Strategic Resource Planning in Healthcare Under Uncertainty

British Columbia Cancer Agency and Vancouver's Downtown Eastside

by

Gregory Robert Werker

BSc, Cornell University, 1998
MEng, Cornell University, 1999

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate Studies

(Business Administration)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2012

Abstract

In two very different healthcare settings we demonstrate the benefits of long term planning using operations research (OR) tools. We present models that handle considerable variability using solutions based on relatively simple approximations.

In the first setting we present a mixed integer program (MIP) with a goal programming (GP) formulation for strategic workforce planning at the British Columbia Cancer Agency (BCCA). Our model considers experience, minimum and maximum durations, and redundancy in staffing to guard against unanticipated employee leaves. We evaluate the model parameters using simulation, and analyze the simulation output with logistic and Poisson regression. The core model can be generalized to other workforce planning applications in healthcare or to other human resource intensive industries; the full BCCA model illustrates a real-world implementation.

This research introduces to the workforce planning literature a technique for building robustness into the plan, together with experience and duration constraints.

In the second setting we study a marginalized population for which myriad organizations provide healthcare and other services in the absence of system-level quantitative planning. We use a queueing network to model clients with complex concurrent disorders (CCD) flowing through services in Vancouver's Downtown Eastside (DTES). We perform sensitivity analysis on the input parameters, validate our solution against a simulation model, and conduct scenario comparisons to evaluate potential procedural and policy changes to the system.

To analyze this network we present a novel approximation technique—called a linearized closed queueing network (LCQN)—for solving closed queueing networks. By using an open queueing network with the fixed population mean (FPM) approach, and by including a trick for dealing with capacitated stations, we create a network representation that is solved with a linear program (LP). This method scales to much larger systems. We derive the approximation ratio between this approximation and the exact solution for a small network, and use simulation to show that this gap is of no practical significance for the full network.

Preface

None of the work in this dissertation has yet been published or submitted for publication.

For the workforce planning research at the BCCA, I created the model, performed the analyses, and prepared Chapters 2 & 3 with suggestions from Dr. Puterman.

For the strategic resource planning research in the DTES, I identified the overall research area and honed the research program with input from Drs. Puterman and Krausz. I created the queueing model, developed the solution approach, and performed the analyses. Dr. Krausz helped identify input parameter values. Marziyeh Emami, working as a research assistant, built the simulation model and helped locate input parameters and relevant literature. I prepared Chapters 4 & 5 with suggestions from Dr. Puterman.

Dr. Atkins provided suggestions and feedback on both topics.

Dr. Maurice Queyranne provided feedback and suggestions for the final version.

Ethics approval for the BCCA research was granted by the UBC BCCA Research Ethics Board (certificate number H06-03843).

Ethics approval for the DTES research was granted by the Behavioural Research Ethics Board (certificate number H12-01268).

Funding support was provided by an NSERC Postgraduate Scholarship D2, by the Shelby L. Brumelle Memorial Graduate Scholarship, by NSERC Discovery Grant 5527 (PI: Martin L. Puterman), and by the CIHR Team in Operations Research for Improved Cancer Care Grant (co-PIs: Scott Tyldesley and Martin L. Puterman).

Table of Contents

Appendices

List of Tables

List of Figures

List of Abbreviations

ACT	assertive community treatment
BCCA	British Columbia Cancer Agency
BCMHA	Burnaby Centre for Mental Health and Addictions
CAD	Canadian dollars
CM	case management
CCD	complex concurrent disorders
CPI	consumer price index
DES	Discrete event simulation
DTES	Downtown Eastside
ED	emergency department
FPM	fixed population mean
GP	goal programming
LoS	length of stay
LCQN	linearized closed queueing network
LP	linear program
MIP	mixed integer program
MMT	methadone maintenance treatment
OR	operations research
QALY	quality adjusted life year
QoL	quality of life
RT	radiation therapist
SD	system dynamics
VCH	Vancouver Coastal Health

Acknowledgements

I would like to recognize the many extraordinary people involved in my education.

Dr. Martin Puterman, thank you for introducing me to operations research healthcare, and for encouraging me to pursue an idea that was new to both of us. Dr. Michael Krausz, thank you for your expertise and guidance, and for making me a part of your team. Dr. Derek Atkins, thank you for providing a big picture view and lots of pertinent and valuable feedback. And thank you Dr. Maurice Queyranne for your extremely useful feedback during the final iteration.

I would like to acknowledge the talented members of the BCCA team, especially Mike Darud, who helped define the requirements for the workforce planning model, and Anthony Slowey, who put it into use. Also thank you John French, Pablo Santibáñez, Vincent Chow, Ruben Aristizabal, and Travis Nordin.

As I worked on the second topic, two longtime friends shared with me their DTES experiences: Thank you Stephen Epp and Saje Crossen for all your wisdom and insight. Also thank you Marziyeh Emami for your work as a research assistant.

To all of the other Sauder folks, a big thank you. In particular, Elaine Cho, thank you for making sure everything that is supposed to happen happens. Thank you Brumelle family for your generous support through the Shelby L. Brumelle Memorial Graduate Scholarship. And Antoine Sauré, thank you for your many ideas and your friendship.

Prior to my PhD studies, I worked at Supply Chain Consultants (now Arkieva). Thank you Drs. Harpal and Bibi Singh—and the rest of the gang—for all you taught me about work and people and operations research. I also want to thank the Cornell OR&IE department for launching me into OR. In particular, thank you Dr. William Maxwell for showing me how to see a problem for what it really is... and for introducing me to Indonesia.

Finally, thank you Mom and Dad for a life-time of love and support, and Eric for being a super brother and role model. And Kim, I thank you wholeheartedly for travelling this road by my side.

Dedication

For Kim, who is imaginative, exuberant, and brilliant.

And for Owen, who is the most amazing little man.

Chapter 1

Introduction

The healthcare industry currently faces many challenges. While doctors and other healthcare workers labour to save lives and cure patients, complex issues are undermining their ability to continue providing the best care possible: Rapidly rising costs, increasingly elaborate and interlinked systems, and ever-changing technologies and treatment techniques. The majority of the quantitative aspects of these complex issues can be addressed using operations research (OR). In the last two decades the use of OR in healthcare has increased dramatically, yet nevertheless, many opportunities remain to tackle medical, logistical, operational, and strategic problems in all forms of healthcare provision.

Considerable OR healthcare research in recent years deals with short term issues, while research on long term healthcare is less common. In the realm of scheduling and planning this disparity is vast. For example, a review of nurse rostering and scheduling [27] includes 83 references, however, only a handful of articles on nurse workforce planning exist [e.g., 62, 64, 108]. We believe that, given this disparity, research with a long term focus is a valuable contribution to the field.

Short term (tactical or operational) models provide insight into staff scheduling, appointment scheduling, treatment provision, and other daily administration. Long term (strategic) models pertain to matters such as human resource planning, capacity planning, and system design. While both foci are vital, in this dissertation we emphasize the long term one. We present two strategic planning models for very different applications that nevertheless share characteristics.

Healthcare systems almost always include variability, whether introduced by patients, workers, or biology. In both of our research topics we study systems with considerable inherent variability. The first—a strategic workforce planning model at the British Columbia Cancer Agency (BCCA)—contends with staff variability in terms of learning rates, vacations, and most importantly, staff leaves (primarily maternity/paternity leave). The second—a strategic model of healthcare provision in Vancouver's Downtown Eastside (DTES)—faces patient and biological variability

in terms of decisions about seeking treatment and the success of various programs.

Many OR techniques exist, and there is great variation within each technique (e.g., within queueing theory there are many types of queueing stations and networks, with different approaches for each one). Choosing the correct tool for a specific problem is essential for success. And when a problem is complex and involves variability, the correct tool may be very complex itself. In both of our research topics we discuss the approaches—which can be quite elaborate—to finding solutions that specifically deal with the type of system and the type of variability at hand. Yet we also show that these elaborate solution approaches are not always needed. In both cases we present simplified modelling techniques that use standard OR tools, albeit perhaps unconventionally. In fact, in both cases we ultimately use mathematical programming to solve very different problems.

These simplified modelling techniques are not exact methods; they do not guarantee completely correct answers. But they are appropriate for these (and other) applications because they provide solutions that are very good approximations. We demonstrate the quality of these approximate techniques: In order to show that the BCCA model performs well we use a simulation model to test it under different levels of variability, and then use regression analysis to interpret the simulation results; for the DTES model we calculate the approximation ratio between our approximate solution and the exact solution for a smaller problem instance, and also use simulation to show that for the full model the resulting gap is of no practical significance. In this way we rely on more elaborate tools to demonstrate that the simple approaches we employ for two problems are, for all intents and purposes, as good as more complex techniques. Yet they have the added benefit of being much easier to explain to healthcare providers, practitioners, and other researchers.

We also discuss how these approximations can be applied to other healthcare problems. In fact, they are applicable to other industries as well. The history of OR in healthcare is relatively brief, and many OR healthcare applications borrow from other industries. But there is also considerable opportunity for other industries to learn from what is being done with OR in healthcare. We demonstrate OR approaches to strategic planning problems and show that some of the forces working against doctors and providers in healthcare can be overcome with the help of models that provide system level insight and well-informed, data-driven decision making.

1.1 Workforce Planning at the BCCA

Amidst talk of increasing healthcare costs, an ageing workforce, and constrained resources, workforce planning is a pertinent topic. Much of the healthcare literature in OR concerns operational decisions—shift scheduling, surgery block allocation, patient flow – but models that go beyond the next week or month are less common. We present a workforce planning model that allows an organization to focus on the strategic issues that will enable it to create and maintain a well-balanced workforce now and several years into the future. Our model includes skill acquisition, minimum and maximum duration constraints, and a mechanism to manage the variability inherent in the planning of individual careers.

We use a mixed integer program (MIP) with a goal programming (GP) formulation employing two different objective functions—one for creating plans and one for making minor adjustments. Variability is handled with an intuitive yet powerful robustness constraint that we evaluate numerically under different assumptions. This unique combination of GP workforce planning incorporating experience, duration, and redundancy constraints is applicable in many healthcare settings as well as in other human resource intensive industries.

One area of healthcare with high costs, continually advancing technology, and limited human resources is cancer care. Of the many skilled workers involved in cancer prevention, treatment, and rehabilitation, we focus on radiation therapy treatment, and specifically, on the radiation therapist (RT) workforce. This group of employees demonstrates a wide array of skills through their involvement in almost all aspects of radiation therapy.

The radiation therapy department at the BCCA, described in Chapter 3, was the motivation for this work. The human resource challenges faced by this department are by no means unique to cancer care: Skills are diverse and wide-ranging, with mastery requiring months or years of experience; employee departures are unpredictable, whether for maternity/paternity leave, relocation, or retirement; job satisfaction is associated with retention, and can be influenced by workforce planning activities. Moreover, prior to implementing our model the department had a process for scheduling RTs for one month but only a rudimentary planning framework for looking beyond that time frame.

Considerable literature is devoted to staffing at the operational level. Staff

scheduling, rostering, shift scheduling, and other related daily or weekly staffing models are studied in many settings. Ernst *et al.* [42] provide a comprehensive overview of the various scheduling approaches in use across a number of industries. Within the health care industry, many examples of nurse scheduling exist from older models [74] up through more advanced approaches [94]. Nurses are not the only health care workers needing schedules; Topaloglu [109] presents a scheduling model for emergency medicine residents and Brunner *et al.* [17] assign physicians to flexible shifts using constraints similar to our duration constraints.

At the strategic level, Lavieri and Puterman [64] use a linear program (LP) to determine how many nurses to train in aggregate in B.C. over a 20-year planning horizon. We are concerned, however, with planning individual career trajectories, which necessitates a horizon that is shorter than 20 years but longer than the typical scheduling time frame. In this vein, Franz and Miller [44] assign medical residents to rotations, including preferences for the number of residents in different areas, but they do not consider durations or experience gained in earlier rotations. Bhadury and Radovilsky [11] study job rotation using an assignment problem in which the solution is achieved by iterating through the periods; they include a form of sequencing that is similar to our requirement for minimum durations, however, they also do not consider experience gained. Li and King [65] explore personnel planning at a health clinic where task substitution is included in the model as a cushion against variability, but this cushion is very different from our redundancy constraint. An example from the semiconductor industry by Bordoloi and Matsuo [14] includes experience gained by line workers, however, this model is built on a production scheduling and inventory planning approach, and we believe it doesn't extend to healthcare.

Our model, introduced in Chapter 2, brings to the literature the novel combination of minimum durations, maximum durations, experience requirements, and a redundancy constraint. To our knowledge, no staff planning research includes all of these components. The GP approach, with soft constraints and penalties in the objective function, handles trade-offs among different requirements. Using GP for staff planning is not new; Schroeder [99] shows how this approach can be used to balance different objectives in determining staffing levels in a university setting. GP is also used in health care, as demonstrated by Topaloglu [110] for medical resident shift scheduling and by Berrada *et al.* [10] for nurse shift scheduling, the latter

including several GP solution approaches. (For a history of GP, see Aouni and Kettani [7]). The output of our model is a strategic plan—robust against variability in staff availability—detailing RT assignments to management, at the quarterly level, for the next several years.

The BCCA is a sophisticated cancer centre, using many varieties of technology—including OR—to achieve excellent results; cancer survival rates in B.C. are some of the best anywhere [33]. Given this environment it is not surprising that our strategic planning project at the BCCA was fairly well-defined from the start, and that it was focused on a specific problem with clear objectives. Our other line of research demonstrates that this is not always the case.

1.2 Strategic Resource Planning in Vancouver's DTES

The DTES in Vancouver, B.C., is a vibrant neighbourhood with an active residents association and a sense of community. Unfortunately, it also has much more than its share of problems. The population of around 18,000 [21] has an average adult income (excluding subsidies) of only $6,282 (2009 CAD) [16] and grapples with many social and health issues including high rates of drug use, disease, crime, and prostitution. Many organizations support the residents of the DTES. Like the BCCA for cancer patients, these organizations offer a wide range of treatments and paths for clients with mental health and addictions issues. Yet the two systems couldn't be more different; the former is centralized, well-funded, and comprehensive while the latter is decentralized, arguably inadequately funded, and full of gaps as well as overlaps in services. With an eye on the role that OR can play within a well-run organization, we now turn to this very different healthcare system to see how OR might be employed to help provide aid to a particularly vulnerable population.

The number of people living with complex concurrent disorders (CCD) in the DTES is staggering. This population with mental health and addiction issues endures additional challenges including homelessness, physical illness, high rates of hepatitis and HIV infection, histories of trauma, suicidal behaviour, and difficulty accessing help. While myriad organizations provide health, social, housing, and criminal justice amenities in and around the DTES, no high-level quantitative strategic approach exists to create efficiency within the system using existing resources or

to optimally introduce and position new services and programs.

For background information on the DTES, the book "A Thousand Dreams" by former Vancouver mayor Campbell, criminologist Boyd, and journalist Culbert discusses the history of this neighbourhood, the reasons for its current state, and possible solutions [21]. Two reports from the Vancouver Police Department give a thorough overview of the mental health and addiction issues from a police perspective, and include portraits of several individuals [107, 121]. A brief overview of the DTES as well as a survey of quantitative literature on mental health and addictions can be found in [68].

Existing research involving this community has focused on epidemiological topics such as describing population characteristics and risks [35, 111, 122], identifying causal relationships [123, 124], quantifying costs and services used [53, 84], and evaluating individual treatment options and interventions [98]. Because of this rich literature we know a great deal about the residents, the available services, and the outcomes of some of the interventions. What we don't know, however, is how the system works and interacts as a whole, and where improvements can be made to benefit residents and streamline the delivery of services.

In other words, considerable resources are being directed toward a plethora of very important, but piecemeal solutions—supportive housing, primary care, treatment teams. In order to address system-wide problems, one must be able to examine the whole system, and this is where OR comes into play. The literature contains relevant examples of OR healthcare models that are specific to addictions or mental health issues, for instance, Earnshaw *et al.* [40] use an LP to determine the allocation of funds for HIV prevention, and Caulkins *et al.* [25] use a Markov chain approach to explore cocaine use in the US, including changing rates of initiation informed by data. Caulkins *et al.* [26] use a dynamic compartmental model to study drug use trends in Australia, and Richter and Loomis [93] employ a similar approach to study an HIV intervention in substance users.

In our research we use a queueing network to model clients flowing through various services. Our approach involves a new method which we refer to as a linearized closed queueing network (LCQN); we have obtained exact bounds for a small example and we demonstrate bounds computationally for larger instances, however, we haven't yet obtained rigorous bounds for those larger instances. We describe our queueing theory approach in detail in Chapter 4 and provide a background

on queueing theory literature therein.

Studying a population of substance users, Kaplan and Johri [50] use a queueing approach to model the cycle from abstinence through drug use and into treatment. Koizumi *et al.* [57] use queueing theory to investigate patients flowing through mental health resources in Philadelphia. These last two examples are the most methodologically relevant mental health or addictions articles.

Only a couple articles use OR models to study these issues in the DTES. Bayoumi and Zaric [9] use simulation to evaluate the cost and health impact of Vancouver's safer injection site. Pourbohloul *et al.* [86] use a compartmental model to understand the effects of a large syphilis intervention. However, to our knowledge, there are no quantitative OR models in the literature that focus on many mental health *and* addiction services for an entire population.

We believe this research is important for two reasons. It helps to demonstrate that OR is applicable to the field of mental health and addictions. And it provides a framework for seeking desperately needed answers in a geographic and medical area lacking quantitative strategic planning. These points will become clear when we discuss the results of this research topic in Chapter 5. Decision makers at the various organizational and political levels will be better informed and more prepared to make informed choices using our models of the services in the DTES.

1.3 Brief Overview of Computer Simulation

Simulation is used to evaluate various aspects of both research topics in this dissertation, so we present an overview of this approach. The term *computer simulation* describes a broad class of tools that allows one to model a real-world or hypothetical system by (repeatedly) trying out the significant events, moves, or decisions affecting the entities or individuals that comprise the system. Put simply, it is about *imitating* a system, in order to study it, rather than attempting to find a corresponding (but perhaps more elegant or tractable) representation of that system.

Simulation models, like many other types of models, allow us to study and learn from a virtual system rather than working with a real-world system, often for practical reasons. Sometimes it is prohibitively expensive, unethical, or too time-consuming to try making changes in a real system, and other times such a system doesn't yet exist.

The term *simulation* has another meaning in healthcare that refers to exercises in which health workers physically practice or act out situations such as surgery or disaster response for education, training, or assessment purposes. However, we use the word to refer to *computer simulation* only.

In the last several decades, simulation models have become more powerful. Because simulation is in many ways a brute force approach—using computers to generate numerous random inputs and outcomes in order to get an idea of an average outcome—the field has benefited greatly from increases in computing power.

A number of computer simulation approaches exist, a few of which are common in healthcare and are mentioned here:

Monte Carlo simulation can be particularly useful for calculation purposes. It involves generating random numbers according to distributions for various (often correlated) inputs and then performing a calculation on those inputs.

Agent-based simulation is appropriate for simulating individual people or other entities that make decisions based on interactions with each other and with the system. Each individual is essentially represented by an algorithm making random decisions that are influenced by other outcomes in the overall model.

Hybrid models are useful when a pure approach is not sufficient. These models combine different types of simulation approaches or they pair simulation with other tools. For instance, in Chapter 3, simulation combined with MIP enables us to evaluate the workforce planning model as it is subjected to random staff leaves.

Discrete event simulation (DES) is typically useful for describing objects or people moving through different stages in a system; it is described in more detail below.

For examples of simulation in healthcare, see these prefaces to special issues on healthcare simulation by Anderson [5] and Anderson and Merode [6] and the rest of the articles contained in those issues. An example of DES in healthcare can be found in Werker *et al.* [118].

1.3.1 Discrete Event Simulation

DES models have several components: Events, random-number generators, and time. Events happen at a specific time, and often lead to other events being scheduled. Simulated time is essentially measured continuously (actually in very small discrete units, e.g., in simulated seconds or milliseconds). The software that performs the simulation keeps track of an event list with all scheduled future events, moving down the list and processing each event in turn. Processing an event means updating the event list accordingly, e.g., if the event "client shows up at hospital" is being processed, it could generate a random service time and then schedule another event for "client departs hospital" at the appropriate future time. Throughout the running of the model, statistics are tracked for any measures of interest. Often another component, stations, is part of these models. E.g., the hospital may be a station with an associated capacity and distribution for service times.

A single replication of a DES model proceeds until some stopping condition. Typically this condition is that a certain amount of simulated time has passed (this condition is easy to model—when the model begins, an event is immediately scheduled at the specified future time that causes the model to halt). After the replication is over, the statistics are tallied and the results can be analyzed. However, because these results are based on random numbers, they may not represent typical system performance. Therefore, DES usually involves performing many replications (with independent streams of random numbers) in order to infer the average performance of the system. With enough replications, and relying on statistical theory, it is easy to calculate not just an average but also a confidence interval for each measure tracked. The end result is that we can describe the system to any level of accuracy, as long as we are willing to run enough replications to achieve confidence intervals with our desired width.

Many software packages simplify the work involved in creating a DES model. Because it is commonly used in healthcare and other industries, and because of prior familiarity, we use Arena to create and run the DES model. (For more information on simulation, on DES in particular, and on the Arena software, see [51].)

1.4 Queueing Review

Queueing theory is used in Chapters 4 and 5 to model healthcare and other services in the DTES. The queueing approach employed builds on basic theory which we review herein.

A typical queueing model consists of entities arriving, possibly waiting, and then getting served by a server before exiting the system. (Henceforth we refer to these entities as "clients" in order to use the terminology most common in the DTES). Arrivals can be classified in different forms including deterministic, general, or Markovian. This last form describes random arrivals that follow a Poisson process, or in other words, the number of arrivals in any particular period length follows a Poisson distribution and correspondingly, the time between two consecutive arrivals follows an exponential distribution. One need not assume the arrival process is stationary (unchanging over time) but we make that assumption in order to simplify analyses. This arrival process and the stationarity assumption are discussed in more detail in §5.1.1.

Service times can also follow any of the same distributions; when service times are exponentially distributed we say that they are also Markovian. (A service time can also be referred to as a length of stay (LoS); a service rate, the reciprocal of LoS, is often used to specify the parameter value.) The model may include a single server, multiple (finite number of) servers, or an infinite number of servers. Several simple models are reviewed below.

For information on these simple models, or on the more common extensions to these models, any queueing theory text can be helpful (Bhat [12], Gross *et al.* [47]). Many papers go beyond what is found in the textbooks with extensions such as a general distribution for service times, s servers, and additional waiting spaces [120] or queues with a last-in-first-out discipline [4], however because we limit our analyses in several ways—Markovian arrivals and service times, single class queues, a first-in-first-out discipline, and stationary arrival processes—the models described below provide sufficient background to introduce our model.

1.4.1 Queueing Notation

The following notation is used to introduce simple queueing models.

n = Number of clients in the system, i.e., the state of the system.

λ = Arrival rate; λ_n is the state-specific arrival rate when the system is in state n.

μ = Service rate; μ_n is the service rate when the system is in state n. Note that LoS—length of stay—is the reciprocal of service rate.

ρ = Traffic intensity; ρ_n is the traffic intensity corresponding to state n.

C_n = Multiplier corresponding to state n (this term simplifies some of the other expressions).

P_0 = Probability there are 0 clients in the system in steady state.

P_n = Probability there are n clients in the system in steady state.

L = Expected number of clients in the system in steady state.

W = Expected waiting time for a client upon arrival to the system in steady state.

1.4.2 Birth-and-Death Processes

The birth-and-death process is a special case of a continuous time Markov chain, and is the foundation for the analyses of the various queueing models. For details, see any queueing text, [e.g., 47, 73]. The basic idea, using the notation from Hillier and Lieberman [48], is as follows:

A birth-and-death process only allows transitions from n to $n+1$ (for $n \geq 0$) or to $n-1$ (for $n > 0$). The former transition is called a "birth" and the latter is called a "death", representing the arrival or departure of a customer from the queueing station, respectively. The time until the next birth (arrival) is exponentially distributed with parameter λ_n and the time until the next death (departure) is exponentially distributed with parameter μ_n.

Assuming, for now, that the steady state exists, we can analyze the queue by recognizing that in the long run the entering rate equals the leaving rate for any state, where the state is the number of customers in the system. (Conditions for the existence of steady state are discussed below for specific queues.) Given all birth transition rates, λ_n, and all death transition rates, μ_n, we can write the balance equation for each state for $n = 0, 1, \ldots$ and then solve this system of equations to determine the steady-state probabilities, denoted P_n. We use the fact that all of

these probabilities must sum to 1 to express the P_n terms as a function of the birth and death rates (introducing the term C_n to simplify the notation):

$$C_0 = 1 \tag{1.1}$$

$$C_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1}, \quad \text{for } n = 1, 2, \ldots \tag{1.2}$$

$$P_0 = \left(\sum_{n=0}^{\infty} C_n\right)^{-1} \tag{1.3}$$

$$P_n = C_n P_0, \quad \text{for } n = 0, 1, 2, \ldots \tag{1.4}$$

In the following subsections we summarize the analyses of several queueing models using this birth-and-death result as a starting point.

1.4.3 $M/M/1$, $M/M/s$, and $M/M/\infty$ Queues

One of the most basic queueing models consists of Markovian arrivals, Markovian service times (with a different parameter), and a single server. Using Kendall's notation ([52], [104]), we refer to this system as an $M/M/1$ queue. The first M represents the Markovian arrivals, the second M represents the Markovian service times, and the 1 represents the number of servers. We could also talk about an $M/M/s$ model with a finite number of servers, $s > 1$. In some situations it makes sense to assume an $M/M/\infty$ model with an unlimited number of servers (representing the situation in which arriving entities are served immediately with the same service rate no matter how busy the system is).

For each of these queueing models it is easy to determine the probability distribution for the number of clients in the queue, where the queue is described as those waiting for service plus those being served. We can also determine the average number of clients in the queue and the expected waiting time.

The $M/M/1$ queue has the same arrival rate regardless of which state, n, the queue is in. We therefore write $\lambda_n = \lambda$. Service time is $\mu_n = \mu$. We define the traffic intensity to be $\rho = \lambda/\mu$. Using the solution approach for a birth-and-death process we determine $C_n = (\lambda/\mu)^n = \rho^n$. If $\rho < 1$ the queue will reach steady state and we can solve for the probability of having zero clients in the queue, P_0; the probability of having n clients, P_n; the average length of the queue (including those waiting and

those being served), L; and the expected waiting time, W:

$$P_0 = \left[\sum_{n=0}^{\infty} C_n\right]^{-1} = 1 - \rho \tag{1.5}$$

$$P_n = (1-\rho)\rho^n, \qquad \text{for } n = 1, 2, \ldots \tag{1.6}$$

$$L = \sum_{n=0}^{\infty} nP_n = \frac{\lambda}{\mu - \lambda} \tag{1.7}$$

$$W = \frac{1}{\mu - \lambda} \tag{1.8}$$

The $M/M/s$ queue is identical to the $M/M/1$ queue with respect to arrivals, however, service time depends on how many clients are in service. The service rate μ_n equals $n\mu$ if up to s clients are in service and $s\mu$ if s or more clients are in service (i.e., the servers are full). Traffic intensity is $\rho = \lambda/s\mu$, and as with a single server, ρ must be less than one to achieve a steady state. We can derive the same quantities as above:

$$C_n = \begin{cases} \frac{\lambda^n}{\mu^n}\frac{1}{n!}, & \text{for } n = 1, 2, \ldots, s \\ \frac{\lambda^n}{\mu^n}\frac{1}{s!s^{n-s}}, & \text{for } n = s, s+1, \ldots \end{cases} \tag{1.9}$$

$$P_0 = \left[1 + \sum_{n=1}^{s-1}\frac{\lambda^n}{\mu^n}\frac{1}{n!} + \sum_{n=s}^{\infty}\frac{\lambda^n}{\mu^n}\frac{1}{s!s^{n-s}}\right]^{-1} = \left[\sum_{n=0}^{s-1}\frac{\lambda^n}{\mu^n}\frac{1}{n!} + \frac{\lambda^s}{\mu^s}\frac{1}{s!}\frac{1}{1-\lambda/(s\mu)}\right]^{-1} \tag{1.10}$$

$$P_n = \begin{cases} \frac{\lambda^n}{\mu^n}\frac{1}{n!}P_0, & \text{for } n = 1, 2, \ldots, s \\ \frac{\lambda^n}{\mu^n}\frac{1}{s!s^{n-s}}P_0, & \text{for } n = s, s+1, \ldots \end{cases} \tag{1.11}$$

$$L = \frac{P_0\rho}{s!(1-\rho)^2}\frac{\lambda^s}{\mu^s} + \frac{\lambda}{\mu} \tag{1.12}$$

$$W = \frac{P_0\rho}{s!(1-\rho)^2}\frac{\lambda^{s-1}}{\mu^s} + \frac{1}{\mu} \tag{1.13}$$

The $M/M/\infty$ queue allows one to model many servers yet has a very simple formula for the average queue length, L. Arrivals are the same as above, and service times are easily seen by letting s in the $M/M/s$ model go to infinity: $\mu_n = n\mu$ for all values of n. The system can handle unlimited arrivals and is therefore guaranteed to achieve steady state. The following equations are easy to derive for this model:

$$C_n = \frac{\lambda^n}{\mu^n}\frac{1}{n!} \tag{1.14}$$

$$P_0 = \left[\sum_{n=0}^{\infty} C_n\right]^{-1} = e^{-\lambda/\mu} \tag{1.15}$$

$$P_n = \frac{\lambda^n}{\mu^n}\frac{1}{n!}e^{-\lambda/\mu} \tag{1.16}$$

$$L = \sum_{n=0}^{\infty} nP_n = \lambda/\mu \tag{1.17}$$

$$W = 1/\mu \tag{1.18}$$

The simple formula for W makes intuitive sense. If all clients begin receiving service immediately upon arrival then the average time a client spends in the system is the reciprocal of the service rate. Using Little's law [69] we can easily see that $L = \lambda W = \lambda/\mu$. This result is very important; in our model, we take advantage of this simple formula for the average number of clients in the queue.

1.4.4 $M/M/s/s/N$ and $M/M/N/N/N$ Queues

We introduce two relevant queues corresponding to $M/M/s$ and $M/M/\infty$ queues, with finite waiting capacity and a finite calling population.

An $M/M/s$ queue with limited space in the system for K clients with $K > s$ is referred to as an $M/M/s/K$ queue (again using Kendall's notation). If there is no additional waiting space beyond the space in the s servers, we have $K = s$ and we refer to the system as an $M/M/s/s$ queue (also known as the Erlang-B model). Furthermore, if the population using this queue is not infinite (as per the assumptions in the previous section), we say it has a finite calling population of size N, and refer to the system as an $M/M/s/s/N$ queue. The arrival rate, l, has a different interpretation than the previous arrival rates: It represents the per client rate, i.e., the rate at which any individual client in the population not currently in the queue arrives at the queue. The size of this waiting population is $N - n$; the total arrival rate, λ, depends on the size of this population thus: $\lambda_n = (N-n)l$. The following equations describe this model (assuming $N > s$):

$$\lambda_n = \begin{cases} (N-n)l & \text{for } n = 0, 1, \ldots, s-1 \\ 0 & \text{for } n \geq s \end{cases} \tag{1.19}$$

$$\mu_n = n\mu \quad \text{for } n = 1, 2, \ldots, s \tag{1.20}$$

$$C_n = \begin{cases} \frac{l^n}{\mu^n}\binom{N}{n} & \text{for } 0 \le n \le s \\ 0 & \text{for } n > s \end{cases} \tag{1.21}$$

$$P_n = C_n P_0 = \begin{cases} \frac{l^n}{\mu^n}\binom{N}{n} P_0 & \text{for } 0 \le n \le s \\ 0 & \text{for } n > s \end{cases} \tag{1.22}$$

$$P_0 = \left[\sum_{n=0}^{s} \frac{l^n}{\mu^n}\binom{N}{n}\right]^{-1} \tag{1.23}$$

$$L = P_0 \sum_{n=0}^{s} n \frac{l^n}{\mu^n}\binom{N}{n} \tag{1.24}$$

This model, as mentioned above, is a modification of the $M/M/s$ queue. We now modify the $M/M/\infty$ queue to consider a finite calling population. Such a model could be described as $M/M/\infty/\infty/N$, however, given the finite population, there could never be more than N busy servers or N clients in the queue. We therefore refer to this system as an $M/M/N/N/N$ queue. The model is described thus:

$$\lambda_n = \begin{cases} (N-n)l & \text{for } n = 0, 1, \ldots, N-1 \\ 0 & \text{for } n = N \end{cases} \tag{1.25}$$

$$\mu_n = n\mu \quad \text{for } n = 1, 2, \ldots, N \tag{1.26}$$

$$C_n = \begin{cases} \frac{l^n}{\mu^n}\binom{N}{n} & \text{for } 0 \le n \le N \\ 0 & \text{for } n > N \end{cases} \tag{1.27}$$

$$P_n = C_n P_0 = \begin{cases} \frac{l^n}{\mu^n}\binom{N}{n} P_0 & \text{for } 0 \le n \le N \\ 0 & \text{for } n > N \end{cases} \tag{1.28}$$

$$P_0 = \left[\sum_{n=0}^{N} \frac{l^n}{\mu^n}\binom{N}{n}\right]^{-1} = \frac{1}{(1 + l/\mu)^N} \tag{1.29}$$

$$L = P_0 \sum_{n=0}^{s} n \frac{l^n}{\mu^n}\binom{N}{n} = \frac{1}{(1 + l/\mu)^N} \sum_{n=0}^{s} n \frac{l^n}{\mu^n}\binom{N}{n} = \frac{N(l/\mu)(1 + l/\mu)^{N-1}}{(1 + l/\mu)^N}$$
$$= \frac{Nl}{l + \mu} \tag{1.30}$$

The equations for P_0 and L are elegantly simple, especially compared to the same measures from the seemingly similar $M/M/s/s/N$ queue. In order to derive these terms we use a couple of tricks: Equation (1.29) utilizes the generating function for the binomial coefficient; equation (1.30) involves rewriting the factorial terms based

on $N-1$ instead of N.

As far as we know, deriving these values by solving the steady state probabilities has not been mentioned in the literature. However, it is a fairly trivial result that is also possible to find in other ways. For instance, the effective arrival rate, λ_e, is equal to the per capita arrival rate, l, times the average number of clients not in the queue: $L_0 = N - L$. Using Little's law, we know that $L = \lambda_e W$. Further, the average waiting time is simply $W = 1/\mu$, because there is no delay for service. Putting these pieces together gives $L = l(N-L)(1/\mu)$. Solving for L yields equation (1.30).

1.4.5 Open Jackson Networks

In 1963, Jackson showed that performance measures could be easily calculated for certain networks of queues [49]. Rather than consider a queueing model that consists of a single queue of clients waiting for one service (with one or more servers), an open Jackson network consists of multiple services, each with its own queue. Clients complete service at one queue and transition to another queue or out of the system according to routing probabilities. These networks assume Markovian arrivals and service times, albeit in a more generalized fashion than illustrated in the single-queue examples above; both arrival and service rates are allowed to depend on the number of customers in the corresponding queue. An open Jackson network has an elegant closed-form solution, but one must assume that the clients arriving at the network are coming from an infinite "calling population". The implication of this assumption is that external arrival rates are unaffected by what goes on in the network because there are always more clients who show up for service.

Two additional assumptions (hereafter referred to as "Jackson network assumptions") must be made for a network to be a Jackson network: 1) Every station must be visited by some sample path with positive probability; and 2) every client must exit the system with probability 1.0.

The notation introduced in §1.4.1 is expanded upon to represent networks of queues:

j = $1, 2, \ldots, J$ stations (we also use i for stations).

$\mathscr{J}$ = the set of all stations.

$\bar{\mathscr{J}}$ = the set of capacitated stations.

λ_j = aggregate arrival rate at station j.

ζ_j = external arrival rate at station j.

z_j = external per-person arrival rate at station j (for use with a finite calling population where the input parameter is of this form and ζ_j is calculated from the model solution).

μ_j = service rate at station j. Note that $\mu_j = 1/LoS_j$.

p_{ij} = routing probability for transitioning from station i to station j.

L_j = expected number of clients at station j.

N = number of people in the total population (calling population).

A key result for the Jackson network is that each queue, or service, can be analyzed in isolation once its aggregate arrival rate has been determined. Determining the aggregate arrival rates of all stations involves solving a system of equations based on the fact that, in steady-state, each service's aggregate arrival rate equals its aggregate departure rate. The resulting set of equations is:

$$\lambda_j = \zeta_j + \sum_{i \in \mathscr{J}} p_{ij} \lambda_i, \quad \forall j \tag{1.31}$$

If the above network is a Jackson network (in which each station is able to reach steady-state) then there exists a unique equilibrium probability distribution (by Theorem 4.5 from [49]). Conversely, every such equilibrium probability distribution induces a positive solution to equation (1.31). The existence of a unique positive solution to the λ values follows from this result.

This system is typically solved using matrix algebra, where P is the square J-dimensional routing probability matrix (and P is clearly invertible given that there is a unique positive solution):

$$\boldsymbol{\lambda} = (I - P')^{-1} \boldsymbol{\zeta} \tag{1.32}$$

With the calculated aggregate arrival rates, λ_j, the individual queues (stations) can be analyzed using the simple models discussed above, or using variations of these models.

1.4.6 Finite Calling Populations

Assuming an infinite calling population is fairly typical in many applications, such as call centres (Koole and Mandelbaum [58]) or manufacturing settings (Bitran and Dasu [13]). However, in the context of the DTES the population is clearly not infinite. If, for instance, a policy change resulted in a drastic increase in the number of clients in treatment, then fewer clients would be outside the system generating "external" arrivals at the entry points, and these external rates would decrease.[1] However, were we to assume an infinite calling population, this feedback would not be represented in the model. We therefore must represent the total population as a finite calling population, thereby necessitating external arrival rates based on the number of clients not in the system, which is in turn based on the number of clients in the system and on the total population, N.

Using a fixed, finite population conveys the idea that no clients enter or leave the DTES, which is not true. However, this assumption is reasonable for the purpose of these analyses if we assume that clients who leave are replaced by similar clients who enter the system. (Modelling the flows in and out of the population is a task for a future version of the model, and will be explored in more detail at that time; it is beyond the scope of this dissertation.)

1.4.7 Closed Queueing Networks

Ignoring the computational challenges, the most straightforward manner for representing a network with a finite calling population is with a closed queueing network. The subsequent section addresses approximating closed networks with open networks, which turn out to require significantly less computation. But first we briefly review the history of closed queueing networks and discuss ways to analyze them. Some of these approaches extend to multi-server stations, and some also include multiclass clients (i.e., different populations moving through the same network)

In 1967, Gordon and Newell [45] formalized and extended Jackson's work to closed queueing systems with exponential service times. They presented the product

[1] This rate decrease is based on the assumption that if there are fewer clients in a population periodically using a service, then the use of that service by that population would decrease proportionally to a decrease in the population. The converse, that somehow clients would compensate for an overall drop in the external population size by using services more frequently, is hard to justify.

form solution analogous to that of the open queueing network. This product form solution includes a normalization constant, $G(N)$, which ensures that the sum of the probabilities across all possible states equals one. However, no efficient way to calculate the normalization constant was introduced at the time, meaning that a brute-force approach involving the enumeration of the entire state space was implied. Given J stations and a population of N, the number of terms to be evaluated in the enumeration is $\binom{J+N-1}{N}$ [20]. This need for enumeration meant that only the smallest problem instances could be solved.

In due course, more efficient approaches for dealing with closed networks began to appear in the literature. Five methodologies are briefly discussed: The convolution algorithm, mean-value analysis, the integral approach, Monte Carlo summation, and using generating functions for the normalization constant. For each methodology, the seminal paper and select others are cited.

Convolution Algorithm The first computational advance following the work by Gordon and Newell came in 1973 from Buzen [20]. This paper included an approach, called the convolution algorithm, for calculating the normalization constant by recursively determining $G(1)$ through $G(N)$ for one station, and then solving for each additional station until all stations are included. As such, the convolution algorithm only requires that a $J \times N$ table be calculated (using very simple operations). Furthermore, only a single N-dimensional column must be stored at each step of the process. In its basic form this algorithm assumes $M/M/1$ stations and a single class of clients.

Mean-Value Analysis The next major idea for working with closed networks followed a different tack, and came in 1980 from Reiser and Lavenberg [92]. Rather than try to calculate the normalization constant, the mean-value analysis approach directly calculates several performance measures of the network in equilibrium, including the mean number of customers in each queue and the mean waiting time. It accomplishes this recursively, through the fact that a customer arriving at the system sees a system in equilibrium with one fewer customer. The mean-value analysis algorithm also extends to multiclass networks; the algorithm requires a nested loop for each class of customers.

The Integral Approach Bell Labs produced considerable research on queueing networks including the integral approach, which was packaged as the PANACEA software in the early 1980s, as described by Ramakrishnan and Mitra [90]. This approach replaces summations with integrals and uses these asymptotic expansions to recursively solve for measures such as mean queue length and expected waiting time. It makes use of pseudo-networks, which are related to the original network but have different parameters and no infinite-capacity servers. Multiclass networks are able to be dealt with, however, the earlier versions could not handle load-dependent service rates which means that multi-server stations were not initially included. As well, assumptions are made about "normal usage" which is defined as utilizations less than 0.85. (When the load is "heavy", the bounds returned by the algorithm can be quite large). Later related work included load-dependent servers [77].

Monte Carlo Summation In 1993, Ross and Wang introduced an approximate approach for analyzing large, multiclass, closed queueing networks: Monte Carlo summation [95]. This method used sampling techniques to derive performance measures, based on a combination of summations and integral representations. Instead of returning bounds, like PANACEA, confidence intervals are calculated. And unlike PANACEA, heavy loads are handled effectively. Initially, multi-server stations were excluded, but a revised version in 1997 [96] added this functionality.

Generating Functions for the Normalization Constant Choudhury *et al.* introduced another approach in 1995 that provides an efficient way of calculating the normalization constant [29]. This method replaces the normalization constants (for the full network and associated networks for which performance measures are desired) with analogous generating functions and then numerically inverts these generating functions in order to calculate the normalization constant(s). This approach is demonstrated on multi-chain closed networks with single-server stations and/or infinite-capacity stations. Although more complex, the approach can be extended to load-dependent cases such as for multi-server stations.

Additional methodologies that have not been discussed here come in various flavours, including extensions to the above approaches (e.g., Lam and Lien [63]), hybrid methods (e.g., Reiser [91]), and others (e.g., Casale [24]).

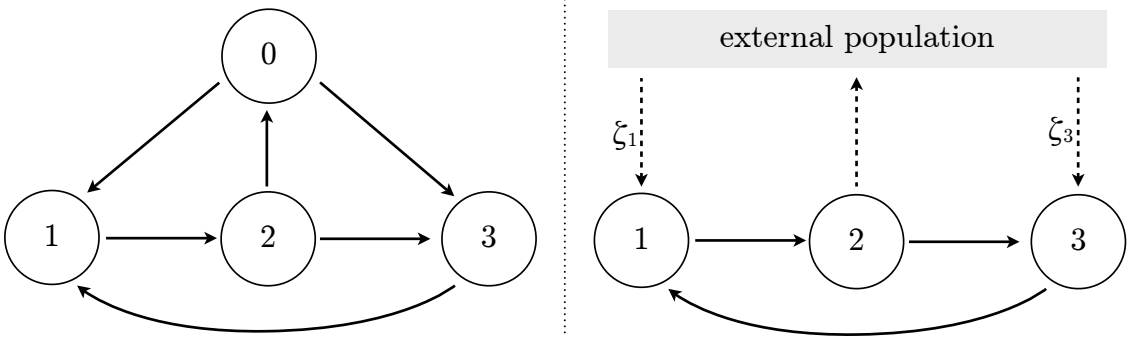
All of the methods discussed thus far work with the product form of a closed queueing network and calculate or approximate either 1) the normalization constant in order to derive performance measures or 2) the performance measures directly. Yet another set of approaches does not involve the product form of a closed network: Approximating a closed network with an open queueing network.

1.4.8 Approximating a Closed Queueing Network with an Open Queueing Network

The difference between an open and a closed queueing network is the size of the population; the former is infinite and the latter is finite. With an infinite population, there are always new customers entering the network, whereas with a finite population the same customers cycle back through the network indefinitely. In the latter case, the activity in the network affects the rate at which customers cycle back, or re-enter, the network.

With some rearrangement, however, a closed network can be made to look like an open network. By removing a single station, it essentially becomes the external population. In Figure 1.1 station 0 has been removed. Flows previously entering station 0 now enter the external population and flows previously leaving station 0 now leave the external population and are termed external arrivals. However, the external arrival rates, ζ_j, must approximate the service times of the removed station if the open network is to approximate a closed network. These rates are calculated by taking the service time of the removed station and multiplying by the corresponding routing probability.

Figure 1.1: Converting a closed network *(left)* to an open network *(right)*: Station = oval; route = arrow; route to/from external population = dashed arrow.



The main approach that uses an open network to approximate a closed one is the fixed population mean (FPM) method formalized by Whitt in 1984 [119]. This approach sets the external arrival rates so that they are proportional to the mean size of the "external" population in equilibrium. To determine the mean size of the external population one takes the total calling population, N, and subtracts the sum of all mean queue lengths, $\sum_{i=1}^{J} L_i$ in the network (where queue lengths include clients in service). The mechanics of the FPM method are explained in §4.1.3, and the quality of the approximation is discussed in §4.2.1.

Chapter 2

Workforce Planning Model

2.1 Basic Model

In this chapter and the next we treat the BCCA model and application. We begin by describing a simplified version of the model.

The basic model is a multi-period assignment problem with side constraints and a GP objective function. The side constraints, enforcing minimum and maximum durations, redundancy, and minimum experience levels, make the problem non-trivial to solve. We use GAMS software to represent the model and CPLEX to find the MIP solution—an assignment of agents to tasks for each period of the planning horizon. In a minor departure from the pure assignment problem, certain tasks in our model require multiple agents. Additionally, we are unconcerned with agent–task assignment costs; we simply wish to meet all task requirements.

Formulating the model in the language of the radiation therapy department, "agents" are RTs, "tasks" are referred to as areas (an area may comprise several daily tasks), and "periods" are quarters (so that a two-year model has a horizon of eight quarters). We designate $i \in I$ RTs, $j \in J$ areas, and $t = 1, \ldots, T$ periods. In some constraints we also use k to represent preceding areas and u to represent time periods. We represent the assignments with binary variables x:

$$x_{i,j,t} = \begin{cases} 1 & \text{if RT } i \text{ is assigned to area } j \text{ in period } t \\ 0 & \text{otherwise} \end{cases} \qquad \forall i, j, t$$

In order to model the duration constraints, we require binary variables y to track the beginning of sequences, where a sequence is defined as a continuous block of one or more periods during which an RT works in one area:

$$y_{i,j,t} = \begin{cases} 1 & \text{if RT } i \text{ begins a sequence in area } j \text{ in period } t \\ 0 & \text{otherwise} \end{cases} \qquad \forall i, j, t$$

We model experience with binary variables z to track when RTs have the necessary experience in one area in order to work in another area (using $k \overset{\text{Exp}}{\Rightarrow} j$ to denote that experience is required in area k in order to be assigned to area j):

$$z_{i,j,k,t} = \begin{cases} 1 & \text{if RT } i \text{ has experience in area } k \text{ to work in } j \text{ by } t \\ 0 & \text{otherwise} \end{cases} \qquad \forall i,j,k,t : k \overset{\text{Exp}}{\Rightarrow} j$$

Finally, we require continuous variables $v^{(1)}$ and $v^{(2)}$ to track and penalize violations of the two soft constraints—maximum duration and redundancy. These two variables appear in the objective function, allowing us to define the optimal solution as one that best balances these two different goals.

$v^{(1)}_{i,j,t}$ = violation of the maximum duration constraint ≥ 0 $\qquad \forall i,j,t$

$v^{(2)}_{j,k,t}$ = violation of the redundancy constraint ≥ 0 $\qquad \forall j,k,t : k \overset{\text{Exp}}{\Rightarrow} j$

The model, shown here, is described in more detail below:

$$\min C^{(1)} \sum_{i,j,t} v^{(1)}_{i,j,t} + C^{(2)} \sum_{j,k,t} v^{(2)}_{j,k,t} \tag{2.1}$$

$$s.t. \sum_{j} x_{i,j,t} \leq 1 \qquad \forall i,t \tag{2.2}$$

$$\sum_{i} x_{i,j,t} \geq D_{min}(j,t) \qquad \forall j,t \tag{2.3}$$

$$x_{i,j,t} - x_{i,j,t-1} \leq y_{i,j,t} \qquad \forall i,j,t \tag{2.4}$$

$$\sum_{u=t}^{t+S_{min}(j)-1} x_{i,j,u} \geq S_{min}(j) \cdot y_{i,j,t} \qquad \forall i,j,t \tag{2.5}$$

$$\sum_{u=t}^{t+S_{max}(j)} x_{i,j,u} \leq S_{max}(j) + v^{(1)}_{i,j,t} \qquad \forall i,j,t \tag{2.6}$$

$$E_{req}(j,k) \cdot z_{i,j,k,t} \leq E_{st}(i,k) + \sum_{u=1}^{t} x_{i,k,u} \qquad \forall i,j,k,t : k \overset{\text{Exp}}{\Rightarrow} j \tag{2.7}$$

$$x_{i,j,t} \leq z_{i,j,k,t} \qquad \forall i,j,k,t : k \overset{\text{Exp}}{\Rightarrow} j \tag{2.8}$$

$$\sum_{i} z_{i,j,k,t} \geq f \cdot D_{min}(j,t) - v^{(2)}_{j,k,t} \qquad \forall j,k,t : k \overset{\text{Exp}}{\Rightarrow} j \tag{2.9}$$

$$x,y,z \in \{0,1\}; \quad v^{(1)}, v^{(2)} \geq 0$$

Constraints (2.2) and (2.3) form the core of the model. Constraint (2.2) states

that each RT must be assigned to at most one area in a period; the inequality allows for RTs to be left unassigned, which is important once staff leaves are included. Constraint (2.3) requires that the number of RTs assigned to an area cannot be less than the minimum staffing demand, D_{min}; the inequality allows the model to assign more RTs than required because, in practical applications, all available staff are typically given assignments. If our model consisted of only these two constraints, the x variables, and a constant objective function, we would have the feasibility version of the multi-period assignment problem with the distinction that some tasks require multiple agents. Without assignment costs, or any constraints that span multiple periods, we could simply take a single-period solution and repeat it for all periods $t = 1, \ldots, T$. The remaining constraints, while making the problem more compelling and more applicable, clearly prevent this approach.

Constraints (2.4) through (2.6) deal with minimum and maximum durations of work sequences. The minimum duration concept is motivated by two aspects of RT work: 1) Most task areas require a short reorientation period to allow RTs to refresh their knowledge and to learn about new technology or recent changes; and 2) RTs report a desire to not move areas too frequently because they find it disruptive. The maximum duration concept is motivated similarly: 1) Some task areas can lead to repetitive stress injuries; and 2) RTs report a desire to not work in one area for too long in order to more easily gain exposure to other areas and to increase job satisfaction. The minimum and maximum sequence lengths for an area j are given by the constants $S_{min}(j)$ and $S_{max}(j)$, respectively.

Minimum durations are described by the first two of these constraints: equation (2.4) forces y to equal one at the beginning of a sequence; equation (2.5) forces all x variables in a sequence to equal one, which means that if an RT begins a sequence, he or she must be assigned to that area for the minimum duration of that sequence. In a model that includes just the core multi-period assignment problem as well as the minimum duration constraints, it is still possible to find a solution for the first period and then repeat it for all periods. However, maximum durations make such an approach impossible. Additionally, in practice, maternity leaves and other departures by RTs will ensure that the single-period solution cannot simply be repeated.

Unlike our approach to modelling minimum durations, we have chosen to model maximum durations as a soft constraint—a restriction that can be violated at a

cost. The costs of violating soft constraints are adjustable to reflect the relative importance of the various soft constraints and are included in the GP objective function. Constraint (2.6) specifies that for a given RT i and area j, every possible continuous block of periods one period longer than the maximum duration must have at least one period in which the appropriate x variable equals zero, or else a violation occurs and is penalized via $v^{(1)}$ in (2.1).

For many applications, the multi-period assignment problem with minimum and maximum durations may be sufficient. In radiotherapy, however, the acquisition of skills is a key part of the staff planning process. Without tracking experience, the model is likely not sufficiently realistic to be of use in an application.

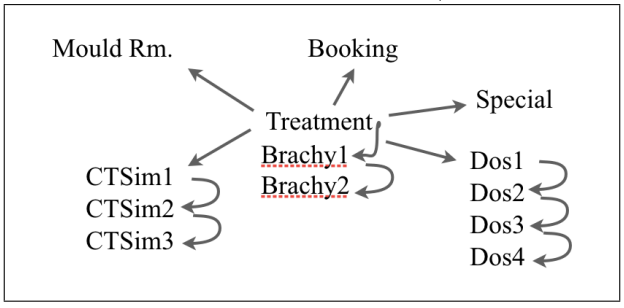
2.2 Experience and Redundancy

The most interesting component of the basic model is the experience requirement. To our knowledge, our model is the first workforce planning model to include *both* duration and experience requirements. In practice at the Vancouver Centre, RTs will typically receive close to five years of experience in the treatment area, and then subsequently in simpler planning areas, before being assigned to the most complex planning area.

2.2.1 Experience Constraints

The z variables and related constraints are only defined for areas k and j where experience in k is required for j. Figure 2.1 shows the experience precedence relationship for the BCCA application.

Figure 2.1: Experience precedence requirements *(does not show times required)*.



The experience constraints work as follows: Constraint (2.7) forces z to equal

zero when the RT does not have the required experience; constraint (2.8) only allows assignment of RTs to areas when the necessary experience exists. These constraints depend on E_{st}, the starting experience (at time $t = 0$) for each RT i in each area k, and E_{req}, the number of periods of experience required in area(s) k to work in area j.

Adding so many binary variables to our model may seem imprudent. In fact, a more concise formulation of the experience requirement uses continuous (non-negative) variables to track the number of periods of experience each RT has in each area, and with some minor rearranging, these continuous variables can be folded into the formulation to eliminate them. However, not including the binary z variables causes problems when we add the redundancy constraint. (Please see Appendix A for more discussion on the alternate formulation of the experience requirements).

2.2.2 Redundancy Constraints

Creating a workforce plan for the next two to three years necessitates flexibility. In practice, it is impossible to adhere to a plan no matter how good it is; RTs take leaves, gain experience at different rates, or end up needing to be reassigned for other reasons. In order to meet our original goal of creating a plan that is robust against these types of variability, we must encourage the model to create plans that are robust against unforeseen circumstances. In practice we find that maternity leaves and early retirements cause the most serious planning problems, but we have nevertheless chosen a broad approach that can handle other types of uncertainty as well.

When RTs take unanticipated leaves, the department must move other RTs into the vacant areas. To retain RTs with experience who are able to fill in these vacancies, the model must encourage redundancy in all of the more complex areas. We achieve this redundancy by adding a redundancy constraint (2.9) to the formulation. This soft constraint uses the $v^{(2)}$ variables to track redundancy violations, thereby encouraging the model to find solutions in which the number of RTs with experience in any particular area is a factor, $f \geq 1$, times the staffing demand requirement D_{min}. Finding the best value for f is investigated through simulation in §3.2.

2.2.3 Objective Function

The model objective (2.1) is to obtain the smallest combined violation of the maximum duration constraint and the redundancy constraint. We use coefficients $C^{(1)}$ and $C^{(2)}$ to weight the two different types of violations captured by $v^{(1)}$ and $v^{(2)}$ and then minimize this weighted sum. $C^{(1)}$ and $C^{(2)}$ are set using trial and error (actually, one can be standardized to 1.0 while the other is changed relative to it); there are no data that suggest a formal way to calculate relative costs of the two different types of violations. In practice, we find the solutions are not very sensitive to these two weights.

Two techniques we have chosen not to employ deserve mention: Robust optimization and stochastic programming. Robust optimization finds the optimal solution for bounded (and often worst-case) inputs. Because employee leaves involve binary values—either an employee is on leave or she isn't—it is difficult to create inputs for a robust optimization model without getting extremely conservative results. Lin *et al.* [67] demonstrate the use of robust optimization in a scheduling context, but they only examine variability in continuous inputs: Task durations, market demand, and material prices. Stochastic programming relies on probability distributions and would be more fitting to our model, however, given that our MIP is already quite large, turning it into a stochastic model would yield excessive solution times. For instance, Punnakitikashem *et al.* [88] solve a stochastic nurse assignment model in 30 minutes (not to optimality) for a much smaller instance than we require that assigns two to four nurses to roughly twenty patients over eight one-hour periods. Denton *et al.* [38] compare robust optimization, stochastic programming, and a heuristic approach in a surgery block scheduling model, but they also only introduce variability into continuous inputs.

Chapter 3

Workforce Planning Application and Simulation

3.1 BCCA Application

Prior to implementing our model at the Vancouver Centre of the BCCA, the manager of the RT workforce created daily schedules for the upcoming 30 days using a simple Excel spreadsheet but did almost no longer-term planning. This manager, whose title is "chief RT", is responsible for scheduling and planning. Using a two- to three-year time horizon in our model, he can create staff plans for approximately 90 RTs working in more than a dozen areas. The planning process consists of creating a new plan for the upcoming two calendar years every November. Additionally, the plan can be rebalanced every May for July through December. Each plan incorporates the first several quarters from the previous plan with little or no changes while making substantial changes further out.

One motivation is clearly to enable the chief RT to create better plans in less time. Another motivating factor came from the RT staff; in 2007, RTs throughout B.C. engaged in a visioning exercise—including a survey exploring job satisfaction—to determine strategic directions for the RT group. One of the findings was that RTs desired more visibility into the department workings and more involvement in their own career path development. The human resources department commissioned each centre's chief RT to implement processes to address these goals. To support this effort, and recognizing that certain RT skills take several years to acquire, we have created a quarterly planning model that spans two to three years in order to deliver a feasible one-year plan that abides by the longer-term learning constraints.

Building upon the model described in Chapter 2, we now present the system that has been implemented at the Vancouver Centre of the BCCA. This system includes additional hard and soft constraints, a more complex objective function, and perhaps most importantly, a stand-alone application to simplify use. We have

strived to describe the basic model so that it will be generally applicable across a wide range of industries and situations, however, we also recognize that adjustments and additions, such as those described in this section, are necessary to adapt our model to a real-world situation.

3.1.1 Extensions

Several extensions were helpful in implementing our model. An initial position constraint allows us to anchor the beginning of the time horizon to current assignments. The initial position denotes the area to which each RT is assigned at the start of the planning horizon (or, in a less formal sense, it is the area in which each RT has been most involved recently). This constraint assigns each RT to this area for an initial period, $t = 0$, so that the model can take advantage of that assignment for minimum durations.

Because the model doesn't account for personnel issues (e.g., personalities that do not work well together), different learning styles, or a variety of other realistic impediments to creating viable staff plans, we have added a forced-assignment feature. The application allows management to specify certain assignments that must be achieved. A soft constraint enforces these assignments with a fairly high penalty such that the model will essentially obey all forced assignments unless they would lead to infeasibility. Along similar lines, the application allows the user to easily block RTs from certain areas during some or all periods.

The Vancouver Centre uses a pool of more than a dozen "casual" RTs who are brought in on a part-time basis to perform tasks, such as treatment, that require less experience. We include a soft constraint to discourage using this pool of casuals and a hard constraint to prohibit them from working in certain higher-skilled areas. Because these casuals are included in daily scheduling, it makes sense to allow casuals to be assigned to entry-level areas in our staff plan. Yet if too much reliance were placed on them, the permanent staff members might not be able to build the experience necessary for the more difficult areas.

Additional extensions include a soft constraint that encourages all RTs to work at least one quarter in the treatment area each year and a penalty on exceeding the staffing coverage for certain areas. The former constraint achieves plans with a better balance between hands-on treatment experience and some of the more difficult planning and imaging areas; the latter ensures that extra staff—above the demand

requirements—are allocated to the areas in which the chief RT feels they would be most useful.

Changes to the basic model to incorporate part-time RTs are included in constraints (2.3), (2.7), and (2.9). $0 < P(i) < 1$ is an input that specifies the proportion of time each RT works. For example, an RT working 0.75 time would only contribute three-quarters as much to the minimum staffing demand requirements and would attain experience more slowly.

3.1.2 Two Objective Functions

The BCCA application includes two different objective functions. The first, which finds an initial plan, is a typical GP objective that seeks to violate the soft constraints as little as possible while also aiming for redundancy. The penalties for violating the various soft constraints can be tweaked to change their importance, as can the penalty for violating the robustness constraint.

The second objective function is used to "adjust" a plan rather than to create a new one; it induces a solution as close to the previous one as possible:

$$\min\ C^{(a)} \sum_{i,j,t} [x_{i,j,t} + X(i,j,t) - 2x_{i,j,t}X(i,j,t)] \tag{3.1}$$

The deviations (over and under) from the previous plan, $X(i,j,t)$, are given by $(1 - X(i,j,t))x_{i,j,t}$ and $X(i,j,t)(1 - x_{i,j,t})$. The sum of these deviations is:

$$\sum_{i,j,t} [(1 - X(i,j,t))x_{i,j,t} + X(i,j,t)(1 - x_{i,j,t})] \tag{3.2}$$

which can be simplified to yield the objective function in equation (3.1).

In practice, this simple function is extremely useful to the chief RT because he tends to make minor changes to the inputs (for instance, to force-assign a particular RT for one or two quarters) and then rerun the model. Without this feature, a new solution would be optimal yet would not resemble the previous solution, causing major disruptions to the staff planning process by drastically changing the planned work paths that may have already been discussed with staff members. In this situation, the chief RT prefers a slightly less-than-optimal solution that only includes a few modifications to accommodate the changes. This idea of iterating between a mathematical objective function and user inputs has been introduced previously—

see Cohn *et al.* [32] for a healthcare shift scheduling application—but our approach with two different GP objective functions appears to be new to healthcare.

Appendix B presents the full BCCA model.

3.1.3 The Application

The BCCA RT staff planning application includes not just the MIP model but also the interface and the system infrastructure on which everything runs. This tool was originally created as an Excel prototype for the Vancouver Centre. Since then, other members of our team, The CIHR Team in Operations Research for Improved Cancer Care, have created the stand-alone application from the original prototype. The main user of the application is the chief RT, working with several of the RT resource therapists—those responsible for helping to create the daily schedules and for providing input into the longer-term plans.

The workforce planning tool is designed to be simple and informative. It contains input tables plus a dashboard that allows the user to run the model, save scenarios, and examine the resulting plans. Other windows display solution reports. Behind the scenes, the application calls the GAMS model to solve the MIP.

Figure 3.1 is a screen shot of the first few rows of the workforce plan generated by the application. The different areas show up as different colours (shown here in grayscale), allowing a user to quickly glimpse the entirety of the schedule without reading each cell. The names of the individual RTs have been blurred.

Figure 3.1: Sample workforce plan *(first few rows)*.

Therapist Name	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08
[illegible]	Treat	Treat	Treat	Treat	Treat	Treat	Treat	CBCT
[illegible]	CTSim2	Treat	CTSim3	CTSim3	CTSim3	Treat	Treat	MR
[illegible]	Treat	TruBeam	TruBeam	Dos4	Dos4	Treat	Treat	Brachy2
[illegible]	Treat	Treat	Treat	CBCT	CBCT	CTSim1	Treat	CBCT
[illegible]	Dos4	Dos4	Dos4	Treat	Dos3	Dos3	Dos3	Treat
[illegible]	Dos1	Dos1	Dos1	Dos1	Dos2	Dos2	CTSim1	Treat
[illegible]	Treat	Treat	Treat	Treat	Treat	Treat	Treat	TruBeam
[illegible]	Brachy1	Brachy1	**Brachy2**	**Brachy2**	**Brachy3**	**Brachy3**	Brachy3	Treat
[illegible]	CBCT	CBCT	CBCT	Treat	Treat	**TruBeam**	**TruBeam**	**TruBeam**
[illegible]	CBCT	CBCT	Treat	Treat	Treat	Treat	Treat	TruBeam
[illegible]	Dos2	Dos2	Dos2	Treat	Treat	CBCT	CBCT	Dos2
[illegible]	***Dos3***	**Dos3**	**Dos3**	Treat	Treat	CTSim2	Dos2	Dos2
[illegible]								

3.2 Simulation

The purpose of performing simulation is to investigate how the redundancy factor impacts robustness under various leave indexes (where a higher index results in higher likelihoods of RTs going on leave) and workforce sizes. In other words, we use it to test the impact of uncertainty on the quality of staff plans, where the uncertainty arises from not knowing which staff members will go on leave and when they decide to do so.

The redundancy constraint described in §2.2 uses a factor of $f = 2$ to encourage robustness in the BCCA application; we found that in our model this parameter value returns solutions that our users judge to have an appropriate level of redundancy. However, we have also sought to determine if there is a value for the parameter f that provides more robustness without negative consequences. Using simulation to iteratively generate employee leaves and re-solve the model, we have tested several parameter values under different settings. Specifically, we evaluate the extent of understaffing for three levels of redundancy as we also vary the employee leave rate and the RT staff size so that our results can be generalized to other situations.

3.2.1 Simulation Approach and Validation

In practice, the workforce planning application is used to reshuffle plans periodically when leaves are realized. We employ a similar approach in the simulation that combines random number generation and many replications with re-solving of the MIP model, using the second objective function described in §3.1.2. (For a brief overview of simulation, see §1.3).

The approach is this: A workforce plan is created with the desired level of redundancy; some (or none) of the RTs then randomly begin leaves in period 1 (for simplicity, all leaves last for four periods unless a particular RT is randomly assigned a leave while already on leave, in which case it will continue for four more periods from that point); the plan is re-solved using the "adjust" objective function to minimize disruptions to the staff. These last two steps—generating leaves and re-solving the plan—are repeated until eight periods have passed. In each period, the horizon is rolled forward one period so that eight-period plans are always generated. In this way, the redundancy constraint is evaluated for two years' worth of plans.

The simulation model may seem difficult to validate, on the one hand, because there are no workforce plans to compare this system against (there was no system in place for staff planning prior to implementing this one). On the other hand, the only aspect of this situation being "simulated" is staff leaves. The rest of the simulation approach involves solving the MIP and rolling the time horizon forward. Therefore this generation of staff leaves is all that requires validation.

Unfortunately we do not have access to historical staff leave data. What we do have is an anecdotal description of the pattern of leaves. We have been told that, on average, four RTs are on leave at any one time. And that leaves typically last one year. We have also been shown examples of leaves that lasted longer than one year. Our simulation varies the leave index such that an average of four concurrent leaves in a pool of 80 – 90 RTs corresponds to a value toward the lower end of the range tested. In this way we cover the likely amount of staff leaves but also test a lower level and a much higher level of staff leaves. Our simulation removes staff for leaves of four quarters. In this way it corresponds to the typical leave length reported to us. Yet occasionally our simulation model randomly chooses to send an already-on-leave RT on leave for an additional four quarters, thereby capturing the occasionally longer leaves reported to us. Although we are not able to validate the entire simulation framework by comparing results to actual data, we have shown that the random elements of this framework compare accurately to anecdotal reports from the workforce.

3.2.2 Experimental Design

In order to evaluate different parameter values for the redundancy constraint we have designed a full factorial experiment with three factors (see [61] for information on experimental design). The quantity f takes on the values 1, 2, or 3, corresponding to no redundancy, some redundancy, and considerable redundancy, respectively. At the same time we vary the leave index, li, and the size of the RT pool, rt. The leave index can take on values from 0.25 to 2.0 in increments of 0.25; in the implementation we have chosen a leave index of 1.0 corresponds roughly to one out of ten RTs beginning a leave every quarter. The RT pool consists of 10, 20, or 30 staff members. Table 3.1 shows the factors and levels for the experiment.

The outcome of each replication can be summarized as two different means: 1) The proportion of time that a simulation replication results in understaffing; and

Table 3.1: Design of experiment for BCCA simulation.

Factor	Levels
li	0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00
f	1, 2, 3
rt	10, 20, 30

2) the number of person-periods of understaffing. The first is a summary of a binary outcome—either there is some understaffing or there is no understaffing. The second is a count—the sum of understaffing (in RTs) across all periods. Both means have practical uses and are included herein.

The experiment consists of all 72 combinations of levels from the three factors, and we simulate each combination with 500 replications. Preliminary trials with 100, 200, 500, and 1000 replications all provided results suitable for analysis, however with only 100 and 200 replications some of the plots of the aggregated results were jagged. We chose 500 because it was a large enough number of replications to provide smoother results, that is, to get a monotonically increasing/decreasing response as any one factor was increased.

3.2.3 Statistical Analysis

Plots from the simulation, displayed in §3.3, appear to show clear results. With the volume of replications, and given that there are clear outcomes we wish to link to the factors in our experiment, regression is the logical framework within which to perform hypothesis testing. These analyses, presented in the next section, were all conducted using R.

The predictor variables are the three factors from the experiment: The robustness factor f, the leave index li, and the size of the RT pool rt. The response variable encodes the extent of understaffing, once again either as a binary outcome or as a count. The binary outcome fits into a logistic regression paradigm while the count fits with Poisson regression. Both approaches are presented.

This choice of logistic and Poisson models for this data is appropriate and also necessary; if one attempted to use multiple regression models with binary or non-negative (count) explanatory variables the results would be based on the wrong assumptions and could therefore exhibit incorrect inferences (e.g., incorrect p-values) or nonsensical predictions (e.g., a probability of understaffing greater than one or a

count less than zero). (See [43, 71], or any statistics reference on generalized linear models, for more information.)

The factor levels are all quantitative. However, in the regression models it simplifies analyses and allows for better fits to treat some of these quantitative inputs as categorical variables. The factor f, taking on the values 1, 2, or 3, may be (and in fact is) highly nonlinear in its effect on the response variables. We therefore create dummy variables for $f = 2$ and $f = 3$ denoted $f2$ and $f3$, respectively. The rt factor, taking on the values 10, 20, or 30, is nonlinear (but not dramatically nonlinear like f). Arguments can be made for treating it as a quantitative variable requiring a second-order term or as a categorical variable; we have opted for the latter to simplify interpretation of the coefficients for rt and any interaction terms involving this variable. The dummy variables for $rt = 20$ and $rt = 30$ are $rt20$ and $rt30$, respectively. The li factor is also somewhat non-linear in its effect on the response variables, but we nevertheless represent it quantitatively. If we were concerned with precision, the model would include li terms up to at least the fourth-order (all are very significant), however, for the purpose of drawing conclusions about relationships among the factors and the direction of those relationships, the first-order li term is sufficient.

In the logistic regression model we regress the simulation inputs on the binary response variable, y, which equals 1 if there is any understaffing during the model's horizon and 0 otherwise. The regression equation is shown below in terms of p, the proportion of replications with understaffing, and does not explicitly include the interaction terms for simplicity:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 li + \beta_2 f2 + \beta_3 f3 + \beta_4 rt20 + \beta_5 rt30 + \textit{interaction terms} \tag{3.3}$$

In the Poisson regression model the simplest approach would be to regress the simulation inputs on the count of understaffing, y (measured in person-periods). However, the amount of understaffing should increase as the RT staff size increases. In fact, if there is no effect, then the amount of understaffing should increase proportionally with the staff size. We therefore include the rate, rather than the count, of understaffing. Rate equals count divided by exposure, where exposure is the size of the staff pool, rt. The Poisson regression model, shown with only the

first-order terms, is written thus:

$$\log\left(\frac{y}{rt}\right) = \beta_0 + \beta_1 li + \beta_2 f2 + \beta_3 f3 + \beta_4 rt20 + \beta_5 rt30 + \textit{interaction terms} \tag{3.4}$$

We rewrite the model so that the response variable y is separate from the rt term, with the latter being treated as an offset—a separate term with the coefficient fixed at 1:

$$\log(y) - \log(rt) = \beta_0 + \beta_1 li + \beta_2 f2 + \beta_3 f3 + \beta_4 rt20 + \beta_5 rt30 + \textit{interaction terms} \tag{3.5}$$

The following hypotheses will be used to investigate the effects of the three factors in both models; "understaffing" refers to the *probability* of understaffing in the logistic regression model and to the *rate* of understaffing in the Poisson regression model:

1. H_0: $\beta 1 = 0$. Understaffing is not affected by the leave index.

   H_A: $\beta 1 > 0$. Understaffing increases as the leave index increases.

2. H_0: $\beta 2 = \beta 3 = 0$. Understaffing is not affected by the level of robustness.

   H_A: At least one of $\beta 2, \beta 3 < 0$. Understaffing decreases when robustness is introduced.

3. H_0: $\beta 4 = \beta 5 = 0$. Understaffing is not affected by the RT pool size.

   H_A: At least one of $\beta 4, \beta 5 < 0$. Understaffing decreases with the RT pool size.

4. H_0: βs for the li and f interaction terms = 0. The level of robustness affects understaffing comparably, regardless of the leave index.

   H_A: At least one of these βs $\neq 0$. As the leave index varies, changes in the level of robustness affect understaffing differently (and vice versa).

5. H_0: βs for the li and rt interaction terms = 0. The RT pool size affects understaffing comparably, regardless of the leave index.

   H_A: At least one of these βs $\neq 0$. As the leave index varies, changes in the RT pool size affect understaffing differently (and vice versa).

6. H_0: βs for the f and rt interaction terms = 0. The RT pool size affects understaffing comparably, regardless of the robustness level.

H_A: At least one of these βs $\neq 0$. As the robustness level varies, changes in the RT pool size affect understaffing differently (and vice versa).

The simulation results and the analysis of these hypotheses are discussed below.

3.3 Results

The workforce planning application (including earlier iterations) was used for three consecutive years to create RT workforce plans at the BCCA. (The chief RT is experimenting with a combined scheduling/planning approach during the current planning cycle). Anecdotal reports are very favourable both in terms of the quality of plans created and the time saved by using this tool. These plans tend to move RTs to different areas more frequently than was previously accomplished, which is generally appreciated by the workforce. Additionally, they achieve a level of robustness that has enabled the department to easily deal with variability thus far.

The solution time for a two-year plan to 1% optimality is one to two minutes, running on a 2.2 GHz dual core CPU with ample RAM (solution times on a multi-processor server are not drastically faster). The model contains over 100,000 constraints, 50,000 variables, and 400,000 non-zero entries. Almost 80% of the variables are binary. Partially due to some of the practicalities of the application that allow more detailed specification of which RTs can work in what areas at what times, the pre-solve process is able to reduce the problem to around 15,000 constraints and 10,000 variables.

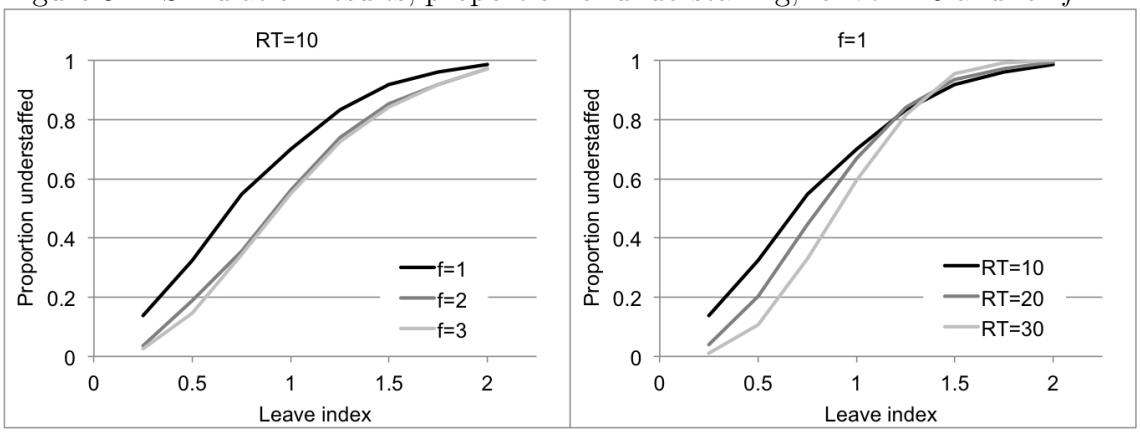
To quantitatively evaluate the workforce plans and the appropriate level of robustness we now turn to simulation and regression analyses.

3.3.1 Simulation Results

Figure 3.2 expresses the results of the simulation as a binary outcome. Observe that the proportion of simulation replications involving understaffing increases as the leave index increases. We include results for all three redundancy levels when the RT pool is fixed at 10 staff members; we also include all three RT pool sizes when the redundancy is fixed at $f = 1$ (no redundancy).

As mentioned, in the Poisson regression we divide the count by the RT pool size in order to get a rate of understaffing. For instance, a replication resulting in a total

Figure 3.2: Simulation results, proportion of understaffing, for $rt = 10$ and for $f = 1$.



shortage of 20 person-periods with $rt = 10$ would have an understaffing rate equal to 2. The average rates from all replications are shown in Figure 3.3; as before we first fix $rt = 10$ and then fix $f = 1$. Some of the differences appear very small when we look at the rates so we also plot the decrease in rate compared with a baseline of $f = 1$. Figure 3.4 shows that an increase in redundancy from $f = 1$ to $f = 2$ corresponds to a rate difference of approximately 0.03. By multiplying this rate by the pool size of 10 to get a count, we see approximately one-third of a person-period less understaffing. In plain terms, the added redundancy leads to the very practical benefit of one less person of understaffing for about a month.

Figure 3.3: Simulation results, rate of understaffing, for $rt = 10$ and for $f = 1$.

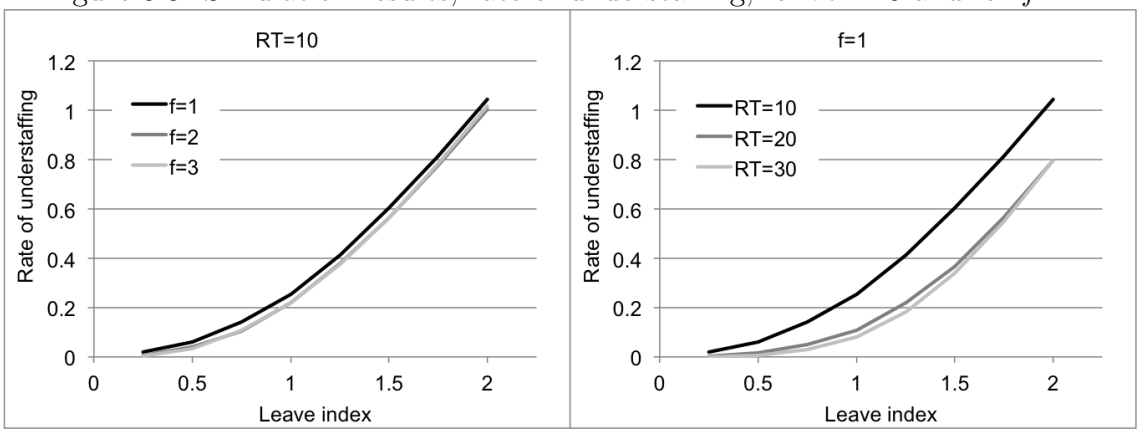
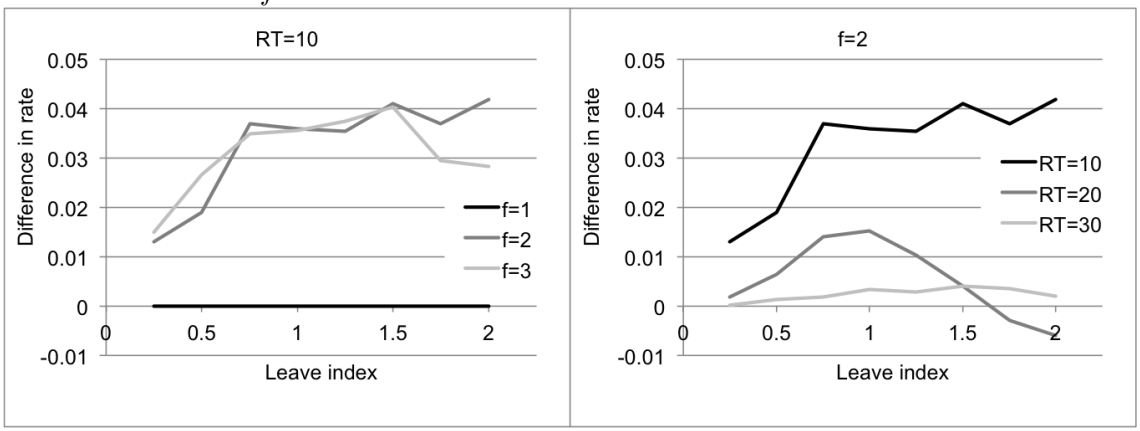


Figure 3.4: Simulation results, difference in rate of understaffing compared to $f = 1$, for $rt = 10$ and for $f = 2$.



3.3.2 Statistical Results

Based on these plots, we expect to find that the level of robustness has a statistically significant effect on the extent of understaffing (hypothesis 2. H_A). In fact, both regression models lead to very clear rejection of the null hypotheses for tests 1 through 5, with barely significant conclusions for some of the terms in test 6.

For the sake of parsimony we remove the coefficients associated with test 6 from both regression models. Table 3.2 shows the output for the logistic regression model and table 3.3 shows the output for the Poisson regression model. All regression analyses were performed in R.

Table 3.2: Logistic regression results.

Variable	Estimate	Std. Error
Intercept	-2.4744***	0.0820
li	3.2092***	0.0833
$f2$	-0.7628***	0.1042
$f3$	-0.9636***	0.1063
$rt20$	-1.1580***	0.0995
$rt30$	-2.2589***	0.1168
$li : f2$	0.2481*	0.1036
$li : f3$	0.3326**	0.1044
$li : rt20$	1.0886***	0.0968
$li : rt30$	2.1576***	0.1148

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.
Based on z-test.

Table 3.3: Poisson regression results.

Variable	Estimate	Std. Error
Intercept	-3.2035***	0.0442
li	1.6854***	0.0262
$f2$	-0.1984***	0.0485
$f3$	-0.3290***	0.0496
$rt20$	-1.2293***	0.0510
$rt30$	-1.6365***	0.0486
$li : f2$	0.1002***	0.0281
$li : f3$	0.1492***	0.0287
$li : rt20$	0.4835***	0.0300
$li : rt30$	0.6785***	0.0285

$^{***}p < 0.001$.
Based on t-test.

The coefficients from both models are statistically significant. We use the signs of these coefficients to discuss the direction of the relationships. First we note that the signs are consistent between the logistic and Poisson models: In both cases, an increase in the leave index li results in an increase in the extent of understaffing; as redundancy f increases, or as staff size rt increases, the extent of understaffing decreases; finally, the interaction terms show that an increase in li coupled with either the addition of redundancy or an increase in staff size results in an increase in understaffing (i.e., an increase in li overpowers any additional redundancy or any increase in staff size).

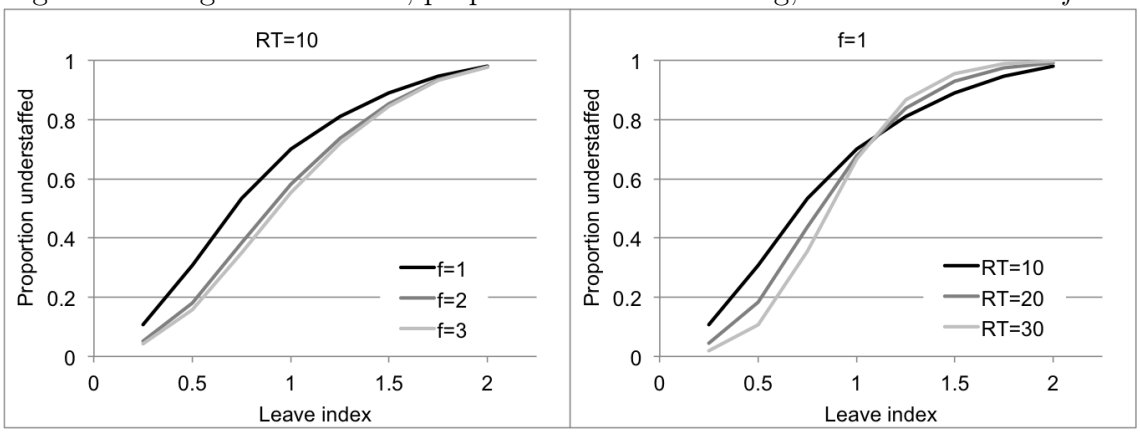
The coefficients from the logistic regression equation can be used to determine odds ratios. For instance, the coefficient for li is 3.20918. Therefore the change in odds associated with a unit increase in li is $e^{3.20918} = 24.76$. In other words, increasing the leave index by 1.0 increases the odds of encountering understaffing at some time during the two year horizon by almost 25 times. Redundancy is also interesting to examine; increasing the redundancy factor from $f = 1$ (baseline) to $f = 2$ corresponds to a change in odds of $e^{-0.76281} = 0.466$ and increasing to $f = 3$ gives us $e^{-0.96355} = 0.382$. Put simply, adding some redundancy *decreases* the odds of encountering understaffing by about 53% while adding considerable redundancy *decreases* the odds by about 62% (or by a further 9% when taken cumulatively). The other coefficients can be interpreted similarly.

To interpret the coefficients from the Poisson regression equation we use incidence rate ratios. The coefficient for li is 1.68537. The change in rate is then $e^{1.68537} = 5.39$, meaning that increasing the leave index by 1.0 increases the rate of understaffing by over five times. Increasing f from 1 to 2 *decreases* the rate of understaffing by about 18% while increasing f from 1 to 3 *decreases* this rate by about 28%. Again, the other coefficients can be interpreted likewise.

To visualize this output we graph the regression lines transformed by the appropriate link function in order to view plots that correspond to the simulation output. (Because these plots are so similar to the simulation output, we graph them separately rather than attempt to display too many lines virtually on top of each other.) Figure 3.5 shows the results of the logistic regression. As above, we include results for the three redundancy levels when $rt = 10$ and also for the three RT pool sizes when $f = 1$. The latter plot, when $f = 1$, shows how the interaction terms mimic the simulation output whereby the probability of understaffing actually *increases* as

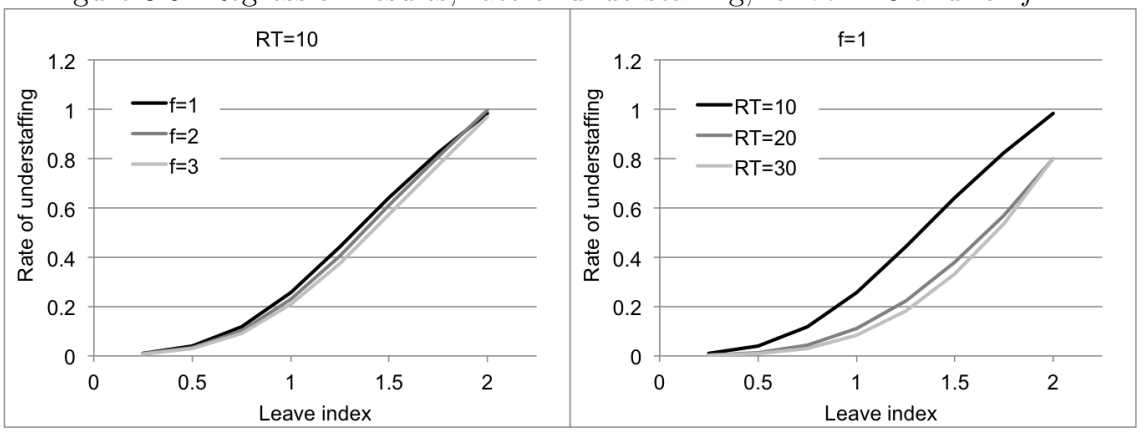
rt increases, but only if li is greater than about 1.2.

Figure 3.5: Regression results, proportion of understaffing, for $rt = 10$ and for $f = 1$.



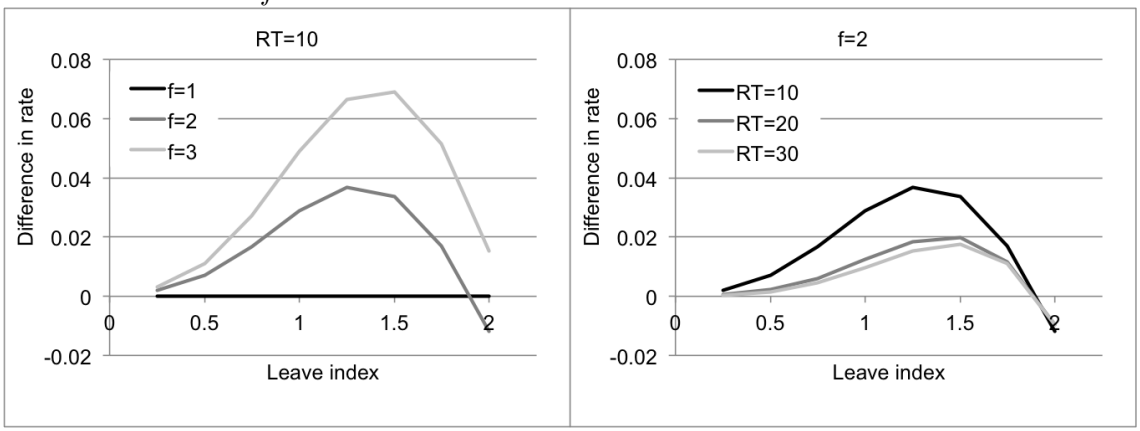
The Poisson regression results are shown in figures 3.6 and 3.7, with the latter displaying the decrease in exposure compared with a baseline of $f = 1$. It is noted that transforming the regression equation by the inverse of the link function introduces bias because the averages determined in the regression model have been manipulated non-linearly. Nevertheless, this bias is very small and the plots are much more informative in this form.

Figure 3.6: Regression results, rate of understaffing, for $rt = 10$ and for $f = 1$.



The conclusions for this research are presented in Chapter 6. The following two chapters address the DTES research.

Figure 3.7: Regression results, difference in rate of understaffing compared to $f = 1$, for $rt = 10$ and for $f = 2$.



Chapter 4

DTES Queueing Model

As described in the Introduction (Chapter 1), we explore two strategic planning topics in this dissertation. This chapter and the following one address the second topic: strategic planning in Vancouver's DTES.

We model services for CCD clients in the DTES with a queueing network. Clients from a finite population arrive at a service, spend some amount of time there, and then transition to another service or back into the Non-treatment Population according to routing probabilities. From the perspective of an individual client, paths through this network tend to be cyclical; a client may arrive at the emergency department (ED), transition to Inpatient care, transition to a long term treatment program, spend time in the Non-treatment Population, and eventually end up back at the ED.

We develop an approximation method for analyzing a closed queueing network. We describe the approximation ratio introduced by this approximation for a simple example—a network with one station and a finite population. We then use simulation to investigate the quality of the approximation for the full network.

For a partial overview of the fundamentals of queueing theory and of queueing networks please see §1.4.

4.1 A Queueing Network Model of DTES Services

There are hundreds of health, housing, social, and criminal justice organization in the DTES. We limit our model to the main health and criminal justice services—those that are most costly and/or most likely to affect health outcomes. These are: Police, Criminal Justice, ED, Acute Care, Inpatient, methadone maintenance treatment (MMT), case management (CM), assertive community treatment (ACT), and Family Practice. We also include a generic entry point to the

system called Other Entry. As mentioned, clients that are in the population but are not in any of these services make up the Non-treatment Population. (In future versions of our model we will include housing and social services, and also revisit the list of included services.) We refer to each of these services using the queueing term "station", noting that the "Police station" is the entire service provided by the Vancouver Police Department and not just the physical building.

4.1.1 The Queueing Model

The DTES queueing model is shown in Figure 4.1. Stations with short-term stays (hours or days) are drawn with ovals; medium term stations (weeks or months) with boxes, and long term stations (months or years) with arrow-boxes (to emphasize the ongoing aspect of these treatments). Capacitated stations have a double outline. Arrows denote paths with non-zero routing probabilities; dashed arrows show paths emanating from or returning to the Non-treatment Population.

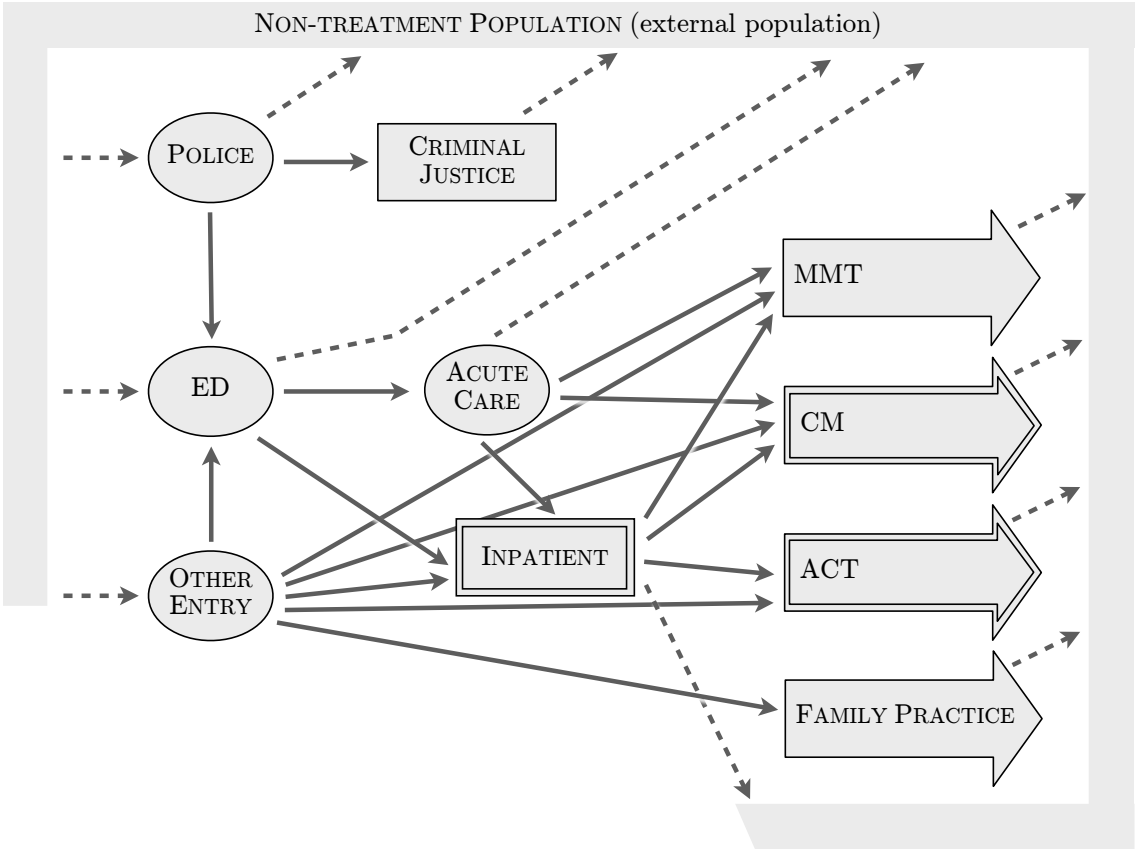
We assume the clients in the DTES comprise a homogeneous population even though we suspect we could group clients demographically or otherwise and observe different arrival rates, routing probabilities, and service rates. In subsequent research we will address this issue by investigating and including sub-populations with different parameter values. Some results, particularly around services tailored toward specific sub-populations, such as ACT, should be interpreted cautiously given this homogeneity assumption.

Another assumption is that an individual client is in exactly one station (including Non-treatment Population) at a time. In order to achieve reasonable results given this simplification we assume that clients receiving multiple treatments are represented in the model as receiving their primary treatment. In this way we ensure no double counting occurs. But we also likely under-estimate MMT and Family Practice usage. In the future we will investigate better ways to address this issue.

4.1.2 Overview of Approximation Approach

Our model of DTES services is based on a closed queueing network which we approximate with an open network using the FPM constraint (the fixed population mean approach was introduced in §1.4.8). This representation reflects the idea that

Figure 4.1: Queueing model of DTES services. Short LoS: oval; medium LoS: box; long LoS: arrow-box; capacitated station: double outline; route: arrow; route to/from Non-treatment Population: dashed arrow.



the population is not completely closed—clients do occasionally enter or leave—but that the population size is roughly constant. Furthermore we assume all $M/M/\infty$ stations so that the number of clients in each station can be computed as a linear function, $L_j = \lambda_j/\mu_j$. These functions (one for each station $j \in J$) ensure the FPM constraint is linear, therefore analyzing the system entails solving a system of linear equations.

A closed network would include a station for the Non-treatment Population. In our FPM approximation, we remove this station so that it becomes the "external" part of the system. This station is the logical choice for several reasons: 1) All services are represented in the model (if we didn't remove the non-treatment population we would have to remove a service-providing station, and it would be more difficult to explain why a station was missing to non-OR audiences); 2) the

input parameters support this arrangement in their given form (we have service rates for all stations, and external arrival rates from the non-treatment population)[2]; and 3) we can calculate bounds in §4.2 on the throughput of this removed station, and it makes more sense to discuss the rate of clients entering or leaving the Non-treatment Population than entering/leaving any other single station; and 4) if we wish to use a semi-open network (an approach that has characteristics of both closed and open networks) for a future version of this model in order to capture a changing non-treatment population, this arrangement is the most appropriate.

For stations that are capacitated, we propose a new approach for approximating these capacities using linear equalities and inequalities; each capacitated station is replaced by three stations, one of which has a bounded aggregate arrival rate. We believe that this combination of the FPM approach, infinite-server stations, and approximated-capacity stations is a new contribution to the literature on closed queueing networks that potentially simplifies the analysis. We refer to this approximate model as a linearized closed queueing network (LCQN).

The following steps outline the transformation of a closed queueing network to the corresponding LCQN:

0. Start with a closed queueing network.

1. Apply the FPM approximation: Remove one station and replace it with an "external population".

2. Represent all remaining stations as $M/M/\infty$ queues, and identify those that are capacitated.

3. Split each of these capacitated stations into three stations using the approximation approach below (§4.1.4).

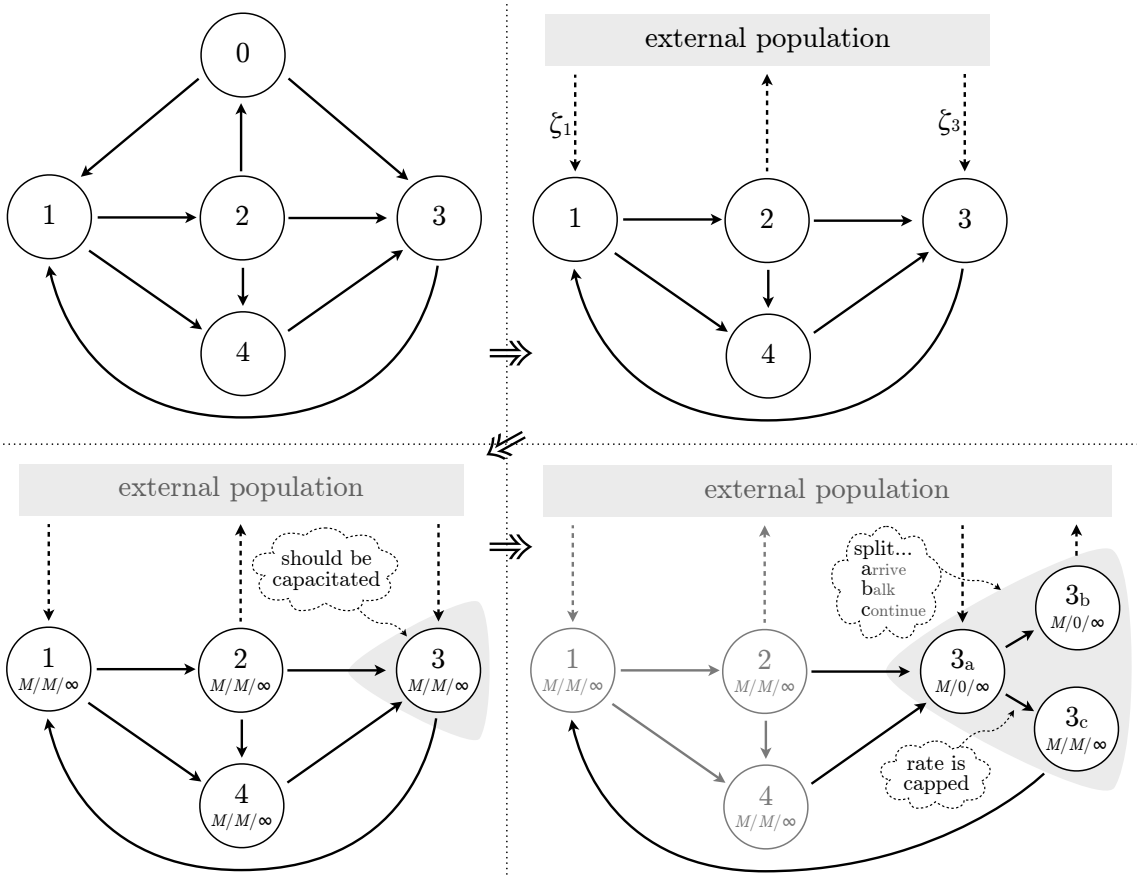
Figure 4.2 illustrates the progression through these steps.

4.1.3 Applying FPM Approach with $M/M/\infty$ Stations

Clients from the (external) Non-treatment Population arrive at station j according to the external per capita arrival rates, $z_j \geq 0 \ \forall j$ with at least one

[2]We could transform the input parameters to enable removal of any station, but because such a step can be skipped it simplifies the explanation of the model.

Figure 4.2: Converting a closed queueing network to an LCQN: Closed network *(top left)*; apply FPM *(top right)*, use $M/M/\infty$ stations but identify stations to be capacitated *(bottom left)*; and split capacitated stations *(bottom right)*. The $M/0/\infty$ queue is defined in §4.1.4.



$z_j > 0$. (Unlike in an open Jackson network, the ζ_j's are now calculated based on the z_j's.) Because all stations are infinite-server stations, service always begins immediately with mean service rate μ_j. The interarrival times and the service rates are assumed to be exponentially distributed, as per the simple models discussed in §1.4. Upon completion of service (where completion includes departure for other reasons besides *completing* treatment), clients transition to another service according to the routing probabilities p_{ij}. With probability $1-\sum_i p_{ij}$ a client transitions back to the Non-treatment Population. We define L_0 as the unknown size of the Non-treatment Population so that we can easily express the total population size as $N = \sum_{i=0}^{J} L_i$. The uncapacitated closed network represented as an open network, including the FPM constraints (equations (4.2) and (4.3)), calculation of the non-

treatment population size, and calculation of the external rates ζ_j is represented with these linear equalities:

$$\lambda_j = \zeta_j + \sum_{i \neq j} p_{ij}\lambda_i, \qquad \forall j \quad (4.1)$$

$$z_j L_0 = \zeta_j, \qquad \forall j \quad (4.2)$$

$$L_0 = N - \sum_i \frac{\lambda_i}{\mu_i} \quad (4.3)$$

This model is based on Whitt [119]. The only difference is that we include only $M/M/\infty$ stations in the open network version, so that the entire system of equations is linear. Examples in the literature that use the FPM approach tend to include single or finite server stations, however, doing so makes the calculation of the L_j, $j = 1, 2, \ldots, J$ terms highly non-linear, and limits the problem size and solution method. (In the general case, equation (4.3) is written $L_0 = N - \sum_i L_i$, where the average number of clients at each station is given by a nonlinear function $L_i = f_i(\lambda_i, \mu_i, s_i, N)$, and a fixed point solution would likely be achievable.[3])

With an eye to using an LP to solve the system of equations, we define $p_{ii} = -1$ and rearrange the above system so that it is clear that λ_j, L_0, and ζ_j ($j = 1, \ldots, J$) are the decision variables:

$$\sum_i p_{ij}\lambda_i + \zeta_j = 0, \qquad \forall j \quad (4.4)$$

$$z_j L_0 - \zeta_j = 0, \qquad \forall j \quad (4.5)$$

$$\sum_i \frac{\lambda_i}{\mu_i} + L_0 = N \quad (4.6)$$

Theorem 4.1. *An open Jackson network with $M/M/\infty$ stations and the FPM constraints has a unique solution with the following property:*

$$\lambda_j > 0 \ \forall j,\ L_0 > 0,\ \zeta_j \geq 0 \ \forall j, \text{ and at least one } \zeta_j > 0. \quad \text{(P1)}$$

Proof. We established the existence of a unique positive solution to the λ variables when the ζ_j's are nonnegative and at least one $\zeta_j > 0$ (see §1.4.5). We require $z_j \geq 0 \ \forall j$, with at least one $z_j > 0$, therefore if $L_0 > 0$ and if equations (4.4), (4.5), and (4.6) have a unique solution, then that solution has property (P1).

[3] See §§1.4.3 & 1.4.4 for examples of functions for L.

We show via linear independence that equations (4.4) and (4.5) have a unique solution by eliminating the ζ variables, yielding:

$$\sum_i p_{ij}\lambda_i \qquad + z_j L_0 = 0, \qquad \forall j \tag{4.7}$$

Yet we already know this system has a unique solution, as long as L_0 is strictly positive. We next consider equation (4.6) which clearly cannot be linearly dependent with equation (4.7) because of the positive constant, N, on the right hand side. Thus, if $L_0 > 0$, the system of equations including the FPM constraints has a unique solution.

It remains to be shown that L_0 is strictly positive. We rewrite the equations in matrix form and solve for L_0:

$$L_0 = N - \left(\frac{1}{\boldsymbol{\mu}}\right)^T \boldsymbol{\lambda} \tag{4.8}$$

$$\boldsymbol{\lambda} = \left(\boldsymbol{I} - \boldsymbol{P}^T\right)^{-1} \boldsymbol{\zeta} = \left(\boldsymbol{I} - \boldsymbol{P}^T\right)^{-1} \boldsymbol{z} L_0 \tag{4.9}$$

$$L_0 = N - \left(\frac{1}{\boldsymbol{\mu}}\right)^T \left(\boldsymbol{I} - \boldsymbol{P}^T\right)^{-1} \boldsymbol{z} L_0 \tag{4.10}$$

$$L_0 = \frac{N}{1 + \left(\frac{1}{\boldsymbol{\mu}}\right)^T \left(\boldsymbol{I} - \boldsymbol{P}^T\right)^{-1} \boldsymbol{z}} \tag{4.11}$$

We have already established that $\left(\boldsymbol{I} - \boldsymbol{P}^T\right)^{-1} \boldsymbol{z}$ is a strictly positive vector, for any nonnegative (and nonzero) vector $\boldsymbol{z}$ when we established this result for the ζ values. The vector $1/\boldsymbol{\mu}$ is strictly positive, therefore the denominator equals 1 plus a strictly positive number. N is also strictly positive, so L_0 must be strictly positive. Furthermore, the denominator is bounded below by 1, giving a range of $0 < L_0 < N$. □

On its own, this model is useful for studying clients flowing through a network in which there is abundant capacity in all stations.

4.1.4 Adding Capacitated Stations—the LCQN Approach

Representing specific stations as uncapacitated, such as POLICE, is a reasonable assumption. Even though a client may incur some small amount of waiting during busy times, all clients can still proceed through service and move on to the next

station or return to the non-treatment population. However, several stations must be modelled with capacities: INPATIENT, CM, and ACT. For other stations we can use the $M/M/\infty$ queue to determine the distribution of L_j, the number of clients in that station, in order to further investigate the uncapacitated assumption.

We present what we believe is a new approach for modelling capacitated stations with a simple yet effective approximation that splits each of these stations into several dummy stations and introduces inequalities to cap arrivals in order to approximate the arrival rate of the corresponding finite-capacity queues.

First we introduce some new notation which will be useful for our approximation: The $M/0/\infty$ queue, with Markovian arrivals, deterministic service time equal to zero, and an unspecified positive number of servers (we denote it as ∞ servers so that it is similar to the other stations in the network). With the LoS equal to zero, clients simply pass right through it, so specifying the precise number of servers is not necessary.

In this new method, each capacitated station $j \in \bar{\mathscr{J}}$ is split into three (a, b, c) stations—j_a for all clients who **arrive** at the station, j_b for clients who observe the station as full and **balk** (i.e., are lost and return to NON-TREATMENT POPULATION), and j_c for clients who **continue** on to receive the service. We model the "a" and "b" stations as $M/0/\infty$ queues, while the "c" station is an $M/M/\infty$ queue.[4]

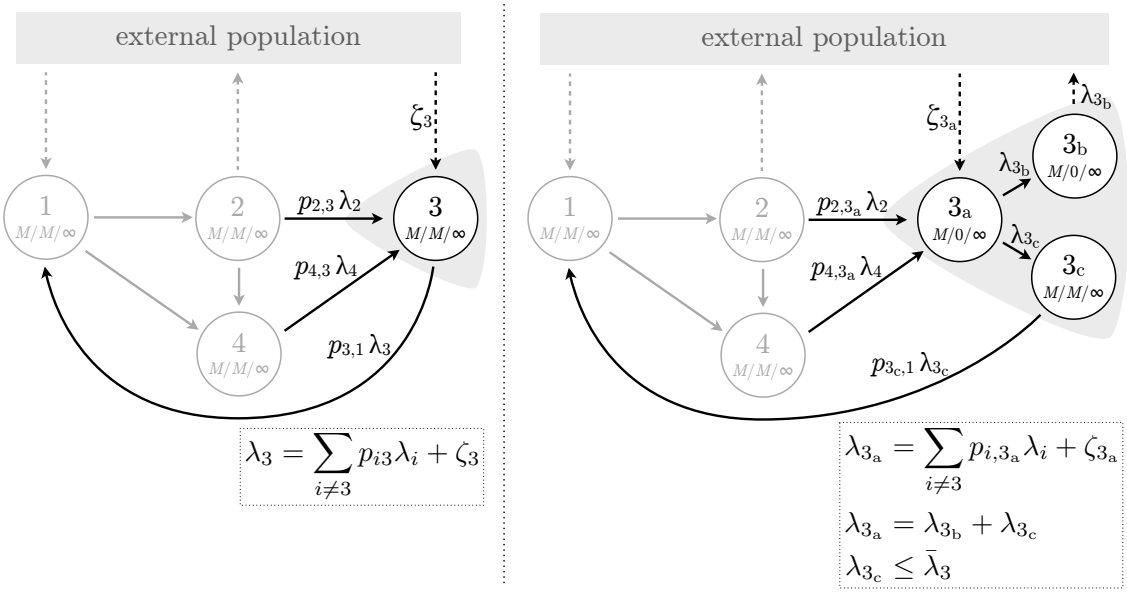
Figure 4.3 illustrates the splitting of a capacitated station ($j = 3$). The uncapacitated version is on the left and the capacitated version is on the right. Associated constraints are shown with each network. The necessary changes to the input data in order to accommodate this split for each such station are:

For each $j \in \bar{\mathscr{J}}$:

- Replace station j with three stations, j_a, j_b, and j_c;
- for each i, rename p_{ij} as p_{i,j_a} and rename p_{ji} as $p_{j_\mathrm{c},i}$;
- set $p_{j_\mathrm{a},j_\mathrm{a}} = -1$ (like the other p_{ii} coefficients), $p_{j_\mathrm{b},j_\mathrm{b}} = 0$, and $p_{j_\mathrm{c},j_\mathrm{c}} = 0$;
- rename ζ_j as ζ_{j_a};
- rename z_j as z_{j_a};
- rename μ_j as μ_{j_c};
- introduce $\bar{\lambda}_j$ to represent the capacitated arrival rate (discussed below);

[4]The "b" stations could be removed from the model, however, we believe the LCQN approach is more clearly illustrated with them present.

Figure 4.3: Example of capacitated station 3 *(left)* split into three stations *(right)* with maximum arrivals capped at $\bar{\lambda}$.



- set all other p, ζ, and z terms involving these three stations to zero[5].

It is also convenient to define several other subsets of $\mathscr{J}$, in addition to $\bar{\mathscr{J}}$, the set of capacitated stations:

$\mathring{\mathscr{J}}^{\infty}$ = the set of uncapacitated stations, i.e., stations that have not been split.

$\mathscr{J}_{\mathbf{a}}$ = the set of "a" stations.

$\mathscr{J}_{\mathbf{b}}$ = the set of "b" stations.

$\mathscr{J}_{\mathbf{c}}$ = the set of "c" stations.

With the aim of ensuring that as few clients as possible balk (travel to the "b" station and then immediately to the Non-treatment Population) we introduce an objective function: $\min \sum_{j\in\bar{\mathscr{J}}} \lambda_{j_b}$. This objective aims to ensure that as much capacity as possible, at capacitated stations, is used. With the requirement that the balking variables be non-negative, we have an LP, the solution of which is an approximate solution to our closed queueing network. The FPM constraint is not new; limiting all stations to $M/M/\infty$ queues and using split stations in conjunction with the FPM approach is, to the best of our knowledge, the primary innovation.

[5]Note that μ_{ja} and μ_{jb} do not enter into the LP because "a" and "b" are $M/0/\infty$ queues.

The complete LP including the equations for split stations is:

$$\min \sum_{j \in \bar{\mathscr{J}}} \lambda_{j_\mathrm{b}} \tag{4.12}$$

$$\sum_{i \in \mathscr{J} \setminus \mathscr{J}_\mathrm{b}} p_{ij}\lambda_i + \zeta_j = 0, \qquad \forall j \in \overset{\infty}{\mathscr{J}} \cup \mathscr{J}_\mathrm{a} \tag{4.13}$$

$$\lambda_{j_\mathrm{a}} - \lambda_{j_\mathrm{b}} - \lambda_{j_\mathrm{c}} = 0, \qquad \forall j \in \bar{\mathscr{J}} \tag{4.14}$$

$$\lambda_{j_\mathrm{c}} \le \bar{\lambda}_j, \qquad \forall j \in \mathscr{J}_\mathrm{c} \tag{4.15}$$

$$z_j L_0 - \zeta_j = 0, \qquad \forall j \in \overset{\infty}{\mathscr{J}} \cup \mathscr{J}_\mathrm{a} \tag{4.16}$$

$$\sum_{i \in \overset{\infty}{\mathscr{J}} \cup \mathscr{J}_\mathrm{c}} \frac{\lambda_i}{\mu_i} + L_0 = N \tag{4.17}$$

$$\lambda_{j_\mathrm{b}} \ge 0 \tag{4.18}$$

Using J for the number of stations, and introducing $\bar{J} = |\bar{\mathscr{J}}| \le J$ for the number of capacitated stations and $Z = \sum_j \mathbf{1}_{\{z_j > 0\}} \le J$ for the number of stations with strictly positive external arrivals (and eliminating the variables ζ_j and constraints (4.16) for which $z_j = 0$), it can be seen that the full LP has $J + 2\bar{J} + Z + 1$ variables and $J + 3\bar{J} + Z + 1$ constraints. There are $\bar{J}$ nonnegativity constraints, therefore the number of remaining constraints is the same as the number of variables. If all stations are capacitated and have external arrivals, we have $4J + 1$ variables and $5J + 1$ constraints.

A result describing the solution form for the full network would be ideal; in lieu of that we present two results—one that describes the solution form for a small example and one that provides a result concerning the solution for any network. Theorem 4.2 describes the form of the optimal solution for a network with a *single* station and a finite calling population; Theorem 4.3 shows that the LP for any network has an optimal solution. (Theorem 4.2 uses notation consistent with that already introduced, except that the subscript j's are omitted because there is only one station.)

Theorem 4.2. *The optimal solution of the LP representing a single station network with the FPM approximation—that is, a network with a single queue and a finite calling population—has one of the following two forms:*

Case 1: *There is no balking and the station is not at capacity (i.e., $\lambda_c < \bar{\lambda}$). The variable values are:*

$$\lambda_b = 0 \tag{4.19}$$

$$\lambda_c = \frac{zN}{1 + \frac{z}{\mu}} < \bar{\lambda} \tag{4.20}$$

$$\lambda_a = \lambda_c > 0 \tag{4.21}$$

$$L_0 = \frac{N\mu}{\mu + z} > 0 \tag{4.22}$$

$$\zeta = zL_0 > 0 \tag{4.23}$$

Case 2: *There is a nonnegative amount of balking and the station is at capacity (i.e., $\lambda_b \geq 0$ and $\lambda_c = \bar{\lambda}$). The variable values are:*

$$\lambda_b = zN - \left(1 + \frac{z}{\mu}\right)\bar{\lambda} \geq 0 \tag{4.24}$$

$$\lambda_c = \bar{\lambda} > 0 \tag{4.25}$$

$$\lambda_a = zN - \frac{z}{\mu}\bar{\lambda} > 0 \tag{4.26}$$

$$L_0 = N - \frac{\bar{\lambda}}{\mu} > 0 \tag{4.27}$$

$$\zeta = zL_0 > 0 \tag{4.28}$$

Proof. The LP for a single station network takes inputs $\bar{\lambda}, z, \mu, N > 0$ and is as follows:

$$\min \quad \lambda_b \tag{4.29}$$

$$-\lambda_a + \zeta = 0 \tag{4.30}$$

$$\lambda_a - \lambda_b - \lambda_c = 0 \tag{4.31}$$

$$\lambda_c \leq \bar{\lambda} \tag{4.32}$$

$$zL_0 - \zeta = 0 \tag{4.33}$$

$$\frac{1}{\mu}\lambda_c + L_0 = N \tag{4.34}$$

$$\lambda_b \geq 0 \tag{4.35}$$

Combining (4.30) and (4.33) gives:

$$-\lambda_a + zL_0 = 0 \tag{4.36}$$

Substituting $\lambda_a = \lambda_b + \lambda_c$ (4.31) and $L_0 = \left(N - \frac{1}{\mu}\lambda_c\right)$ (4.34) yields:

$$-\lambda_b - \lambda_c + z\left(N - \frac{1}{\mu}\lambda_c\right) = 0 \tag{4.37}$$

Solving for λ_c and using (4.32) gives:

$$\lambda_c = \frac{zN - \lambda_b}{1 + \frac{z}{\mu}} \leq \bar{\lambda} \tag{4.38}$$

Solving for λ_b yields:

$$\lambda_b \geq zN - \left(1 + \frac{z}{\mu}\right)\bar{\lambda} \tag{4.39}$$

As a result of the objective function, $\lambda_b = \max\{zN - \left(1 + \frac{z}{\mu}\right)\bar{\lambda}, \quad 0\}$.

Case 1 occurs when $\lambda_b = 0$. Equation (4.20) follows from (4.38); equation (4.21) follows from (4.31). Equation (4.22) comes from using the value for λ_c from (4.20) in (4.34) and then simplifying, and from the fact that z, N, and μ are all strictly positive.

Case 2 occurs when $\lambda_b = zN - (1 + \frac{z}{\mu})\bar{\lambda}$. Equation (4.25) comes from using this value for λ_b in (4.38) and then simplifying, while (4.26) combines $\lambda_a = \lambda_b + \lambda_c = \lambda_b + \bar{\lambda}$ with the value for λ_b. Equation (4.27) follows directly from (4.34) and $\lambda_c = \bar{\lambda}$. □

This theorem shows that, for the simple example, there is either balking or unused capacity. It is possible for there to be both (if the capacity, $\bar{\lambda}$, is exactly what is needed), but it is not possible for there to be unused capacity in conjunction with no balking. A more desirable result would have been to show this situation exists for each station in any network at the optimal solution, however, we have not done so. Instead, we are able to show that any network has an optimal solution (below).

Theorem 4.3. *The LP for the LCQN approximation of any closed network with the Jackson network assumptions described in §1.4.5 has an optimal solution.*

Proof. The LP has a feasible solution: Treat all stations as capacitated and take, for instance, $\lambda_{j_c} = 0$ for all j. (Any uncapacitated station can be represented, without loss of generality, as a capacitated station with $\bar{\lambda}_j$ sufficiently large.) This solution has the following form:

$$\lambda_{j_c} = 0 \qquad \forall j \in \bar{\mathscr{J}} = \mathscr{J} \quad (4.40)$$

$$\lambda_{j_a} = \lambda_{j_b} = \zeta_j = z_j L_0 \qquad \forall j \in \bar{\mathscr{J}} = \mathscr{J} \quad (4.41)$$

$$L_0 = N - \sum_{j \in \mathscr{J}_c} \frac{\lambda_{j_c}}{\mu_j} = N \quad (4.42)$$

Existence of a feasible solution to the LP implies that, if the optimal objective function value is bounded, there is an optimal solution. To show the solution is bounded we consider the objective function: $\min \sum_{j \in \bar{\mathscr{J}}} \lambda_{j_b}$. Because the λ_{j_b}'s are all nonnegative (as per the nonnegativity constraint), the optimal objective function value is bounded below by 0. □

For capacitated stations, the LP redirects all external and internal arrivals to the "a" station (4.13), then sends all clients from this station directly to either the "b" or the "c" station (4.14), and caps all arrivals at the "c" station with a maximum aggregate arrival rate $\bar{\lambda}_j$ (4.15).

The parameters $\bar{\lambda}_j$ must allow the model to achieve an appropriately accurate approximation of a multi-server station with s_j servers. We propose a value to use for these parameters, along with an alternate maximum aggregate arrival rate, $\hat{\lambda}_j$. The proposed value, $\bar{\lambda}_j$, is based on the idea that the aggregate arrivals to that station need to be capped such that the traffic intensity, ρ_j, is close to 1.0, as it will be for any busy station that needs capacitating. In fact, the more servers a station has, the closer the traffic intensity can be to 1.0, so we use a simple way to include this asymptotic behaviour with $\rho_j = 1 - 1/s_j$. The second proposed value, $\hat{\lambda}_j$, equates the capacitated arrival rate to the effective arrival rate of the multi-server station we wish to approximate. It does this by recognizing that the effective arrival rate equals the raw arrival rate multiplied by the probability a client is not blocked. It might seem that by matching a station's maximum arrival rate to the effective arrival rate of the station we wish to approximate we would achieve an exact result, however, the distribution of (and average of) the number of clients in that queue would still differ. Both metrics—effective arrival rate and average number of clients—are investigated

in §4.2.2.

Parameters $\bar{\lambda}_j$ and $\hat{\lambda}_j$ are total arrival rates and are therefore appropriate for a finite population (closed) network model:

1. $\bar{\lambda}_j = \rho_j s_j \mu_j$, with $\rho_j = 1 - 1/s_j$, which gives:
   $$\bar{\lambda}_j = (s_j - 1)\mu_j \tag{4.43}$$
2. $$\hat{\lambda}_j = \lambda_j \cdot (1 - P\{\text{blocking}\}) \tag{4.44}$$

The value of $\bar{\lambda}_j$ in equation (4.43) is bounded above by the theoretical arrival limit. Equation (4.44) is based on matching the effective arrival rate, which equals the total arrival rate times one minus the probability of blocking, which we calculate below in §4.2.2.

4.2 Approximation Bounds and Ratios

4.2.1 FPM

Whitt [119] (§3.3) discusses bounds on the FPM approach for balanced and unbalanced networks. The former refers to networks in which all stations have the same traffic utilizations; the latter refers to networks with stations having different traffic utilizations, which is certainly the case for our model. A bound that compares the network throughput—the rate of flow through the removed station—can be conservatively estimated. For instance, using Whitt's result, an unbalanced network with single-server stations that is otherwise the same size as our model would have an approximate throughput for the closed network that is 1.00013 times higher than the throughput for the corresponding open network with the FPM approximation (with the caveat that severely unbalanced networks might result in a slightly looser bound, something we check in §5.4.2). This number comes from comparing the open throughput, θ^o, with the approximate closed throughput, θ^c_{approx}, using $N = 7500$ clients and $J = 11$ stations (notation has been changed from Whitt's for consistency with out terminology):

$$\theta^c_{\text{approx}} = \theta^o(J + N)/(J + N - 1)$$

Our model does not have single-server stations, but the above equation can also be

used as a bound on networks with multi- and infinite-server stations because the single-server version results in a larger (more conservative) bound.

4.2.2 Approximate Capacitated Arrival Rate

In our model we use $M/M/\infty$ stations (split into a, b, c) to approximate $M/M/s/s$ stations (stations with s servers, and clients who balk if all servers are busy). However, rather than evaluate the quality of this approximation within the FPM paradigm (itself an approximation), we wish to evaluate it on its own. We do so for a system with a finite external population and only one station, which is analogous to that considered in Theorem 4.2.

The single station analogue for an $M/M/s/s$ station under FPM is the $M/M/s/s/N$ queue, which includes the finite calling population without requiring an approximation (i.e., no FPM constraint). (See §1.4.4 for the description and notation of the $M/M/s/s/N$ and $M/M/N/N/N$ queues.)

The single station analogue for the capacitated station in our approximation is the $M/M/N/N/N$ queue, which also incorporates the finite calling population without the FPM approximation. Even though the capacitated station is split into three stations, we can replace it with this single $M/M/N/N/N$ queue which essentially represents the "c" station with capped arrivals (assuming high traffic intensity). Clients who do not make it to "c" simply pass through "a" and "b" then immediately return to the finite calling population, so they can therefore be ignored in this setting. Figure 4.4 helps explain the single station analogues for both situations.

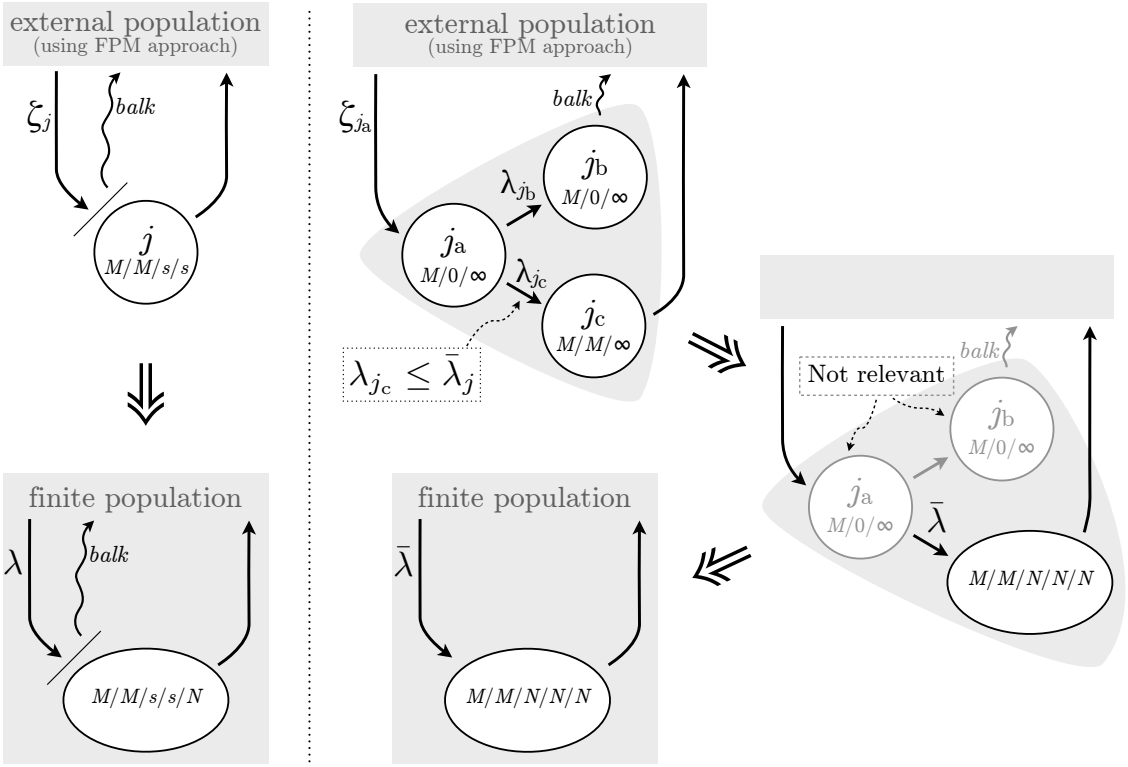
In the larger model we use $\bar{\lambda}_j$ for the capacitated arrival rate, however, with the single station analoque we must use a per capita arrival rate, and therefore use a corresponding capacitated per capita arrival rate, $\bar{l}$. The corresponding l for some λ is given for any state, n, by:

$$l_n = \lambda_n / \mathscr{L}_n \tag{4.45}$$

where $\mathscr{L}_n$ is the number of clients in the queue when the system is in state n.

We only consider high traffic intensities because these are the types of stations we choose to capacitate in our model. (In fact, we are interested in traffic intensities at, or even well above, $\rho = 1$, which can clearly exist when balking occurs.) Because

Figure 4.4: Rationale for $M/M/s/s/N$ and $M/M/N/N/N$ comparison of single station systems. The finite-server station with FPM is converted into an exact finite-server station with a finite calling population *(left)*; the capacitated split station with FPM constraint (single station LCQN) is converted into an exact infinite-server station with restricted arrivals $\bar{\lambda}$ *(right)*.



this small example may not necessarily extend to a larger network, we also address the quality of this approximation in §5.4.2 by using a simulation model to validate the network performance.

The comparison is shown for two performance measures, effective per capita arrival rate, l_e, and average queue length, L, for the $M/M/s/s/N$ and $M/M/N/N/N$ systems (hereafter referred to as the (s) case and the (N) case, respectively). First we give the measures of interest:

$$\begin{aligned} l_e^{(s)} &= l\left(1 - P\{\text{blocking}\}\right) \\ &= l\left(1 - P_s^{(s)}\right) \\ &= l\left(1 - \frac{\frac{l^s}{\mu^s}\binom{N}{s}}{\sum_{n=0}^{s}\frac{l^n}{\mu^n}\binom{N}{n}}\right) \end{aligned} \tag{4.46}$$

$$l_e^{(N)} = l \tag{4.47}$$

$$L^{(s)} = \sum_{n=0}^{s} \frac{n\frac{l^n}{\mu^n}\binom{N}{n}}{\sum_{m=0}^{s}\frac{l^m}{\mu^m}\binom{N}{m}} \tag{4.48}$$

$$\begin{aligned} L^{(N)} &= \sum_{n=0}^{N} \frac{n\frac{l^n}{\mu^n}\binom{N}{n}}{\sum_{m=0}^{N}\frac{l^m}{\mu^m}\binom{N}{m}} \\ &= \frac{Nl}{l+\mu} \end{aligned} \tag{4.49}$$

These measures come from the analysis of $M/M/s/s/N$ and $M/M/N/N/N$ queues in §1.4.4. Equation (4.46) results from (1.22) and (1.23), equation (4.48) combines (1.24) and (1.23), and equation (4.49) combines (1.30) and (1.29).

The approximation ratio for model throughput (where throughput only includes arrivals that receive service) comes directly from (4.46) and (4.47):

$$\frac{l_e^{(s)}}{l_e^{(N)}} = 1 - \frac{\frac{l^s}{\mu^s}\binom{N}{s}}{\sum_{n=0}^{s}\frac{l^n}{\mu^n}\binom{N}{n}} \tag{4.50}$$

If an approximation is perfect, the approximation ratio will be 1.0. While it is difficult to interpret terms involving ratios of exponents and binomial coefficients, we can see that if s is very small then the numerator will be closer to the denominator, and consequently the ratio will be larger. Numerical examples below explore this further.

The approximation ratio for queue length comes from (4.48) and (4.49):

$$\frac{L^{(s)}}{L^{(N)}} = \frac{l+\mu}{Nl}\sum_{n=0}^{s} \frac{n\frac{l^n}{\mu^n}\binom{N}{n}}{\sum_{m=0}^{s}\frac{l^m}{\mu^m}\binom{N}{m}} \tag{4.51}$$

We now turn to examples for insight.

First we discuss how $\bar{\lambda}$ relates to $\bar{l}$ (and how $\hat{\lambda}$ relates to $\hat{l}$). On average, the number of clients outside the station, based on the effective arrival rate at the station, λ_e, is given by $L_0 = N - \lambda_e/\mu$. Using the elementary result in (4.45) we now use $\bar{l} = \bar{\lambda}/L_0 = \bar{\lambda}/(N - \bar{\lambda}/\mu)$ for the following comparisons (similarly for $\hat{l}$ using $\hat{\lambda}$). (In the few cases where $\bar{l} > l$, we use $\bar{l} = l$ because these situations would be uncapacitated in the LP.)

We explore numerical examples for the approximation ratio for throughput, l_e,

and for average queue length, L. These ratios are shown for a variety of s and N values in Table 4.1 with incoming traffic intensity $\rho = 1.0$. To achieve this traffic intensity, we set the per capita offered load $l/\mu = \rho s/N$. (Dividing by the total population, N, rather than the average size of the external population, L_0, makes this offered load a conservative value.) Because the capacitated stations in our model tend to have more incoming traffic than they can handle, we also show the same results for $\rho = 1.5$ (Table 4.2) and $\rho = 2.0$ (Table 4.3).

Table 4.1: Approximation ratios for a multi-server station vs. infinite-server capacitated station with incoming traffic intensity = 1.

s	N	l/μ	$l_e^{(s)}/l_e^{(N)}$	$L^{(s)}/L^{(N)}$	$L^{(s)}/L^{(N)}(\bar{l})$	$L^{(s)}/L^{(N)}(\hat{l})$
2	100	0.02	0.60241	0.61036	1.19679	1.00530
10	100	0.1	0.82070	0.83863	0.84710	1.00519
20	100	0.2	0.91998	0.93599	0.93599	1.00383
10	500	0.02	0.79248	0.79663	0.86779	1.00115
50	500	0.1	0.94076	0.94669	0.94669	1.00088
100	500	0.2	0.99321	0.99457	0.99457	1.00023
20	1000	0.02	0.84959	0.85259	0.87987	1.00058
100	1000	0.1	0.96973	0.97276	0.97276	1.00036
200	1000	0.2	0.99930	0.99944	0.99944	1.00002

Table 4.2: Approximation ratios for a multi-server station vs. infinite-server capacitated station with incoming traffic intensity = 1.5.

s	N	l/μ	$l_e^{(s)}/l_e^{(N)}$	$L^{(s)}/L^{(N)}$	$L^{(s)}/L^{(N)}(\bar{l})$	$L^{(s)}/L^{(N)}(\hat{l})$
2	100	0.03	0.47309	0.48363	1.40863	1.00659
10	100	0.15	0.63022	0.66720	0.96695	1.00761
20	100	0.3	0.72589	0.78071	0.94823	1.00749
10	500	0.03	0.59748	0.60553	0.97982	1.00159
50	500	0.15	0.70493	0.73444	0.97751	1.00176
100	500	0.3	0.79694	0.83755	0.97617	1.00171
20	1000	0.03	0.62955	0.63696	0.97643	1.00085
100	1000	0.15	0.72050	0.74845	0.98610	1.00092
200	1000	0.3	0.81225	0.84980	0.98546	1.00090

Each of these three tables shows the ratio of the effective arrival rates, using uncapped arrivals, under the (s) and (N) models (as stated earlier, the former is the exact calculation and the latter is the approximation). These ratios, $l_e^{(s)}/l_e^{(N)}$, are clearly quite far from 1.0 for the examples with smaller s and N values. As such,

Table 4.3: Approximation ratios for a multi-server station vs. infinite-server capacitated station with incoming traffic intensity = 2.

s	N	l/μ	$l_e^{(s)}/l_e^{(N)}$	$L^{(s)}/L^{(N)}$	$L^{(s)}/L^{(N)}(\bar{l})$	$L^{(s)}/L^{(N)}(\hat{l})$
2	100	0.04	0.38700	0.39926	1.53560	1.00736
10	100	0.2	0.49870	0.54883	1.01635	1.00857
20	100	0.4	0.57367	0.65894	0.99088	1.00872
10	500	0.04	0.46900	0.47962	1.02483	1.00176
50	500	0.2	0.54063	0.58657	0.99756	1.00190
100	500	0.4	0.61181	0.68945	0.99487	1.00191
20	1000	0.04	0.48693	0.49719	1.00647	1.00092
100	1000	0.2	0.54770	0.59293	0.99820	1.00097
200	1000	0.4	0.61809	0.69447	0.99709	1.00098

the ratios of the queue lengths, $L^{(s)}/L^{(N)}$, are also quite far from 1.0 for the same examples.

With our goal of devising a good approximation for the queue length, L, we investigate the ratio when the (N) case uses capacitated arrival rates, $\bar{l}$ or $\hat{l}$ (while the (s) case continues to use the uncapped arrival rate). The ratio for queue lengths using the first capacitated arrival rate, $L^{(s)}/L^{(N)}(\bar{l})$, is close to 1.0 for most of the examples, especially those with higher traffic intensities. The ratio using the second rate, $L^{(s)}/L^{(N)}(\hat{l})$, is very close to 1.0.

In our network model we must use total arrival rates rather than per capita arrival rates (i.e., λ's rather than l's). We use the first approximation, $\bar{\lambda}$, even though it is less accurate than the second one. Our network model determines the aggregate (uncapacitated) arrival rates with linear equations, and $\bar{\lambda}$ can be calculated without knowing the arrival rate. On the other hand, $\hat{\lambda}$ requires knowing the aggregate arrival rate, and furthermore, cannot be calculated linearly. Consequently we can neither calculate it before running the model nor in the model. The only option would be to calculate it after running the model and then rerun the model, in the hopes of getting closer to the solution.

Regardless, these approximation ratios and the numerical examples show that, at least for a network with a single station, our LCQN approximation is quite good for larger values of s and N. As we will see, the parameters we use for the DTES model do indeed include such larger values.

4.2.3 Summary of Single Station Network Comparisons

In the previous section we compared the (s) case to the (N) case. We now discuss how these two cases relate to the LCQN approach from §4.1.4, and specifically to the analagous single station network LP introduced in Theorem 4.2, hereafter denoted (LP).

The (s) case represents this small network exactly, while the (N) case approximates the capacitated station with the split station approach. Both the (s) and (N) cases represent the finite calling population exactly. The (LP) case is similar to the (N) case, in that it uses the split station approach, however, it also includes an approximation for the finite calling population in the form of the FPM approach. We see that the average number of clients in the queue is the same for these latter two cases, as long as we adjust for the fact that the (N) case uses a per capita arrival rate, $\bar{l}$, while the (LP) case uses a total arrival rate, $\bar{\lambda}$. To compare these cases we assume traffic intensities over 1.0, in other words, the single station is capacitated. Table 4.4 summarizes the characteristics of each of the three cases.

Table 4.4: Comparison of single station network models with capacitated arrivals: (s), (N), and (LP).

Case	(s)	(N)	(LP)
Description	$M/M/s/s/N$	$M/M/N/N/N$	LCQN
Finite pop.	exact	exact	FPM
Capacity	exact	split station	split station
Eff. arr. rate	$l_e^{(s)} = l\left(1 - \frac{\frac{l^s}{\mu^s}\binom{N}{s}}{\sum_{n=0}^{s}\frac{l^n}{\mu^n}\binom{N}{n}}\right)$	$l_e^{(N)} = \bar{l} = \frac{\bar{\lambda}}{(N - \bar{\lambda}/\mu)}$	$\lambda_e^{(LP)} = \bar{\lambda}$
Avg. # clients	$L^{(s)} = \sum_{n=0}^{s} \frac{n\frac{l^n}{\mu^n}\binom{N}{n}}{\sum_{m=0}^{s}\frac{l^m}{\mu^m}\binom{N}{m}}$	$L^{(N)} = \frac{N\bar{l}}{\bar{l}+\mu} = \frac{N\frac{\bar{\lambda}}{(N-\bar{\lambda}/\mu)}}{\frac{\bar{\lambda}}{(N-\bar{\lambda}/\mu)}+\mu} = \bar{\lambda}/\mu$	$L^{(LP)} = N - L_0 = \bar{\lambda}/\mu$

It is clear from this table that, while the effective arrival rates for the (N) case

and the (LP) are given in different forms—the former is a per capita rate and the latter is a total rate—the average number of clients in each case, $L^{(N)}$ and $L^{(LP)}$ respectively, is the same.

4.3 Justification for a Queueing Network Model

Many techniques exist for working with this type of model including other closed network analytical or approximate procedures, various simulation approaches, and system dynamics (SD). We believe the approach we have chosen—the FPM approximation of a closed network with some capacitated stations—is the most appropriate way to understand our model of the DTES.

4.3.1 LCQN vs. Other Queueing Approaches

The differences among the exact approaches and various approximate approaches, including ours, are very small for a model of the size considered here (where size refers to the population size and number of stations). However, the inaccuracies introduced by a number of the model parameters are orders of magnitude larger. Given this discrepancy between the data quality and the approximation quality, the choice of technique, with respect to approximation quality, is arguably unimportant. We have therefore opted for a technique with fast solution time that can scale well and can potentially be extended to more complex future versions of our model (including multiple client classes and additional services).

4.3.2 Advantages and Disadvantages of Simulation vs. Queueing Networks

There are several reasons to use a queueing model over a simulation model, but they don't negate the benefits of doing the latter alongside the former. We discuss a particular type of simulation—DES—in the next chapter as a means of validating certain assumptions about the queueing approach. As the main method of addressing our research question, however, queueing models are the most appropriate.

Our queueing model will give the same answer every time. In contrast, a DES model generates many random numbers over many replications in order to determine the "typical" behaviour of the system. The resulting solutions will differ from run

to run, which is a drawback when discussing and comparing multiple scenarios. To address this issue, DES software produces confidence intervals, which help us interpret these differences, yet it is simpler if we avoid this issue to the extent possible. Furthermore these intervals are often rather large, unless we are able to take advantage of variance reduction techniques or are willing to run the model for a very long time.

When precision in the results is important, speed is clearly a factor. The queueing model produces an answer in well under a second, yet the DES model takes many minutes, hours, or days to run (depending on how small we wish the confidence intervals to be). This speed difference becomes magnified when we wish to run multiple scenarios. As we will explain in detail in Chapter 6, this difference means we can perform all desired analyses on the queueing model in under a minute, yet it would take months to do the same set of analyses with our DES model.

We can also sometimes find interesting structural policies with a queueing model. With a DES model we cannot. However, the DES approach can be very useful for some models because it is much more flexible (e.g., for operational or tactical models). With DES we can represent queue priorities, different arrival or service time distributions, complex interactions, object attributes, and many other details that typically have to be ignored with queueing models.

A big advantage of simulation is that users can watch the animation, and often have more confidence in this approach because it is easier to grasp.

One less tangible advantage of queueing theory is as follows: In the healthcare field many non-OR practitioners know about simulation but seem to believe it is the only tool for conducting such analyses. We feel that introducing additional techniques allows us to act as educators, teaching practitioners about the many benefits of using an OR approach and demonstrating that the OR toolkit is capable of addressing myriad healthcare issues.

4.3.3 Advantages and Disadvantages of System Dynamics

The advantage of using our approach over SD is best demonstrated with a comparison between DES, which is very useful in validating our model, and SD, which could be used for similar purposes.

SD is a modelling approach based on “stocks” and “flows”, and is continuous. In contrast, DES represents discrete clients flowing through a system. SD is con-

sequently deterministic; by working with flows it is incapable of handling different distributions like DES (and to a lesser extent like queueing theory).

The benefits of SD are twofold: It can handle qualitative inputs and it shows how a system changes over time. The qualitative input idea seems tempting, at first, however, the ability to produce precise output values is severely affected by using qualitative (imprecise) inputs. An example of a qualitative input one might want to include is the positive impact on enrolment in a new program or service from word-of-mouth communication among potential clients. This feature is useful for developing models of higher-level policy implications but doesn't give us the detail we require. For the DTES model we wish to investigate policy issues, but with a sound quantitative approach.

The ability to show how a system changes over time may be useful for future work on this topic, but for now we wish to study the system in steady state, and SD does not excel at this type of analysis. One idea is to use it to show how the steady-state system currently in existence changes over time when a particular scenario is implemented, and we may pursue this idea at some point.

DES has several benefits over SD. The ability to show how a system reaches steady state is also built into DES models, as is the ability to represent the system once it is in steady state. As well, time step issues that appear in SD models (particularly when the service times are so dissimilar) are not an issue in DES. DES can also handle many different distributions for input parameters (such as service times and arrival rates). Finally, DES offers more flexibility in general, allowing us to validate any assumptions we make in the queueing model.

For an example of SD applied in healthcare see [39]. Comparisons of SD and DES can be found in [15, 79].

Chapter 5

A Model of Services in the DTES

The purposes of our modelling research on CCD client services in the DTES are: 1) To better understand the existing system; and 2) to be able to quantify the projected outcomes from potential improvements to the system. The model described in Chapter 4 helps us understand the structure of the system as well as the limitations of our approach. Now we apply this model to the DTES.

The first part of this chapter covers the inputs to the model. We then discuss sensitivity analysis on those inputs to contribute to our understanding of the system by illustrating which inputs tend to affect which outputs. Our main results are presented toward the end of this chapter when we present a comparison with a simulation model and then discuss scenario comparison in order to evaluate potential procedural or policy changes to the system.

Our model produces two primary outputs: Total system cost and health outcomes. We can examine each of these measures for each station—cost is an estimate of the average cost per client for that station, while the health outcome is an estimate of the quality adjusted life years (QALYs), on average, for clients in that station. Ultimately we are interested in the total cost and the total average QALYs, over the population being studied, for the entire system, so that we can compare these two measures across scenarios.

The model also produces other outputs, such as the average number of clients in each station or in the Non-treatment Population. In some cases we discuss some of these outputs as well as the two primary ones in order to lend additional insight.

Based on the input parameters described in the following section, the total cost of the system is about $208 billion per year, or $75.93 per person per day. This figure, as discussed in more detail below, includes the cost of running all services as well as crime costs (including intangible costs). The average QALYs per person is 18.06. For perspective, the average Canadian consumes public health resources at $15.92 per day [30] and policing resources at $1.06 per day [103].

5.1 Model Parameters

The parameters for this model come from published articles, publicly available reports, expert interviews, and secondary analysis of existing data. In some cases we were unable to find an exact figure and thus used estimates. For these cases, and in fact for all inputs, we describe extensive sensitivity analysis in §5.3.

We tried to find exact sources for all parameters (i.e., numbers specifically pertaining to the services and population being studied). When that was not possible, we used proxies from similar situations in Canada or the US, and failing that, we used the best source we could find. All cost figures have been converted to Canadian dollars (CAD), and adjusted to 2011 levels based on the consumer price index (CPI) (see Appendix C for details).

The inputs, costs, and health measures (Tables 5.1, 5.2, and 5.3, respectively) are described below.

5.1.1 External Arrival Rates

We model the external arrival rates in two different ways depending on the data available—as a total external daily arrival rate or as a per-person external daily arrival rate. A per-person daily rate can be used directly in the model, whereas the total daily rates require special consideration in the base case and are then converted to per-person rates prior to scenario analyses. Use of per-person arrival rates results in an optimal solution (as shown in Theorem 4.3), however, use of total arrival rates can result in infeasible solutions. We have checked our parameter values to ensure we avoid such solutions. These inputs are calculated thus:

Police external arrival rate is the sum of the non-criminal and the criminal arrival rates. Police involvement for non-criminal activity averages 17 contacts per day [113] and for criminal activity is 4.7 contacts per day, calculated from several statistics and assumptions: 7248 crimes are committed per year (assuming half of all District 2 crimes are attributable to the population being studied) [116]; about 23.8% of crimes result in criminal justice involvement (there were 16,897 new cases at the Vancouver Provincial Court out of 70,898 calls for police with priority 1, 2, or 3 during the Downtown Community Court's pre-evaluation period) [3]; the ratio of crimes to cases for the overall

Table 5.1: Inputs for DTES model. A subjective rating of each parameter value's quality is given, where 1 star is low quality and 3 stars is high quality.

Parameter	Units	Value	Quality	Source(s)
External Arrival Rate				
Police	/day	21.7	★★★	[113]; [116]; [3]
ED	/person/day	0.0091	★★★	[106]
Other	/day	10.7	★★☆	[19]
Routing Probability				
Police→CrimJst		0.22	★★★	[113]; [116]; [3]
Police→ED		0.05	★☆☆	
ED→Acute		0.14	★★★	[105]
ED→Inpatnt		0.05	★☆☆	
Other→ED		0.11	★★☆	[19]
Other→Inpatnt		0.11	★★☆	[19]
Other→MMT		0.04	★★☆	[19]
Other→CM		0.10	★★☆	[19]
Other→ACT		0.01	★☆☆	
Other→Family		0.08	★★☆	[19]
Acute→Inpatnt		0.05	★☆☆	
Acute→MMT		0.05	★☆☆	
Acute→CM		0.05	★☆☆	
Inpatnt→MMT		0.05	★☆☆	
Inpatnt→CM		0.05	★☆☆	
Inpatnt→ACT		0.01	★☆☆	
Length of Stay				
Police	hours	2.6	★★★	[121]
CrimJst	days	63	★★★	[3]
ED	hours	4	★★☆	
Other	-	-		
Acute	days	12	★★★	[115]
Inpatnt	days	89	★★★	[112]; [100]; [18]; [114]
MMT	days	338	★★★	[81]
CM	days	1275	★★☆	[23]
ACT	days	3464	★★☆	[80]
Family	days	1275	★★☆	[23]
Population				
CCD population	people	7500	★★☆	[72]
Station Capacity				
Inpatnt	beds	162	★★★	[112]; [18]; [114]
CM	people	1400	★★★	[60]
ACT	people	90	★★★	[60]

Vancouver population is assumed to hold for DTES crimes. Another way to arrive at the number of crimes per year suggests 8070, which means our estimate is probably somewhat conservative. This alternate figure is calculated based on the following statistics: The average opioid user not in treatment commits 1.65 crimes per year [117]; there are approximately 4000 non-treatment clients in the model committing crimes at about this rate; there are approximately 2700 outpatient (treatment) clients committing crimes at 33% of this rate [46]. Note also that the 1.65 and 33% figures almost exactly match similar statistics from another source that measured police contacts before and after introducing supportive housing to a homeless population in Portland, Maine [78].

ED aggregate per capita arrival rate is based on the average ED usage per person for the cohort population from the At Home Study [106], which is a very good proxy for the population in this model. Using the pre-trial results, the total person-years for the cohort is 382 and the total ED visits is 962. We adjust the calculated rate by a factor of 1.333 in order to include Vancouver General Hospital visits on top of the St. Paul's Hospital visits accounted for in the study to get an average rate of 0.0092 visits per person per day. This figure includes arrivals from the other two entry points; the model determines the external arrival rate by subtracting any other ED arrivals.

Other Entry external arrival rate is a catch-all station for the various additional ways clients can enter the system, such as via Vancouver's safer injection site, Insite. We use Insite referrals as a starting point [19, Table D7], and then make the assumption that this number can be doubled to get total other entries. The average daily arrival rate comes from Insite referrals to all programs other than housing for 2005, times two, which equals 10.7 arrivals per day.

These arrivals all assume that the arrival processes are stationary, or unchanging over time. In practice, arrivals are affected by factors such as time of year and time within the two-week welfare cheque period. These cyclic behaviours will likely be investigated in a later version of this model, but we decided that in keeping with our goal of creating a fairly simple model with limited scope, this issue could be postponed.

5.1.2 Routing Probabilities

Routing probabilities describe the likelihood that an individual at a particular service moves to each other service upon leaving. For example, clients in the ED transition to Acute Care 14% of the time, to Inpatient 5% of the time, and back into the Non-treatment Population the rest of the time. Of all the inputs to the model, this data was the most difficult to find. Therefore, many of these routing probabilities are simply estimates that we perform sensitivity analysis on to study how the outcome behaves over a range of values. The figures that come from reliable sources are:

Police→Criminal Justice is based on the two rates discussed above: 21.7 police arrivals per day, out of which 4.7 lead to criminal justice. 4.7/21.7 = 0.217

ED→Acute Care comes from the preliminary At Home Study results [105].

Other Entry→ED comes from Insite referrals to "Hospital Emergency" [19, (Table D7)].

Other Entry→Inpatient uses Insite referrals to "Detoxification bed" as a proxy [19, (Table D7)].

Other Entry→MMT comes from Insite referrals to "Methadone" [19, (Table D7)].

Other Entry→CM uses Insite referrals to "Community services" as a proxy [19, (Table D7)].

Other Entry→Family Practice uses Insite referrals to "Community clinics", divided by two as an estimate of the clinics of the family practice variety [19, (Table D7)].

The remaining routing probabilities are estimates for which sources were unavailable.

5.1.3 Lengths of Stay

LoSs are given for all services. For the services with the shortest durations (Police, ED), the precise input value is unimportant. We nevertheless use the available data,

but concentrate our efforts on the longer LoSs. One service, Other Entry, has an LoS equal to zero days[6] because it is simply a pass-through entry point that represents clients arriving at the various treatment options through other means. Average LoSs range from several hours to a number of years. The numbers come from the following sources:

Police LoS is simply the average contact time, which is assumed to be 2.6 hours [121], the city-wide average for all calls. If an estimate were available for calls involving this population, it would likely be higher.

Criminal Justice LoS is a weighted average using some statistics from the Downtown Community Court [3]. For clients involved in the Downtown Community Court: 27% were in pre-trial jail for an average of 16 days and 73% were in pre-trial supervision for an average of 33 days (assuming all clients not in jail were in supervision); 45% were in after-trial jail for an average of 22 days and 55% were in supervision for an average of 45 days (again assuming those not in jail were under supervision). The average LoS works out to 63 days.

ED LoS is estimated at 4 hours, though this number is entirely insignificant to the model outputs.

Acute Care LoS is 12 days [115] (average LoS in mental health/psych at St. Paul's Hospital).

Inpatient LoS is a weighted average of the LoS at the three inpatient services considered in this model: Burnaby Centre for Mental Health and Addictions (BCMHA) (100 beds [112], 135 day LoS [100]); Vancouver and Cordova Detox Centres (53 beds, 6 day LoS [18]); Community Transition Care Team (9 beds, 59 day LoS which is based on 56 clients served in one year and assuming 100% utilization [114]). The average LoS is 89 days.

MMT LoS is based on the 12-month retention rate [81], calculated assuming exponentially distributed LoS.

CM LoS comes from a study of case management and outreach [23].

[6] We use an $M/0/\infty$ queue for this service, as described in §4.1.4.

ACT LoS is based on the B.C. standard for flow-through rate of 10% per annum turnover [80], calculated assuming exponentially distributed LoS.

Family Practice LoS is assumed to be the same as CM.

5.1.4 Population Size

One of the reasons for creating this model and the accompanying research on CCD clients in the DTES is to better understand this population about which many basic facts are not known or understood. For instance, no one seems to know the actual size of the population. The entire population of the DTES is roughly between 16,000 and 18,000, depending on the source [70], [21]. The population of *harmed individuals*[7] is estimated to be about 15,000, with 50 to 70% having concurrent disorders [72]. A fairly conservative estimate of the total CCD population is therefore 7,500 (50% of 15,000).

Another way to arrive at an estimate (involving very rough estimates) is to start with the estimated 5000 injection drug users [54], and then double this number to include those struggling with other substances (including alcohol) and those who are not actively using but nevertheless have a substance addiction. We then need to determine what portion of people with an addiction issue have mental health issues, which is difficult because so many mental health issues are undiagnosed or under-reported. We start with the estimate of 35% of a sample of homeless people who reported having received a mental health diagnosis at some point [55], and then double this number to account for two factors: 1) Undiagnosed or unreported illnesses, and 2) higher concurrence specifically among people with addictions than among homeless individuals. Recognizing that this estimate is far from precise, we nevertheless arrive at the similar figure of 7000 clients.

This input, the size of the DTES population with CCD, can substantially alter model outcomes. In queueing terminology, this population is referred to as the finite calling population. We perform considerable sensitivity analysis on this parameter in order to observe the outcomes over a range of inputs and to investigate whether or not the overall policy implications are robust with respect to changes in inputs.

[7] *Harmed individuals* refers to people with some form of addiction issue and any of the following: Mental illness, history of trauma/abuse, suicidality, significant physical illness.

5.1.5 Service Capacities

Three of the services in our model have capacities that are restrictive enough to clearly cause clients to balk: Inpatient, CM, and ACT. These capacities are derived as follows:

Inpatient capacity is the total number of beds across all facilities grouped into this service – BCMHA, Detox (Vancouver and Cordova), and Community Transition Care Team (see Inpatient LoS, above, for sources).

CM capacity is based on the typical number of active clients at the Strathcona Mental Health Team [60], because they are the main provider of this type of treatment and appear to be operating close to maximum capacity.

ACT capacity is based on the ten-member team's mandate to provide service at a 9:1 ratio [60].

5.1.6 Station Costs

The two primary outputs of the model are costs and health outcomes. The former is intended to include all costs – direct costs for providing the service and indirect costs to society, businesses, or individuals from activities such as crime. Costs for stations come in two forms: The cost for Police and ED are given per client served, whereas all others are given per client per day. See Table 5.2. (From the client's perspective, contact with Police or Criminal Justice might not seem like they are being "served", however, we use the word in a general sense to describe the provision of any resource in our model.)

Crime costs must also be considered. We use the estimate of $77 for the average daily cost of crime for members of the non-treatment population, and then apply a multiplier of 0.33 [46] to calculate the average daily cost of crime for clients in long term treatment programs. We assume clients in other stations do not commit crimes while in those stations. The average daily cost of crime, and the station costs, adjusted to 2011 CAD, are calculated thus:

Police cost per contact is based on the Vancouver Police Department estimate of 2.6 hours per call, with two officers responding at an annual cost of $100,000 each (based on 90 full-time officers equalling a $9 million cost) [121]. It also

Table 5.2: Cost inputs in 2011 CAD, including a subjective rating of each parameter value's quality. The cost of each service is listed without and also with crime costs factored in.

Station	Units	Service Cost	Crime Cost	Combined Cost	Quality	Source(s)
Police	$/contact	642		642	★★★	[121]; [41]
CrimJst	$/day	54	20	74	★★☆	[3]; [76]
ED	$/contact	387		387	★★★	[105]; [59]; [85]
Other		-		-		
Acute	$/day	647		647	★★★	[85]
Inpatnt	$/day	289		289	★★☆	[112]; [18]; [114]; [100]
MMT	$/day	17	25	42	★★★	[125]
CM	$/day	22	25	47	★★★	[60]
ACT	$/day	46	25	71	★★★	[60]
Family	$/day	2	25	27	★★☆	[87]
Non-tr	$/day	0	77	77	★★★	[116]; [89]

assumes hourly overhead costs ($40 in 1993) and 1560 active patrol duty hours per year (from a study of Canadian police forces published in 1994) [41].

Criminal Justice cost per day is based on a weighted average of the costs of the components. Using the same averages and LoS described in the Criminal Justice LoS calculation above, as well as an adjusted per diem jail cost of $199.75 and per diem supervision cost of $11.55 [76], the per day average criminal justice cost is $53.91.

ED cost per visit comes from the cost of a hospital emergency ($195.08 in 2002) [59] plus 35.9% [105] times the cost of an ambulance ride ($396 in 2007) [85].

Acute Care cost per day is from [85].

Inpatient cost per day uses the same weighted average as the Inpatient LoS calculation, with the following costs: BCMHA and Community Transition Care Team per diem cost is $350 (2011 CAD) (unpublished per diem rate for tertiary care from Vancouver Coastal Health (VCH)); detox per diem cost is $125.42 (2002 CAD)[59].

MMT daily cost is from [125] (this cost matches, or is somewhat higher than, anecdotal figures I was told by local sources).

CM cost per day is $22, calculated based on yearly operating costs [60].

ACT cost per day is $46, calculated based on yearly operating costs [60].

Family Practice cost per day is estimated at $2 based on two monthly visits and the fee-for-service figure of $30.06 for a "visit in office (age 2 – 49)" [87].

Non-treatment Population daily cost is assumed to include only the crime cost of $77. By not including other costs to society (lost business, lost productivity, etc.) this cost can be considered conservative.

Cost Non-tr crime is calculated based on 2011 Vancouver Police Department crime statistics [116], assuming half of all District 2 crimes are committed by this population, and on crime cost estimates not including criminal justice system costs but including estimates for intangible costs using a conservative valuation for the intangible cost of a homicide [89].

5.1.7 Health Outcomes

Health outcomes are most useful in showing the extent of the effect on client health of changes to resources, procedures, or policies. Health outcomes are included in this model by way of quality of life (QoL) estimates for the various stations, as well as an overall estimate of average life years remaining for this population. Research (cited below) shows that different treatment programs, or a lack of treatment, are associated with different QoL. By multiplying the QoL for each service by the number of people in that service and by the average remaining life years, LY, we can estimate the expected QALY for the entire population for any scenario (for brevity, we define the set of all stations containing clients and therefore necessary in the QALY calculation as $\mathscr{J}_{\mathrm{Q}} = \overset{\infty}{\mathscr{J}} \cup \mathscr{J}_{\mathrm{c}} \cup \{0\}$, where $\{0\}$ is the Non-treatment Population):

$$E[\text{QALYs per person}] = \frac{1}{N} \sum_{j \in \mathscr{J}_{\mathrm{Q}}} \left[\mathrm{QoL}_j \cdot \mathrm{LY}_j \cdot L_j \right] \tag{5.1}$$

Table 5.3 shows the estimates, which are derived from these sources and assumptions:

Non-treatment Population QoL is the weighted average of the "before" measure for CM and ACT from [31].

CM QoL comes from [31].

ACT QoL also comes from [31].

Police, ED QoL is assumed to be 10% worse than QoL for the Non-treatment Population. However, because the LoS is very short, this figure is inconsequential to model outcomes.

Criminal Justice QoL is assumed to be the same as QoL for Non-tr.

Acute Care, Inpatient QoL is based on the CM QoL, but adjusted up by 8% to incorporate the improvement associated with inpatient vs. outpatient treatment [34].

MMT QoL is an 18.6% improvement [83] over the non-treatment QoL.

Family Practice QoL is assumed to be the same as CM.

Table 5.3: Quality of Life inputs, inclding a subjective rating of parameter quality.

Station	QoL	Quality	Source(s)
Police	0.526	★☆☆	
CrimJst	0.584	★★☆	
ED	0.526	★☆☆	
Other	-		
Acute	0.702	★★★	[31]; [34]
Inpatnt	0.702	★★★	[31]; [34]
MMT	0.693	★★★	[31]; [83]
CM	0.650	★★★	[31]
ACT	0.660	★★★	[31]
Family	0.650	★★☆	
Non-tr	0.584	★★★	[31]

The average remaining life years, LY, is estimated to be 29.4 for all services. This calculation is based on: The proportion of males to females in the DTES (60:40) and the Vancouver health area life expectancies for males and females (79.1, 84.2) [36]; the lost life years due to severe mental illness for males and females (14.1, 5.7) [37]; and the mean age of this population, estimated at 41 from the VIDUS cohort [2].

A Note on QALYs The use of QALYs is well established in the medical decision making and cost-effectiveness literatures [97], and there are many arguments for and against this approach (see Knapp and Mangalore [56] for arguments against using such measures or Chisholm *et al.* [28] for a discussion of QALYs in mental health).

Given that we felt it valuable to report health outcomes in some form, we chose practicality over perfection and went with the approach with the most readily available data. Many studies attempt to quantify QoL for clients in situations similar to those in our model. We considered using the Disability Adjusted Life Year (DALY) measure, but had more difficulty acquiring data. We therefore report QALY outcomes, with the caveat that this measure is not perfect but does allow us to make population-level comparisons. We also stress that we are only applying the QALY methodology at the population level and not at the individual level.

5.2 Model Output

We mainly discuss the two primary outputs—cost and QALYs. Other outputs, such as the average number of clients in a station or in the Non-treatment Population or the balking rate at a capacitated station, are discussed to a lesser extent.

The next section delves into sensitivity analysis to address the less precise inputs, however, a few simple results stand regardless of the exact input parameter values: 1) The Non-treatment Population accounts for roughly the same proportion of overall costs as it does of total population (56% and 55%), yet a large cost of this group is in the increased chance of contact with the most expensive stations—Police, ED, Acute Care, and Inpatient; 2) conversely, the least expensive stations are MMT, CM, Family Practice, and Criminal Justice (with the last due to the lower cost of supervision), so changes to the system that result in more clients entering and/or remaining in these services will result in lower costs[8]; 3) lastly, the QoL rates tend to be higher for the lower cost stations, with the exception of Acute Care and Inpatient, therefore these two stations may help explain policies that differ the most between cost and health outputs.

Though the model produces precise results, it is important to remember that

[8]We are not saying more clients *should* be in Criminal Justice; this is simply a statement about cause and effect.

the accuracy of those results is limited by the accuracy and quality of the inputs (and to a much lesser extent, by the accuracy of the queueing approximations). Because in some cases the input quality is limited due to data availability, the exact cost or QALY output values are perhaps less important than the comparisons across different scenarios. With these limitations in mind, the model clearly has two important uses:

1. Calculator for planners. Procedural or policy changes can be evaluated from a cost and health perspective; sensitivity analysis must be included in these evaluations to understand the limitations corresponding to specific inputs (§5.3). DES could also be used to develop confidence intervals on model outcomes (though we do not demonstrate this use of DES herein).

2. Scenario comparison. Many different scenarios can be compared in order to choose several good alternatives for additional study. Again sensitivity analysis helps provide more robust results (§5.5.2).

5.2.1 Model Validation

The scope of this model is limited to the main health and criminal justice components that the CCD population in the DTES encounters. As well, several other limitations—clients are only in one station at a time; clients in long term treatment cannot depart briefly for Police, ED, or other involvement and then return; and the population is homogeneous—potentially skew the results. Yet we still must validate that the outputs are correct insofar as they apply to this limited portion of the real world. We do this by examining several other results from the model that are neither inputs (which we seek to indirectly validate) nor primary outputs (which measure unknown quantities), asking whether our results are reasonable given the model limitations. The results we discuss are: Size of the Non-treatment Population, average usage levels of the uncapacitated stations, and amount of balking at the capacitated stations.

The mean size, m, of the Non-treatment Population is 4201. This figure is very difficult to validate because it is a count of the people generally not counted by scientific studies. When compared with anecdotal reports we find it is in the right range. We can also compare it to numbers from the literature that, while

not measuring the same population, measure similar groups or overlapping groups. For instance, from March 2004 to April 2005, 4764 clients registered with Insite [1]. Clearly this population is not the same as the non-treatment population, but there is probably considerable overlap.

The average usage levels of the stations that are interesting to compare are Criminal Justice (297), Acute Care (64), MMT (265), and Family Practice (1091). We do not have data on the total number of clients in the criminal justice system or in acute care at either hospital. For MMT, we know that the DTES rate is 35 people per 1000 general population [81]. This rate suggests 560 – 630 clients should be on MMT. However, one limitation of our model is that only the main treatment modality is captured. We know that clients in CM, ACT, Family Practice, and even Criminal Justice are also on MMT, yet have not accounted for them in this version. Given this limitation, the result from our model of 265 clients with MMT as their main treatment course seems reasonable. The Family Practice number of 1091 clients is also difficult to validate; we know that primary care access is fairly high—among injection drug users in one study, 78% accessed some form of primary care within the past year [53]. However we also know that most of the primary care available does not involve clients seeing their own family doctor (which we are trying to model with this station). If we apply 78% to the total population of 7500, we see that about 5850 clients access primary care. Given that most of these clients saw a nurse rather than their own family doctor at a drop-in clinic, the result from our model is not unreasonable.

The capacitated stations are all at capacity, which in itself provides some validation. We can also look at the amount of balking: For Inpatient, one patient balks for every patient served; for CM, one patient balks for every three patients served; and for ACT, three patients balk for every patient served.

These balking rates are also difficult to validate, but we can present some comparisons. One of the components of the Inpatient treatment is detoxification. Over a one-year period, 35% of clients referred to Vancouver Detox dropped out before treatment began [66]. This statistic suggests our balking rate is a bit too high, however, it is quite conceivable that the balking rates at the other inpatient treatment components are higher because of the longer treatment times, resulting in an overall inpatient treatment balking level on par with what our model suggests.

Our model suggests a believable rate for CM—one in four patients referred to

CM (including self-referrals) is lost. We do not yet have data to validate this number, however, we plan to obtain it for the next version of the model.

For ACT we also do not have data to validate the balking rate of over three patients per patient served. Yet in this case such a high rate is sensible given the limited capacity and high LoS associated with this service. Furthermore, an incorrect rate in this case would mean the routing probabilities going into the ACT station are incorrect, yet if that is true, then we're simply over- or under-estimating the number of clients who balk and then return to the Non-treatment Population compared to the number of clients who do not get referred and therefore return to the Non-treatment Population. Suffice it to say that as long as all ACT capacity is being used in the real world, our model seems to represent this station adequately.

Additional model validation is discussed in §5.3 when we perform sensitivity analysis on the model, which serves the purpose of testing it under extreme input values, and in §5.4 when we compare the results to a simulation model, which serves to show that the results are similar given a different solution methodology.

Further validation will not be performed for this version of the model. In future versions that include housing and that address some of the other limitations, we will conduct additional validation.

Having created a model of the services in the DTES, we use sensitivity analysis to examine the inputs in more detail in order to identify limitations in the conclusions we can draw from these inputs. We then discuss the DES model and its utility in validating certain assumptions about our queueing model. Finally, we compare a number of scenarios in the queueing model to see what types of changes result in the most favourable cost and health outcomes.

5.3 Sensitivity Analysis

Sensitivity analysis involves varying one or more inputs and observing the resulting variations in the outputs or in other measures of interest. It is usually performed for three reasons: 1) When inputs are not known exactly, sensitivity analysis tells us if it is important to try to find a precise value or if a lack of precision will not have a meaningful effect on the outputs; 2) when inputs are known but might change at some point in the future, sensitivity analysis allows us to describe how robust

the solution is to future parameter fluctuations; and 3) when trying to understand similar problems, sensitivity analysis provides insight into the relationship between inputs and outputs.

When working with LPs there is a special type of sensitivity analysis that relies on the solution structure to determine a range within which each objective coefficient or right-hand-side value can change while still maintaining essentially the same LP solution. However, even though we use an LP to solve our problem, the input parameters we wish to investigate are not the ones included in this type of sensitivity analysis. We therefore use the more basic approach of solving the model repeatedly—once for each input parameter value we wish to investigate.

5.3.1 Sensitivity to a Single Parameter

To test sensitivity to single parameter changes, we use an Excel macro that loops through all input parameters, solving the model for each step within a specified range for each parameter (while holding all other inputs at their base value), and recording the outputs corresponding to each solution. Table 5.4 shows the specifications; for instance, the first row corresponds to trying the values 4000, 5000, 6000, 7000, 8000, and 9000 for the total population size.

The range over which each parameter is varied was chosen to represent plausible values. We used literature sources, interviews, and our judgement to determine the more extreme values that could be possible for each parameter.

The detailed results of the sensitivity analysis include the cost and QALY estimates—as well as the mean queue lengths and aggregate arrival rates for all stations, in addition to the Non-treatment Population size—for each parameter value. This report is far too large to include in its entirety, but we have summarized the results in Table 5.5 by showing the range over which each parameter was varied as well as the ranges of the two primary outputs—cost and QALYs. We shade those outputs displaying a higher sensitivity, however, this shading must be interpreted prudently because it depends on the input parameter ranges.

QALYs are much less sensitive to input changes compared to costs. This insensitivity is (at least partially) due to our conservative approach toward incorporating health outcomes into the model. Specifically, we assume everyone in the population has the same life expectancy, so the only change among stations is in QoL. This issue will be revisited in the next version when we extend the model to include

Table 5.4: Sensitivity analysis—single parameter input specifications.

Parameter	Min	Max	Step	Base Value
Population	4000	9000	1000	7500
Arrivals Police	20	30	1	21.7
Arr. /capita ED	0.005	0.015	0.001	0.0091
Arrivals Other	5	15	2	10.7
Cap Inpatnt	90	190	50	162
Cap CM	1100	2000	300	1400
Cap ACT	60	100	10	90
LoS Police	0.1	0.3	0.1	0.108
LoS CrimJst	50	80	10	63
LoS ED	0.1	0.5	0.1	0.167
LoS Acute	8	16	2	12
LoS Inpatnt	60	120	10	89
LoS MMT	180	720	60	338
LoS CM	365	2190	365	1275
LoS ACT	2000	5000	1000	3464
LoS Family	365	2190	365	1275
Cost coef Police	400	800	50	642
Cost CrimJst	50	125	25	54
Cost coef ED	300	600	100	387
Cost Acute	500	1000	100	647
Cost Inpatnt	200	500	100	289
Cost MMT	8	18	2	17
Cost CM	18	28	2	22
Cost ACT	30	60	5	46
Cost Family	0.5	10	0.5	2
Cost Non-tr	0	50	10	0
Cost Non-tr crime	50	90	10	77
Police→CrimJst	0.2	0.4	0.05	0.217
Police→ED	0.01	0.1	0.01	0.05
CrimJst→Inpatnt	0	0.1	0.05	0
ED→Acute	0.1	0.2	0.01	0.141
ED→Inpatnt	0.01	0.1	0.01	0.05
Other→ED	0.05	0.15	0.02	0.11
Other→Inpatnt	0.05	0.15	0.02	0.11
Other→MMT	0.01	0.1	0.01	0.04
Other→CM	0.05	0.2	0.05	0.1
Other→ACT	0.005	0.02	0.005	0.01
Other→Family	0.05	0.13	0.02	0.08
Acute→Inpatnt	0.01	0.1	0.01	0.05
Acute→MMT	0.01	0.1	0.01	0.05
Acute→CM	0.01	0.1	0.01	0.05
Inpatnt→MMT	0.01	0.1	0.01	0.05
Inpatnt→CM	0.01	0.1	0.01	0.05
Inpatnt→ACT	0.005	0.02	0.005	0.01
Crime treat ratio	0.2	1	0.2	0.33

Table 5.5: Sensitivity analysis—single parameter results summarized. *Cost ranges > $5 and QALY ranges > 0.5 are shaded.*

Parameter	Range	Costs	QALYs
Population	4000 – 9000	$64.39 – $77.94	18.7 – 17.9
Arrivals Police	20 – 30	$75.83 – $76.42	18.1 – 18.1
Arr. /capita ED	0.005 – 0.015	$73.16 – $79.77	18.0 – 18.1
Arrivals Other	5 – 15	$82.39 – $72.03	17.8 – 18.2
Cap Inpatnt	90 – 190	$74.09 – $76.65	18.0 – 18.1
Cap CM	1100 – 2000	$77.56 – $73.83	18.0 – 18.1
Cap ACT	60 – 100	$76.00 – $75.91	18.0 – 18.1
LoS Police	0.1 – 0.3	$75.93 – $75.88	18.1 – 18.1
LoS CrimJst	50 – 80	$76.05 – $75.78	18.1 – 18.1
LoS ED	0.1 – 0.5	$75.96 – $75.79	18.1 – 18.1
LoS Acute	8 – 16	$74.35 – $77.50	18.0 – 18.1
LoS Inpatnt	60 – 120	$75.84 – $75.98	18.1 – 18.1
LoS MMT	180 – 720	$76.69 – $74.16	18.0 – 18.2
LoS CM	365 – 2190	$80.58 – $75.93	17.8 – 18.1
LoS ACT	2000 – 5000	$75.93 – $75.93	18.1 – 18.1
LoS Family	365 – 2190	$82.23 – $69.60	17.9 – 18.2
Cost coef Police	$400 – $800	$75.23 – $76.39	18.1 – 18.1
Cost CrimJst	$50 – $125	$75.77 – $78.74	18.1 – 18.1
Cost coef ED	$300 – $600	$75.50 – $77.00	18.1 – 18.1
Cost Acute	$500 – $1000	$74.69 – $78.92	18.1 – 18.1
Cost Inpatnt	$200 – $500	$74.02 – $80.46	18.1 – 18.1
Cost MMT	$8 – $18	$75.61 – $75.97	18.1 – 18.1
Cost CM	$18 – $28	$75.19 – $77.05	18.1 – 18.1
Cost ACT	$30 – $60	$75.74 – $76.10	18.1 – 18.1
Cost Family	$0.5 – $10	$75.71 – $77.10	18.1 – 18.1
Cost Non-tr	$0 – $50	$75.93 – $103.44	18.1 – 18.1
Cost Non-tr crime	$50 – $90	$57.43 – $84.84	18.1 – 18.1
Police→CrimJst	0.2 – 0.4	$75.98 – $75.45	18.1 – 18.1
Police→ED	0.01 – 0.1	$75.93 – $75.93	18.1 – 18.1
CrimJst→Inpatnt	0 – 0.1	$75.93 – $75.93	18.1 – 18.1
ED→Acute	0.1 – 0.2	$74.70 – $77.66	18.0 – 18.1
ED→Inpatnt	0.01 – 0.1	$75.93 – $75.93	18.1 – 18.1
Other→ED	0.05 – 0.15	$75.93 – $75.93	18.1 – 18.1
Other→Inpatnt	0.05 – 0.15	$75.93 – $75.93	18.1 – 18.1
Other→MMT	0.01 – 0.1	$76.59 – $74.61	18.0 – 18.1
Other→CM	0.05 – 0.2	$77.26 – $75.93	18.0 – 18.1
Other→ACT	0.005 – 0.02	$75.93 – $75.93	18.1 – 18.1
Other→Family	0.05 – 0.13	$79.24 – $70.42	18.0 – 18.2
Acute→Inpatnt	0.01 – 0.1	$75.93 – $75.93	18.1 – 18.1
Acute→MMT	0.01 – 0.1	$76.37 – $75.40	18.0 – 18.1
Acute→CM	0.01 – 0.1	$75.93 – $75.93	18.1 – 18.1
Inpatnt→MMT	0.01 – 0.1	$76.08 – $75.75	18.0 – 18.1
Inpatnt→CM	0.01 – 0.1	$75.93 – $75.93	18.1 – 18.1
Inpatnt→ACT	0.005 – 0.02	$75.93 – $75.93	18.1 – 18.1
Crime treat ratio	0.2 – 1	$71.83 – $97.08	18.1 – 18.1

heterogeneous sub-populations.

We also show the detailed results for two of the several dozen parameters: Population size in Table 5.6 and the routing probability for Acute→MMT in Table 5.7. The sensitivity analysis for population size shows how drastically the model outputs can change when a key input such as this one is varied over a large range. Additionally, it shows how the queue lengths are affected: Capacitated services essentially remain at capacity with a smaller (and certainly with a larger) population; ED queues increase with Non-tr because this arrival rate is expressed as a per capita rate; and the services fed by ED also increase with Non-tr. If the population were made much smaller we would get an infeasible solution with Non-tr becoming negative. This infeasible solution, as explained in §5.1.1, is a result of fixing one or more ζ values rather than letting the LP choose these values based on the z per capita arrival rates.

As an example of an insensitive parameter, the sensitivity analysis for the routing probability (Acute→MMT) shows how little the model is affected by some inputs; the main outputs are virtually unaffected by changes to this routing probability, while only the queue length for MMT is substantially affected. In fact, most of the input parameters have similarly small effects on the outcomes over the ranges we tested. From this analysis we can see that the exact parameter estimates for many of the inputs are less important than for key inputs such as population size. Nevertheless, as the MMT queue length in the last example shows, attention should be paid to the inputs that do substantially affect any outputs we wish to examine in detail.

5.3.2 Sensitivity to Groups of Parameters

The above analysis shows that many of the parameters have little effect on the outputs, as long as they are varied in isolation. However, if we want to ask questions about the effects of multiple parameter changes on the model outputs we must perform more involved sensitivity analysis. For instance, cost and QALYs are insensitive to changes in all of the routing probabilities—except Other→Family and perhaps ED→Acute—when investigated one at a time. But the result likely differs if two or more of these inputs are changed simultaneously.

Ideally we will see nothing of interest in this analysis. We would like to know

Table 5.6: Sensitivity analysis—single parameter example for population size.

Value	4000	5000	6000	7000	8000	9000
Cost	$64.39	$69.89	$72.91	$75.07	$76.69	$77.94
QALYs	18.71	18.44	18.25	18.11	18.01	17.93
Queue Length						
Police	2.4	2.4	2.4	2.4	2.4	2.4
CrimJst	296.7	296.7	296.7	296.7	296.7	296.7
ED	1.2	2.6	4.1	5.5	7.0	8.4
Acute	12.0	26.5	41.3	56.1	70.9	85.8
Inpatnt	140.8	161.0	161.0	161.0	161.0	161.0
MMT	188.3	212.5	233.4	254.3	275.1	296.0
CM	1399.0	1399.0	1399.0	1399.0	1399.0	1399.0
ACT	89.0	89.0	89.0	89.0	89.0	89.0
Family	1091.4	1091.4	1091.4	1091.4	1091.4	1091.4
Non-tr	779.4	1719.0	2681.8	3644.7	4607.5	5570.3
Aggregate Arr.						
Police	21.70	21.70	21.70	21.70	21.70	21.70
CrimJst	4.71	4.71	4.71	4.71	4.71	4.71
ED	7.09	15.64	24.40	33.17	41.93	50.69
Other	10.70	10.70	10.70	10.70	10.70	10.70
Acute	1.00	2.21	3.44	4.68	5.91	7.15
Inpatnt arr.	1.58	2.07	2.57	3.07	3.57	4.07
Inpatnt balk	0.00	0.26	0.76	1.26	1.76	2.26
Inpatnt cont.	1.58	1.81	1.81	1.81	1.81	1.81
MMT	0.56	0.63	0.69	0.75	0.81	0.88
CM arr.	1.20	1.27	1.33	1.39	1.46	1.52
CM balk	0.10	0.17	0.24	0.30	0.36	0.42
CM cont.	1.10	1.10	1.10	1.10	1.10	1.10
ACT arr.	0.12	0.13	0.13	0.13	0.13	0.13
ACT balk	0.10	0.10	0.10	0.10	0.10	0.10
ACT cont.	0.03	0.03	0.03	0.03	0.03	0.03
Family	0.86	0.86	0.86	0.86	0.86	0.86
External Arr.						
Police ext.	21.70	21.70	21.70	21.70	21.70	21.70
ED ext.	4.83	13.38	22.14	30.90	39.67	48.43
Other ext.	10.70	10.70	10.70	10.70	10.70	10.70

Table 5.7: Sensitivity analysis—single parameter example for Acute→MMT routing probability.

Value	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
Cost	$76.37	$76.26	$76.15	$76.04	$75.93	$75.82	$75.72	$75.61	$75.50	$75.40
QALYs	18.03	18.03	18.04	18.05	18.06	18.06	18.07	18.08	18.08	18.09
Queue Length										
Police	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4
CrimJst	296.7	296.7	296.7	296.7	296.7	296.7	296.7	296.7	296.7	296.7
ED	6.4	6.3	6.3	6.3	6.3	6.2	6.2	6.2	6.2	6.1
Acute	64.6	64.3	64.1	63.8	63.5	63.3	63.0	62.7	62.5	62.2
Inpatnt	161.0	161.0	161.0	161.0	161.0	161.0	161.0	161.0	161.0	161.0
MMT	193.4	211.5	229.4	247.1	264.7	282.2	299.5	316.6	333.6	350.5
CM	1399.0	1399.0	1399.0	1399.0	1399.0	1399.0	1399.0	1399.0	1399.0	1399.0
ACT	89.0	89.0	89.0	89.0	89.0	89.0	89.0	89.0	89.0	89.0
Family	1091.4	1091.4	1091.4	1091.4	1091.4	1091.4	1091.4	1091.4	1091.4	1091.4
Non-tr	4196.2	4178.4	4160.8	4143.4	4126.1	4108.9	4091.9	4075.0	4058.3	4041.7
Aggregate Arr.										
Police	21.70	21.70	21.70	21.70	21.70	21.70	21.70	21.70	21.70	21.70
CrimJst	4.71	4.71	4.71	4.71	4.71	4.71	4.71	4.71	4.71	4.71
ED	38.19	38.02	37.86	37.70	37.55	37.39	37.24	37.08	36.93	36.78
Other	10.70	10.70	10.70	10.70	10.70	10.70	10.70	10.70	10.70	10.70
Acute	5.38	5.36	5.34	5.32	5.29	5.27	5.25	5.23	5.21	5.19
Inpatnt arr.	3.36	3.35	3.34	3.33	3.32	3.31	3.30	3.29	3.28	3.28
Inpatnt balk	1.55	1.54	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.47
Inpatnt cont.	1.81	1.81	1.81	1.81	1.81	1.81	1.81	1.81	1.81	1.81
MMT	0.57	0.63	0.68	0.73	0.78	0.83	0.89	0.94	0.99	1.04
CM arr.	1.43	1.43	1.43	1.43	1.43	1.42	1.42	1.42	1.42	1.42
CM balk	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.32	0.32	0.32
CM cont.	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10
ACT arr.	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13
ACT balk	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
ACT cont.	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Family	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
External Arr.										
Police ext.	21.70	21.70	21.70	21.70	21.70	21.70	21.70	21.70	21.70	21.70
ED ext.	35.92	35.76	35.60	35.44	35.29	35.13	34.97	34.82	34.67	34.52
Other ext.	10.70	10.70	10.70	10.70	10.70	10.70	10.70	10.70	10.70	10.70

that there are no surprises in which individually sensitive inputs become much more sensitive, or in which insensitive inputs become sensitive, when taken jointly. We proceed with this exercise to provide additional confidence in our later results.

The idea behind sensitivity analysis on groups of input parameters is simple: We change several parameters, solve the model, observe the outputs, and repeat. In practice, however, it is too easy to become swamped in data. If we were to try ten different values for each of the seventeen routing probabilities, and were interested in all permutations, we would have to somehow make sense of 10^{17} sets of results. Even if we only looked at two different values for each of these inputs, we would still face $2^{17} \approx 130,000$ sets of outputs. We clearly need to carefully choose which inputs we investigate in tandem.

As with single parameter analyses, we use an Excel macro to adjust the parameters and record the results. This macro loops through a list of parameter groups. For each group of two or more parameters (each with a specified range and step size), the macro tries all permutations and records the two primary outputs. This process is best illustrated with several examples.

Table 5.8 shows the specifications for the macro. Each group of parameters is separated by a blank line, and results in its own output table.

Family Practice sensitivity analysis. The first group of parameters describes sensitivity analysis to be performed simultaneously on all three inputs involving family practice: LoS Family has four steps (800, 1200, 1600, 2000), Cost Family has five steps, and Other→Family has three steps, so the resulting output table will have $4 \times 5 \times 3 = 60$ rows. For brevity, the first few and the last few rows are shown in Table 5.9. The results are also shown (in their entirety) in Figure 5.1.

The left chart shows that LoS and routing probability (Other→Family) have larger effects on total cost when varied than does the cost of the Family Practice station. We can also see that total cost decreases with higher LoS and/or higher routing probability. Furthermore, the sensitivity analysis shows greater sensitivity when LoS is high and/or when routing probability is low. For instance, the effect of changing the cost of Family Practice from \$0 to \$8 is small regardless of the other two parameter values, but it is smallest when they are also small.

The right chart displays the same information for QALYs instead of cost. Note that the cost of Family Practice is omitted—this is because it has no effect on

Table 5.8: Sensitivity analysis—multiple parameter input specifications.

Parameter	Min	Max	Step	Base Value
LoS Family	800	2000	400	1275
Cost Family	0	8	2	2
Other→Family	0.05	0.15	0.05	0.08
LoS MMT	180	720	60	338
LoS CM	365	2190	365	1275
Cap ACT	90	180	45	90
LoS ACT	2000	4000	1000	3464
Cost ACT	30	70	20	46
Other→ACT	0.01	0.02	0.01	0.01
Inpatnt→ACT	0.01	0.03	0.01	0.01

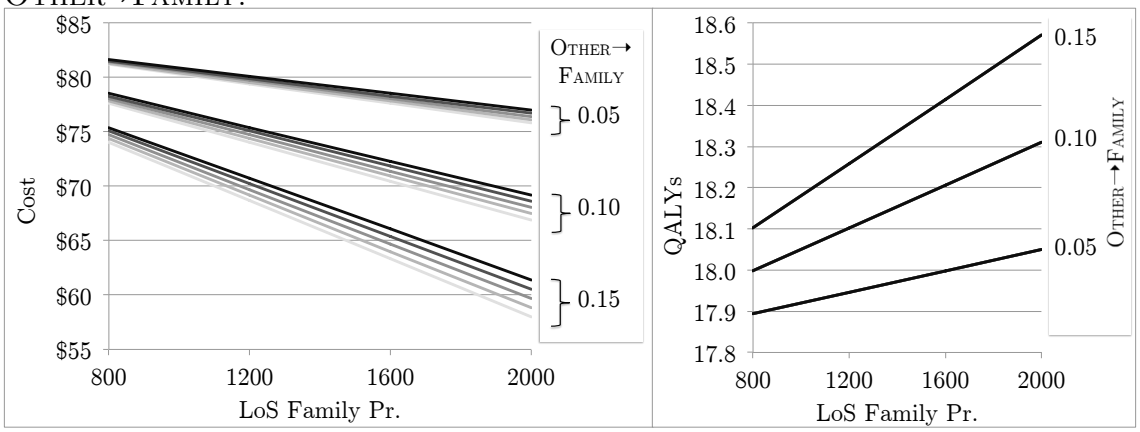
Table 5.9: Sensitivity analysis—multiple parameter example for Family Practice *(12 out of 60 rows shown).*

LoS Family	Cost Family	Other→Family	Cost	QALYs
800	0	0.05	$ 81.18	17.9
1200	0	0.05	$ 79.39	17.9
1600	0	0.05	$ 77.61	18.0
2000	0	0.05	$ 75.82	18.1
800	2	0.05	$ 81.29	17.9
1200	2	0.05	$ 79.56	17.9
1600	2	0.05	$ 77.83	18.0
2000	2	0.05	$ 76.10	18.0
800	4	0.05	$ 81.41	17.9
⋮	⋮	⋮	⋮	⋮
1200	8	0.15	$ 70.73	18.3
1600	8	0.15	$ 66.06	18.4
2000	8	0.15	$ 61.38	18.6

QALY outcomes in our model. Here we see that the effect on QALYs is opposite to the effect on cost: Higher LoS and/or higher routing probability correspond to *higher* QALYs. Again, the output is more sensitive when the LoS is high and/or routing probability is low.

This example shows purely linear relationships among the inputs and the outputs, however, this is often not the case. If changes to input parameters cause the model to cross a capacitated station threshold, the resulting outputs will exhibit

Figure 5.1: Cost *(left)* and QALYs *(right)* under different values of three input parameters: LoS Family Practice, Cost of Family Practice, and Other→Family.



piecewise linearity. Furthermore, many analyses will show nonlinear results for other reasons. Knowing that inputs will often be linear means we can restrict the steps to a small number (e.g., 2) when conducting sensitivity analysis on groups of parameters.

This example also illustrates an inverse relationship between cost and QALYs, however, that will not always be true. In some cases, input changes will cause both outputs to increase (or decrease) simultaneously.

We also mention an obvious limitation concerning the plots of results from multiple parameter sensitivity analysis. In the Family Practice example we only examine three parameters; plotting more than three parameters is difficult to do in a decipherable manner.

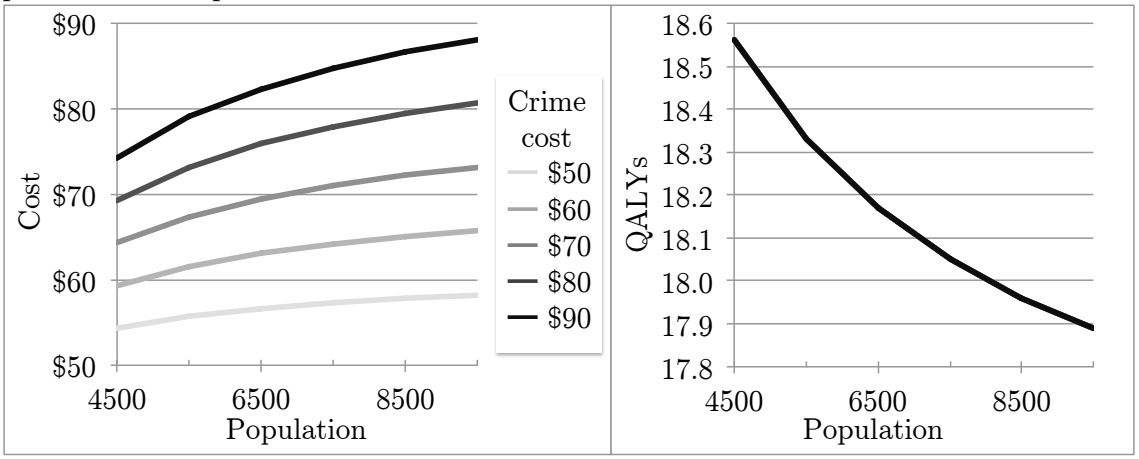
We include two more examples of multiple parameter analyses, and then discuss the entire sensitivity exercise below.

ACT sensitivity analysis Five parameters deal with the ACT station: Capacity of ACT, LoS of ACT, Cost of ACT, Other→ACT, and Inpatnt→ACT. This analysis produces interesting results in that these parameters, simultaneously varied over their respective ranges, have almost no effect on the main outputs. Across all 162 permutations the cost varies from $75.28 to $76.23 and the QALYs vary from 18.05 to 18.08. This group of parameters likely has little effect on the output for several reasons: The ACT station tends to fill to capacity regardless of what input

values we use (without making drastic parameter changes); the cost, including crime cost for clients in ACT, is very similar to the average per capita cost; and the station can only handle a small number of clients (compared to MMT, CM, or Family Practice). This result does not mean ACT is unimportant in an overall DTES strategy, however, it may simply highlight a limitation of this model. We believe a future version of the model that includes sub-populations may show ACT to be an important treatment option for certain types of clients.

Population and crime cost sensitivity analysis Two parameters that have a large effect on outputs when analyzed individually are population size and cost of crime in Non-tr (the estimate of crime costs per capita for Non-tr that is also used at a reduced level to estimate crime costs for clients in long term treatment). The sensitivity analysis of these two parameters shows that these effects intensify when both are varied simultaneously. Figure 5.2 shows the cost and QALY outputs as these two input parameters vary; the right chart does not include cost Non-tr crime because it has no effect on QALYs.

Figure 5.2: Cost *(left)* and QALYs *(right)* under different values of two input parameters: Population and cost Non-tr crime.



Additional parameter groups Some of the other parameter groups analyzed in the same manner but not discussed in detail are:

Crime parameters: Cost Non-tr crime, Police→CrimJst, and Crime treat ratio. The higher each of these inputs is, the higher the cost (up to $119 for extreme input values). QALYs are virtually unaffected.

Sensitive parameters 1: Several parameters that display high sensitivity individually are Population, Arrivals OTHER, LoS CM, and LoS FAMILY. As a group, these parameters only account for moderately larger fluctuations than they do individually. However, two instances result in infeasible solutions (when Population is small and LoS's are large, the external arrival rates are too high to sustain a positive NON-TREATMENT POPULATION).

Sensitive parameters 2: Another group of individually sensitive parameters are Population, OTHER→FAMILY, and Crime treat ratio. These parameters result in large fluctuations (cost ranges from $45 to $103 and QALYs range from 17.8 to 18.9) but the more extreme values correspond to the extreme input examples only.

The multiple parameter sensitivity analysis capability is useful for examining groups of inputs that require special consideration because it is clearly not possible to examine all groups. We have shown that it is easy to perform these analyses on any small group of inputs and analyze the resulting effects on cost and QALYs. With this framework in place we can address any concerns that arise regarding values of the estimates in our model.

One other form of sensitivity analysis is probabilistic sensitivity analysis [e.g., 82]. This approach involves treating parameters as random variables with various distributional assumptions. The analysis consists of repeatedly drawing different values for the parameters and measuring model outputs, then drawing inferences on the aggregate output. We do not use this approach herein.

The main inference from all of the sensitivity analyses—both single and multiple parameter—is that there are a few key inputs that have a potentially large effect on the main outputs. From a cost perspective, these key inputs are Population, Arrivals OTHER, LoS FAMILY, Cost INPATNT, Cost NON-TR, Cost NON-TR crime, and OTHER→FAMILY. From a QALY perspective, Population has the largest effect, though the range is within one QALY. All of these inputs are considered to be of high or medium quality, regardless, we recognize that there are likely inaccuracies in these data that do have an effect on the model outcomes. Nevertheless, the effects of changing the model to accommodate different scenarios later in this chapter will be fairly robust to input inaccuracies for two reasons: 1) The more sensitive inputs are of reasonable quality, and 2) the direction and magnitude of changes across different

scenarios will likely be affected only slightly by inexact inputs, because those inputs will be consistent across scenarios.

The inputs that we identified in the group parameter analyses as potentially problematic are LoS FAMILY and OTHER→FAMILY. Based on what we have observed in single and group parameter sensitivity analysis, will be addressing the quality of Population, Arrivals OTHER, LoS FAMILY, OTHER→FAMILY, and Cost INPATNT in the next iteration of our model (incidentally all are ★★☆; those that are ★★★ do not need additional investigation, and those that are ★☆☆ do not appear to exhibit sensitivity.).

5.4 Discrete Event Simulation

The closed queueing network approach we have chosen as the main tool for analyzing the system of services in the DTES provides us with accurate output values contingent on several questions:

- Is the model valid?
- Are the input parameters correct?
- Are the bounds established by the approximation approach reasonable?
- Are the mathematical modelling assumptions appropriate?

We discussed the validity of the model in §5.2.1, and addressed the fact that it has limited scope and is therefore limited by the features included in this version; we also showed that the error introduced by the LCQN approach is minimal for a small version of our network. In the previous section, we discussed the precision of the input parameters. Now we turn to the modelling assumptions, and also to confirming that the approximation approach introduces only minimal errors for the full model.

Because of the limitations of the queueing approach, the modelling assumptions and approximation approach are necessary. DES doesn't have this limitation; simulation models are more flexible in many ways, so even though there are drawbacks to using them (as discussed in §4.3.2), DES is a useful tool for justifying the modelling assumptions and approximations.

5.4.1 DES Model

The DES model is similar to the queueing model: It contains the same stations, the Non-treatment Population, and clients flowing through the system. However, rather than working with averages to calculate arrival rates and queue lengths which in turn allow us to calculate cost and health outcomes, the DES model accounts for the individual clients moving from station to station, and in and out of the Non-treatment Population.

The queueing models discussed in Chapter 4 describe a system in steady-state, however, simulation models usually require a warm-up period to reach an equilibrium[9]. For each replication we therefore run the model for a warm-up period during which we let the system settle into an equilibrium, and then we collect statistics after this time.

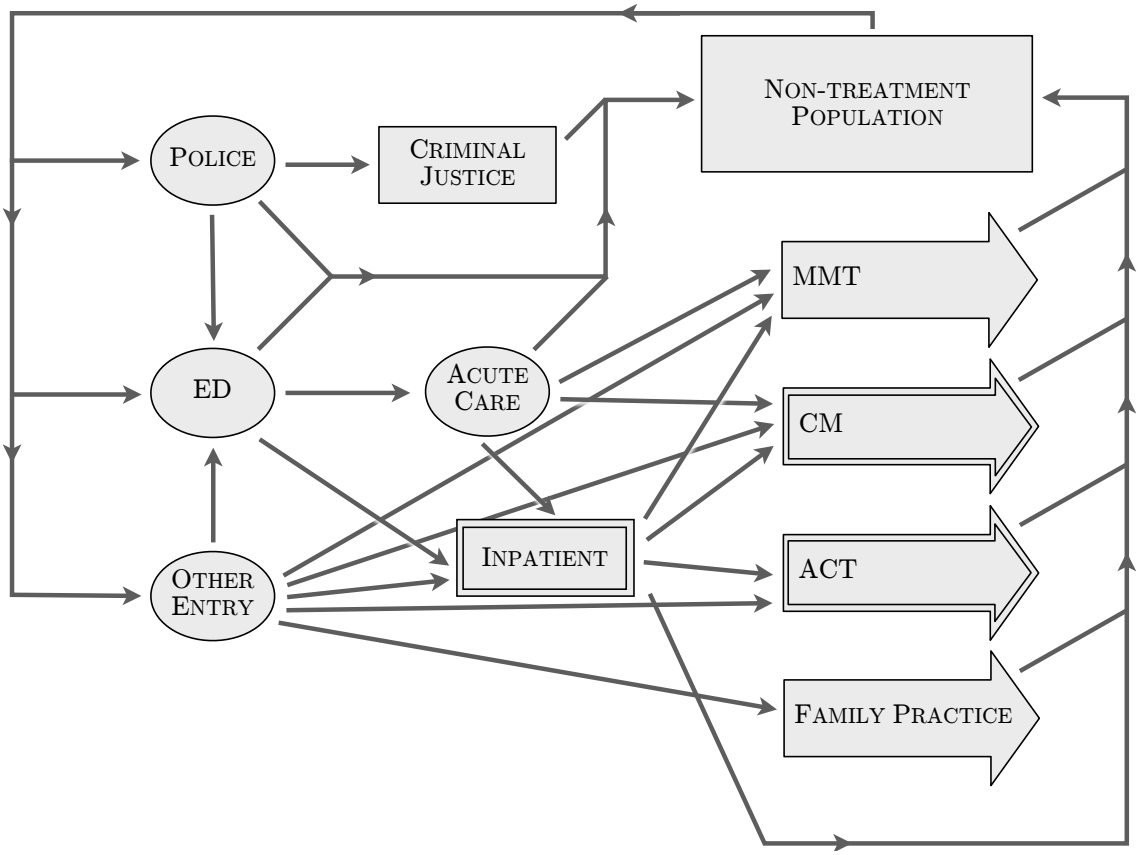
The easiest way to initialize the system would be to start with all clients in the Non-treatment Population, and let the stations fill up and equilibrate during the warm-up. Not surprisingly, this method could take a long time to reach an equilibrium. Instead we initialize the system using the suggested queue lengths from the queueing model so that the simulation reaches steady-state much more quickly. Because of this trick, the warm-up period can be fairly short, and even if the suggested starting point from the queueing model is not quite right, it is bound to be much closer to the simulation's equilibrium than an empty system. For each replication's warm-up period we use an ample 365 days. After the warm-up, each replication runs for 20 years. We use 30 replications (and a long run time) in order to obtain very narrow confidence intervals on the simulation output. Experimentation with different values shows that shorter warm-ups and runs provide slightly less consistent results, longer warm-ups do not provide a better steady-state starting point, and longer runs do not provide narrower confidence intervals.

During the simulation, clients randomly enter stations, depart stations, and move to the next station or to the Non-treatment Population. We provide a simple account of what happens when a client enters ED from Non-tr:

- The data for ED is updated to reflect a new arrival;
- the data for Non-tr is updated to reflect a departure;
- a service time is randomly chosen, based on the ED service time distribution;

[9] Queueing models that describe "transient" systems also exist.

Figure 5.3: DES model of DTES services: Short term = oval; medium term = box; long term = arrow-box; capacitated station = double outline; route = arrow.



- an event is scheduled in the software's event list for a client to leave ED after the service time has elapsed;
- the rates for external arrivals are all adjusted according to this new external (Non-tr) population size;
- the time for the next entry to ED is randomly chosen based on the external arrival rate distribution;
- and this next entry is scheduled for the random future time.

When a client leaves ED:

- The data for this station is updated to reflect a departure;
- the next destination for that client is randomly chosen, based on the ED routing probabilities...
  - an arrival is created for the chosen station *or*

– if the next destination is Non-treatment Population then an arrival is created for Non-tr and the external arrival rates are updated.

In this manner the simulation model proceeds, generating many of these events for each "day" in each replication.

The DES model is shown in Figure 5.3. Much of the diagram as the same is Figure 4.1 in Chapter 4, however the Non-treatment Population that was managed through the FPM constraint in the queueing model has been replaced by a station that functions the same as all of the other stations in the system except that departures from Non-tr are generated to reflect the external arrival processes to Police, ED, and Other Entry. In this way the DES model is an exact representation of a closed system (with finite population) rather than an approximation of one.

5.4.2 Testing Model Assumptions

In §4.2 we discussed bounds and approximation ratios associated with a simplified version of our approximated network. And in §4.1 we mention the distribution assumption—that the interarrival times and the service times are all exponentially distributed, i.e., they each follow a Poisson process. These are the two types of assumptions we explore with the DES model. Specifically, we investigate each of the following—running 30 replications (one year for warm-up plus twenty years for simulation) each.

Simulation Models:

1. **FPM**: We have shown the FPM approximation is extremely accurate for this setting, but we nevertheless compare the *uncapacitated* queueing network with the corresponding *uncapacitated* DES model using Markovian arrivals and service times. This model also serves to validate the simulation approach by showing that we obtain the same output measures as with a queueing model.

2. **Capacitated stations:** We activate the *capacities* for Inpatient, CM, and ACT in both models to see if the approximation in the queueing network performs similarly to the *capacitated stations* in the DES model (in which individual clients balk if they observe the station to be full, thereby more accurately representing the real world).

3. **Service time distributions:** We compare the exponentially distributed *service times* in the queueing model with other distributions in the DES model to see how sensitive the outputs are to distributional assumptions.

4. **Arrival process distributions:** We compare the exponentially distributed *interarrival times* in the queueing model with other distributions in the DES model.

One way to compare the queueing network with the DES model is to look at the average number of clients in each station and in the Non-treatment Population. We can see that the bounds and approximation ratios related to our approximations are indeed practically insignificant for the level of precision we require: Table 5.10 shows a comparison for the uncapacitated models and Table 5.11 shows the same comparison for the capacitated models.

Table 5.10: Simulation—comparison of average number of clients in each station for the *uncapacitated* queueing model and the corresponding simulation model. 95% confidence intervals (CI) are also shown for the DES results.

Station	Queue Uncap.	DES Uncap.	DES Uncap. CI	% Dif.
Police	2.1	2.1	(2.1, 2.1)	–0.1%
CrimJst	259.6	259.0	(258.0, 259.9)	–0.2%
ED	5.5	5.5	(5.5, 5.5)	–0.1%
Acute	55.6	55.8	(55.6, 56.0)	+0.4%
Inpatnt	258.5	258.5	(257.4, 259.6)	+0.0%
MMT	254.0	254.5	(252.4, 256.7)	+0.2%
CM	1674.3	1678.7	(1670.0, 1687.3)	+0.3%
ACT	424.9	422.6	(417.9, 427.3)	–0.5%
Family	955.0	956.9	(951.4, 962.5)	+0.2%
Non-tr	3610.6	3606.5	(3601.0, 3611.9)	–0.1%

The uncapacitated queueing model is very similar to the uncapacitated simulation model. The difference in the average number of clients in each station is statistically insignificant at the 0.05 significance level[10], illustrating (for this specific set of parameter values) that the FPM approximation is certainly accurate enough for our purposes.

[10]Confidence intervals are created using a t-statistic with $(30 - 1)$ degrees of freedom and the averages from the 30 independent replications.

This comparison also verifies that the simulation model is working as designed, in other words, there are no bugs or modelling mistakes.

Table 5.11: Simulation—comparison of average number of clients in each station for the *capacitated* queueing model and the capacitated simulation model (with $1/\mu \sim$ Exp and $1/z_j L_0 \sim$ Exp). Again, 95% confidence intervals are given for the DES results.

Station	Queue Base	DES Cap.	DES Cap. CI	% Dif.
Police	2.4	2.4	(2.3, 2.4)	+0.1%
CrimJst	296.7	297.4	(296.7, 298.2)	+0.3%
ED	6.3	6.3	(6.2, 6.3)	−0.0%
Acute	63.5	63.6	(63.4, 63.8)	+0.2%
Inpatnt	161.0	160.8	(160.8, 160.8)	−0.1%
MMT	264.7	264.0	(262.3, 265.7)	−0.3%
CM	1399.0	1396.7	(1396.5, 1396.8)	−0.2%
ACT	89.0	89.7	(89.7, 89.8)	+0.8%
Family	1091.4	1089.6	(1082.9, 1096.4)	−0.2%
Non-tr	4126.1	4129.5	(4123.4, 4135.5)	+0.1%

The capacitated queueing model (LCQN), despite including both the FPM approximation and our capacitated station approximation, clearly shows very little difference from the corresponding simulation model. Though the differences are statistically significant (several queueing values fall outside the DES confidence intervals), they are practically insignificant. This comparison validates the use of these approximations, at least for this specific set of input parameters. We also include a comparison of the same capacitated queueing and simulation models with a very different set of input parameters in Appendix E, which again shows very little difference between the two sets of outputs.

To investigate the effect of different service time distributions we modify the capacitated simulation model to use distributions other than exponential. Instead of exponential service times with $1/\mu_j \sim \text{Exp}(1/\mu_j)$ we test two other distributions: Uniform with $1/\mu_j \sim \mathcal{U}\big(0.9/\mu_j, 1.1/\mu_j\big)$; and log-normal[11] with $1/\mu_j \sim \ln\mathcal{N}\big(1/\mu_j, 2/\mu_j\big)$. The uniform distribution is intended to test fairly consistent times (we have parameterized the variance to be quite small—about 6% of that given by the corresponding exponential distribution), and the log-normal distribution is for testing very right

[11]The log-normal distribution is often described with its own special parameters μ (log-scale) and σ^2 (shape); for simplicity, and to be consistent with Arena's notation, we describe it with the mean and standard deviation of the resulting random variable.

skewed times with the same expected value (we have parameterized the variance to be four times that given by the corresponding exponential distribution).

The results are shown in Table 5.12. With this set of input parameters, the uniform service times increase the number of clients in (uncapacitated) long term treatments and Criminal Justice while decreasing the number in the Non-treatment Population. The log-normal service times have a small effect—decreasing the number in Family Practice and increasing the Non-treatment Population. The size of the effects on Family Practice and Non-tr are large enough to warrant further exploration, though both of the distributions we tested represent extremes that are unlikely to be observed for the actual service times at these stations. Nevertheless, we plan on examining service time data in the next iteration of our research.

Table 5.12: Simulation—comparison of average number of clients in each station for the capacitated queueing model and simulation model under different *service time* distributions.

Station	Queue Base	DES $1/\mu_j \sim$Exp	DES $1/\mu_j \sim \mathcal{U}$	DES $1/\mu_j \sim \ln\mathcal{N}$	CI for $1/\mu_j \sim \mathcal{U}$	CI for $1/\mu_j \sim \ln\mathcal{N}$
Police	2.4	2.4	2.3	2.3	(2.3, 2.4)	(2.3, 2.4)
CrimJst	296.7	297.4	297.0	295.6	(296.4, 297.7)	(294.4, 296.9)
ED	6.3	6.3	6.6	6.9	(6.6, 6.6)	(6.9, 6.9)
Acute	63.5	63.6	67.1	69.9	(67.0, 67.3)	(69.6, 70.2)
Inpatnt	161.0	160.8	160.8	160.9	(160.8, 160.8)	(160.9, 161.0)
MMT	264.7	264.0	269.9	258.9	(268.7, 271.1)	(256.5, 261.3)
CM	1399.0	1396.7	1323.9	1380.5	(1323.1, 1324.8)	(1378.6, 1382.5)
ACT	89.0	89.7	89.0	89.6	(88.9, 89.1)	(89.6, 89.6)
Family	1091.4	1089.6	1176.7	944.6	(1170.9, 1182.4)	(937.9, 951.3)
Non-tr	4126.1	4129.5	4106.5	4290.7	(4100.7, 4112.2)	(4283.5, 4297.8)

Finally we investigate different interarrival time distributions by modifying the capacitated simulation model (using the original exponentially distributed service times). In our queueing model, the time between arrivals—interarrival times—is assumed to be exponentially distributed, i.e., $1/(z_j L_0) \sim \text{Exp}(1/(z_j L_0))$. Note that the interarrival time is the reciprocal of the per capita rate times the calling population L_0. We test the same two distributions as above: Uniform with $1/(z_j L_0) \sim \mathcal{U}\big(0.9/(z_j L_0), 1.1/(z_j L_0)\big)$; and log-normal with $1/(z_j L_0) \sim \ln\mathcal{N}\big(1/(z_j L_0), 2/(z_j L_0)\big)$.

The results are displayed in Table 5.13. Changing the arrival processes has almost no effect on the DES model outcomes. The most plausible explanation is

that the model is much less sensitive to the arrival process than it is to the service times, because arrivals happen many times a day while service times can be months or years in duration. And changing the way one or two dozen arrivals happen throughout the day is not noticeable at the time scale of the longer term stations.

Table 5.13: Simulation—comparison of average number of clients in each station for the capacitated queueing model and simulation model under different *interarrival time* distributions.

Station	Queue Cap.	DES $1/(z_j L_0)$~Exp	DES $1/(z_j L_0)$~$\mathcal{U}$	DES $1/(z_j L_0)$~ln$\mathcal{N}$	CI for $1/(z_j L_0)$~$\mathcal{U}$	CI for $1/(z_j L_0)$~ln$\mathcal{N}$
Police	2.4	2.4	2.3	2.4	(2.3, 2.4)	(2.3, 2.4)
CrimJst	296.7	297.4	297.2	296.8	(296.3, 298.0)	(295.5, 298.1)
ED	6.3	6.3	6.3	6.3	(6.2, 6.3)	(6.3, 6.3)
Acute	63.5	63.6	63.7	63.8	(63.5, 63.8)	(63.6, 64.0)
Inpatnt	161.0	160.8	160.9	160.7	(160.8, 160.9)	(160.7, 160.8)
MMT	264.7	264.0	264.8	262.7	(263.2, 266.4)	(260.8, 264.5)
CM	1399.0	1396.7	1396.8	1396.3	(1396.7, 1396.9)	(1396.2, 1396.5)
ACT	89.0	89.7	89.7	89.7	(89.7, 89.7)	(89.7, 89.7)
Family	1091.4	1089.6	1094.0	1093.4	(1088.4, 1099.7)	(1087.7, 1099.2)
Non-tr	4126.1	4129.5	4124.4	4127.8	(4119.5, 4129.2)	(4122.3, 4133.4)

Note that in our queueing model, L_0 is the calculated average size of the Non-treatment Population, however, in the simulation model it is continually updated to reflect the simulated size of this population as the model progresses. The arrival rates are therefore constantly fluctuating.

In future versions of the DTES model it will be necessary to test additional assumptions related to sub-populations, clients in long term treatment who depart temporarily to visit other stations (e.g., Police, ED), and clients concurrently receiving multiple treatments. We have shown that the DES model provides a mechanism to test such assumptions, and that despite the shortcomings of this approach, it is a valuable tool for verifying the correctness of our queueing model.

5.5 Scenarios

Scenario analysis is the process of comparing different scenarios to the "base case" and to each other in order to draw conclusions about which system modification or set of modifications is most likely to achieve the best outcomes. It is frequently used in many modelling applications, including those based on LPs and on simulation.

Out of the many possibly interesting scenarios to study we limit our discussion to a handful that represent the breadth and form of the procedural and policy changes this model is able to explore. When we mention "long term treatments" we mean MMT, CM, ACT, and Family Practice. We study each scenario in isolation (not cumulatively).

Scenarios:

1. **Increased referrals**: From the ED and Acute Care to Inpatient and long term treatments.
2. **Lower turnover**: By increasing LoS in long term treatments.
3. **Expanded Inpatient capacity**: Inpatient capacity is doubled.
4. **Expanded long term capacity**: CM and ACT capacities are both increased.
5. **Urgent Response Centre**: A facility is added that accepts clients from the Police and refers them on to other stations.
6. **Decreased ED use**: ED per capita total arrival rate is cut in half, to approximate the effect of a very successful drop-in clinic program or of other interventions that reduce the need for this service.
7. **Decreased Crime**: Crime costs and Criminal Justice rates are reduced to see how much the criminalization of this population affects outcomes.
8. **Uncapacitated model**: Capacities on Inpatient, CM, and ACT are removed so that the entire network is free of capacity restrictions (this scenario is the same as the uncapacitated queueing network compared to the first DES model discussed in §5.4.2, but with the FPM constraint).

Combination Scenarios

A. **Increased referrals** *and* **Expanded long term capacity**

B. **Increased referrals** *and* **Lower turnover**

C. **Expanded Inpatient capacity** *and* **Expanded long term capacity**

D. **Increased referrals**, **Lower turnover**, **Expanded long term capacity**, **Decreased ED use**, *and* **Decreased Crime**

E. **All scenarios** (excluding Uncapacitated)

The framework we use for creating scenarios is quite simple: For each scenario we create a copy of the model that links back to the base model for all parameter values. We then adjust only the desired inputs and/or model structure accordingly. The one additional adjustment is that the scenarios all use the per capita external arrival rates (calculated in the base case), not just for ED (this one is provided as an input), but also for POLICE and OTHER ENTRY. In this way we can realistically investigate changes to the system that increase or decrease the size of the NON-TREATMENT POPULATION, based on the assumption that these arrival rates do occur in the DTES on a per client basis.

Creating a copy of the model for each scenario that links back to our input parameters makes it easy to run sensitivity analysis on these scenarios. We use the same framework described in §5.3, but record the outputs from the scenario(s) of interest in addition to the base case outputs.

5.5.1 Scenario Analysis Results

The scenarios we test vary in their effect on cost and QALYs. Clearly we would like to identify a scenario that would be inexpensive, politically feasible, and procedurally straightforward to implement and that includes the largest cost decrease and QALY increase. However, to actually implement any of these scenarios would require at least some investment in new resources, system changes, or new programs. At this point we restrict our discussion to the outputs captured in our model, with the caveat that the cost output does not represent the entire cost.

Table 5.14 lists each scenario, its description, and its predicted per capita cost and QALYs. (As with other solution approaches, scenario analysis produces many additional outputs.) Of note is that all scenarios reduce the cost (compared to the base case) except for Scenario 2—Expanded INPATNT capacity. This increase is due to the high cost of the INPATIENT station. The scenario with the largest cost decrease is Combo D—a combination of Scenarios 1, 2, 4, 6, and 7. The predicted per capita cost is $55.29, which represents an annual population cost of $151 million (compared to $208 million for the base case). This substantial savings corresponds

Table 5.14: Scenario analysis—costs and QALYs for individual scenarios and combinations of scenarios.

Scenario	Description	Cost	QALYs
Base	Base case	$75.93	18.1
Scenario 1	Inreased referrals	$72.20	18.2
Scenario 2	Lower turnover	$72.14	18.2
Scenario 3	Expanded INPATNT capacity	$79.40	18.1
Scenario 4	Expanded long term capacity	$74.67	18.1
Scenario 5	Urgent Response Centre	$75.23	18.1
Scenario 6	Decreased ED use	$72.94	18.0
Scenario 7	Decreased crime	$65.25	18.1
Scenario 8	Uncapacitated	$77.25	18.2
Combo A	Increased referrals & expanded long term cap.	$70.18	18.3
Combo B	Inreased referrals & lower turnover	$67.79	18.3
Combo C	Expanded INPATNT & long term capacities	$77.38	18.2
Combo D	Scenarios 1, 2, 4, 6, 7	$55.42	18.4
Combo E	All scenarios (excluding "uncapacitated")	$56.70	18.5

to a predicted per capita lifetime QALY increase of 0.3, which is better than with most of the other scenarios.

This savings, compared to the current situation, ignores the costs associated with implementing programs that would bring about the increased referrals, lower turnover, etc. (if such changes are even possible). But it also emphasizes that the best scenario need not involve doing *all* of the things that seem like good ideas (note that Combo E—All scenarios—is inferior in terms of cost but not in terms of QALYs). Furthermore if shows how several changes made in chorus can have a smaller impact than the sum of each considered individually: The per person cost savings of Combo D is about $21, whereas the sum of the savings of the individual scenarios comprising it is about $23. However sometimes the reverse is true. For instance, Combo B produces more savings than the combined savings of Scenarios 1 and 2. These two examples illustrate the extent to which the results of such a system can be nonlinear, and why it is crucial to model all of the interactions rather than examining individual programs and services.

As in the sensitivity analysis, the QALYs are fairly consistent, likely because of the conservative approach we use in modelling health outcomes.

Several other outputs are also interesting to compare across scenarios. Table 5.15 lists the projected average NON-TREATMENT POPULATION size, as well as the

Table 5.15: Scenario analysis—Non-treatment Population size and average client counts for certain stations.

Scenario	Non-tr	CrimJst	Inpatnt	MMT	CM	ACT	Family
Base	4,126	297	161	265	1,399	89	1,091
Scenario 1	3,704	266	161	346	1,399	89	1,470
Scenario 2	3,695	266	161	360	1,399	89	1,466
Scenario 3	4,019	289	288	283	1,399	89	1,063
Scenario 4	3,798	273	161	246	1,682	269	1,005
Scenario 5	4,054	291	161	295	1,399	89	1,140
Scenario 6	4,180	301	161	225	1,399	89	1,106
Scenario 7	4,169	239	161	267	1,399	89	1,103
Scenario 8	3,611	260	258	254	1,674	425	955
Combo A	3,208	231	161	304	1,999	269	1,273
Combo B	3,203	230	161	455	1,399	89	1,906
Combo C	3,690	265	264	260	1,711	269	976
Combo D	2,844	163	161	344	1,999	269	1,693
Combo E	2,749	158	215	382	1,999	269	1,702

projected average number of clients in Criminal Justice and each of the medium and long term treatments. From this table it is clear that Combo D is the lowest cost scenario because it shifts clients out of Non-tr and into the long term treatments. But we can also see that this combination scenario would require more Family Practice resources than are currently estimated to be in use.

The last set of outputs we compare across scenarios are the balking rates for the three capacitated stations. These rates are calculated as the "balk" station arrival rate divided by the "continue" station arrival rate for each of Inpatient, CM, and ACT. Table 5.16 shows how many of the scenarios actually *increase* the balking rates as they divert more clients to, or decrease turnover at, these stations. A hyphen signifies the arrivals at that station, in that scenario, are low enough that we can assume no clients balk.

The specific parameter and/or structure changes introduced in each scenario could be adjusted; different values would no doubt result in different outcomes. Table 5.17 describes these specifications for each scenario including the factors used to increase or decrease various inputs. A single table containing all of the outputs discussed above is included in Appendix F.

Table 5.16: Scenario analysis—rate of balking ($\lambda_{jb}/\lambda_{jc}$) for capacitated stations, i.e., number of clients who balk for each client served.

Scenario	INPATNT λ_b/λ_c	CM λ_b/λ_c	ACT λ_b/λ_c
Base	0.8	0.3	3.9
Scenario 1	1.5	0.7	5.3
Scenario 2	0.6	0.8	5.6
Scenario 3	-	0.3	4.3
Scenario 4	0.7	-	0.5
Scenario 5	0.9	0.4	5.9
Scenario 6	0.3	0.2	3.9
Scenario 7	0.9	0.3	3.9
Scenario 8	-	-	-
Combo A	1.1	0.1	0.8
Combo B	1.1	1.2	7.3
Combo C	-	-	0.6
Combo D	0.3	0.3	1.5
Combo E	-	0.3	2.2

5.5.2 Sensitivity Analysis of Scenarios

The sensitivity analysis we performed on the base case is intended to show that even though the model may be sensitive to certain inputs, we can still rely on the quality of the outputs, especially if we are aware of which input parameters we must be mindful of. Performing sensitivity analysis on each scenario is often considered unnecessary, especially given that we are most interested in comparing the main outputs across scenarios that will all tend to be affected in the same direction, and to a similar magnitude, given slight changes in the input parameter values. As with the sensitivity analysis of groups of parameters, we proceed with this exercise hoping to find no surprises.

We perform single parameter sensitivity analysis (as per §5.3) on each scenario. The results of these analyses are presented in Appendix F (each scenario occupies a full-page table). There are indeed no surprises; the same parameters that have a larger impact on the base case also tend to have a large impact on the scenarios. For instance, the cost range associated with varying Cost NON-TR is fairly large (from the base value of \$0 up to \$50). Similarly the cost range associated with varying the Crime treat ratio is also somewhat large (from its base value of 0.33 up to a value of 1.0, representing equal crime rates for clients in long term treatment as for clients in the NON-TREATMENT POPULATION). The population size also has a large effect

Table 5.17: Scenario analysis—detailed description of parameter and model structure changes for each scenario.

Scenario	Parameter and structure changes
Base	-
Scenario 1	Increase routing probabilities from {ED, OTHER, ACUTE} to {INPATNT, MMT, CM, ACT, FAMILY} by a factor of 1.5.
Scenario 2	Increase LoS for {INPATNT, MMT, CM, ACT, FAMILY} by a factor of 1.5.
Scenario 3	Increase INPATNT capacity by a factor of 2.
Scenario 4	Increase capacity for CM to 2000 and for ACT to 270.
Scenario 5	Divert half of POLICE→ED clients to an urgent response centre station for an average of one hour, and then on to {INPATNT (0.4), MMT (0.2), CM (0.2), ACT (0.1), FAMILY (0.1)}.
Scenario 6	Adjust ED arrivals by a factor of 0.5.
Scenario 7	Adjust CRIMJST arrivals (via POLICE, leaving noncrime arrivals as is) and adjust crime cost, both by a factor of 0.8.
Scenario 8	Remove capacities on {INPATNT, CM, ACT}.
Combo A	Scenario 1 & Scenario 4
Combo B	Scenario 1 & Scenario 2
Combo C	Scenario 3 & Scenario 4
Combo D	Scenarios 1, 2, 4, 6, 7
Combo E	Scenarios 1 – 7

on both cost and QALY outputs.

We do not present results for multiple parameter sensitivity analysis of scenarios. Given the sheer number of permutations that might be considered we reserve this capability for addressing specific concerns if and when they arise.

Overall, the same considerations discussed with respect to the sensitivity analysis results for the base case also hold for the scenarios. This information is useful in interpreting the scenario results, and will also be useful as we refine input values for the next version of this model.

5.6 Discussion

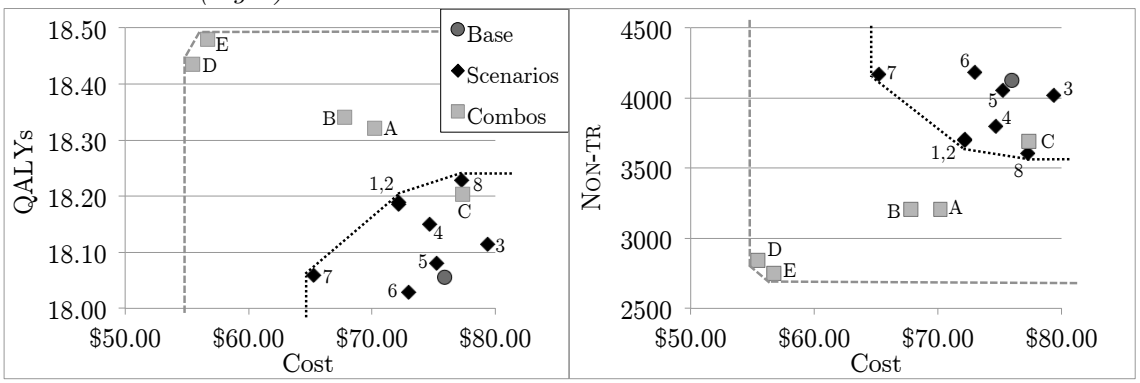
The most important implication of this chapter is that there are ways to save money *and* improve health outcomes. This claim requires two caveats: 1) Costs of system changes were not included and 2) the scope of this model is limited. Regardless, we can see that the cost of not changing anything is very high. If a combination of these scenarios were implemented—for instance, increased capacities in long term treatment stations, programs to decrease turnover, an urgent response centre for the police, and higher routing probabilities to specific programs—clients in the DTES would surely benefit from more and longer treatment stays and from fewer trips to the ED. Consequently, the strain on the ED and Acute Care would be lessened a little bit. Depending on the costs of these changes, the overall cost to the system could decrease while health outcomes increased.

One clear driver of the cost output is the crime cost. The sensitivity analysis shows that this input affects the outputs, and Scenario 7 explores reducing this and related parameters. The cost of crime is therefore important as an input, but it is also important as a factor when we explore scenarios. Clearly, any procedural or policy changes that can reduce this cost will have an impact on the overall system cost.

One way to explore which scenarios are most beneficial is by creating efficient frontiers. Figure 5.4 shows cost plotted against two different measures: QALYs and Non-treatment Population size. We observe that Non-tr is an almost perfect proxy for QALYs—the left and right charts are almost exact mirror images of each other in which higher QALYs correspond to lower Non-tr. This relationship suggests that any positive change in health outcomes will correspond to a decrease in the Non-treatment Population.

Because it is difficult to quantify the trade-off between cost and QALYs, it is useful to examine all scenarios that aren't dominated by other scenarios (i.e., are closest to the optimal corner of the chart using an arbitrary weighting for the two axes). The black dotted line shows the efficient frontier for individual scenarios and the grey dashed line shows it for combination scenarios. We observe that, when considering the individual scenarios, there are four that stand out: S1, S2, S7, and S8. For the combination scenarios the two that sit on the efficient frontier are CD and CE.

Figure 5.4: Scenario results and efficient frontiers for scenarios (black) and combination scenarios (grey). Cost vs. QALYs *(left)*; cost vs. NON-TREATMENT POPULATION *(right)*.



These charts also help us explore trade-offs between scenarios. For instance Scenarios 7 and 8 exemplify the competing goals of reducing costs and increasing QALYs. The former reduces costs with almost no change in QALYs, while the latter increases QALYs with a minor increase in cost. The ability to examine various options in this manner is one of the strengths of this type of model.

Even with the limitations mentioned above, the sensitivity analysis demonstrates that policy implications we draw from this model are robust against sensitive inputs. In other words, while we may not know the exact routing probabilities or LoS, and we consequently cannot predict the exact cost and QALY outcomes, the best scenarios are generally still the best choices across a large range of inputs.

In subsequent research we will incorporate estimates for the costs of implementation into these (and other) scenarios. At this point it is clear that improvements to the system likely involve increased capacities, lower turnover rates, ways to deal with crime rates and crime costs, and reduced ED use. There are no surprises in this list, however, as has been mentioned before, our model provides a method for quantifying the effects of each and all of these changes in order to determine where to expend resources and what to implement first.

Chapter 6

Conclusion

Our research illustrates how strategic planning provides benefits in two very different healthcare systems. In one system we plan career paths for skilled healthcare workers for the next couple of years; in the other we assess and plan services and resources for a marginalized population. We measure success in the first system by the soft constraints we are able to meet and by the plan's ability to handle variability in staff leaves. These goals are indirectly related to better resource usage and improved ability to deliver quality service, therefore they are also indirectly related to cost and health outcomes. In the second system variability is incorporated into the methodology, and we measure success directly in terms of predicted cost and health outcomes.

In both applications we have chosen simple approaches for representing complex aspects of the respective systems: The BCCA workforce planning model handles variability in an MIP via a soft constraint that encourages redundancy in workforce experience; the DTES queueing network is solved using two different approximations that, when combined, allow for an LP formulation to achieve solutions almost instantaneously, even though closed queueing networks are typically hard to solve.

6.1 Workforce Planning at the BCCA

For strategic workforce planning, our multi-period assignment problem with side constraints and a GP solution approach incorporates the primary components necessary in a radiation therapy setting. Our basic model—with duration, experience, and redundancy constraints—is extended in our application to approximately 90 RTs, 14 areas, and 8 periods. We also provide guidance, informed by a simulation model and regression analysis, on setting the redundancy parameter; at the Vancouver Centre of the BCCA we have opted to use $f = 2$ to achieve a good balance between robustness and over-training. The combination of constraints is a new contribution to the workforce planning literature.

The tool our team developed was used for three years and is considered valuable for two main reasons: 1) It provides plans that are feasible, robust, and appealing to the RTs; and 2) it saves the chief RT and resource therapists considerable time.

Providing a plan that achieves all of the goals defined by management has been a great success for the Vancouver Centre. According to the chief RT, "The introduction of a Radiation Therapy Staff Planning Model provides the chief therapists the opportunity to plan the direction of each individual staff member over a period of years. The advantages of this lie in the ability to use resources more effectively, to enhance career development, to anticipate potential shortages in specialty areas and train staff to fill those gaps before they appear, and to provide staff with a clear map of their future direction within the workplace. The staffing model provides chief therapists with a clear record of the progress and development of any staff member and allows for personnel changes within the department to be effectively managed." Additionally, providing a rapid solution means the tool can help the department look beyond the next few weeks. "Historically, the scheduling of radiation therapists has required extensive amounts of time spent by chief therapists in planning the daily and monthly schedules, a complex and often tedious process that has focused on the coming month." [101]

Although we have run the model beyond two years while returning timely solutions, we find the two-year horizon is satisfactory. Because the casual staff tend to fill areas that require less experience, most of the permanent staff already have three or more years of experience. Given that it takes five to six years to acquire experience at the highest level, the horizon is long enough to track this entire trajectory. Furthermore, extending the model beyond two or three years could introduce more variability, particularly from retirements. Nevertheless, rounding heuristics [75], a looser solution gap, or additional constraints on staff eligibility (to further restrict the solution space) could be useful should we need to lengthen the planning horizon.

6.2 Strategic Planning in the DTES

Our queueing network model of health and criminal justice services in Vancouver's DTES demonstrates that a high-level quantitative approach can be very useful in this setting. Containing ten of the most important and/or costly services and a

finite population of 7,500 clients, this model produces cost and QALY outputs that can be used to compare potential procedural or policy changes to the system.

This research provides two main contributions. The first contribution, to the queueing network literature, is the LCQN approximation; the second contribution, to the field of operations research healthcare, is the systems modelling approach for a network of mental health and addiction services for a needy population.

Even with the caveats mentioned in Chapter 5 that the model is limited in scope and doesn't include all costs, we note several important policy implications: 1) The Non-treatment Population must be reduced in size to decrease overall system costs; 2) such a reduction must come from expanded long term treatment capacities and probably from increased referrals to such programs; 3) the changes we explore can help reduce ED usage, but only so far (i.e., additional programs to reduce ED usage need to be explored); and 4) a *combination* of system improvements is required to bring about meaningful cost and health improvements.

We created the model in Excel so that it is easy to share with other researchers and decision-makers. Even though it is limited in scope, it nevertheless provides useful insight into how the current system works and what effects might be felt from different scenarios.

The model also demonstrates our new methodology—the LCQN approach— for approximating closed queueing networks with certain characteristics. Extremely fast solutions can be found for much larger networks, assuming some infinite-server stations, some finite-server stations with many servers, and a large finite population that is impatient (won't wait but rather chooses to balk when services are full). We demonstrate that the quality of the approximate solutions, given appropriate network characteristics, are very close to the exact solutions.

Much of the value of this application is that it is a first step toward building a more comprehensive model of the services in the DTES. On its own it demonstrates the types of procedural and policy changes we can explore, and the potential benefits from various changes. It supports the general consensus that treatment is less expensive—and leads to better health outcomes—than not doing anything. It also introduces new tools to this community. Finally it demonstrates that we can quantify the outcomes of the various solutions that researchers, providers, and politicians have proposed for the people of the DTES.

6.3 Approximation Bounds and Ratios; Solution Times

Solutions to the workforce planning model typically take one–two minutes. One reason we are able to find solutions so quickly is that we are satisfied with near-optimality. The solver stops when the solution is within 1% of optimality, though it is worth noting that finding an optimal solution typically only doubles the total solution time. Because the objective function involves arbitrarily-determined trade-offs, we find that a 1% solution gap is perfectly acceptable. Furthermore, in light of user adjustments (e.g., via forced-assignments), our solution gap is probably narrower than it needs to be.

Solutions to the DTES queueing network take a fraction of a second, and because the model uses an LP, the approach scales very well to much larger networks. We did not include comparisons to exact approaches for solving large closed queueing networks, though they would be slower. But we demonstrated that our new approximation approach of linearizing a closed queueing network is simple, fast, and provides small gaps.

Compared to DES our queueing network algorithm is many times faster. Our Excel model finds the solution for the base case as well as 13 additional scenarios, for 300 sensitivity analysis instances, in under 40 seconds. That works out to over 100 solutions per second. In contrast, with the warm-up period of one year and a further 20 years of simulation in each of 30 replications, a single solution from the DES model takes about ten hours. If we were to do all of the scenario and sensitivity analyses using this simulation model it would take $14 \times 300 \times 10$ hours, or over four years. Even using variance reduction techniques, a shorter run time, and fewer replications, we would still need days, weeks, or months to perform these analyses.

6.4 Lessons Learned and Challenges

These applications were not without challenge. At the BCCA we found that ease of use and intuitive interfaces can be more important than the details of objective costs or side constraints. As well, we discovered that the capability for the user to "adjust" the resulting plan—using the second objective function—was indispensable.

Figuring out how to represent certain rules of thumb as mathematical constraints also required a lot of discussion in which we sometimes learned that the rule of thumb was unnecessary, or at the least, misleading, given our modelling approach.

With the DTES model we learned a lot about the system while defining the research question. Initial versions of the model were either too abstract to be applicable or too detailed to be generalizable. The main challenges we faced were in defining an appropriate scope and then finding suitable data inputs to populate the model.

6.5 Future Work

We are working on sharing the workforce planning tool with other BCCA centres around the province, and our team—the CIHR Team in Operations Research for Improved Cancer Care—has recently created a daily scheduling model that will interface, in the future, with this planning model in order to replace the manual monthly process of scheduling the workforce.

The next version of the DTES queueing model will include additional health, housing, and social services. We will also incorporate sub-populations, removing the homogeneity assumption. This step will allow us to model different demographic groups as well as clients with different types of needs and service usage patterns.

This next version will also remove some of the limitations around client flow, so that a client could be in multiple treatments simultaneously (rather than in their primary treatment only), or they could have Police, ED, or Criminal Justice involvement while in long term treatment and then return to that treatment.

Future work on the queueing approximation (LCQN) involves several directions: 1) Establishing an appropriate result regarding the LP approximation for the full network; 2) finding bounds for the entire network, and not only for a single station (as in §4.2); 3) determining how to extend the queueing approximation to the multi-class case, so that sub-populations can be represented; and 4) extending the network to include clients concurrently in more than one station.

The overarching goal of the next version of the DTES model is to not only create high-level policy recommendations, but to be able to use the model for more detailed planning. To do so, costs must be all-inclusive, the scope must include all major services, and population must include more detail so that different types of clients

who interact with the system in unique ways are better understood and represented.

6.6 Applicability to Healthcare and Other Industries

Beyond the BCCA centres, other healthcare workers—physicians, nurses, or technicians—as well as other industries—education, hospitality, military, or others—would benefit from the workforce planning methodology. Our basic model provides a starting point for a staff planning application, and our BCCA implementation illustrates some of the features and challenges one can expect to encounter.

The DTES model is also generalizable. It is immediately applicable to other finite populations consuming healthcare resources within a network, such as other marginalized urban populations. The approach would have to be modified if it were applied to populations with different medical conditions where waiting for services is the norm.

The queueing approximation we introduce certainly has applicability to other industries. Any large closed network with high levels of balking and with some stations operating at capacity could be modelled. Busy computer networks and flexible manufacturing systems are prime examples.

Beyond the specifics of the models, we have demonstrated an approach to strategic planning that is applicable to other healthcare systems. This approach involves examining the goals and needs of each system, and also recognizing the appropriate level a solution must fill: In an organization like the BCCA, the understanding of OR tools is high, the system is well-organized, and solutions must fit neatly onto specific problems with generally well-defined outcomes; in a situation like the DTES, few people understand quantitative tools, the system is very decentralized, and problems require a great deal of clarification before they can even be defined. Regardless, these and other OR tools are able to address strategic planning issues in healthcare under uncertainty.

Bibliography

[1] *Findings from the evaulation of Vancouver's pilot medically supervised safer injecting facility—Insite*. British Columbia Centre for Excellence in HIV/AIDS, June 2009.

[2] *Drug situation in Vancouver*. Urban Health Research Initiative of the British Columbia Centre for Excellence in HIV/AIDS, November 2009.

[3] *Downtown Community Court in Vancouver: Interim evaluation report*. Ministry of Attorney General, Justice Services Branch; Ministry of Public Safety and Solicitor General, Corrections Branch, August 2010.

[4] J Abate, G L Choudhury, and W Whitt. Calculating the M/G/1 busy-period density and LIFO waiting-time distribution by direct numerical transform inversion. *Operations Research Letters*, 18(3):113–119, 1995.

[5] J G Anderson. Preface: Special issue of simulation in health care management. *Health Care Management Science*, 5(2):73–73, 2002.

[6] J G Anderson and G G van Merode. Special issue on health care simulation. *Health Care Management Science*, 10(4):309–310, 2007.

[7] B Aouni and O Kettani. Goal programming model: A glorious history and a promising future. *European Journal of Operational Research*, 133(2): 225–231, 2001.

[8] Bank of Canada. Monthly and annual average exchange rates - Bank of Canada, June 2012. URL `http://www.bankofcanada.ca/rates/exchange/`.

[9] A M Bayoumi and G S Zaric. The cost-effectiveness of Vancouver's supervised injection facility. *Canadian Medical Association Journal*, 179(11): 1143–1151, 2008.

[10] I Berrada, J A Ferland, and P Michelon. A multi-objective approach to nurse scheduling with both hard and soft constraints. *Socio-Economic Planning Sciences*, 30(3):183–193, 1996.

[11] J Bhadury and Z Radovilsky. Job rotation using the multi-period assignment problem. *International Journal of Production Research*, 44(20): 4431–4444, 2006.

[12] U N Bhat. *An introduction to queueing theory*. Birkhäuser Boston, Boston, 2008.

[13] G R Bitran and S Dasu. A review of open queueing network models of manufacturing systems. *Queueing Systems*, 12(1):95–133, 1992.

[14] S Bordoloi and H Matsuo. Human resource planning in knowledge-intensive operations: A model for learning with stochastic turnover. *European Journal of Operational Research*, 130:169–189, 2001.

[15] S C Brailsford. System dynamics: What's in it for healthcare simulation modelers. *Winter Simulation Conference 2008*, pages 1478–1483, 2008.

[16] P Brethour. Exclusive demographic picture. *The Globe and Mail*, February 2009.

[17] J O Brunner, J F Bard, and Rainer Kolisch. Flexible shift scheduling of physicians. *Health Care Management Science*, 12(3):285–305, 2009.

[18] J Buxton. *Vancouver drug use epidemiology*. Canadian Community Epidemiology Network on Drug Use, June 2005.

[19] J Buxton, A Mehrabadi, E Preston, and A Tu. *Vancouver drug use epidemiology*. June 2007.

[20] J P Buzen. Computational algorithms for closed queueing networks with exponential servers. *Communications of the ACM*, 16(9):527–531, 1973.

[21] L Campbell, N Boyd, and L Culbert. *A thousand dreams: Vancouver's Downtown Eastside and the fight for its future*. Douglas & McIntyre, 2009.

[22] CanadianForex. Yearly average exchange rates - CanadianForex, June 2012. URL `http://www.canadianforex.ca/forex-tools/historical-rate-tools/yearly-average-rates`.

[23] J Carver. *Dufferin mental health and addictions planning project—backgrounder for the dufferin plan*. Central West LHIN Mental Health and Addictions Core Action Group, July 2008.

[24] G Casale. An efficient algorithm for the exact analysis of multiclass queueing networks with large population sizes. *SIGMetrics/Performance*, 34(1): 169–180, 2006.

[25] J P Caulkins, D Behrens, C Knoll, and G Tragler. Markov chain modeling of initiation and demand: The case of the US cocaine epidemic. *Health Care Management Science*, 7(4):319–329, 2004.

[26] J P Caulkins, P Dietze, and A Ritter. Dynamic compartmental model of trends in Australian drug use. *Health Care Management Science*, 10(2): 151–162, 2007.

[27] B Cheang, H Li, A Lim, and B Rodrigues. Nurse rostering problems–a bibliographic survey. *European Journal of Operational Research*, 151(3): 447–460, 2003.

[28] D Chisholm, A Healey, and M Knapp. QALYs and mental health care. *Social Psychiatry and Psychiatric Epidemiology*, 32(2):68–75, 1997.

[29] G L Choudhury, K K Leung, and W Whitt. Calculating normalization constants of closed queuing networks by numerically inverting their generating functions. *Journal of the Association for Computing Machinery*, 42(5):935–970, 1995.

[30] CIHI. *National health expenditure trends, 1975 to 2011.* Canadian Institute for Health Information, October 2011.

[31] R E Clark, G B Teague, S K Ricketts, P W Bush, H Xie, T G McGuire, R E Drake, G J McHugo, A M Keller, and M Zubkoff. Cost-effectiveness of assertive community treatment versus standard case management for persons with co-occurring severe mental illness and substance use disorders. *Health Services Research*, 33(5 Pt 1):1285–1308, 1998.

[32] A Cohn, S Root, C Kymissis, J Esses, and N Westmoreland. Scheduling medical residents at Boston University School of Medicine. *Interfaces*, 39(3): 186–195, 2009.

[33] M P Coleman, D Forman, H Bryant, J Butler, B Rachet, C Maringe, U Nur, *et al.* Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995–2007 (the International Cancer Benchmarking Partnership): An analysis of population-based cancer registry data. *The Lancet*, 377(9760): 127–138, 2011.

[34] Connecticut Department of Mental Health and Addiction Services. *Quality of life assessment pilot report.* June 2009.

[35] K J P Craib, P M Spittal, E Wood, N Laliberte, R S Hogg, K Li, K Heath, M W Tyndall, M V O'Shaughnessy, and M T Schechter. Risk factors for elevated HIV incidence among Aboriginal injection drug users in Vancouver. *Canadian Medical Association Journal*, 168(1):19–24, 2003.

[36] M Dawar. *DTES community health profile.* VCHA, April 2011.

[37] B P Dembling, D T Chen, and L Vachon. Life expectancy and causes of

death in a population treated for serious mental illness. *Psychiatric Services (Washington, D.C.)*, 50(8):1036–1042, 1999.

[38] B T Denton, A J Miller, H J Balasubramanian, and T R Huschka. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations Research*, 58(4-Part-1):802–816, 2010.

[39] M S Desai, M L Penn, Sally Brailsford, and M Chipulu. Modelling of Hampshire adult services—gearing up for future demands. *Health Care Management Science*, 11(2):167–176, 2008.

[40] S Earnshaw, K Hicks, A Richter, and A Honeycutt. A linear programming model for allocating HIV prevention funds with state agencies: A pilot study. *Health Care Management Science*, 10(3):239–252, 2007.

[41] D Edmonds and D McCready. Costing and pricing of police services. *International Journal of Public Sector Management*, 7(5):4–14, 1994.

[42] A Ernst, H Jiang, M Krishnamoorthy, and D Sier. Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research*, 143(1):3–27, 2004.

[43] J J Faraway. *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Texts in Statistical Science. Taylor & Francis, 2006.

[44] L S Franz and J L Miller. Scheduling medical residents to rotations: Solving the large-scale multiperiod staff assignment problem. *Operations Research*, 41(2):269–279, 1993.

[45] W J Gordon and G F Newell. Closed queuing systems with exponential servers. *Operations Research*, 15(2):254–265, 1967.

[46] M Gossop, J Marsden, D Stewart, and A Rolfe. Reductions in acquisitive crime and drug use after treatment of addiction problems: 1-year follow-up outcomes. *Drug and Alcohol Dependence*, 58(1-2):165–172, 2000.

[47] D Gross, J F Shortle, J M Thompson, and C M Harris. *Fundamentals of queueing theory*. John Wiley & Sons Inc, 2011.

[48] F S Hillier and G J Lieberman. *Introduction to operations research*. McGraw-Hill Higher Education, New York, 8th edition, 2005.

[49] J R Jackson. Jobshop-like queueing systems. *Management Science*, 10(1): 131–142, 1963.

[50] E Kaplan and M Johri. Treatment on demand: An operational model.

Health Care Management Science, 3(3):171–183, 2000.

[51] W D Kelton, R P Sadowski, and D T Sturrock. *Simulation with Arena*. McGraw-Hill Higher Education, 2009.

[52] D G Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics*, 24(3):338–354, 1953.

[53] T Kerr, E Wood, E Grafstein, T Ishida, K Shannon, C Lai, J S G Montaner, and M W Tyndall. High rates of primary care and emergency department use among injection drug users in Vancouver. *Journal of Public Health*, 27 (1):62–66, 2005.

[54] T Kerr, W Small, W Peeace, D Douglas, A Pierre, and E Wood. Harm reduction by a "user-run" organization: A case study of the Vancouver Area Network of Drug Users (VANDU). *International Journal of Drug Policy*, 17 (2):61–69, 2006.

[55] E Khandor and K Mason. *The street health report 2007: Community-based research for social change*. Street Health and Wellesley Institute, 2007.

[56] M Knapp and R Mangalore. "The trouble with QALYs...". *Epidemiologia e Psichiatria Sociale*, 16(4):289–293, 2007.

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

[58] G Koole and A Mandelbaum. Queueing models of call centers: An introduction. *Annals of Operations Research*, 113(1):41–59, 2002.

[59] L Kopala, G Smith, and A Malla. Resource utilization in a Canadian national study of people with schizophrenia and related psychotic disorders. *Acta Psychiatrica Scandinavica*, 113(s430):29–39, 2006.

[60] M Krausz. Expert interview, July 2012.

[61] M H Kutner, C J Nachtsheim, J Neter, and W Li. *Applied linear statistical models*. McGraw-Hill/Irwin, 2005.

[62] M Lagarde and J Cairns. Modelling human resources policies with Markov models: An illustration with the South African nursing labour market. *Health Care Management Science*, 15(3):270–282, 2012.

[63] S S Lam and Y L Lien. A tree convolution algorithm for the solution of queueing networks. *Communications of the ACM*, 26(3):203–215, 1983.

[64] M Lavieri and M Puterman. Optimizing nursing human resource planning in British Columbia. *Health Care Management Science*, 12(2):119–128, 2009.

[65] L L Li and B E King. A healthcare staff decision model considering the effects of staff cross-training. *Health Care Management Science*, 2(1):53–61, 1999.

[66] X Li, H Sun, A Puri, D C Marsh, and A H Anis. Medical withdrawal management in Vancouver: Service description and evaluation. *Addictive behaviors*, 32(5):1043–1053, 2007.

[67] X Lin, S L Janak, and C A Floudas. A new robust optimization approach for scheduling under uncertainty: I. bounded uncertainty. *Computers & Chemical Engineering*, 28(6-7):1069–1085, 2004.

[68] I Linden, M Mar, G Werker, K Jang, and M Krausz. Research on a vulnerable neighbourhood—The Vancouver Downtown Eastside from 2001 to 2011. Manuscript submitted for publication. August 2012.

[69] J D C Little. A proof for the queuing formula. *Operations Research*, 9: 383–387, 1961.

[70] H Ma. 2005/06 Downtown Eastside community monitoring report, July 2006. URL `http://vancouver.ca/commsvcs/planning/dtes/pdf/2006MR.pdf`.

[71] P McCullagh and J A Nelder. *Generalized linear models, second edition*. Monographs on Statistics and Applied Probability. Taylor & Francis, 1989.

[72] Bill McEwan. *Psychosis in the DTES light*. DTES Conference, 2007.

[73] J Medhi. *Stochastic models in queueing theory*. Elsevier Science, 2009.

[74] H Miller, W Pierskalla, and G Rath. Nurse scheduling using mathematical programming. *Operations Research*, 24(5):857–870, 1976.

[75] J L Miller and L S Franz. Binary-rounding heuristic for multi-period variable-task-duration assignment problems. *Computers & Operations Research*, 23(8):819–828, 1996.

[76] Ministry of Public Safety and Solicitor General. *A profile of B.C. corrections: Protect communities, reduce reoffending*. Ministry of Public Safety and Solicitor General, September 2010.

[77] D Mitra and J McKenna. Asymptotic expansions for closed Markovian networks with state-dependent service rates. *Journal of the Association for Computing Machinery*, 33(3):568–592, 1986.

[78] M Mondello, A B Gass, T McLaughlin, and N Shore. *Cost of homelessness: Cost analysis of permanent supportive housing*. Corporation for Supportive Housing, MaineHousing, Maine Department of Health and Human Services, September 2007.

[79] J Morecroft and S Robinson. Explaining puzzling dynamics: Comparing the use of system dynamics and discrete-event simulation. *Proceedings of the 23rd International Conference of the System Dynamics Society*, pages 17–21, 2005.

[80] I Musgrave. Expert interview—assertive community treatment, July 2012.

[81] B Nosyk, H Sun, S Sizto, D C Marsh, and A H Anis. *An evaluation of methadone maintenance treatment in British Columbia: 1996-2007*. Centre for Health Evaluation & Outcome Sciences and B.C. Ministry of Health, Institute of Health Living & Sport, 2009.

[82] J E Oakley and A O'Hagan. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):751–769, 2004.

[83] Z Padaiga, E Subata, and G Vanagas. Outpatient methadone maintenance treatment program. Quality of life and health of opioid-dependent persons in Lithuania. *Medicina (Kaunas, Lithuania)*, 43(3):235–241, 2007.

[84] A Palepu, M W Tyndall, H Leon, J Muller, M V O'Shaughnessy, M T Schechter, and A H Anis. Hospital utilization and costs in a cohort of injection drug users. *Canadian Medical Association Journal*, 165(4):415–420, 2001.

[85] M Patterson, J Somers, K McIntosh, A Shiell, C J Frankish, and G Van der Leer. *Housing and support for adults with severe addictions and/or mental illness in British Columbia*. Centre for Applied Research in Mental Health and Addiction, Simon Fraser University Burnaby, Canada, 2008.

[86] B Pourbohloul, M Rekart, and R Brunham. Impact of mass treatment on syphilis transmission: A mathematical modeling approach. *Sexually Transmitted Diseases*, 30(4):297–305, 2003.

[87] Province of British Columbia. *MSP fee-for-service payment analysis fee items*. September 2011.

[88] P Punnakitikashem, J M Rosenberger, and D Buckley Behan. Stochastic programming for nurse assignment. *Computational Optimization and Applications*, 40(3):321–349, 2007.

[89] A S Rajkumar and M T French. Drug abuse, crime costs, and the economic benefits of treatment. *Journal of Quantitative Criminology*, 13(3):291–323, 1997.

[90] K G Ramakrishnan and D Mitra. An overview of PANACEA, a software package for analyzing Markovian queueing networks. *Bell System Technical Journal*, 61:2849–2872, 1982.

[91] M Reiser. Mean-value analysis and convolution method for queue-dependent servers in closed queueing networks. *Performance Evaluation*, 1(1):7–18, 1981.

[92] M Reiser and S S Lavenberg. Mean-value analysis of closed multichain queuing networks. *Journal of the Association for Computing Machinery*, 27 (2):313–322, 1980.

[93] A Richter and B Loomis. Health and economic impacts of an HIV intervention in out of treatment substance abusers: Evidence from a dynamic model. *Health Care Management Science*, 8(1):67–79, 2005.

[94] E Rönnberg and T Larsson. Automating the self-scheduling process of nurses in Swedish healthcare: A pilot study. *Health Care Management Science*, 13 (1):35–53, 2010.

[95] K W Ross and J Wang. Asymptotically optimal importance sampling for product-form queuing networks. *ACM Transactions on Modeling and Computer Simulation*, 3(3):244–268, 1993.

[96] K W Ross and J Wang. Implementation of Monte Carlo integration for the analysis of product-form queueing networks. *Performance Evaluation*, 29(4): 273–292, 1997.

[97] F Sassi. Calculating QALYs, comparing QALY and DALY calculations. *Health Policy and Planning*, 21(5):402–408, 2006.

[98] M T Schechter, S A Strathdee, P G A Cornelisse, S Currie, D M Patrick, M L Rekart, and M V O'Shaughnessy. Do needle exchange programmes increase the spread of HIV among injection drug users?: An investigation of the Vancouver outbreak. *Aids*, 13(6):F45–F51, 1999.

[99] R G Schroeder. Resource planning in university management by goal programming. *Operations Research*, 22(4):700–710, 1974.

[100] C Schutz. Interview with Burnaby Centre management, July 2012.

[101] A Slowey. Interview with BCCA Vancouver Centre chief RT, March 2010.

[102] Statistics Canada. Consumer Price Index, historical summary - Statistics Canada, March 2012. URL `http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/econ46a-eng.htm`.

[103] Statistics Canada. Police resources in Canada: Table 8 — total expenditures on policing, 2010, August 2012. URL `http://www.statcan.gc.ca/pub/85-225-x/2011000/t009-eng.htm`.

[104] H A Taha. *Operations research: An introduction*. Prentice Hall, 2010.

[105] J tan de Bibiana. Vancouver At Home Study—Preliminary results. Technical report, March 2012.

[106] J tan de Bibiana, A Palepu, E Grafstein, C J Frankish, and A Kazanjian. *Housing first and emergency department use among homeless individuals with mental illness in Vancouver*. May 2012.

[107] S Thompson. *Policing Vancouver's mentally ill: The disturbing truth*. VPD, September 2011.

[108] G Tomblin Murphy, S Birch, A MacKenzie, R Alder, L Lethbridge, and L Little. Eliminating the shortage of registered nurses in Canada: An exercise in applied needs-based planning. *Health Policy*, 105(2-3):192–202, 2011.

[109] S Topaloglu. A multi-objective programming model for scheduling emergency medicine residents. *Computers & Industrial Engineering*, 51:375–388, 2006.

[110] S Topaloglu. A shift scheduling model for employees with different seniority levels and an application in healthcare. *European Journal of Operational Research*, 198(3):943–957, 2009.

[111] M W Tyndall, S Currie, P M Spittal, K Li, E Wood, M V O'Shaughnessy, and M T Schechter. Intensive injection cocaine use as the primary risk factor in the Vancouver HIV-1 epidemic. *Aids*, 17(6):887, 2003.

[112] VCH. *About the Burnaby Centre for Mental Health and Addiction*. Vancouver Coastal Health, June 2008.

[113] VCH. Integrated street response pilot project. Technical report, Vancouver Coastal Health and Vancouver Police Department Integrated Response Working Group, April 2009.

[114] VCH. Services in the Downtown Eastside, June 2012. URL `http://www.vch.ca/about_us/news/media_contacts/fact_sheets/services_in_the_downtown_eastside`.

[115] VCH. *MHA Vancouver roadmap*. Vancouver Coastal Health, January 2012.

[116] VPD. *Vancouver Police Department crime incident statistics*. Vancouver Police Department, January 2012.

[117] R Wall, J Rehm, B Fischer, B Brands, L Gliksman, J Stewart, W Medved, and J Blake. Social costs of untreated opioid dependence. *Journal of Urban Health*, 77(4):688–722, 2000.

[118] G Werker, A Sauré, J French, and S Shechter. The use of discrete-event simulation modelling to improve radiation therapy planning processes. *Radiotherapy and Oncology*, 92(1):76–82, 2009.

[119] W Whitt. Open and closed models for networks of queues. *AT&T Bell Lab Tech J*, 63(9):1911–1979, 1984.

[120] W Whitt. Engineering solution of a basic call-center model. *Management Science*, 51(2):221–235, 2005.

[121] F Wilson-Bates. *Lost in transition: How a lack of capacity in the mental health system is failing vancouver's mentally ill and draining police resources*. Vancouver Police Department, February 2008.

[122] E Wood, M W Tyndall, P M Spittal, K Li, T Kerr, R S Hogg, J S G Montaner, M V O'Shaughnessy, and M T Schechter. Unsafe injection practices in a cohort of injection drug users in Vancouver: Could safer injecting rooms help? *Canadian Medical Association Journal*, 165(4): 405–410, 2001.

[123] E Wood, J S G Montaner, K Chan, M W Tyndall, M T Schechter, D Bangsberg, M V O'Shaughnessy, and R S Hogg. Socioeconomic status, access to triple therapy, and survival from HIV-disease since 1996. *Aids*, 16 (15):2065–2072, 2002.

[124] E Wood, E Lloyd-Smith, K Li, S A Strathdee, W Small, M W Tyndall, J S G Montaner, and T Kerr. Frequent needle exchange use and HIV incidence in Vancouver, Canada. *The American Journal of Medicine*, 120(2): 172–179, 2007.

[125] G A Zarkin, L J Dunlap, and G Homsi. The substance abuse services cost analysis program (SASCAP): A new method for estimating drug treatment services costs. *Evaluation and Program Planning*, 27(1):35–43, 2004.

Appendix A

Alternate Formulation of Experience Variables

The formulation of the experience variables in the basic model uses binary variables z to denote when each RT i has the required level of experience in area k for area j by period t. Constraint (2.7) forces z to be zero when the experience is not sufficient. Constraint (2.8) only allows assignment if the experience is sufficient. However, this restriction can be represented with continuous variables instead of binary variables, which would be preferable were it not for the robustness constraint also requiring the binary form of z.

Let z' denote a continuous (positive) variable that can replace z:

$$z'_{i,k,t} = \text{amount of experience RT } i \text{ has in area } k \text{ by period } t \tag{A.1}$$

$$z'_{i,k,t} \geq 0 \qquad \forall i,k,t \tag{A.2}$$

The following constraints then replace (2.7) and (2.8):

$$z'_{i,k,t} = E_{st}(i,k) + \sum_{u=1}^{t} x_{i,k,u} \qquad \forall i,k,t \tag{A.3}$$

$$E_{req}(j,k) \cdot x_{i,j,t} \leq z'_{i,k,t} \qquad \forall i,j,k,t : k \xrightarrow{\text{Exp}} j \tag{A.4}$$

It is easy to see that (A.3) and (A.4) could actually be combined, and the variables z' removed entirely. As long as these variables weren't required elsewhere in the model, the entire experience requirement could be captured in a flow constraint represented thus:

$$E_{req}(j,k) \cdot x_{i,j,t} \leq E_{st}(i,k) + \sum_{u=1}^{t} x_{i,k,u} \qquad \forall i,j,k,t : k \xrightarrow{\text{Exp}} j \tag{A.5}$$

This alternative formulation leads to solution time improvements of about 15%.

Appendix B

Full BCCA Model

The full BCCA model uses the same variables as the basic model, but includes several additional ones as well. Additional violation variables are required: $v^{(0)}$ penalizes the system for exceeding the minimum staffing demand requirements, ; $v^{(1)}$ and $v^{(2)}$ are unchanged; $v^{(3)}$ tracks violations of the forced assignments; $v^{(4)}$ tracks violations of the requirement that each RT should work in the treatment area at least once per year.

The model includes all of the original constraints and some new ones as well. The original constraints that have been modified have been renumbered with an "a" or "b":

$$s.t. \sum_j x_{i,j,t} \le 1 \qquad \forall i,t \quad (2.2a)$$

$$\sum_i P(i) \cdot x_{i,j,t} = D_{min}(j,t) + v^{(0)}_{j,t} \qquad \forall j,t \quad (2.3a)$$

$$\sum_i P(i) \cdot x_{i,j,t} \le D_{max}(j,t) \qquad \forall j,t \quad (2.3b)$$

$$x_{i,j,t} - x_{i,j,t-1} \le y_{i,j,t} \qquad \forall i,j,t \quad (2.4)$$

$$\sum_{u=t}^{t+S_{min}(j)-1} x_{i,j,u} \ge S_{min}(j) \cdot y_{i,j,t} \qquad \forall i,j,t \quad (2.5)$$

$$\sum_{u=t}^{t+S_{max}(j)} x_{i,j,u} \le S_{max}(j) + v^{(1)}_{i,j,t} \qquad \forall i,j,t \quad (2.6)$$

$$E_{req}(j,k) \cdot z_{i,j,k,t} \le E_{st}(i,k) + P(i) \cdot \sum_{u=1}^{t} x_{i,k,u} \qquad \forall i,j,k,t : k \overset{\text{Exp}}{\rightarrow} j \quad (2.7a)$$

$$x_{i,j,t} \le z_{i,j,k,t} \qquad \forall i,j,k,t : k \overset{\text{Exp}}{\rightarrow} j \quad (2.8)$$

$$\sum_i P(i) \cdot z_{i,j,k,t} \ge f \cdot D_{min}(j,t) - v^{(2)}_{j,k,t} \qquad \forall j,k,t : k \overset{\text{Exp}}{\rightarrow} j \quad (2.9a)$$

$$x_{i,j,0} = 1 \qquad \forall i,j : i \text{ is initially in } j \quad (B.1)$$

$$x_{i,j,t} \ge 1 - v^{(3)}_{i,j,t} \qquad \forall i,j,t \in F(i,j,t) \quad (B.2)$$

$$x_{i,j,t} \le A(i,j,t) \qquad \forall i,j,t \quad (B.3)$$

$$\sum_{u=t}^{t+3} x_{i,\text{treat},u} \geq 1 - v_{i,t}^{(4)} \qquad \forall i, t \in \{1, 5, 9, \dots\} \quad \text{(B.4)}$$

$$x, \bar{x}, \underline{x}, y, z \in \{0, 1\}; \quad v^{(0)}, v^{(1)}, v^{(2)}, v^{(3)}, v^{(4)} \geq 0$$

The first objective function in the BCCA application balances five types of violations against each other:

$$\min\ C^{(0)} \sum_{j,t} v_{j,t}^{(0)} + C^{(1)} \sum_{i,j,t} v_{i,j,t}^{(1)} + C^{(2)} \sum_{j,k,t} v_{j,k,t}^{(2)} + C^{(3)} \sum_{i,j,t} v_{i,j,t}^{(3)} + C^{(4)} \sum_{i,t} v_{i,t}^{(4)} \qquad \text{(B.5)}$$

The second objective function, introduced in §3.1.2, allows the user to make minor changes and then find a similar solution. Because the previous solution was found using violation penalties, these penalties can now be ignored. However, violations of forced assignments, $v^{(3)}$, must be retained so that when the user adjusts the plan—primarily through forced assignments—these adjustments can be realized:

$$\min\ C^{(a)} \sum_{i,j,t} \left[x_{i,j,t} + X(i,j,t) - 2x_{i,j,t}X(i,j,t) \right] + C^{(b)} \sum_{i,j,t} v_{i,j,t}^{(3)} \qquad \text{(B.6)}$$

The additional constants in the full BCCA model are $P(i)$, a ratio between 0 and 1 to handle part-time staff, where a value of 1 represents a full-time RT; $D_{max}(j,t)$, the upper bound on staffing demand; and $A(i,j,t)$, a matrix with 0-1 entries representing all eligible assignments that is generated from several input tables in the Excel application. An alternate approach would be to restrict *each* constraint to the eligible assignments, however, we have chosen this simple notation for clarity in the formulation.

Appendix C

Converting to 2011 Costs

In order to compare apples to apples, all cost figures taken from the literature were converted to 2011 Canadian dollars.

Costs used were either presented in Canadian dollars (CAD) or United States dollars (USD), so those in the latter currency were converted to the former currency using historical yearly averages from the Bank of Canada (1997 - 2011) [8] and from CanadianForex (1992 - 1996) [22]. Bank of Canada only provides annual exchange rates as far back as 1997, however, the figures from CanadianForex match those of Bank of Canada during that time.

To convert costs to 2011 dollars we used the Canadian CPI for all items, from Statistics Canada [102]. Table C.1 shows the exchange rate and index by year along with the calculated index conversion factor.

Table C.1: Exchange rates and CPI used to adjust cost figures to 2011 CAD.

Year	USD-CAD Rate	CPI	CPI Factor
1992	1.209	84	1.427
1993	1.290	85.6	1.401
1994	1.366	85.7	1.399
1995	1.373	87.6	1.369
1996	1.364	88.9	1.349
1997	1.385	90.4	1.326
1998	1.483	91.3	1.313
1999	1.486	92.9	1.291
2000	1.485	95.4	1.257
2001	1.548	97.8	1.226
2002	1.570	100	1.199
2003	1.402	102.8	1.166
2004	1.302	104.7	1.145
2005	1.212	107	1.121
2006	1.134	109.1	1.099
2007	1.075	111.5	1.075
2008	1.066	114.1	1.051
2009	1.142	114.4	1.048
2010	1.030	116.5	1.029
2011	0.989	119.9	1.000

Appendix D

Excel Queueing Model

D.1 Solving the LP

We use Excel to formulate and solve the LP model of the queueing network because the problem is well within the size limitations of Excel's built-in solver. This solver, an add-in from Frontline Systems, has a limit of 200 decision variables; our model only has 20—one for each λ (16, including split stations), one for each z (3), and one for the L_0 variable.

Because we use macros to run sensitivity analysis, it is important that the solver be accessible from Visual Basic for Applications (VBA), the programming language of Excel macros. However, even though it is accessible, it ends up being fairly slow. Additionally, on a Mac the solver add-in runs as a separate application, adding additional overhead and making VBA calls more complex. We therefore developed a different way to solve the LP.

Most of the constraints in the LP are equality constraints, so these must be binding at any solution. We can therefore easily examine all possible extreme points by looking at all binding/non-binding permutations of the remaining inequality constraints. As discussed in §4.1.4, there are $J+2\bar{J}+Z+1$ variables and $J+3\bar{J}+Z+1$ constraints. Each extreme point must have at least as many binding constraints as variables: Specifically, there are $J+\bar{J}+Z+1$ equality constraints so (at least) $\bar{J}$ out of the total $2\bar{J}$ inequalities must be at equality in any extreme point solution. We simply evaluate all $\binom{2\bar{J}}{\bar{J}} = \binom{6}{3} = 20$ permutations to see which feasible one has the lowest objective function value.

It turns out to be much faster to examine all of these extreme points using matrix functions in a spreadsheet than to call the solver. To do so, we have 20 square matrices lined up vertically below the base case (and below each scenario) that cover all permutations of these inequality constraints. For each permutation the spreadsheet computes the objective function value and checks for feasibility. The feasible extreme point with the best objective function value is chosen as the

optimal solution.

Because this solution is found entirely using spreadsheet functions, it happens almost instantaneously for a model of this small size. It is in this way that we are able to average about 100 solutions per second when running the sensitivity analysis macro.

D.2 Other Excel Model Notes

The external arrival inputs are supposed to represent only external arrivals, i.e., only those that come from the Non-treatment Population but not those that come from other stations within the network. The ED per capita arrival rate we identified as a data input actually represents the aggregate arrival rate. We therefore modified the corresponding constraint to allow the model to use this rate directly. Rather than use $z_{ED}L_0 - \zeta_{ED} = 0$, we use $z_{ED}L_0 - \lambda_{ED} = 0$.

External arrival rates for Police and Other Entry are given in absolute terms (e.g., $\zeta_{\text{Police}} = 21.7$). We use these absolute arrival numbers in the base case, but calculate the per capita rates based on L_0. These per capita rates are used throughout the scenarios and also in the DES.

Figure D.1 is a screen shot of the base case as it is set up in Excel. The variables are shown across the top in row 2, and the constraints are shown one per row starting in row 6. Cost and QALY outcomes are calculated at the bottom. The values for the $\bar{\lambda}$ parameters are calculated based on station capacities at the far right.

Figure D.1: Screen shot of Excel queueing model showing the base case.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	**Base case**	Police	Crim/Just	ED	Other entry	Acute	Inpatient	Inpatient served	Inpatient balked	MMT	CM	CM served	CM balked	ACT	ACT served	ACT balked	Family practice	non-treat	Police raw	ED raw	Other raw				
2	vars	21.70	4.71	37.55	10.70	5.29	3.32	1.81	1.51	0.78	1.43	1.10	0.33	0.13	0.03	0.10	0.86	4126.09	21.70	35.29	10.70				
3							Balked/Served:		0.83				0.30			3.87									
4	obj. func.								1				1			1						1.93739388			
5																						LHS	=	RHS	
6	Police	-1																	1			0.00	=	0	
7	Crim/Just	**0.217**	-1																			0.00	=	0	
8	ED	**0.05**		-1	**0.11**															1		0.00	=	0	
9	Other entry				-1																1	0.00	=	0	
10	Acute			**0.141**		-1																0.00	=	0	
11	Inpatient		**0**	**0.05**	**0.11**	**0.05**	-1															0.00	=	0	
12	Inpatient total						1	-1	-1													0.00	=	0	
13	MMT				**0.04**	**0.05**		**0.05**		-1												0.00	=	0	
14	CM				**0.1**	**0.05**		**0.05**			-1											0.00	=	0	
15	CM total										1	-1	-1									0.00	=	0	
16	ACT				**0.01**			**0.01**						-1								0.00	=	0	
17	ACT total													1	-1	-1						0.00	=	0	
18	Family practice				**0.08**												-1					0.00	=	0	
19	***LoS***	0.108333	63	0.16667	0	12	0	89	0	338	0	1275	0	0	3464	0	1275	1				7500.00	=	7500	params
20	Police external																	0	-1			-21.70	=	-21.7	21.7
21	ED external			-1														0.0091				0.00	=	0	
22	Other external																	0			-1	-10.70	=	-10.7	10.7
23	Inpatient cap							1														1.81	<=	1.808988764	162
24	CM cap											1										1.10	<=	1.097254902	1400
25	ACT cap														1							0.03	<=	0.025692841	90
26																									
27	Queue																								
28	mean queue length	2.350833	296.6607	6.25791	0	63.53027	0	161	0	264.70771	0	1399	0	0	89	0	1091.4	4126.0926							
29																		55.0%							
30	**Costs**		0.775															0.33							
31	cost/patient/day		$ 54			$ 647		$ 289		$ 17		$ 22			$ 46		$ 2	$ -	daily per capita cost:			$ 75.93			
32	$ crime/patient/day	$ -	$ 20	$ -						$ 25		$ 25			$ 25		$ 25	$ 77							
33	cost coeficient	$ 642	$ 4,643	$ 387	$-	$ 7,764		$ 25,721		$ 14,335		$ 60,448			$ 247,364		$ 34,948	$ 77							
34	daily cost	$ 13,931	$ 21,862	$14,531	$-	$ 41,104	$ -	$ 46,529	$ -	$ 11,226	$-	$ 66,327	$ -	$ -	$ 6,355	$ -	$ 29,915	$ 317,709	total daily cost:			$ 569,490			
35	% of total	2.45%	3.84%	2.55%		7.22%		8.17%		1.97%		11.65%			1.12%		5.25%	55.79%							
36																			total annual cost:			$ 207,863,786.41			
37																									
38																									
39																									
40	**Health**																								
41	QoL	0.526	0.584	0.526		0.702		0.702		0.693		0.65			0.66		0.65	0.584	Exp. life years (pop):			29.4			
42																			QALYs per capita:			18.06			
43																									
44																			total QALYs:			135,415.78			
45																									

Appendix E

Simulation with Alternate Inputs

In order to further validate the quality of our queueing network approximation approach we compare the queueing model (with FPM and capacitated station approximations) to the DES model (with Markovian arrivals and service times) using a very different set of input parameters. These parameters are not chosen to be realistic. We maintain the same population size, and leave the same three stations capacitated. Beyond these similarities, each arrival rate, routing probability, length of stay, and station capacity is either doubled or halved (based on a virtual coin flip, maintaining feasible routing probabilities if any sums were to exceed 1.0).

Table E.1: Simulation—comparison of capacitated queueing model with the capacitated simulation model alternate other input parameters.

Station	Queue Other	DES Other	DES Other CI	% Dif.
Police	2.4	2.3	(2.3, 2.4)	−0.1%
CrimJst	150.4	150.0	(149.1, 150.9)	−0.2%
ED	8.5	8.6	(8.6, 8.6)	+0.5%
Acute	171.9	172.6	(172.4, 172.8)	+0.4%
Inpatnt	80.0	80.8	(80.8, 80.8)	+1.0%
MMT	586.8	590.7	(589.1, 592.3)	+0.7%
CM	699.0	699.9	(699.9, 699.9)	+0.1%
ACT	44.0	44.7	(44.7, 44.7)	+1.6%
Family	136.4	135.9	(134.3, 137.5)	−0.4%
Non-tr	5620.7	5614.5	(5612.5, 5616.5)	−0.1%

The results of this comparison are shown in Table E.1. The differences are negligible, and even though some are statistically significant, they are all practically insignificant for our purposes. For example, the difference between the two approaches for MMT is only four clients, a number that is smaller than the presumed error in related inputs.

Table E.2 shows the original and alternate input parameters for this comparison.

Table E.2: Alternate instance of input parameters for queueing model and DES.

Parameter	Units	Base Value		Alt. Value
External Arrival Rate				
Police	/day	21.7	× ½ =	10.85
ED	/person/day	0.0091	× 2 =	0.0182
Other	/day	10.7	× ½ =	5.35
Routing Probability				
Police→CrimJst		0.22	× ½ =	0.11
Police→ED		0.05	× ½ =	0.025
ED→Acute		0.14	× 2 =	0.28
ED→Inpatnt		0.05	× 2 =	0.1
Other→ED		0.11	× ½ =	0.055
Other→Inpatnt		0.11	× ½ =	0.055
Other→MMT		0.04	× 2 =	0.08
Other→CM		0.1	× ½ =	0.05
Other→ACT		0.01	× 2 =	0.02
Other→Family		0.08	× ½ =	0.04
Acute→Inpatnt		0.05	× ½ =	0.025
Acute→MMT		0.05	× 2 =	0.1
Acute→CM		0.05	× 2 =	0.1
Inpatnt→MMT		0.05	× 2 =	0.1
Inpatnt→CM		0.05	× ½ =	0.025
Inpatnt→ACT		0.01	× ½ =	0.005
Length of Stay				
Police	hours	2.6	× 2 =	5.2
CrimJst	days	63	× 2 =	126
ED	hours	4	× ½ =	2
Other	-	-		-
Acute	days	12	× ½ =	6
Inpatnt	days	89	× ½ =	44.5
MMT	days	338	× ½ =	169
CM	days	1275	× 2 =	2550
ACT	days	3464	× ½ =	1732
Family	days	1275	× ½ =	637.5
Population				
CCD population	people	7500		7500
Station Capacity				
Inpatnt	beds	162	× ½ =	81
CM	people	1400	× ½ =	700
ACT	people	90	× ½ =	45

Appendix F

Scenario Analysis—Extras

This appendix provides additional data on the scenario analyses performed in Chapter 5. Table F.1 summarizes all of the scenarios analysis results.

Table F.1: Scenario analysis—costs, QALYs, population and station counts, and balking rates. B0 is the base case; S1–S8 are the scenarios; CA–CD are the combination scenarios.

ID	Cost	QALY	Non	Crim Jst	Inp	mmt	cm	act	Fam	Inp λ_b/λ_c	cm λ_b/λ_c	act λ_b/λ_c
B0	$75.93	18.1	4126	297	161	265	1399	89	1091	0.8	0.3	3.9
S1	$72.20	18.2	3704	266	161	346	1399	89	1470	1.5	0.7	5.3
S2	$72.14	18.2	3695	266	161	360	1399	89	1466	0.6	0.8	5.6
S3	$79.40	18.1	4019	289	288	283	1399	89	1063	-	0.3	4.3
S4	$74.67	18.1	3798	273	161	246	1682	269	1005	0.7	-	0.5
S5	$75.23	18.1	4054	291	161	295	1399	89	1140	0.9	0.4	5.9
S6	$72.94	18.0	4180	301	161	225	1399	89	1106	0.3	0.2	3.9
S7	$65.25	18.1	4169	239	161	267	1399	89	1103	0.9	0.3	3.9
S8	$77.25	18.2	3611	260	258	254	1674	425	955	-	-	-
CA	$70.18	18.3	3208	231	161	304	1999	269	1273	1.1	0.1	0.8
CB	$67.79	18.3	3203	230	161	455	1399	89	1906	1.1	1.2	7.3
CC	$77.38	18.2	3690	265	264	260	1711	269	976	-	-	0.6
CD	$55.42	18.4	2844	163	161	344	1999	269	1693	0.3	0.3	1.5
CD	$56.70	18.5	2749	158	215	382	1999	269	1702	-	0.3	2.2

We performed sensitivity analysis on all of the scenarios. Tables F.2 to F.14 summarize the results of these analyses. In each table the "Base Range" shows the values used in the base case and in the scenarios. When a particular scenario adjusts an input, the "Adjusted Range" is also displayed (otherwise "—" is shown).

Table F.2: Sensitivity analysis of Scenario 1—Inreased referrals.

Cell	Base Range	Adjusted Range	Costs	QALYs
Population	4000 – 9000	—	\$61.65 – \$74.56	18.8 – 18.0
Arrivals Police	20 – 30	—	\$72.11 – \$72.64	18.2 – 18.2
Arr. /capita ED	0.005 – 0.015	—	\$69.77 – \$75.53	18.2 – 18.2
Arrivals Other	5 – 15	—	\$79.04 – \$67.61	18.0 – 18.3
Cap Inpatnt	90 – 190	—	\$70.33 – \$72.92	18.1 – 18.2
Cap CM	1100 – 2000	—	\$73.75 – \$69.52	18.1 – 18.3
Cap ACT	60 – 100	—	\$72.26 – \$72.18	18.2 – 18.2
LoS Police	0.1 – 0.3	—	\$72.20 – \$72.15	18.2 – 18.2
LoS CrimJst	50 – 80	—	\$72.30 – \$72.06	18.2 – 18.2
LoS ED	0.1 – 0.5	—	\$72.22 – \$72.07	18.2 – 18.2
LoS Acute	8 – 16	—	\$70.77 – \$73.61	18.2 – 18.2
LoS Inpatnt	60 – 120	—	\$72.11 – \$72.24	18.2 – 18.2
LoS MMT	180 – 720	—	\$73.09 – \$70.18	18.1 – 18.3
LoS CM	365 – 2190	—	\$76.06 – \$72.20	18.0 – 18.2
LoS ACT	2000 – 5000	—	\$72.20 – \$72.20	18.2 – 18.2
LoS Family	365 – 2190	—	\$80.39 – \$64.87	17.9 – 18.4
Cost coef Police	400 – 800	—	\$71.57 – \$72.61	18.2 – 18.2
Cost CrimJst	50 – 125	—	\$72.05 – \$74.72	18.2 – 18.2
Cost coef ED	300 – 600	—	\$71.81 – \$73.15	18.2 – 18.2
Cost Acute	500 – 1000	—	\$71.08 – \$74.88	18.2 – 18.2
Cost Inpatnt	200 – 500	—	\$70.29 – \$76.73	18.2 – 18.2
Cost MMT	8 – 18	—	\$71.78 – \$72.24	18.2 – 18.2
Cost CM	18 – 28	—	\$71.45 – \$73.32	18.2 – 18.2
Cost ACT	30 – 60	—	\$72.01 – \$72.36	18.2 – 18.2
Cost Family	0.5 – 10	—	\$71.90 – \$73.76	18.2 – 18.2
Cost Non-tr	0 – 50	—	\$72.20 – \$96.89	18.2 – 18.2
Cost Non-tr crime	50 – 90	—	\$54.69 – \$80.63	18.2 – 18.2
Police→CrimJst	0.2 – 0.4	—	\$72.24 – \$71.78	18.2 – 18.2
Police→ED	0.01 – 0.1	—	\$72.20 – \$72.20	18.2 – 18.2
CrimJst→Inpatnt	0 – 0.1	—	\$72.20 – \$72.20	18.2 – 18.2
ED→Acute	0.1 – 0.2	—	\$71.13 – \$73.68	18.2 – 18.2
ED→Inpatnt	0.01 – 0.1	0.015 – 0.15	\$72.20 – \$72.20	18.2 – 18.2
Other→ED	0.05 – 0.15	—	\$72.20 – \$72.20	18.2 – 18.2
Other→Inpatnt	0.05 – 0.15	0.075 – 0.225	\$72.20 – \$72.20	18.2 – 18.2
Other→MMT	0.01 – 0.1	0.015 – 0.15	\$72.99 – \$70.66	18.1 – 18.3
Other→CM	0.05 – 0.2	0.075 – 0.3	\$72.63 – \$72.20	18.2 – 18.2
Other→ACT	0.005 – 0.02	0.0075 – 0.03	\$72.20 – \$72.20	18.2 – 18.2
Other→Family	0.05 – 0.13	0.075 – 0.195	\$76.38 – \$65.77	18.1 – 18.4
Acute→Inpatnt	0.01 – 0.1	0.015 – 0.15	\$72.20 – \$72.20	18.2 – 18.2
Acute→MMT	0.01 – 0.1	0.015 – 0.15	\$72.72 – \$71.57	18.1 – 18.2
Acute→CM	0.01 – 0.1	0.015 – 0.15	\$72.20 – \$72.20	18.2 – 18.2
Inpatnt→MMT	0.01 – 0.1	—	\$72.34 – \$72.02	18.2 – 18.2
Inpatnt→CM	0.01 – 0.1	—	\$72.20 – \$72.20	18.2 – 18.2
Inpatnt→ACT	0.005 – 0.02	—	\$72.20 – \$72.20	18.2 – 18.2
Crime treat ratio	0.2 – 1	—	\$67.51 – \$96.34	18.2 – 18.2

Table F.3: Sensitivity analysis of Scenario 2—Lower turnover.

Cell	Base Range	Adjusted Range	Costs	QALYs
Population	4000 – 9000	—	\$58.79 – \$74.51	18.8 – 18.1
Arrivals Police	20 – 30	—	\$72.05 – \$72.57	18.2 – 18.2
Arr. /capita ED	0.005 – 0.015	—	\$69.71 – \$75.46	18.2 – 18.2
Arrivals Other	5 – 15	—	\$78.96 – \$67.56	18.0 – 18.3
Cap Inpatnt	90 – 190	—	\$70.30 – \$72.85	18.2 – 18.2
Cap CM	1100 – 2000	—	\$73.69 – \$69.46	18.1 – 18.3
Cap ACT	60 – 100	—	\$72.19 – \$72.12	18.2 – 18.2
LoS Police	0.1 – 0.3	—	\$72.14 – \$72.09	18.2 – 18.2
LoS CrimJst	50 – 80	—	\$72.24 – \$72.00	18.2 – 18.2
LoS ED	0.1 – 0.5	—	\$72.16 – \$72.01	18.2 – 18.2
LoS Acute	8 – 16	—	\$70.71 – \$73.54	18.2 – 18.2
LoS Inpatnt	60 – 120	—	\$72.02 – \$72.20	18.2 – 18.2
LoS MMT	180 – 720	270 – 1080	\$73.05 – \$70.06	18.1 – 18.3
LoS CM	365 – 2190	547.5 – 3285	\$75.95 – \$72.14	18.0 – 18.2
LoS ACT	2000 – 5000	3000 – 7500	\$72.14 – \$72.14	18.2 – 18.2
LoS Family	365 – 2190	547.5 – 3285	\$80.31 – \$64.83	18.0 – 18.4
Cost coef Police	400 – 800	—	\$71.51 – \$72.55	18.2 – 18.2
Cost CrimJst	50 – 125	—	\$71.99 – \$74.65	18.2 – 18.2
Cost coef ED	300 – 600	—	\$71.75 – \$73.09	18.2 – 18.2
Cost Acute	500 – 1000	—	\$71.02 – \$74.81	18.2 – 18.2
Cost Inpatnt	200 – 500	—	\$70.23 – \$76.67	18.2 – 18.2
Cost MMT	8 – 18	—	\$71.70 – \$72.18	18.2 – 18.2
Cost CM	18 – 28	—	\$71.39 – \$73.26	18.2 – 18.2
Cost ACT	30 – 60	—	\$71.95 – \$72.30	18.2 – 18.2
Cost Family	0.5 – 10	—	\$71.84 – \$73.70	18.2 – 18.2
Cost Non-tr	0 – 50	—	\$72.14 – \$96.77	18.2 – 18.2
Cost Non-tr crime	50 – 90	—	\$54.65 – \$80.55	18.2 – 18.2
Police→CrimJst	0.2 – 0.4	—	\$72.17 – \$71.72	18.2 – 18.2
Police→ED	0.01 – 0.1	—	\$72.14 – \$72.14	18.2 – 18.2
CrimJst→Inpatnt	0 – 0.1	—	\$72.14 – \$72.14	18.2 – 18.2
ED→Acute	0.1 – 0.2	—	\$71.08 – \$73.62	18.2 – 18.2
ED→Inpatnt	0.01 – 0.1	—	\$71.70 – \$72.14	18.2 – 18.2
Other→ED	0.05 – 0.15	—	\$72.14 – \$72.14	18.2 – 18.2
Other→Inpatnt	0.05 – 0.15	—	\$72.14 – \$72.14	18.2 – 18.2
Other→MMT	0.01 – 0.1	—	\$72.92 – \$70.60	18.1 – 18.3
Other→CM	0.05 – 0.2	—	\$72.57 – \$72.14	18.2 – 18.2
Other→ACT	0.005 – 0.02	—	\$72.14 – \$72.14	18.2 – 18.2
Other→Family	0.05 – 0.13	—	\$76.31 – \$65.73	18.1 – 18.4
Acute→Inpatnt	0.01 – 0.1	—	\$72.14 – \$72.14	18.2 – 18.2
Acute→MMT	0.01 – 0.1	—	\$72.66 – \$71.51	18.2 – 18.2
Acute→CM	0.01 – 0.1	—	\$72.14 – \$72.14	18.2 – 18.2
Inpatnt→MMT	0.01 – 0.1	—	\$72.33 – \$71.90	18.2 – 18.2
Inpatnt→CM	0.01 – 0.1	—	\$72.14 – \$72.14	18.2 – 18.2
Inpatnt→ACT	0.005 – 0.02	—	\$72.14 – \$72.14	18.2 – 18.2
Crime treat ratio	0.2 – 1	—	\$67.44 – \$96.35	18.2 – 18.2

Table F.4: Sensitivity analysis of Scenario 3—Expanded Inpatnt capacity.

Cell	Base Range	Adjusted Range	Costs	QALYs
Population	4000 – 9000	—	\$64.39 – \$81.59	18.7 – 18.0
Arrivals Police	20 – 30	—	\$79.33 – \$79.74	18.1 – 18.1
Arr. /capita ED	0.005 – 0.015	—	\$74.50 – \$84.07	18.0 – 18.2
Arrivals Other	5 – 15	—	\$85.05 – \$76.12	17.9 – 18.2
Cap Inpatnt	90 – 190	180 – 380	\$76.55 – \$79.36	18.1 – 18.1
Cap CM	1100 – 2000	—	\$81.36 – \$76.72	18.0 – 18.2
Cap ACT	60 – 100	—	\$79.50 – \$79.37	18.1 – 18.1
LoS Police	0.1 – 0.3	—	\$79.40 – \$79.35	18.1 – 18.1
LoS CrimJst	50 – 80	—	\$79.58 – \$79.16	18.1 – 18.1
LoS ED	0.1 – 0.5	—	\$79.43 – \$79.24	18.1 – 18.1
LoS Acute	8 – 16	—	\$77.88 – \$80.91	18.1 – 18.1
LoS Inpatnt	60 – 120	—	\$76.82 – \$80.45	18.1 – 18.1
LoS MMT	180 – 720	—	\$80.35 – \$77.20	18.1 – 18.2
LoS CM	365 – 2190	—	\$84.91 – \$79.40	17.9 – 18.1
LoS ACT	2000 – 5000	—	\$79.40 – \$79.40	18.1 – 18.1
LoS Family	365 – 2190	—	\$86.36 – \$72.31	17.9 – 18.3
Cost coef Police	400 – 800	—	\$78.72 – \$79.85	18.1 – 18.1
Cost CrimJst	50 – 125	—	\$79.25 – \$82.14	18.1 – 18.1
Cost coef ED	300 – 600	—	\$78.98 – \$80.44	18.1 – 18.1
Cost Acute	500 – 1000	—	\$78.19 – \$82.31	18.1 – 18.1
Cost Inpatnt	200 – 500	—	\$75.99 – \$87.50	18.1 – 18.1
Cost MMT	8 – 18	—	\$79.06 – \$79.44	18.1 – 18.1
Cost CM	18 – 28	—	\$78.65 – \$80.52	18.1 – 18.1
Cost ACT	30 – 60	—	\$79.21 – \$79.57	18.1 – 18.1
Cost Family	0.5 – 10	—	\$79.19 – \$80.53	18.1 – 18.1
Cost Non-tr	0 – 50	—	\$79.40 – \$106.20	18.1 – 18.1
Cost Non-tr crime	50 – 90	—	\$61.30 – \$88.12	18.1 – 18.1
Police→CrimJst	0.2 – 0.4	—	\$79.47 – \$78.65	18.1 – 18.1
Police→ED	0.01 – 0.1	—	\$79.40 – \$79.40	18.1 – 18.1
CrimJst→Inpatnt	0 – 0.1	—	\$79.40 – \$80.37	18.1 – 18.1
ED→Acute	0.1 – 0.2	—	\$78.08 – \$81.26	18.1 – 18.1
ED→Inpatnt	0.01 – 0.1	—	\$75.95 – \$80.37	18.1 – 18.1
Other→ED	0.05 – 0.15	—	\$79.40 – \$79.40	18.1 – 18.1
Other→Inpatnt	0.05 – 0.15	—	\$77.95 – \$80.35	18.1 – 18.1
Other→MMT	0.01 – 0.1	—	\$80.16 – \$77.87	18.1 – 18.2
Other→CM	0.05 – 0.2	—	\$80.69 – \$79.40	18.1 – 18.1
Other→ACT	0.005 – 0.02	—	\$79.40 – \$79.40	18.1 – 18.1
Other→Family	0.05 – 0.13	—	\$83.07 – \$73.23	18.0 – 18.3
Acute→Inpatnt	0.01 – 0.1	—	\$78.93 – \$79.99	18.1 – 18.1
Acute→MMT	0.01 – 0.1	—	\$79.91 – \$78.78	18.1 – 18.1
Acute→CM	0.01 – 0.1	—	\$79.40 – \$79.40	18.1 – 18.1
Inpatnt→MMT	0.01 – 0.1	—	\$79.69 – \$79.04	18.1 – 18.1
Inpatnt→CM	0.01 – 0.1	—	\$79.40 – \$79.40	18.1 – 18.1
Inpatnt→ACT	0.005 – 0.02	—	\$79.40 – \$79.40	18.1 – 18.1
Crime treat ratio	0.2 – 1	—	\$75.32 – \$100.43	18.1 – 18.1

Table F.5: Sensitivity analysis of Scenario 4—Expanded long term capacity.

Cell	Base Range	Adjusted Range	Costs	QALYs
Population	4000 – 9000	—	\$64.21 – \$76.39	18.8 – 18.0
Arrivals Police	20 – 30	—	\$74.58 – \$75.14	18.2 – 18.1
Arr. /capita ED	0.005 – 0.015	—	\$72.50 – \$77.54	18.1 – 18.2
Arrivals Other	5 – 15	—	\$82.25 – \$70.04	17.9 – 18.3
Cap Inpatnt	90 – 190	—	\$72.95 – \$75.34	18.1 – 18.2
Cap CM	1100 – 2000	2000 – 2000	\$75.23 – \$73.88	18.1 – 18.2
Cap ACT	60 – 100	270 – 270	\$74.73 – \$74.65	18.1 – 18.2
LoS Police	0.1 – 0.3	—	\$74.68 – \$74.63	18.1 – 18.1
LoS CrimJst	50 – 80	—	\$74.77 – \$74.55	18.2 – 18.1
LoS ED	0.1 – 0.5	—	\$74.70 – \$74.54	18.2 – 18.1
LoS Acute	8 – 16	—	\$73.21 – \$76.12	18.1 – 18.2
LoS Inpatnt	60 – 120	—	\$74.42 – \$74.81	18.2 – 18.1
LoS MMT	180 – 720	—	\$75.35 – \$73.09	18.1 – 18.3
LoS CM	365 – 2190	—	\$80.47 – \$73.42	17.9 – 18.2
LoS ACT	2000 – 5000	—	\$74.67 – \$74.67	18.1 – 18.1
LoS Family	365 – 2190	—	\$80.25 – \$68.98	18.0 – 18.3
Cost coef Police	400 – 800	—	\$74.03 – \$75.09	18.1 – 18.1
Cost CrimJst	50 – 125	—	\$74.53 – \$77.26	18.1 – 18.1
Cost coef ED	300 – 600	—	\$74.27 – \$75.66	18.1 – 18.1
Cost Acute	500 – 1000	—	\$73.53 – \$77.43	18.1 – 18.1
Cost Inpatnt	200 – 500	—	\$72.76 – \$79.20	18.1 – 18.1
Cost MMT	8 – 18	—	\$74.38 – \$74.71	18.1 – 18.1
Cost CM	18 – 28	—	\$73.78 – \$76.02	18.1 – 18.1
Cost ACT	30 – 60	—	\$74.10 – \$75.18	18.1 – 18.1
Cost Family	0.5 – 10	—	\$74.47 – \$75.75	18.1 – 18.1
Cost Non-tr	0 – 50	—	\$74.67 – \$99.99	18.1 – 18.1
Cost Non-tr crime	50 – 90	—	\$56.95 – \$83.21	18.1 – 18.1
Police→CrimJst	0.2 – 0.4	—	\$74.71 – \$74.29	18.2 – 18.1
Police→ED	0.01 – 0.1	—	\$74.67 – \$74.67	18.1 – 18.1
CrimJst→Inpatnt	0 – 0.1	—	\$74.67 – \$74.67	18.1 – 18.1
ED→Acute	0.1 – 0.2	—	\$73.80 – \$75.86	18.1 – 18.2
ED→Inpatnt	0.01 – 0.1	—	\$74.37 – \$74.67	18.1 – 18.1
Other→ED	0.05 – 0.15	—	\$74.67 – \$74.67	18.1 – 18.1
Other→Inpatnt	0.05 – 0.15	—	\$74.67 – \$74.67	18.1 – 18.1
Other→MMT	0.01 – 0.1	—	\$75.26 – \$73.50	18.1 – 18.2
Other→CM	0.05 – 0.2	—	\$77.21 – \$73.42	18.0 – 18.2
Other→ACT	0.005 – 0.02	—	\$74.67 – \$74.67	18.1 – 18.1
Other→Family	0.05 – 0.13	—	\$77.61 – \$69.72	18.1 – 18.3
Acute→Inpatnt	0.01 – 0.1	—	\$74.67 – \$74.67	18.1 – 18.1
Acute→MMT	0.01 – 0.1	—	\$75.06 – \$74.20	18.1 – 18.2
Acute→CM	0.01 – 0.1	—	\$75.46 – \$73.76	18.1 – 18.2
Inpatnt→MMT	0.01 – 0.1	—	\$74.81 – \$74.50	18.1 – 18.2
Inpatnt→CM	0.01 – 0.1	—	\$74.96 – \$74.32	18.1 – 18.2
Inpatnt→ACT	0.005 – 0.02	—	\$74.67 – \$74.67	18.1 – 18.1
Crime treat ratio	0.2 – 1	—	\$70.12 – \$98.15	18.1 – 18.1

Table F.6: Sensitivity analysis of Scenario 5—Urgent Response Centre.

Cell	Base Range	Adjusted Range	Costs	QALYs
Population	4000 – 9000	—	\$64.02 – \$77.33	18.7 – 17.9
Arrivals Police	20 – 30	—	\$75.18 – \$75.43	18.1 – 18.1
Arr. /capita ED	0.005 – 0.015	0.0049 – 0.0149	\$72.50 – \$78.99	18.1 – 18.1
Arrivals Other	5 – 15	—	\$81.13 – \$71.40	17.9 – 18.2
Cap Inpatnt	90 – 190	—	\$73.38 – \$75.94	18.0 – 18.1
Cap CM	1100 – 2000	—	\$76.84 – \$72.86	18.0 – 18.2
Cap ACT	60 – 100	—	\$75.29 – \$75.20	18.1 – 18.1
LoS Police	0.1 – 0.3	—	\$75.23 – \$75.18	18.1 – 18.1
LoS CrimJst	50 – 80	—	\$75.34 – \$75.08	18.1 – 18.1
LoS ED	0.1 – 0.5	—	\$75.25 – \$75.08	18.1 – 18.1
LoS Acute	8 – 16	—	\$73.69 – \$76.75	18.1 – 18.1
LoS Inpatnt	60 – 120	—	\$75.14 – \$75.27	18.1 – 18.1
LoS MMT	180 – 720	—	\$76.05 – \$73.31	18.0 – 18.2
LoS CM	365 – 2190	—	\$79.74 – \$75.23	17.9 – 18.1
LoS ACT	2000 – 5000	—	\$75.23 – \$75.23	18.1 – 18.1
LoS Family	365 – 2190	—	\$81.78 – \$68.74	17.9 – 18.3
Cost coef Police	400 – 800	—	\$74.54 – \$75.67	18.1 – 18.1
Cost CrimJst	50 – 125	—	\$75.07 – \$77.98	18.1 – 18.1
Cost coef ED	300 – 600	—	\$74.80 – \$76.26	18.1 – 18.1
Cost Acute	500 – 1000	—	\$74.02 – \$78.12	18.1 – 18.1
Cost Inpatnt	200 – 500	—	\$73.31 – \$79.75	18.1 – 18.1
Cost MMT	8 – 18	—	\$74.87 – \$75.26	18.1 – 18.1
Cost CM	18 – 28	—	\$74.48 – \$76.34	18.1 – 18.1
Cost ACT	30 – 60	—	\$75.04 – \$75.39	18.1 – 18.1
Cost Family	0.5 – 10	—	\$75.00 – \$76.44	18.1 – 18.1
Cost Non-tr	0 – 50	—	\$75.23 – \$102.25	18.1 – 18.1
Cost Non-tr crime	50 – 90	—	\$56.89 – \$84.05	18.1 – 18.1
Police→CrimJst	0.2 – 0.4	—	\$75.27 – \$74.76	18.1 – 18.1
Police→ED	0.01 – 0.1	0.005 – 0.05	\$75.79 – \$74.54	18.1 – 18.1
CrimJst→Inpatnt	0 – 0.1	—	\$75.23 – \$75.23	18.1 – 18.1
ED→Acute	0.1 – 0.2	—	\$74.03 – \$76.90	18.1 – 18.1
ED→Inpatnt	0.01 – 0.1	—	\$75.23 – \$75.23	18.1 – 18.1
Other→ED	0.05 – 0.15	—	\$75.23 – \$75.23	18.1 – 18.1
Other→Inpatnt	0.05 – 0.15	—	\$75.23 – \$75.23	18.1 – 18.1
Other→MMT	0.01 – 0.1	—	\$75.87 – \$73.93	18.0 – 18.2
Other→CM	0.05 – 0.2	—	\$76.17 – \$75.23	18.0 – 18.1
Other→ACT	0.005 – 0.02	—	\$75.23 – \$75.23	18.1 – 18.1
Other→Family	0.05 – 0.13	—	\$78.47 – \$69.81	18.0 – 18.2
Acute→Inpatnt	0.01 – 0.1	—	\$75.23 – \$75.23	18.1 – 18.1
Acute→MMT	0.01 – 0.1	—	\$75.66 – \$74.71	18.1 – 18.1
Acute→CM	0.01 – 0.1	—	\$75.23 – \$75.23	18.1 – 18.1
Inpatnt→MMT	0.01 – 0.1	—	\$75.37 – \$75.04	18.1 – 18.1
Inpatnt→CM	0.01 – 0.1	—	\$75.23 – \$75.23	18.1 – 18.1
Inpatnt→ACT	0.005 – 0.02	—	\$75.23 – \$75.23	18.1 – 18.1
Crime treat ratio	0.2 – 1	—	\$71.02 – \$96.89	18.1 – 18.1

Table F.7: Sensitivity analysis of Scenario 6—Decreased ED use.

Cell	Base Range	Adjusted Range	Costs	QALYs
Population	4000 – 9000	—	\$62.97 – \$74.54	18.7 – 17.9
Arrivals Police	20 – 30	—	\$72.81 – \$73.55	18.0 – 18.0
Arr. /capita ED	0.005 – 0.015	0.0028 – 0.0078	\$71.56 – \$74.87	18.0 – 18.1
Arrivals Other	5 – 15	—	\$79.44 – \$69.41	17.8 – 18.2
Cap Inpatnt	90 – 190	—	\$71.03 – \$73.68	18.0 – 18.0
Cap CM	1100 – 2000	—	\$74.35 – \$71.45	18.0 – 18.1
Cap ACT	60 – 100	—	\$72.98 – \$72.93	18.0 – 18.0
LoS Police	0.1 – 0.3	—	\$72.94 – \$72.89	18.0 – 18.0
LoS CrimJst	50 – 80	—	\$73.01 – \$72.85	18.0 – 18.0
LoS ED	0.1 – 0.5	—	\$72.96 – \$72.86	18.0 – 18.0
LoS Acute	8 – 16	—	\$72.09 – \$73.78	18.0 – 18.0
LoS Inpatnt	60 – 120	—	\$72.31 – \$72.98	18.0 – 18.0
LoS MMT	180 – 720	—	\$73.53 – \$71.54	18.0 – 18.1
LoS CM	365 – 2190	—	\$77.13 – \$72.94	17.8 – 18.0
LoS ACT	2000 – 5000	—	\$72.94 – \$72.94	18.0 – 18.0
LoS Family	365 – 2190	—	\$78.74 – \$67.13	17.8 – 18.2
Cost coef Police	400 – 800	—	\$72.23 – \$73.40	18.0 – 18.0
Cost CrimJst	50 – 125	—	\$72.78 – \$75.78	18.0 – 18.0
Cost coef ED	300 – 600	—	\$72.71 – \$73.51	18.0 – 18.0
Cost Acute	500 – 1000	—	\$72.27 – \$74.55	18.0 – 18.0
Cost Inpatnt	200 – 500	—	\$71.03 – \$77.47	18.0 – 18.0
Cost MMT	8 – 18	—	\$72.67 – \$72.97	18.0 – 18.0
Cost CM	18 – 28	—	\$72.19 – \$74.06	18.0 – 18.0
Cost ACT	30 – 60	—	\$72.75 – \$73.11	18.0 – 18.0
Cost Family	0.5 – 10	—	\$72.72 – \$74.12	18.0 – 18.0
Cost Non-tr	0 – 50	—	\$72.94 – \$100.81	18.0 – 18.0
Cost Non-tr crime	50 – 90	—	\$54.27 – \$81.93	18.0 – 18.0
Police→CrimJst	0.2 – 0.4	—	\$72.97 – \$72.64	18.0 – 18.0
Police→ED	0.01 – 0.1	—	\$72.87 – \$73.03	18.0 – 18.0
CrimJst→Inpatnt	0 – 0.1	—	\$72.94 – \$72.94	18.0 – 18.0
ED→Acute	0.1 – 0.2	—	\$72.29 – \$73.86	18.0 – 18.0
ED→Inpatnt	0.01 – 0.1	—	\$72.28 – \$72.94	18.0 – 18.0
Other→ED	0.05 – 0.15	—	\$72.88 – \$72.98	18.0 – 18.0
Other→Inpatnt	0.05 – 0.15	—	\$72.66 – \$72.94	18.0 – 18.0
Other→MMT	0.01 – 0.1	—	\$73.53 – \$71.77	18.0 – 18.1
Other→CM	0.05 – 0.2	—	\$74.57 – \$72.94	17.9 – 18.0
Other→ACT	0.005 – 0.02	—	\$72.94 – \$72.94	18.0 – 18.0
Other→Family	0.05 – 0.13	—	\$75.99 – \$67.88	17.9 – 18.2
Acute→Inpatnt	0.01 – 0.1	—	\$72.94 – \$72.94	18.0 – 18.0
Acute→MMT	0.01 – 0.1	—	\$73.19 – \$72.64	18.0 – 18.1
Acute→CM	0.01 – 0.1	—	\$72.94 – \$72.94	18.0 – 18.0
Inpatnt→MMT	0.01 – 0.1	—	\$73.07 – \$72.78	18.0 – 18.0
Inpatnt→CM	0.01 – 0.1	—	\$72.94 – \$72.94	18.0 – 18.0
Inpatnt→ACT	0.005 – 0.02	—	\$72.94 – \$72.94	18.0 – 18.0
Crime treat ratio	0.2 – 1	—	\$68.87 – \$93.93	18.0 – 18.0

Table F.8: Sensitivity analysis of Scenario 7—Decreased crime.

Cell	Base Range	Adjusted Range	Costs	QALYs
Population	4000 – 9000	—	\$57.34 – \$66.56	18.7 – 17.9
Arrivals Police	20 – 30	—	\$65.12 – \$65.86	18.1 – 18.1
Arr. /capita ED	0.005 – 0.015	—	\$62.33 – \$69.29	18.0 – 18.1
Arrivals Other	5 – 15	—	\$70.54 – \$61.97	17.8 – 18.2
Cap Inpatnt	90 – 190	—	\$63.24 – \$66.03	18.0 – 18.1
Cap CM	1100 – 2000	—	\$66.48 – \$63.61	18.0 – 18.2
Cap ACT	60 – 100	—	\$65.28 – \$65.24	18.1 – 18.1
LoS Police	0.1 – 0.3	—	\$65.25 – \$65.21	18.1 – 18.1
LoS CrimJst	50 – 80	—	\$65.28 – \$65.20	18.1 – 18.1
LoS ED	0.1 – 0.5	—	\$65.27 – \$65.13	18.1 – 18.1
LoS Acute	8 – 16	—	\$63.61 – \$66.88	18.1 – 18.1
LoS Inpatnt	60 – 120	—	\$65.18 – \$65.29	18.1 – 18.1
LoS MMT	180 – 720	—	\$65.85 – \$63.84	18.0 – 18.2
LoS CM	365 – 2190	—	\$68.77 – \$65.25	17.9 – 18.1
LoS ACT	2000 – 5000	—	\$65.25 – \$65.25	18.1 – 18.1
LoS Family	365 – 2190	—	\$70.57 – \$59.89	17.9 – 18.3
Cost coef Police	400 – 800	—	\$64.57 – \$65.69	18.1 – 18.1
Cost CrimJst	50 – 125	—	\$65.12 – \$67.51	18.1 – 18.1
Cost coef ED	300 – 600	—	\$64.81 – \$66.32	18.1 – 18.1
Cost Acute	500 – 1000	—	\$63.99 – \$68.27	18.1 – 18.1
Cost Inpatnt	200 – 500	—	\$63.34 – \$69.78	18.1 – 18.1
Cost MMT	8 – 18	—	\$64.93 – \$65.28	18.1 – 18.1
Cost CM	18 – 28	—	\$64.50 – \$66.37	18.1 – 18.1
Cost ACT	30 – 60	—	\$65.06 – \$65.41	18.1 – 18.1
Cost Family	0.5 – 10	—	\$65.03 – \$66.43	18.1 – 18.1
Cost Non-tr	0 – 50	—	\$65.25 – \$93.04	18.1 – 18.1
Cost Non-tr crime	50 – 90	40 – 72	\$50.35 – \$72.42	18.1 – 18.1
Police→CrimJst	0.2 – 0.4	0.1811 – 0.1811	\$65.27 – \$64.95	18.1 – 18.1
Police→ED	0.01 – 0.1	—	\$65.26 – \$65.24	18.1 – 18.1
CrimJst→Inpatnt	0 – 0.1	—	\$65.25 – \$65.25	18.1 – 18.1
ED→Acute	0.1 – 0.2	—	\$63.93 – \$67.09	18.0 – 18.1
ED→Inpatnt	0.01 – 0.1	—	\$65.25 – \$65.25	18.1 – 18.1
Other→ED	0.05 – 0.15	—	\$65.25 – \$65.25	18.1 – 18.1
Other→Inpatnt	0.05 – 0.15	—	\$65.25 – \$65.25	18.1 – 18.1
Other→MMT	0.01 – 0.1	—	\$65.77 – \$64.20	18.0 – 18.1
Other→CM	0.05 – 0.2	—	\$66.23 – \$65.25	18.0 – 18.1
Other→ACT	0.005 – 0.02	—	\$65.25 – \$65.25	18.1 – 18.1
Other→Family	0.05 – 0.13	—	\$68.05 – \$60.59	18.0 – 18.2
Acute→Inpatnt	0.01 – 0.1	—	\$65.25 – \$65.25	18.1 – 18.1
Acute→MMT	0.01 – 0.1	—	\$65.60 – \$64.83	18.0 – 18.1
Acute→CM	0.01 – 0.1	—	\$65.25 – \$65.25	18.1 – 18.1
Inpatnt→MMT	0.01 – 0.1	—	\$65.37 – \$65.10	18.0 – 18.1
Inpatnt→CM	0.01 – 0.1	—	\$65.25 – \$65.25	18.1 – 18.1
Inpatnt→ACT	0.005 – 0.02	—	\$65.25 – \$65.25	18.1 – 18.1
Crime treat ratio	0.2 – 1	—	\$62.00 – \$82.00	18.1 – 18.1

Table F.9: Sensitivity analysis of Scenario 8—Uncapacitated.

Cell	Base Range	Adjusted Range	Costs	QALYs
Population	4000 – 9000	—	\$64.43 – \$79.64	18.8 – 18.1
Arrivals Police	20 – 30	—	\$77.18 – \$77.60	18.2 – 18.2
Arr. /capita ED	0.005 – 0.015	—	\$73.41 – \$82.08	18.1 – 18.3
Arrivals Other	5 – 15	—	\$84.72 – \$72.76	17.9 – 18.4
Cap Inpatnt	90 – 190	∞ – ∞	\$77.42 – \$77.19	18.2 – 18.2
Cap CM	1100 – 2000	∞ – ∞	\$77.81 – \$76.47	18.2 – 18.3
Cap ACT	60 – 100	∞ – ∞	\$77.31 – \$77.23	18.2 – 18.2
LoS Police	0.1 – 0.3	—	\$77.25 – \$77.21	18.2 – 18.2
LoS CrimJst	50 – 80	—	\$77.40 – \$77.06	18.2 – 18.2
LoS ED	0.1 – 0.5	—	\$77.28 – \$77.12	18.2 – 18.2
LoS Acute	8 – 16	—	\$75.88 – \$78.61	18.2 – 18.2
LoS Inpatnt	60 – 120	—	\$74.82 – \$79.78	18.2 – 18.3
LoS MMT	180 – 720	—	\$78.05 – \$75.39	18.2 – 18.3
LoS CM	365 – 2190	—	\$84.24 – \$73.13	18.0 – 18.4
LoS ACT	2000 – 5000	—	\$77.40 – \$77.11	18.2 – 18.3
LoS Family	365 – 2190	—	\$83.18 – \$71.09	18.1 – 18.4
Cost coef Police	400 – 800	—	\$76.64 – \$77.65	18.2 – 18.2
Cost CrimJst	50 – 125	—	\$77.11 – \$79.71	18.2 – 18.2
Cost coef ED	300 – 600	—	\$76.87 – \$78.19	18.2 – 18.2
Cost Acute	500 – 1000	—	\$76.16 – \$79.87	18.2 – 18.2
Cost Inpatnt	200 – 500	—	\$74.18 – \$84.52	18.2 – 18.2
Cost MMT	8 – 18	—	\$76.95 – \$77.29	18.2 – 18.2
Cost CM	18 – 28	—	\$76.36 – \$78.59	18.2 – 18.2
Cost ACT	30 – 60	—	\$76.35 – \$78.05	18.2 – 18.2
Cost Family	0.5 – 10	—	\$77.06 – \$78.27	18.2 – 18.2
Cost Non-tr	0 – 50	—	\$77.25 – \$101.32	18.2 – 18.2
Cost Non-tr crime	50 – 90	—	\$60.09 – \$85.52	18.2 – 18.2
Police→CrimJst	0.2 – 0.4	—	\$77.31 – \$76.66	18.2 – 18.2
Police→ED	0.01 – 0.1	—	\$77.25 – \$77.25	18.2 – 18.2
CrimJst→Inpatnt	0 – 0.1	—	\$77.25 – \$78.13	18.2 – 18.2
ED→Acute	0.1 – 0.2	—	\$76.39 – \$78.42	18.2 – 18.3
ED→Inpatnt	0.01 – 0.1	—	\$74.32 – \$80.64	18.2 – 18.3
Other→ED	0.05 – 0.15	—	\$77.25 – \$77.25	18.2 – 18.2
Other→Inpatnt	0.05 – 0.15	—	\$76.02 – \$78.05	18.2 – 18.2
Other→MMT	0.01 – 0.1	—	\$77.90 – \$75.95	18.2 – 18.3
Other→CM	0.05 – 0.2	—	\$80.23 – \$73.15	18.1 – 18.3
Other→ACT	0.005 – 0.02	—	\$77.38 – \$77.01	18.2 – 18.3
Other→Family	0.05 – 0.13	—	\$80.39 – \$71.90	18.2 – 18.4
Acute→Inpatnt	0.01 – 0.1	—	\$76.85 – \$77.75	18.2 – 18.2
Acute→MMT	0.01 – 0.1	—	\$77.69 – \$76.73	18.2 – 18.3
Acute→CM	0.01 – 0.1	—	\$78.22 – \$76.12	18.2 – 18.3
Inpatnt→MMT	0.01 – 0.1	—	\$77.48 – \$76.97	18.2 – 18.2
Inpatnt→CM	0.01 – 0.1	—	\$77.85 – \$76.53	18.2 – 18.2
Inpatnt→ACT	0.005 – 0.02	—	\$77.29 – \$77.18	18.2 – 18.2
Crime treat ratio	0.2 – 1	—	\$72.57 – \$101.39	18.2 – 18.2

Table F.10: Sensitivity analysis of Combo Scenario A—Increased referrals & expanded long term cap..

Cell	Base Range	Adjusted Range	Costs	QALYs
Population	4000 – 9000	—	\$61.10 – \$72.63	18.9 – 18.2
Arrivals Police	20 – 30	—	\$70.10 – \$70.56	18.3 – 18.3
Arr. /capita ED	0.005 – 0.015	—	\$68.22 – \$73.07	18.3 – 18.4
Arrivals Other	5 – 15	—	\$78.45 – \$66.21	18.0 – 18.5
Cap Inpatnt	90 – 190	—	\$68.29 – \$70.92	18.3 – 18.3
Cap CM	1100 – 2000	2000 – 2000	\$70.64 – \$69.52	18.3 – 18.3
Cap ACT	60 – 100	270 – 270	\$70.23 – \$70.17	18.3 – 18.3
LoS Police	0.1 – 0.3	—	\$70.18 – \$70.14	18.3 – 18.3
LoS CrimJst	50 – 80	—	\$70.27 – \$70.07	18.3 – 18.3
LoS ED	0.1 – 0.5	—	\$70.20 – \$70.07	18.3 – 18.3
LoS Acute	8 – 16	—	\$68.95 – \$71.40	18.3 – 18.3
LoS Inpatnt	60 – 120	—	\$70.10 – \$70.23	18.3 – 18.3
LoS MMT	180 – 720	—	\$70.96 – \$68.41	18.3 – 18.4
LoS CM	365 – 2190	—	\$76.05 – \$70.18	18.0 – 18.3
LoS ACT	2000 – 5000	—	\$70.18 – \$70.18	18.3 – 18.3
LoS Family	365 – 2190	—	\$77.28 – \$63.96	18.1 – 18.5
Cost coef Police	400 – 800	—	\$69.64 – \$70.54	18.3 – 18.3
Cost CrimJst	50 – 125	—	\$70.06 – \$72.37	18.3 – 18.3
Cost coef ED	300 – 600	—	\$69.84 – \$71.01	18.3 – 18.3
Cost Acute	500 – 1000	—	\$69.21 – \$72.51	18.3 – 18.3
Cost Inpatnt	200 – 500	—	\$68.27 – \$74.71	18.3 – 18.3
Cost MMT	8 – 18	—	\$69.82 – \$70.22	18.3 – 18.3
Cost CM	18 – 28	—	\$69.12 – \$71.78	18.3 – 18.3
Cost ACT	30 – 60	—	\$69.61 – \$70.68	18.3 – 18.3
Cost Family	0.5 – 10	—	\$69.93 – \$71.54	18.3 – 18.3
Cost Non-tr	0 – 50	—	\$70.18 – \$91.57	18.3 – 18.3
Cost Non-tr crime	50 – 90	—	\$53.85 – \$78.04	18.3 – 18.3
Police→CrimJst	0.2 – 0.4	—	\$70.22 – \$69.82	18.3 – 18.3
Police→ED	0.01 – 0.1	—	\$70.18 – \$70.18	18.3 – 18.3
CrimJst→Inpatnt	0 – 0.1	—	\$70.18 – \$70.18	18.3 – 18.3
ED→Acute	0.1 – 0.2	—	\$69.28 – \$71.47	18.3 – 18.3
ED→Inpatnt	0.01 – 0.1	0.015 – 0.15	\$70.18 – \$70.18	18.3 – 18.3
Other→ED	0.05 – 0.15	—	\$70.18 – \$70.18	18.3 – 18.3
Other→Inpatnt	0.05 – 0.15	0.075 – 0.225	\$70.18 – \$70.18	18.3 – 18.3
Other→MMT	0.01 – 0.1	0.015 – 0.15	\$70.86 – \$68.85	18.3 – 18.4
Other→CM	0.05 – 0.2	0.075 – 0.3	\$72.53 – \$70.18	18.2 – 18.3
Other→ACT	0.005 – 0.02	0.0075 – 0.03	\$70.18 – \$70.18	18.3 – 18.3
Other→Family	0.05 – 0.13	0.075 – 0.195	\$73.80 – \$64.71	18.2 – 18.5
Acute→Inpatnt	0.01 – 0.1	0.015 – 0.15	\$70.18 – \$70.18	18.3 – 18.3
Acute→MMT	0.01 – 0.1	0.015 – 0.15	\$70.64 – \$69.64	18.3 – 18.4
Acute→CM	0.01 – 0.1	0.015 – 0.15	\$70.72 – \$70.18	18.3 – 18.3
Inpatnt→MMT	0.01 – 0.1	—	\$70.32 – \$70.01	18.3 – 18.3
Inpatnt→CM	0.01 – 0.1	—	\$70.18 – \$70.18	18.3 – 18.3
Inpatnt→ACT	0.005 – 0.02	—	\$70.18 – \$70.18	18.3 – 18.3
Crime treat ratio	0.2 – 1	—	\$64.81 – \$97.86	18.3 – 18.3

Table F.11: Sensitivity analysis of Combo Scenario B—Inreased referrals & lower turnover.

Cell	Base Range	Adjusted Range	Costs	QALYs
Population	4000 – 9000	—	\$56.10 – \$70.40	18.9 – 18.2
Arrivals Police	20 – 30	—	\$67.71 – \$68.17	18.3 – 18.3
Arr. /capita ED	0.005 – 0.015	—	\$65.74 – \$70.57	18.3 – 18.4
Arrivals Other	5 – 15	—	\$76.08 – \$62.79	18.1 – 18.5
Cap Inpatnt	90 – 190	—	\$65.90 – \$68.52	18.3 – 18.4
Cap CM	1100 – 2000	—	\$69.21 – \$65.46	18.3 – 18.5
Cap ACT	60 – 100	—	\$67.83 – \$67.77	18.3 – 18.3
LoS Police	0.1 – 0.3	—	\$67.79 – \$67.75	18.3 – 18.3
LoS CrimJst	50 – 80	—	\$67.87 – \$67.67	18.3 – 18.3
LoS ED	0.1 – 0.5	—	\$67.81 – \$67.67	18.3 – 18.3
LoS Acute	8 – 16	—	\$66.55 – \$69.01	18.3 – 18.3
LoS Inpatnt	60 – 120	—	\$67.68 – \$67.84	18.3 – 18.3
LoS MMT	180 – 720	270 – 1080	\$68.77 – \$65.63	18.3 – 18.5
LoS CM	365 – 2190	547.5 – 3285	\$70.83 – \$67.79	18.2 – 18.3
LoS ACT	2000 – 5000	3000 – 7500	\$67.79 – \$67.79	18.3 – 18.3
LoS Family	365 – 2190	547.5 – 3285	\$77.87 – \$59.95	18.1 – 18.6
Cost coef Police	400 – 800	—	\$67.24 – \$68.14	18.3 – 18.3
Cost CrimJst	50 – 125	—	\$67.66 – \$69.97	18.3 – 18.3
Cost coef ED	300 – 600	—	\$67.45 – \$68.61	18.3 – 18.3
Cost Acute	500 – 1000	—	\$66.82 – \$70.11	18.3 – 18.3
Cost Inpatnt	200 – 500	—	\$65.88 – \$72.32	18.3 – 18.3
Cost MMT	8 – 18	—	\$67.24 – \$67.85	18.3 – 18.3
Cost CM	18 – 28	—	\$67.04 – \$68.91	18.3 – 18.3
Cost ACT	30 – 60	—	\$67.60 – \$67.95	18.3 – 18.3
Cost Family	0.5 – 10	—	\$67.41 – \$69.82	18.3 – 18.3
Cost Non-tr	0 – 50	—	\$67.79 – \$89.14	18.3 – 18.3
Cost Non-tr crime	50 – 90	—	\$51.47 – \$75.64	18.3 – 18.3
Police→CrimJst	0.2 – 0.4	—	\$67.82 – \$67.43	18.3 – 18.3
Police→ED	0.01 – 0.1	—	\$67.79 – \$67.79	18.3 – 18.3
CrimJst→Inpatnt	0 – 0.1	—	\$67.79 – \$67.79	18.3 – 18.3
ED→Acute	0.1 – 0.2	—	\$66.91 – \$69.01	18.3 – 18.4
ED→Inpatnt	0.01 – 0.1	0.015 – 0.15	\$67.79 – \$67.79	18.3 – 18.3
Other→ED	0.05 – 0.15	—	\$67.79 – \$67.79	18.3 – 18.3
Other→Inpatnt	0.05 – 0.15	0.075 – 0.225	\$67.79 – \$67.79	18.3 – 18.3
Other→MMT	0.01 – 0.1	0.015 – 0.15	\$68.65 – \$66.15	18.3 – 18.5
Other→CM	0.05 – 0.2	0.075 – 0.3	\$68.29 – \$67.79	18.3 – 18.3
Other→ACT	0.005 – 0.02	0.0075 – 0.03	\$67.79 – \$67.79	18.3 – 18.3
Other→Family	0.05 – 0.13	0.075 – 0.195	\$72.74 – \$60.86	18.2 – 18.5
Acute→Inpatnt	0.01 – 0.1	0.015 – 0.15	\$67.79 – \$67.79	18.3 – 18.3
Acute→MMT	0.01 – 0.1	0.015 – 0.15	\$68.36 – \$67.11	18.3 – 18.4
Acute→CM	0.01 – 0.1	0.015 – 0.15	\$67.79 – \$67.79	18.3 – 18.3
Inpatnt→MMT	0.01 – 0.1	—	\$67.96 – \$67.57	18.3 – 18.4
Inpatnt→CM	0.01 – 0.1	—	\$67.79 – \$67.79	18.3 – 18.3
Inpatnt→ACT	0.005 – 0.02	—	\$67.79 – \$67.79	18.3 – 18.3
Crime treat ratio	0.2 – 1	—	\$62.41 – \$95.49	18.3 – 18.3

Table F.12: Sensitivity analysis of Combo Scenario C—Expanded Inpatnt & long term capacities.

Cell	Base Range	Adjusted Range	Costs	QALYs
Population	4000 – 9000	—	\$64.21 – \$79.81	18.8 – 18.1
Arrivals Police	20 – 30	—	\$77.31 – \$77.74	18.2 – 18.2
Arr. /capita ED	0.005 – 0.015	—	\$73.45 – \$81.67	18.1 – 18.3
Arrivals Other	5 – 15	—	\$84.72 – \$73.01	17.9 – 18.4
Cap Inpatnt	90 – 190	180 – 380	\$75.31 – \$77.32	18.2 – 18.2
Cap CM	1100 – 2000	2000 – 2000	\$77.94 – \$76.59	18.2 – 18.2
Cap ACT	60 – 100	270 – 270	\$77.44 – \$77.36	18.2 – 18.2
LoS Police	0.1 – 0.3	—	\$77.38 – \$77.33	18.2 – 18.2
LoS CrimJst	50 – 80	—	\$77.53 – \$77.19	18.2 – 18.2
LoS ED	0.1 – 0.5	—	\$77.41 – \$77.24	18.2 – 18.2
LoS Acute	8 – 16	—	\$75.98 – \$78.77	18.2 – 18.2
LoS Inpatnt	60 – 120	—	\$74.90 – \$79.18	18.2 – 18.2
LoS MMT	180 – 720	—	\$78.20 – \$75.48	18.2 – 18.3
LoS CM	365 – 2190	—	\$84.59 – \$75.86	18.0 – 18.3
LoS ACT	2000 – 5000	—	\$77.40 – \$77.38	18.2 – 18.2
LoS Family	365 – 2190	—	\$83.45 – \$71.08	18.1 – 18.4
Cost coef Police	400 – 800	—	\$76.75 – \$77.79	18.2 – 18.2
Cost CrimJst	50 – 125	—	\$77.24 – \$79.89	18.2 – 18.2
Cost coef ED	300 – 600	—	\$76.99 – \$78.33	18.2 – 18.2
Cost Acute	500 – 1000	—	\$76.27 – \$80.06	18.2 – 18.2
Cost Inpatnt	200 – 500	—	\$74.25 – \$84.81	18.2 – 18.2
Cost MMT	8 – 18	—	\$77.07 – \$77.42	18.2 – 18.2
Cost CM	18 – 28	—	\$76.47 – \$78.75	18.2 – 18.2
Cost ACT	30 – 60	—	\$76.81 – \$77.88	18.2 – 18.2
Cost Family	0.5 – 10	—	\$77.19 – \$78.42	18.2 – 18.2
Cost Non-tr	0 – 50	—	\$77.38 – \$101.98	18.2 – 18.2
Cost Non-tr crime	50 – 90	—	\$60.03 – \$85.73	18.2 – 18.2
Police→CrimJst	0.2 – 0.4	—	\$77.44 – \$76.78	18.2 – 18.2
Police→ED	0.01 – 0.1	—	\$77.38 – \$77.38	18.2 – 18.2
CrimJst→Inpatnt	0 – 0.1	—	\$77.38 – \$78.29	18.2 – 18.2
ED→Acute	0.1 – 0.2	—	\$76.51 – \$78.57	18.2 – 18.2
ED→Inpatnt	0.01 – 0.1	—	\$74.37 – \$78.92	18.1 – 18.2
Other→ED	0.05 – 0.15	—	\$77.38 – \$77.38	18.2 – 18.2
Other→Inpatnt	0.05 – 0.15	—	\$76.12 – \$78.21	18.2 – 18.2
Other→MMT	0.01 – 0.1	—	\$78.04 – \$76.05	18.2 – 18.3
Other→CM	0.05 – 0.2	—	\$80.46 – \$75.86	18.1 – 18.3
Other→ACT	0.005 – 0.02	—	\$77.38 – \$77.38	18.2 – 18.2
Other→Family	0.05 – 0.13	—	\$80.59 – \$71.91	18.1 – 18.3
Acute→Inpatnt	0.01 – 0.1	—	\$76.97 – \$77.89	18.2 – 18.2
Acute→MMT	0.01 – 0.1	—	\$77.83 – \$76.84	18.2 – 18.2
Acute→CM	0.01 – 0.1	—	\$78.39 – \$76.21	18.2 – 18.2
Inpatnt→MMT	0.01 – 0.1	—	\$77.62 – \$77.09	18.2 – 18.2
Inpatnt→CM	0.01 – 0.1	—	\$78.00 – \$76.64	18.2 – 18.2
Inpatnt→ACT	0.005 – 0.02	—	\$77.38 – \$77.38	18.2 – 18.2
Crime treat ratio	0.2 – 1	—	\$72.81 – \$100.92	18.2 – 18.2

Table F.13: Sensitivity analysis of Combo Scenario D—Scenarios 1, 2, 4, 6, 7.

Cell	Base Range	Adjusted Range	Costs	QALYs
Population	4000 – 9000	—	\$48.65 – \$56.74	19.0 – 18.3
Arrivals Police	20 – 30	—	\$55.33 – \$55.90	18.4 – 18.4
Arr. /capita ED	0.005 – 0.015	0.0028 – 0.0078	\$54.49 – \$56.72	18.4 – 18.5
Arrivals Other	5 – 15	—	\$61.67 – \$52.14	18.1 – 18.6
Cap Inpatnt	90 – 190	—	\$53.31 – \$56.25	18.4 – 18.5
Cap CM	1100 – 2000	2000 – 2000	\$55.80 – \$54.89	18.4 – 18.5
Cap ACT	60 – 100	270 – 270	\$55.46 – \$55.41	18.4 – 18.4
LoS Police	0.1 – 0.3	—	\$55.42 – \$55.40	18.4 – 18.4
LoS CrimJst	50 – 80	—	\$55.43 – \$55.42	18.4 – 18.4
LoS ED	0.1 – 0.5	—	\$55.43 – \$55.38	18.4 – 18.4
LoS Acute	8 – 16	—	\$54.84 – \$56.01	18.4 – 18.4
LoS Inpatnt	60 – 120	—	\$54.82 – \$55.46	18.4 – 18.4
LoS MMT	180 – 720	270 – 1080	\$55.93 – \$54.29	18.4 – 18.6
LoS CM	365 – 2190	547.5 – 3285	\$58.30 – \$55.42	18.2 – 18.4
LoS ACT	2000 – 5000	3000 – 7500	\$55.42 – \$55.42	18.4 – 18.4
LoS Family	365 – 2190	547.5 – 3285	\$62.26 – \$50.18	18.2 – 18.6
Cost coef Police	400 – 800	—	\$54.96 – \$55.73	18.4 – 18.4
Cost CrimJst	50 – 125	—	\$55.34 – \$56.97	18.4 – 18.4
Cost coef ED	300 – 600	—	\$55.26 – \$55.81	18.4 – 18.4
Cost Acute	500 – 1000	—	\$54.97 – \$56.51	18.4 – 18.4
Cost Inpatnt	200 – 500	—	\$53.51 – \$59.95	18.4 – 18.4
Cost MMT	8 – 18	—	\$55.01 – \$55.47	18.4 – 18.4
Cost CM	18 – 28	—	\$54.36 – \$57.02	18.4 – 18.4
Cost ACT	30 – 60	—	\$54.85 – \$55.93	18.4 – 18.4
Cost Family	0.5 – 10	—	\$55.09 – \$57.23	18.4 – 18.4
Cost Non-tr	0 – 50	—	\$55.42 – \$74.39	18.4 – 18.4
Cost Non-tr crime	50 – 90	40 – 72	\$43.02 – \$61.40	18.4 – 18.4
Police→CrimJst	0.2 – 0.4	0.1811 – 0.1811	\$55.45 – \$55.08	18.4 – 18.5
Police→ED	0.01 – 0.1	—	\$55.38 – \$55.48	18.4 – 18.4
CrimJst→Inpatnt	0 – 0.1	—	\$55.42 – \$55.42	18.4 – 18.4
ED→Acute	0.1 – 0.2	—	\$54.99 – \$56.04	18.4 – 18.5
ED→Inpatnt	0.01 – 0.1	0.015 – 0.15	\$54.78 – \$55.42	18.4 – 18.4
Other→ED	0.05 – 0.15	—	\$55.39 – \$55.45	18.4 – 18.4
Other→Inpatnt	0.05 – 0.15	0.075 – 0.225	\$55.20 – \$55.42	18.4 – 18.4
Other→MMT	0.01 – 0.1	0.015 – 0.15	\$55.92 – \$54.48	18.4 – 18.6
Other→CM	0.05 – 0.2	0.075 – 0.3	\$56.33 – \$55.42	18.4 – 18.4
Other→ACT	0.005 – 0.02	0.0075 – 0.03	\$55.42 – \$55.42	18.4 – 18.4
Other→Family	0.05 – 0.13	0.075 – 0.195	\$58.77 – \$50.79	18.3 – 18.6
Acute→Inpatnt	0.01 – 0.1	0.015 – 0.15	\$55.42 – \$55.42	18.4 – 18.4
Acute→MMT	0.01 – 0.1	0.015 – 0.15	\$55.64 – \$55.16	18.4 – 18.5
Acute→CM	0.01 – 0.1	0.015 – 0.15	\$55.42 – \$55.42	18.4 – 18.4
Inpatnt→MMT	0.01 – 0.1	—	\$55.53 – \$55.29	18.4 – 18.5
Inpatnt→CM	0.01 – 0.1	—	\$55.42 – \$55.42	18.4 – 18.4
Inpatnt→ACT	0.005 – 0.02	—	\$55.42 – \$55.42	18.4 – 18.4
Crime treat ratio	0.2 – 1	—	\$50.69 – \$79.81	18.4 – 18.4

Table F.14: Sensitivity analysis of Combo Scenario E—All scenarios (excluding "uncapacitated").

Cell	Base Range	Adjusted Range	Costs	QALYs
Population	4000 – 9000	—	\$48.49 – \$59.04	19.0 – 18.3
Arrivals Police	20 – 30	—	\$56.63 – \$57.04	18.5 – 18.5
Arr. /capita ED	0.005 – 0.015	0.0026 – 0.0076	\$54.53 – \$59.62	18.4 – 18.5
Arrivals Other	5 – 15	—	\$62.02 – \$53.94	18.2 – 18.6
Cap Inpatnt	90 – 190	180 – 380	\$55.69 – \$56.66	18.5 – 18.5
Cap CM	1100 – 2000	2000 – 2000	\$57.02 – \$56.24	18.5 – 18.5
Cap ACT	60 – 100	270 – 270	\$56.73 – \$56.69	18.5 – 18.5
LoS Police	0.1 – 0.3	—	\$56.70 – \$56.67	18.5 – 18.5
LoS CrimJst	50 – 80	—	\$56.73 – \$56.65	18.5 – 18.5
LoS ED	0.1 – 0.5	—	\$56.71 – \$56.65	18.5 – 18.5
LoS Acute	8 – 16	—	\$56.15 – \$57.24	18.5 – 18.5
LoS Inpatnt	60 – 120	—	\$54.53 – \$58.94	18.5 – 18.5
LoS MMT	180 – 720	270 – 1080	\$57.44 – \$55.06	18.4 – 18.6
LoS CM	365 – 2190	547.5 – 3285	\$60.70 – \$56.70	18.3 – 18.5
LoS ACT	2000 – 5000	3000 – 7500	\$56.70 – \$56.70	18.5 – 18.5
LoS Family	365 – 2190	547.5 – 3285	\$65.03 – \$50.24	18.2 – 18.7
Cost coef Police	400 – 800	—	\$56.25 – \$56.99	18.5 – 18.5
Cost CrimJst	50 – 125	—	\$56.61 – \$58.19	18.5 – 18.5
Cost coef ED	300 – 600	—	\$56.55 – \$57.06	18.5 – 18.5
Cost Acute	500 – 1000	—	\$56.27 – \$57.72	18.5 – 18.5
Cost Inpatnt	200 – 500	—	\$54.14 – \$62.75	18.5 – 18.5
Cost MMT	8 – 18	—	\$56.24 – \$56.75	18.5 – 18.5
Cost CM	18 – 28	—	\$55.63 – \$58.30	18.5 – 18.5
Cost ACT	30 – 60	—	\$56.12 – \$57.20	18.5 – 18.5
Cost Family	0.5 – 10	—	\$56.36 – \$58.51	18.5 – 18.5
Cost Non-tr	0 – 50	—	\$56.70 – \$75.02	18.5 – 18.5
Cost Non-tr crime	50 – 90	40 – 72	\$44.53 – \$62.56	18.5 – 18.5
Police→CrimJst	0.2 – 0.4	0.1811 – 0.1811	\$56.72 – \$56.41	18.5 – 18.5
Police→ED	0.01 – 0.1	0.025 – 0.025	\$56.81 – \$56.56	18.5 – 18.5
CrimJst→Inpatnt	0 – 0.1	—	\$56.70 – \$57.35	18.5 – 18.5
ED→Acute	0.1 – 0.2	—	\$56.21 – \$57.38	18.5 – 18.5
ED→Inpatnt	0.01 – 0.1	0.015 – 0.15	\$54.64 – \$59.17	18.5 – 18.5
Other→ED	0.05 – 0.15	—	\$56.61 – \$56.75	18.5 – 18.5
Other→Inpatnt	0.05 – 0.15	0.075 – 0.225	\$55.00 – \$57.81	18.5 – 18.5
Other→MMT	0.01 – 0.1	0.015 – 0.15	\$57.36 – \$55.44	18.4 – 18.6
Other→CM	0.05 – 0.2	0.075 – 0.3	\$57.74 – \$56.70	18.4 – 18.5
Other→ACT	0.005 – 0.02	0.0075 – 0.03	\$56.70 – \$56.70	18.5 – 18.5
Other→Family	0.05 – 0.13	0.075 – 0.195	\$60.63 – \$51.15	18.4 – 18.6
Acute→Inpatnt	0.01 – 0.1	0.015 – 0.15	\$56.41 – \$57.05	18.5 – 18.5
Acute→MMT	0.01 – 0.1	0.015 – 0.15	\$56.96 – \$56.38	18.5 – 18.5
Acute→CM	0.01 – 0.1	0.015 – 0.15	\$56.70 – \$56.70	18.5 – 18.5
Inpatnt→MMT	0.01 – 0.1	—	\$56.89 – \$56.47	18.5 – 18.5
Inpatnt→CM	0.01 – 0.1	—	\$56.70 – \$56.70	18.5 – 18.5
Inpatnt→ACT	0.005 – 0.02	—	\$56.70 – \$56.70	18.5 – 18.5
Crime treat ratio	0.2 – 1	—	\$51.92 – \$81.32	18.5 – 18.5