o\right)$ is dependent only on the disease progression states. I.e., the only parameter effecting its value is $i \in \mathcal{I}$. In all of the following tables, it is assumed that $j = j'$, $k = k'$, $h = h'$. If these conditions are not met, then automatically $q\left(s^{i'}_{j',k',h'}, s^{i}_{j,k,h}, \hat{a}, o\right) = 0$.

Table C.9: $q\left(s^{i'}_{j',k',h'}, s^{i}_{j,k,h}, \hat{a}, o\right)$ Values for All $\hat{a}, j, k, h$ When $o = T$- and $i' \in \{0,1,2,3,4,5\}$

	$s^{0}_{j,k,h}$	$s^{1}_{j,k,h}$	$s^{2}_{j,k,h}$	$s^{3}_{j,k,h}$	$s^{4}_{j,k,h}$	$s^{5}_{j,k,h}$
$s^{0}_{j,k,h}$	$1-d_{\neg poly}(\hat{a})$	$1-d_{\neg poly}(\hat{a})$	0	0	0	0
$s^{1}_{j,k,h}$	0	$1-d_{\neg poly}(\hat{a})$	$1-d_{\neg poly}(\hat{a})$	0	0	0
$s^{2}_{j,k,h}$	0	0	$1-d_{C}$	0	0	0
$s^{3}_{j,k,h}$	0	0	0	$1-d_{\neg poly}(\hat{a})$	$1-d_{\neg poly}(\hat{a})$	0
$s^{4}_{j,k,h}$	0	0	0	0	$1-d_{\neg poly}(\hat{a})$	$1-d_{\neg poly}(\hat{a})$
$s^{5}_{j,k,h}$	0	0	0	0	0	$1-d_{C}$
$s^{6}_{j,k,h}$	0	0	0	0	0	0
$s^{7}_{j,k,h}$	0	0	0	0	0	0
$s^{8}_{j,k,h}$	0	0	0	0	0	0
$s^{9}_{j,k,h}$	0	0	0	0	0	0
$s^{10}_{j,k,h}$	0	0	0	0	0	0

Table C.10: $q\left(s^{i'}_{j',k',h'}, s^{i}_{j,k,h}, \hat{a}, o\right)$ Values for All $\hat{a}, j, k, h$ When $o = T$- and $i' \in \{6, 7, 8, 9, 10\}$

	$s^{6}_{j,k,h}$	$s^{7}_{j,k,h}$	$s^{8}_{j,k,h}$	$s^{9}_{j,k,h}$	$s^{10}_{j,k,h}$
$s^{0}_{j,k,h}$	0	0	0	0	$(0.5 - d_{\neg poly}(\hat{a}))\left(1 - \kappa^{i,j}_{\hat{a},o}\right)$
$s^{1}_{j,k,h}$	0	0	0	0	$(0.5 - d_{\neg poly}(\hat{a}))\left(1 - \kappa^{i,j}_{\hat{a},o}\right)$
$s^{2}_{j,k,h}$	0	0	0	0	$0.5\left(1 - d_{C}\right)\left(1 - \kappa^{i,j}_{\hat{a},o}\right)$
$s^{3}_{j,k,h}$	0	0	0	0	$(0.5 - d_{\neg poly}(\hat{a}))\left(1 - \kappa^{i,j}_{\hat{a},o}\right)$
$s^{4}_{j,k,h}$	0	0	0	0	$(0.5 - d_{\neg poly}(\hat{a}))\left(1 - \kappa^{i,j}_{\hat{a},o}\right)$
$s^{5}_{j,k,h}$	0	0	0	0	$0.5\left(1 - d_{C}\right)\left(1 - \kappa^{i,j}_{\hat{a},o}\right)$
$s^{6}_{j,k,h}$	$1 - d_{\neg poly}(\hat{a})$	$1 - d_{\neg poly}(\hat{a})$	0	0	$(0.5 - d_{\neg poly}(\hat{a}))\left(1 - \kappa^{i,j}_{\hat{a},o}\right)$
$s^{7}_{j,k,h}$	0	$1 - d_{\neg poly}(\hat{a})$	$1 - d_{\neg poly}(\hat{a})$	0	$(0.5 - d_{\neg poly}(\hat{a}))\left(1 - \kappa^{i,j}_{\hat{a},o}\right)$
$s^{8}_{j,k,h}$	0	0	$1 - d_{C}$	0	$0.5\left(1 - d_{C}\right)\left(1 - \kappa^{i,j}_{\hat{a},o}\right)$
$s^{9}_{j,k,h}$	$1 - d_{UCT}$	$1 - d_{UCT}$	0	$1 - d_{UCT}$	$0.5\left(1 - d_{UCT}\right)\left(1 - \kappa^{j}_{UCT}\right)$
$s^{10}_{j,k,h}$	0	0	0	0	1

Table C.11: $q\left(s^{i'}_{j',k',h'}, s^{i}_{j,k,h}, \hat{a}, o\right)$ Values for $\hat{a} = cl$ and All j, k, h When $o = P+$

	$s^{3}_{j,k,h}$	$s^{4}_{j,k,h}$	$s^{5}_{j,k,h}$	$s^{6}_{j,k,h}$	$s^{7}_{j,k,h}$	$s^{8}_{j,k,h}$	$s^{10}_{j,k,h}$
$s^{1}_{j,k,h}$	$1 - d_{poly}(\hat{cl})$	$1 - d_{poly}(\hat{cl})$	0	0	0	0	$(0.5 - d_{poly}(cl))\left(1 - \kappa^{i,j}_{cl,P+}\right)$
$s^{4}_{j,k,h}$	$1 - d_{poly}(\hat{cl})$	$1 - d_{poly}(\hat{cl})$	0	0	0	0	$(0.5 - d_{poly}(cl))\left(1 - \kappa^{i,j}_{cl,P+}\right)$
$s^{7}_{j,k,h}$	0	0	0	$1 - d_{poly}(\hat{cl})$	$1 - d_{poly}(\hat{cl})$	0	$(0.5 - d_{poly}(cl))\left(1 - \kappa^{i,j}_{cl,P+}\right)$

Table C.12: $q\left(s^{i'}_{j',k',h'}, s^{i}_{j,k,h}, \hat{a}, o\right)$ Values for All $\hat{a}, j, k, h$ When $o \in \{C+, SD\}$

	$s^{6}_{j,k,h}$	$s^{7}_{j,k,h}$	$s^{8}_{j,k,h}$	$s^{9}_{j,k,h}$	$s^{10}_{j,k,h}$
$s^{2}_{j,k,h}$	$1 - d_{CT}$	$1 - d_{CT}$	$1 - d_{CT}$	$1 - d_{CT}$	$0.5\left(1 - d_{UCT}\right)\left(1 - \kappa^{j}_{UCT}\right)$
$s^{5}_{j,k,h}$	$1 - d_{CT}$	$1 - d_{CT}$	$1 - d_{CT}$	$1 - d_{CT}$	$0.5\left(1 - d_{UCT}\right)\left(1 - \kappa^{j}_{UCT}\right)$
$s^{8}_{j,k,h}$	$1 - d_{CT}$	$1 - d_{CT}$	$1 - d_{CT}$	$1 - d_{CT}$	$0.5\left(1 - d_{UCT}\right)\left(1 - \kappa^{j}_{UCT}\right)$