The Pennsylvania State University

Harold and Inge Marcus Department of Industrial and Manufacturing Engineering

COST-EFFECTIVENESS OF HEALTH CARE INTERVENTIONS

A Dissertation in

Industrial Engineering and Operations Research

by

Sai Zhang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2015

The dissertation of Sai Zhang was reviewed and approved* by the following:

Paul Griffin
Professor in Industrial Engineering
Dissertation Advisor
Chair of Committee

Guodong Pang
The Harold and Inge Marcus Career Assistant Professor
in Industrial Engineering

Vittal Prabhu
Professor in Industrial Engineering

Zhibiao Zhao
Associate Professor in Statistics

Harriet Nembhard
Professor and Interim Department Head in Industrial Engineering

*Signatures are on file in the Graduate School

ABSTRACT

Over the past decade, cost-effectiveness analysis has increasingly gained popularity as a means to determine the allocation of scarce health resources. It is a tool that assesses the tradeoff between consumed resources and achieved health outcomes. The results provide information that help decision makers determine which policy best serves their programmatic and financial needs. In this dissertation, three examples that fall under the cost-effectiveness framework are modeled and addressed in details. The first example is to determine an optimal breast cancer screening policy in the presence of overdiagnosis. We examine the impact of overdiagnosis on quality adjusted life years (QALYs) and the expected number of mammograms. We find screening intervals that range from 2 to 3 years outperform the current recommendations of annual screening. The second example is a cost-effectiveness analysis of treatments for chronic hepatitis CHC. A series of new drugs recently approved for the treatment CHC are characterized with high cure rates. However, there are corresponding high prices. Cost-effectiveness of these new treatments is determined in comparison with standard care regimens, based on which usage suggestions are provided. The third example is to reduce congestion issues of intensive care units (ICUs). Due to limited resources, it is critical to rationalize the use of medical units and maximize throughput from a cost-effective perspective. In this study, we assume that ICU patients in a less severe critical condition can be stepped down early and designated to lower level care units so that beds can be vacated to accommodate new arrivals. Optimal stepdown policies under different scenarios are analyzed and discussed. We find that reserving capacity in the ICU can improve overall system performance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my advisor, Dr. Paul Griffin, for his guidance, assistance and support during my doctoral study at the Pennsylvania State University. I thank him for involving me in the Health Care project. He is a warm-hearted and patient teacher to me. The things I have learnt from him and the project experience I have had with him place a firm foundation for me to pursue a professional career in the industry.

I would like to express my appreciation to Dr. Guodong Pang, Dr. Vittal Prabhu and Dr. Zhibiao Zhao for their helpful comments and suggestions on this thesis. I would like to give my special thanks to my student colleagues, Yan Renfei, Cai Xinxin. They have helped and supported me in many ways. After all, I am greatly indebted to my parents for their unconditional love, support and guidance.

Chapter 1 Introduction

1.1 Introduction

Cost-Effectiveness (CE) analysis is a common approach for determining the best use of limited resources to achieve desired health outcomes. The method is designed to evaluate the comparative impacts of expenditures on different health interventions. The results provide information that can help decision makers weigh alternatives and decide which best serve their programmatic and financial needs. CE analysis can be done from several perspectives including governments, physicians, and insurance companies. In an ideal world with unlimited resources, any program or intervention with a positive health effect should be adopted. However, limits on health care resources mandate that resource allocation decisions must be guided by considerations of cost in relation to the expected benefits.

As a tool to balance tradeoff between consumed resources and achieved health outcomes, CE analysis is based on the premise that "for any given level of resources available, society ... wishes to maximize the total aggregate health benefits conferred" according to Weinstein and Stason[1]. Thus from the societal perspective, cost-effectiveness analysis should not only consider those who gain health but those who pay for it. Put another way, anyone who is affected by the intervention should be considered no matter whether she or he experiences the outcomes and costs or not.

CE analysis provides a framework where questions regarding health care cost and effect can be raised. For example, when a pharmaceutical company prices their new product, it might want to know how much their medication cost per year of life gained compares to a similar product on the market. Or when a screening program is proposed, a question might be posed on how much the intervention costs for each quality-adjusted life year saved. CE analysis can be conducted to answer such questions. It works by showing the relative values of alternative interventions in terms of increasing life years or improving life quality. In other words, it helps to define and compute opportunity cost of each option. The results are usually summarized by cost-effectiveness ratios, which represent the cost per unit health effect achieved, such as, the cost per year of life gained for medical interventions. In this way, alternative interventions for consideration can be sorted based on their cost-effectiveness ratios; those with the lowest CE ratio achieve the greatest health effect. Examples include preventing the most cases of disease or improving most years of life.

Decisions in practice can be far more complicated. There are other components involved for final decisions, for example that are not captured in CE analysis. Examples include feasibility, equity, or factors outside the health care sector. Therefore, CE analysis should only be considered as one factor in a decision making process.

1.2 Economic Foundations of CE Analysis

A variety of disciplines have contributed to CE analysis development including decision analysis and operations research. As a methodology to enhance societal welfare,

the theoretical foundation of CE analysis falls into the context of welfare economics. Welfare economics is based on the following two assumptions[2]: (1) individuals maximize a preference function (utility function), (2) the overall welfare of society is a function of these individuals preferences. Individuals are assumed to have utility functions that associate their well-being to their levels of consumption of goods and services. In welfare economics, a social utility function is defined as some aggregate of individual utilities. Although there is no consensus about how individual utilities should be aggregated, there is still a way to discuss the effect of reallocation of resources on social welfare. Pareto optimality provides a powerful framework for determining whether a resource reallocation will improve the social welfare. A resource distribution is considered to be Pareto optimal when any change in the distribution would make someone worse off, even if others are better off. It implies that if an allocation is not Pareto optimal, there might be a way to reallocate the resources such that at least one person is better off while no one is worse off. If the reallocation makes at least one person better off without others worse off, it is said to be Pareto improving. Thus even if the social welfare function is unknown, it is possible to check whether social welfare is improved with the information of the effects of a change in intervention on individual utilities.

However, in practice, there are few if any public programs that make everyone better off. Typically funds must be raised by taxes or some other mechanism that imposes costs on some people that exceed the benefits they can expect to receive. Thus a less restrictive standard was proposed to resolve such situations when there exist both gainers and losers from a reallocation, called potential Pareto improvement. In this concept, a program is considered to be welfare improving if the gainers are willing to pay enough for their

gains in order to compensate the losers. The rationale behind this standard is that if there were a mechanism for such payment to occur, the program would result in an actual Pareto improvement. Cost-benefit analysis is directly based on this potential Pareto improvement criterion. It can be shown that if a program is undertaken whose benefits exceed their costs. a potential Pareto improvement will occur. Garber and Phelps[3] describe a set of assumptions under which rankings derived from cost-effectiveness ratios provide the optimal expenditure of health resources. Under those assumptions, it was shown that individuals will optimally set priorities for health care expenditures by selecting those ratios (e.g., cost/QALY less than certain threshold).

1.3 Cost-benefit Analysis

From many perspectives, cost-benefit (CB) analysis is similar to cost-effectiveness analysis. However, CB analysis has a closer connection to welfare economics. In CB analysis, when the benefits of a program exceed its costs, the program is considered to satisfy optimal welfare economic conditions. CB analysis is characterized with expressing benefit of health interventions with monetary terms instead of nonmonetary effectiveness measures[4]. The monetary measure is obtained by estimating how much an individual is willing to pay for life saving and health improving, which implicitly favors the wealthy over the poor. In the medical world, many people find it unacceptable to embrace the idea that monetary value can be placed on a human life. This motivates the use of CE analysis in health care area. Further, it is a huge challenge to convert all health outcomes for (e.g., mortality and morbidity) to dollar values and few researchers would

attempt to do this. However, since CB analysis uses monetary values to represent health effects without differentiating input and output, it helps to overcome this CE analysis limitation that only interventions measured in the same units of effectiveness can be compared. CB analysis can theoretically handle comparisons over a wider scope.

1.4 Approach to Perform CE Analysis

Summarizing the literature[1,5], the approach to perform CE analysis is briefly described as followings. First, define the intervention to be studied and alternative interventions for comparison. The placebo intervention (i.e., do nothing) is often used as one of the alternative interventions. However, a comparison to such an alternative is not always informative because in many circumstances placebo treatment is not a reasonable choice. Instead, comparisons should be conducted on relevant options, for example, two promising treatments physicians or policy makers would consider for a disease. In fact, an intervention inferior to all alternative interventions could still be superior to the placebo treatment. The research question we face is always how the new intervention performs compared to commonly accepted interventions.

After intervention and alternative interventions are determined, we build models to capture the incremental CE ratio. Incremental means the difference between the investigated intervention and the benchmark intervention. The incremental CE ratio is denoted as difference of cost per unit of incremental health effect, calculated as the ratio of difference in costs and difference in effects. When the intervention under study is both more effective and less costly than the alternative, the intervention is said to dominate the

alternative. In this situation, there is no need to calculate a cost-effectiveness ratio. Typically, the intervention is both more effective and costly than the alternative, under this circumstance, intervention with a relatively low CE ratio would be considered to have a higher priority for resources. A decision maker can choose the intervention with the lowest CE ratio and continue down the list until available funds are used. The interventions picked are those that generate the largest possible health effect for the given expenditure.

The denominator of a CE ratio is the incremental health effects of the intervention. The most commonly used measure of the health effect is quality-adjusted life years (QALYs). It differs from life expectancy in the sense that for each additional year of life, a utility weight takes a value between 0 and 1 as a function of health status and is added to QALYs. For example, a person in good health status receives a utility weight of 1 whereas a person who dies from illness receives a weight of 0 for the year. Interventions can increase QALYs by lengthening life as well as improving life quality, thus different interventions are comparable in terms of QALYs. Calculation of QALYs is straightforward, but the data regarding utility weight for each health status is typically difficult to obtain, and also subject to opinion. The numerator of the CE ratio is the incremental cost of the intervention. It represents the net present value of costs. There is no consensus on which specific cost should be included and which should not; it is problem specific. In general, the costs may include direct medical and health care costs, such as drugs, hospitalization, office visits, and laboratory services. Some also include costs associated with adverse side effects of treatment and indirect costs such as time costs of treatment. Both costs and QALYs can be discounted.

In order to calculate the incremental CE of several alternative interventions, it is not necessary to calculate every pair of alternative interventions since when many interventions are considered, the number could be large. Instead, we can rank each alternative by the health effect result (e.g. QALYs achieved) first. When there is no strictly dominating intervention, we can then calculate the incremental CE ratios between two adjacent interventions. Furthermore, interventions with extended dominance can be detected and eliminated. Extended dominance[6] is illustrated as follows. Suppose there are 3 alternative interventions, A, B, and C. The costs and outcomes of intervention A are both greater than B, likewise costs and outcomes associated with B are both greater than C. Thus there is no strict dominance intervention. The incremental CE ratio of B versus C is $50,000/QALY, and incremental CE ratio of A versus B is $20,000/QALY. Under this circumstance, if intervention B is chosen over C, implying that an additional QALY gain is worth $50,000, then it must be true that A is more preferable than B, because with only $20,000 cost, one additional QALY will be gained. Here, B is called extended dominance and it should be ruled out of consideration. Different from this stepwise comparison, another approach is to compare CE ratios of alternative interventions to the same intervention. However, it is typically impossible to detect either strict dominance or extended dominance with this approach. Thus this approach is not recommended.

After the incremental CE ratios are calculated, the next step is to choose among the interventions. Sometimes, a benchmark value is used to make decisions. However, this is more of a CB approach, wherein all terms are converted to monetary value. CE analysis provides an alternative way to avoid monetary valuation of health benefits.

1.5 Sensitivity Analysis

Almost all cost-effectiveness problems have uncertainty, which can stem from several sources. For example, the health-specific utility weight for each life-year is based on somewhat unreliable sources such as self-reported data. Assumptions of underlying transition probabilities for disease progression have to be made for analysis considering that the natural course of disease is unobservable. The variation around point estimates may be also large. As a result, sensitivity analysis is necessary when conducting CE analysis. It manifests the effect of variation in uncertain parameters on the final results, and thus it can increase the level of confidence in some decisions while suggesting areas where further research may be valuable in guiding others. There are two ways to conduct sensitivity analysis, one is "one-way" sensitivity analysis and the other is "multi-way" sensitivity analysis. In one-way sensitivity analysis, one parameter is varied at one time, and the other parameters are kept constant, and we check the effect of that one parameter's variation on the results. In multi-way sensitivity analysis, multiple parameters are varied simultaneously to one direction at one time, the challenge with multi-way sensitivity analysis is that the results are difficult to present and compare. To overcome the limitations when it comes to investigate uncertainty in multiple parameters, more advanced statistical approaches[7,8,9] have been developed mostly involving calculation of confidence regions around CE ratios, such as delta method, bootstrap, etc. However, in many cases such approaches suffer from poor computational complexity, and statistical theory is usually not well developed. This limits the application of such approaches in practice.

1.6 Use of CE Analysis

Currently CE analysis is widely conducted in the pharmaceutical area. In countries such as Australia and Canada, CE analysis is explicitly required by government for drug price regulation. In European countries, CE analysis is implicitly promoted because a drug needs to be proven to be of sufficient value to justify its price during price negotiations. Before hospitals and medical agencies purchase a newly developed medication, they require information on the cost-effectiveness of the drug. Pharmaceutical companies strive to demonstrate that their products are more cost-effectiveness compared with other available drugs, thus sufficient funding flows into this pot for either in-house or outsourcing CE analysis.

CE analysis also plays an important role in other health care service in a less formal way. It is often used as evidence of a particular program, or intervention in order to promote or discourage its use. It is often commonly believed that, preventive interventions should be provided to all people for whom they are effective because they seem to save more money than they cost. However, CE analysis shows that depending on how they are applied, some effective forms of prevention come with significant costs. For example, screening for cervical cancer can help detect cancer at an early stage thereby reducing mortality rate, but screening too frequently e.g. annually will cost all payers more than $1 million per year of life saved compared to screening every 2 years whereas health gain is minimal according to Eddy[10]. Thus CE analysis is also a useful tool to determine cost-effective screening frequency for preventive interventions.

1.7 Three Examples

This dissertation includes three topics, all of which fall into cost-effectiveness analysis in a general sense. The first topic is to determine an optimal breast cancer screening policy in the presence of overdiagnosis (cancer which will not develop if left undetected), wherein health cost is life-time expected number of screening and health effect is measured as expected quality-adjusted life years (QALYs). Existing literature in determining the impact of overdiagnosis, which might lead to inappropriate screening recomming. The resulting consequence could be that more overdiagnosis cases are detected, and overtreatment is conducted. Our research remedies this neglect and provides cost-effective screening policies considering overdiagnosis. The second topic is cost-effectiveness analysis of treatment for hepatitis C disease. A series of new drugs such as sofosbuvir have been recently approved for the treatment of hepatitis C virus (HCV). They are characterized with high cure rates as well as high prices. Questions have been raised regarding whether these new treatments' benefits would justify their additional costs. In this project, cost-effectiveness of these new treatments is determined in comparison with standard care regimens. The third topic is aimed to help resolve congestion issues of intensive care units (ICUs). Due to limited resources, it is critical to effectively utilize medical units and maximize throughputs from a cost-effectiveness perspective. In this study, we assume that ICU patients in a less severe condition can be stepped down to medical-surgical unit units so that beds can be vacated to accommodate new arrivals. Under a range of scenarios stepdown policies are developed, analyzed and discussed.

1.7.1 Breast Cancer Screening Policies in the Presence of Overdiagnosis

Breast cancer is a major cause of death among women all over the world. In the US, one in eight women will develop breast cancer in their lifetime [32]. Although there is no guaranteed way to prevent breast cancer, the disease is more likely to be cured when detected early. Mammography screening contributes to the decrease of breast cancer mortality by increasing likelihood of detecting breast cancers at its preclinical stage. On average, it detects cancer 1.7 years [39] before a woman can feel the lump and several years before physical symptoms develop. Many countries have carried out nation-wide breast cancer screening programs. The central issue is how frequently screening should be conducted. This is important because mass screening is a costly service. Appropriate screening frequency has considerable impact on both cost and effect. Further, there is an illusion that the more frequent the screening, the better it is if screening cost is neglected. However, recent research has shown a serious risk associated with screening, namely "overdiagnosis" (i.e. tumors detected by mammogram, that otherwise would never develop into symptoms or increased mortality). According to a study in Bleyer et al.[21], over the past decade the introduction of mammography screening programs has doubled the number of early-stage breast cancer detection, whereas the number of following late-stage breast cancer cases has shown no corresponding drop. This imbalance implies that the newly detected breast cancers were not necessarily destined to progress if left undetected. There are further studies indicating that overdiagnosis rate can be as large as 50% [22-27]. Considering the significant presence of overdiagnosis, the original screening policy which has been officially suggested may no longer be cost-effective, as less

frequent screening not only reduces associated costs, but also will reduce the number of overdiagnosis detections. The goal of the research presented in Chapter 2 is to discuss optimal breast cancer screening policy considering overdiagnosis. A Markov-based model is applied to describe the progression of breast cancer wherein the overdiagnosis state is also captured. QALYs associated with a specific screening policy are used as health care effects and lifetime expected number of screenings is used as the heath care cost. Optimal screening policies are enumerated and compared under assumptions of different overdiagnosis rates.

1.7.2 Cost-effectiveness of Sofosbuvir-based Treatments for Chronic Hepatitis C in US

Hepatitis C is a contagious liver disease caused by the infection of the HCV. It is categorized by acute illness which lasts a few weeks and a chronic life long illness that can lead to long-term health problems including liver damage, liver failure, liver cancer or even death. It is estimated by the Centers for Disease Control (CDC) that 3.2 million persons in the US are infected with chronic Hepatitis C. Because it can be asymptomatic for many years, most people are unaware of their infection. The disease can be detected during routine blood tests to measure liver function and liver enzyme level. There are several medications available to treat chronic Hepatitis C (CHC). The standard care of treatment is interferon plus ribavirin (plus protease inhibitor for genotype 1), and is effective in 50% to 70% of patients with CHC of all genotypes. Recently, several new treatments including Harvoni, Olysio + Sovaldi, Viekira Pak (for genotype 1) and Sofosbuvir-based regimens (for all genotypes) characterized with potent inhibitors have

been approved by the Food and Drug Administration (FDA), providing more options for CHC patients. Trials have shown that the new treatments increased the average rate to 80% to 95%, though with substantial increases in the costs. In particular, current market pricing of a 12-week course of sofosbuvir is approximately $84,000. Thus, there is concern that the expenses of new treatments may not justify their clinical benefits. In this study, a Markov simulation model of CHC disease progression is used to evaluate the cost-effectiveness of different treatment strategies. The model calculates the expected lifetime medical costs and quality adjusted life years (QALYs) of hypothetical cohorts of identical patients, 52-year old, 64% male, treatment-native who have CHC with or without cirrhosis. Cost-effectiveness of new treatments is compared to the standard care of treatments.

1.7.3 ICU Stepdown Policies

ICUs are facilities providing care for the sickest and most unstable patients in a hospital. They typically provide the highest level of care with one nurse for every one or two patients. These units are very expensive to operate and typically require 20% of hospital operating costs despite only consisting of 10% of the beds[104]. Consequently, these units are often operated at or above capacity. Especially in recent years, hospitals increasingly have engaged in mergers, affiliations, downsizings, closings[103]. As a result, approximately 25% in the number of hospital beds nationwide has been reduced during last 20 years according to statistics from American Hospital Association[40], which makes limited resources even more scarce. Delays in receiving intensive care have many

adverse outcomes, for example, physiologic deterioration, longer lengths of stay etc. Hospitals have developed a number of approaches to deal with ICU congestion. For instance, ICU congestion can result in discharging current patients preemptively[87,88,92], blocking new patients via ambulance diversion[105] or rerouting patients to different units[93,97]. Our work proposes ICU early stepdown policies where ICU patients with less severe conditions are stepped down earlier to lower level care units (e.g., medical surgical units) to provide room for new arrivals. The service system is modeled as birth-death process and steady state distribution can be obtained by solving balanced equations, with which specific questions regarding when and how to early step down patients from the ICU to lower level care units under different scenarios are answered.

Chapter 2 Markov-based Breast Cancer Screening Policies in the Presence of Overdiagnosis

2.1 Introduction

Breast cancer is the most prevalent form of cancer among U.S. women and the second leading cause of cancer death.[11-13] Screening programs have been effective at reducing mortality[14] and in early stage detection of breast cancer.[15,16] A key issue for overall effectiveness of screening is the length of the screening interval. There is no consistency among agencies, however, in their recommendations for this length. The American Cancer Society recommends annual mammograms beginning at age 40.[17] The US Preventive Services Taskforce (USPSTF) recommends biannual screening mammography for women between the age of 50 and 74 years.[18]

A further complication for effective screening programs is overdiagnosis. This occurs when the diagnosed cancer progresses at a rate that does not lead to mortality[19] or when the cancer itself is not aggressive.[20] Overdiagnosis can lead to unnecessary treatment and subsequent post-surgery distress. A recent study estimated that after adjusting for the effect of hormone therapy, overdiagnosis accounted for 31% of breast cancers diagnosed in US women in 2008.[21] Several other studies, primarily based on randomized controlled trials, have estimated overdiagnosis anywhere from 1% to 50%.[22-27]

In the absence of better differentiation of detected breast cancer, overdiagnosis could impact the effectiveness of screening program recommendations. In this study, the impact of overdiagnosis on recommended screening intervals is conducted using a Markov model. As there is no agreement on the magnitude of overdiagnosis, quality adjusted life

years (QALYs) and expected number of mammograms over a lifetime are determined for policies for various levels of overdiagnosis.

2.2 Literature Review

In recent years, there has been little agreement about the value of mammography. Undoubtedly, mammography contributes to a significant reduction of breast cancer mortality, but it can also lead to overtreatment where women go through grueling therapies -- surgery, radiation, chemotherapy, that may not be necessary. The problem occurs because mammography may overdiagnose breast cancer, meaning that some cancers it finds would never progress or threaten a patient's life. However, the magnitude of overdiagnosis is based on theory since medical technology cannot tell whether a cancer is overdiagnosed or not.

The common approach to estimate the extent of overdiagnosis is to conduct randomized trials and compare incidence from screened and unscreened populations. During the early years of programs, due to the lead time effect of breast cancer, a lower incidence in the unscreened group is expected. Over time, this initial decrease would be fully compensated by a similar increase in later periods for the unscreened group if small tumors in the group are really life-threatening. They would have to grow large enough to be noticed or cause symptoms. However, in several studies, the incidence in the unscreened group never caught up with that of the screened group. Thus, researchers concluded that there must be women in the unscreened group who had cancer that was never diagnosed and never progressed.

The fact that overdiagnosis stands as a population-based statistic makes the precise estimation of over-diagnosis rate impossible. In the literature, overdiagnosis was estimated based on three randomized trials[23,44,45] including the Malmo trial and 2 Canadian trials. A meta-analysis[47] based on these 3 trials estimated that 19% of all cancers diagnosed during the screening period were overdiagnosed among women. In the most recent update of the Canadian studies[42], it was still found that there was a residual excess of 106 invasive cancers in the mammography group after 15 years, and it amounted to 22% of the 484 invasive cancers found by mammography. Due to different populations, assumptions and measurement methods, overdiagnosis rates have been estimated to fall in a rather wide range of 5-50%[22-27]. A recent study[21] based on SEER incidence and survival trends using historical incidence rates as a comparison reported that 31% of all breast cancers diagnosed in the United States represented overdiagnosis. There are also studies based on screening service and statistical modeling. A population based cohort study[27] of incidence of breast cancer during the introduction of nationwide screening programs in Norway and Sweden show that one third of invasive breast cancers in the age group 50-69 would not have been detected in the patients' lifetime. In addition, a Swedish study[25] of increasing incidence of invasive breast cancer after the introduction of screening showed a 21-54% excess incidence depending on age. Evaluation of the Nijmegen program in 1989, which used geographically distinct controls showed an excess of 11% of breast cancer over a 12 year period[37]. A study[22] based on two trails (Swedish two-county and Gothenburg trials) suggest a much lower rate of overdiagnosis, 1%. Also based on randomized trials, Gotzsche and Nielsen[38] estimate that screening leads to a reduction in breast cancer mortality of 15% and an increase in over-diagnosis

rate of 30%, which means that out of 2000 women invited for screening throughout 10 years, one will have her life saved, whereas 10 healthy women who otherwise will not be diagnosed and receive unnecessary treatments and more than 200 women experience psychological distress because of false positive findings. Zahl et al. [20] compared cumulative breast cancer incidence in age-matched cohorts of women aged 50-64 years residing in 4 Norwegian counties before and after the initiation of biennial mammography. Because the cumulative incidence among controls never reached that of the screened group, it appears that some breast cancers detected by repeated mammographic screening would not persist to be detectable by a single mammogram at the end of 6 years. This raises the possibility that the natural course of some screen-detected invasive breast cancers is to spontaneously regress. It is well known that many cases of carcinoma *in situ* in the breast do not develop into potentially lethal invasive disease. In contrast, screening for breast cancer also leads to over-diagnosis of invasive cancer.

Despite the debate about the actual over-diagnosis rate, the existence of over-diagnosis itself is no longer a negligible issue. To the best of our knowledge, among the ample literature addressing breast cancer screening policies, none has taken it into consideration. The ignorance of over-diagnosis may explain the bias of real data and screening policies proposed in literature. Therefore, the purpose of our research is to investigate the impact of over-diagnosis on screening policies and propose unbiased screening policies by adding over-diagnosis as a state in addition to the early-stage breast cancer state and late-stage breast cancer state widely adopted by literature.

Of the existing studies that provide efficient mammography screening recommendations, two types are involved. One is a personalized screening polices, where the policy is tailored to each specific patient's characteristics. For instance, Ayer et al.[32] proposed a personalized mammography screening policy based on prior screening history and personal risk characteristics of women. Clinically it is impossible to tell which patient's tumor is over-diagnosis, thus over-diagnosis exists in a statistical sense and it can only be discussed in the population based scope. The literature our paper directly connects to is Mailart et al.[28], wherein a partially observed Markov chain model was proposed with age-based dynamics and imperfect sensitivity and specificity. In this work, a broad range of policies were enumerated and measured with two metrics, lifetime mortality and expected number of mammograms. Analysis was performed to determine a set of efficient policies. In our paper, we revise the model to resolve the over-diagnosis problem by adding an over-diagnosis state, though a mechanism must be developed for estimating the transition probabilities. Further, instead of using mortality rate as a metric, we use QALYs because mortality does not capture the downside of screening; the more screenings prescribed, the higher morbidity is. Thus, to capture the costs of overdiagnosis and overtreatment, we resort to QALYs where cost of unnecessary screenings are recognized. As there is no consensus to the extent of over-diagnosis rate, we conduct a sensitivity analysis of 5%-50%. We perform numerical analysis to produce a menu of policies that efficiently balance QALYs and policy effort under different over-diagnosis rates, and make comparisons among them.

2.3 Methods

Maillart et al.[28] developed a partially observable Markov chain model with age-based dynamics and imperfect sensitivity and specificity. In their model, a broad range of policies were enumerated based on the metrics of lifetime mortality and expected number of mammograms. In the study presented here, our model is modified by including an overdiagnosis state. Further, since mortality does not capture the harm of unnecessary treatment in the presence of overdiagnosis, QALYs are used as a primary metric. Transition probabilities for the Markov chain were taken from the literature.

The problem is formulated as a Markov chain decision making process. Breast cancer progression is categorized to follow 5 states: no breast cancer, early stage invasive breast cancer, overdiagnosis state, advanced stage breast cancer, breast cancer induced death, and other cause induced death. During each time period the states can transit to each other following a transition probability matrix. At the beginning of each period, a mammogram may be prescribed. Based on the results, the patient's new state distribution will be updated. We allow the existence of false positives and false negatives. Specifically, if a mammogram result is positive and a perfect test (e.g. biopsy) confirms that the patient does have cancer, she will receive treatment and exits the model by accruing the amount of QALYs that depends on which state is being exited. If the perfect test shows that the patient does not have cancer, the patient will be updated to be in no breast cancer state. If the mammogram result is negative, no further procedures will be conducted to reveal the true status of the patient, and there is a Bayesian update on the patient's states. We assume that the decision process starts at age 25, which is generally

considered as cancer free age and it ends at age 100. For a screening policy, two measures will be computed, one is the expected number of mammograms during a lifetime, the other is corresponding expected QALYs.

Since there is no consensus to the extent of overdiagnosis, a sensitivity analysis over the range of values of 0%-50% was conducted. A menu of policies that efficiently balance QALYs and screening effort was determined, and comparisons among them were made. The policy includes the age that screening should start and the screening interval. The model allows for a time (switching age) at which a different screening interval may be used. A policy can therefore be defined by a vector. For example [40,1,50,2] would represent a policy with a starting age of 40 years, a screening interval of 1 year from 40 years to 49 years of age, and a screening interval of 2 years for age 50 years and above.

2.3.1 Markov Model Description

The American Joint Committee on Cancer[100] categorizes breast cancer progression from Stage 0 to IV based on the tumor size and degree of metastasis, where Stage 0 is *in situ* breast cancer and Stage I to IV are invasive breast cancers. A discrete time Markov chain model is applied to describe breast cancer progression with the following 5 states: no breast cancer (state 0), early stage invasive breast cancer (state 1), overdiagnosis state (state 1'), advanced stage breast cancer (state 2), breast cancer induced death (state 3), other cause induced death (state 4). Specifically state 1 includes Stage 0, I, and II without lymph node involvement; and state 2 include Stage III and IV with lymph node involvement. We choose to cluster in situ and early stage breast cancer to be consistent

with practice, as the distinction in terms of treatment and prognosis is most related to whether there is involvement of a lymph node or not. Overdiagnosis can occur to both *in situ* and early invasive breast cancer patients. Although *in situ* cancer might occur at a higher rate than early stage cancer, there is no way to differentiate between them, which is another reason we cluster these two states.

Since it is not possible clinically to determine if a patient's tumor is overdiagnosis, state 1' is not directly observable. Since most of overdiagnosis occurs in early stage breast cancer (state 1), a fraction of the original state 0 to state 1 transitions is determined for state 0 to state 1' transitions. This fraction is the overdiagnosis rate. The feature that distinguish state 1' from state 0 and state 1 is that a patient in state 1' is essentially healthy, but clinical tests including mammograms and more advanced procedures see no difference between the overdiagnosed tumor and early stage invasive cancer. Therefore, they are detected with the same probability as state 1 cancer and tend to be treated for disease similarly. A diagram of the Markov chain is shown in Figure 2.1. This model is a modification of that presented in Maillart et al., who do not consider overdiagnosis.[28]

The problem is formulated as a Markov chain decision making process. During each time period the states can transit to each other following the transition probability matrix. At the beginning of a period, a mammogram may be prescribed. The result can be positive or negative with probability as a function of patient's true disease status, which implies the existence of false positives and false negatives. If the result is negative, no further procedures will be conducted to reveal the true status of the patient. If the result is positive, we assume a perfect test (e.g. biopsy) will be subsequently conducted to reveal the true state of the patient. This assumption is reasonable as the literature reports that

biopsies are a reliable procedure.[48] If the patient does have cancer, she will be receive treatment and exits the model. If the perfect test shows that the patient does not have cancer, the patient will be updated to be in state 0. For a screening policy, two measures will be computed, one is expected number of mammograms during a lifetime, the other is corresponding expected QALYs. The reason we adopt QALYs as a measure instead of mortality like is that mortality metric fails to capture the harms of over-diagnosis and overtreatment.

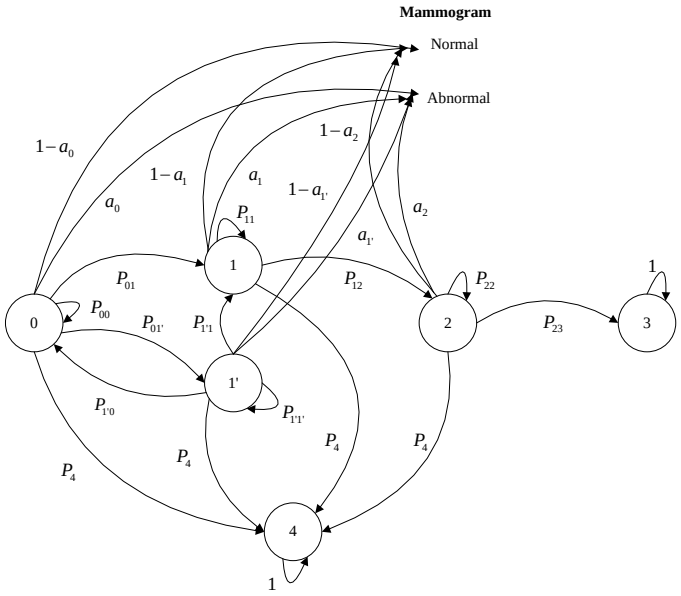

Figure 2.1 Markov chain transition diagram

2.3.2 Model Formulation

The notation for the model is as follows.

- *State space*: the patient's health is categorized to 5 states, no breast cancer (state 0), early-stage breast cancer including in situ and invasive early-stage breast cancer (state

1), over-diagnosis(state 1'), advanced invasive breast cancer (state 2), breast cancer induced death (state 3), other cause induced death (state 4).

- *Time horizon*: $n = 0,1,2,\cdots,T$. As we consider half year transition cycling, each epoch represents half a nominal year. The decision process starts at age 25, because this age is generally considered as cancer free and it ends at age 100 ($T = 150$). Time epoch n corresponds to age α_n .
- $p_{ij}(\alpha_n)$: transition probability from state i to state j for patient at age α_n , $i, j \in \{0,1,1',2,3,4\}$. In particular, the probability that the state goes back to itself is $p_{ii}(\alpha_n)$.
- $p_4(\alpha_n)$: probability of death from other causes than breast cancer (state 4) at age α_n . Given an individual is alive at the current state, she is subject to the same probability of death from other causes, $p_{04}(\alpha_n) = p_{14}(\alpha_n) = p_{1'4}(\alpha_n) = p_{24}(\alpha_n)$, thus we use $p_4(\alpha_n)$ to denote them uniformly for simplicity.
- $a_j(\alpha)$: probability that the mammography result is positive if the patient at age α_n is at state j, specifically, $a_0(\alpha)$ denotes the false positive probability that mammogram result is positive when actually the individual is cancer free, for $j = 1,1',2$; $a_j(\alpha)$ is the sensitivity (true positive).
- $1 - a_j(\alpha)$: probability of obtaining a negative mammogram result at state j, $1 - a_0(\alpha)$ is the specificity (true negative) because it is the probability that mammogram result is negative when the individual in fact is cancer free, whereas for $j = 1,1',2$, $1 - a_j(\alpha)$ is a false negative.
- $R_j(\alpha_n)$: lump sum QALYs which are received when the patient of age α_n at stage j is sent for treatment and exits the model (accruing an amount of QALYs depending on which state is being exited). $R_j(\alpha_n)$ can also be interpreted as a measurement of effective treatments. Intuitively, $R_j(\alpha_n)$ will be higher if the technology level is at a high level, because the patient's QALY will be lengthened accordingly.
- c^- : intermediate cost of QALYs provided mammogram result is false positive.
- c^+ : intermediate cost of QALYs provided mammogram result is true positive.
- c : intermediate cost of QALYs for a negative mammogram.
- l : permanent percentage loss of QALYs from treatment.
- $r_{n,j}$: intermediate QALY reward gained for period n provided the patient starts the current period with state j.

- $W_n(\pi)$: expected aggregate QALYs for an individual at time period n given the current state distribution $\pi = [\pi_0, \pi_1, \pi_{1'}, \pi_2]$, with time period n corresponding to age α_n. When no mammogram is conducted during this period, $W_n(\pi)$ is alternatively written as $DN_n(\pi)$, whereas denoted as $M_n(\pi)$ when there is mammogram prescribed in period n. $W_n(\pi)$ can be recursively represented as a function of next period's reward $W_{n+1}(\pi)$.
- $\pi'(\pi)$: updated state distribution given that starting in the last period with distribution π the patient does not receive a mammogram and survives the last period. The realization of this transition takes one time period.
- $\pi''(\pi)$: immediate updated state distribution based on the mammogram result given the initial distribution π. We assume a mammogram occurs at the beginning of a time period. Based on the test result, the patient's health distribution is updated.

The problem is formulated as a Markov chain dynamic program and a recursive method is developed to solve the problem. For simplicity, α_n is omitted when writing the formula. The case when no mammogram is prescribed at current period *n* is discussed first. Given the initial state distribution $\pi = [\pi_0, \pi_1, \pi_{1'}, \pi_2]$, the QALY rewards gained during period n are expressed as $\pi_0 r_{n,0}, \pi_1 r_{n,1}, \pi_{1'} r_{n,1'}, \pi_2 r_{n,2}$ corresponding to initial state $0, 1, 1', 2$. The probability that the patient dies from breast cancer and other causes during this period are $\pi_2 p_{23}$ and $(\pi_0 + \pi_1 + \pi_{1'} + \pi_1) p_4$ respectively. If the patient lives, future expected aggregate QALY $W_{n+1}(\pi'(\pi))$ are recognized.

$$
\begin{aligned}
W_n(\pi) &= DN_n(\pi) \\
&= \pi_0 r_{n,0} + \pi_1 r_{n,1} + \pi_{1'} r_{n,1'} + \pi_2 r_{n,2} + [1 - \pi_2 p_{23} - (\pi_0 + \pi_1 + \pi_{1'} + \pi_1) p_4] W_{n+1}(\pi'(\pi))
\end{aligned}
$$

Bayes' rule is applied to update the state distribution $\pi'(\pi)$ provided that the patient survives the last period.

$$\pi'(\pi)_0 = \frac{\pi_0 p_{00} + \pi_{1'} p_{1'0}}{1 - \pi_2 p_{23} - p_4},$$

$$\pi'(\pi)_1 = \frac{\pi_1 p_{11} + \pi_0 p_{01} + \pi_{1'} p_{1'1}}{1 - \pi_2 p_{23} - p_4},$$

$$\pi'(\pi)_{1'} = \frac{\pi_{1'} p_{1'1'} + \pi_0 p_{01'}}{1 - \pi_2 p_{23} - p_4},$$

$$\pi'(\pi)_2 = \frac{\pi_2 p_{22} + \pi_1 p_{12}}{1 - \pi_2 p_{23} - p_4}$$

Next, the case when the mammogram is prescribed at the beginning of period *n* is considered. If the mammogram result is positive, it is assumed that a biopsy is performed which will reveal the true state of the patient. $\pi_0 a_0, \pi_1 a_1, \pi_{1'} a_{1'}, \pi_2 a_2$ denotes the probabilities of having a positive mammogram result when true states are 0,1,1',2 respectively. Given a patient's true state is cancer free, the biopsy will update the process to $e_0 = [1, 0, 0, 0]$. If a patient is diagnosed as having cancer by the mammogram and confirmed by the biopsy, the patient will be sent for treatment and exits the model and a lump sum QALY R_j will be attainted. A loss of QALY c^+ is recognized if the mammogram result is a true positive, whereas a loss of QALY c^- is recognized if the mammogram result is a false positive. A negative result incurs a cost *c*. A 6% permanent loss of QALYs from treatment is assumed for l. In the case of a negative result, the state distribution will be updated.

$$\begin{aligned} W_n(\pi) &= M_n(\pi) \\ &= \pi_0 a_0 (DN_n(e_0) - c^-) + \pi_1 a_1 (R_1 - c^+) + \pi_{1'} a_{1'} (R_{1'} - c^-) + \pi_2 a_2 (R_2 - c^+) \\ &\quad + [\pi_0 (1 - a_0) + \pi_1 (1 - a_1) + \pi_{1'} (1 - a_{1'}) + \pi_2 (1 - a_2)](DN_n(\pi''(\pi)) - c) \end{aligned}$$

Similarly, $\pi''(\pi)$ is updated according to Bayes' rule, provided that the individual survives period *n*.

$$\pi''(\pi)_j = \frac{\pi_j(1-a_j)}{\pi_0(1-a_0)+\pi_1(1-a_1)+\pi_{1'}(1-a_{1'})+\pi_2(1-a_2)}, \quad j = 0,1,1',2$$

In terms of the boundary condition, it is assumed at the end of the time horizon that $W_T(\pi) = 0$. $W_0(\pi)$ can be obtained by using dynamic programming. The expected number of mammograms during a lifetime can be computed as:

$$\sum_{j:W_j(\pi)=M_j(\pi)} \prod_{n=1}^{j} \frac{p_{00}(\alpha_n)}{1-p_{01}(\alpha_n)-p_{01'}(\alpha_n)}$$

2.3.3 Parameter inputs

Data regarding the age based transition probability matrix and the specificity and sensitivity rates is from Maillart et al.[28] The sensitivity analysis of overdiagnosis rate used is 0% to 50%. The original state 1 from Maillart et al[28] include both patients with progressive cancer as well as those that are overdiagnosed. We partition these two groups into two states (1,1'). For those who transit to state 1', we assume that they are absorbed in this overdiagnosis state and they are considered to be an aggregate fraction that will not advance to a later stage cancer. Since it is clinically impossible to differentiate these two groups, we cannot observe how they are partitioned directly; rather we will utilize overdiagnosis rate definition and reverse engineer to approximate the partition. Suppose that for the transitions from state 0 to original state 1, x percent transits to state 1 and (1-x) percent transits to state 1', we call this x as a proxy. In the literature, overdiagnosis rate is

estimated within randomized trial framework and is generally computed as the excess number of cancers detected between the control group and the study group normalized by the total numbers of screen detected cancers in the study group[23]. For example, women aged 40-59 were randomly assigned to five annual mammography screen program and no mammography control program, after 15 years of follow-up, a residual excess of 106 cancers was observed in the mammography arm attributable to over-diagnosis, accounting for 22% of overall screen detected cancers[49]. We mimic the randomized trial with the Markov model and pursue an x which renders the correct overdiagnosis rate. The procedure is to first assume a partition x, an identical cohort of women at 25 years old free from cancer follow natural course of breast cancer progression without screening until an age between 40-59, then annual mammogram is performed for 5 consecutive years, during which number of detected overdiagnosis cases and overall screening detected cancers will be revealed, furthermore, model based overdiagnosis rate can be calculated, if it equals to observed overdiagnosis rate from literature, that corresponding x is the correct proxy, otherwise, we adjust x until it matches the overdiagnosis rate from literature. Thus, in this way, given an overdiagnosis rate, we can compute the corresponding x to be used as an input to the Markov chain. When overdiagnosis rate is assumed from 0% to 50%, corresponding x is given in Table 2.1 below.

Overdiagnosis rate	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
x	0%	4%	7.5%	11.5%	15.5%	20%	24%	28.5%	32.5%	37%	42%

Table 2.1. Proxy x value corresponding to overdiagnosis rate

The lump-sum rewards are obtained using age-specific mortality rates[28] for patients under cancer treatment based on the method described in Arias.[31] c^-, c^+, c can be estimated according to Ayer et al.[32] The intermediate QALY reward, $r_{n,j}$, gained for period *n* provided the patient starts the current period in state *j*, naturally the it is assumed the patient lives a full life (QALY=0.5) in a cancer free condition, whereas half of 0.5 years is assigned if the patient dies during the epoch for whatever reason. The states are formulated following the widely employed half-cycle correction method.[33]

$$r_{n,j} = 0.5 \cdot P(\text{live through current period} \mid \text{current state is j}) + 0.25 \cdot P(\text{die in current period} \mid \text{current state is j})$$

For example, the probability of death during period *n* starting at state 2 is $p_{23}(\alpha_n) + p_{24}(\alpha_n)$. Hence:

$$r_{n,2} = 0.5(1-(p_{23}(\alpha_n) + p_{24}(\alpha_n))) + 0.25(p_{23}(\alpha_n) + p_{24}(\alpha_n)) = 0.5 - 0.25(p_{23}(\alpha_n) + p_{24}(\alpha_n))$$

Similarly:

$$r_{n,0} = 0.5 - 0.25 \cdot p_4(\alpha_n),\ r_{n,1} = 0.5 - 0.25 \cdot p_4(\alpha_n),\ r_{n,1'} = 0.5 - 0.25 \cdot p_4(\alpha_n)$$

Parameter	Estimation	Source	Sensitivity Analysis
$p_{ii}(\alpha_n)$	—	Maillart et al[28]	N
$a_j(\alpha_n)$	—	Maillart et al.[28]	Y
$R_j(\alpha_n)$	—	Maillart et al.[28] Arias[31]	Y
c^-	$4/26 = 0.1538$	Ayer et al.[32]	Y
c^+	$2/26 = 0.0769$	Ayer et al.[32]	Y
c	$1/365 = 0.0027$	Ayer et al.[32]	Y
$r_{n,j}$	$0.5P(\text{live through current period} \mid \text{state}$ $+0.25P(\text{die in current period} \mid \text{state j})$	half-cycle correction[33]	N

Table 2.2 Input parameters

For a given screening policy, the model will output two performance measures: expected number of mammograms and QALYs. Screening policies are enumerated as characterized by five elements: starting age, first screening interval, switching age, second screening interval and ending age. The starting age is assumed to be 25 years, which is a commonly accepted as a cancer free age. The ending age is assumed to be 100 years. For example, [30, 4, 50, 1, 100] corresponds to policy starting at age 30 years, screen every 4 years until age 50, then screen annually. The constraints for the five elements are listed in the Table 2.3.

starting screening age	30-60 years in 5-year increments
first screening interval	0.5 year, 1 year, 2 years, 3 years, 4 years
switching age	40-60 years in 5-year increments
second switching interval	0.5 year, 1 year, 2 years, 3 years, 4 years
ending screening age	35-100 years in 5-year increments

Table 2.3 Requirements on screening policies

Sensitivity analysis of overdiagnosis rate ranging from 0% to 50% with a 5% interval was conducted. A comparison of the optimal policies under different overdiagnosis rates was performed. The full results are shown in Table 2.6.

2.4 Results

The results for screening policies when there is no overdiagnosis are shown in Figure 2.2. The points in the figure form a scatter plot of all enumerated policies. The annual screening policy suggested by American Cancer Society[17] is denoted by the green

triangle. For a given expected number of mammograms, the upper bound of the scatter plot (Pareto frontier) dominates the points below it. Only points on the increasing part of the Pareto frontier are efficient (represented by the red blocks) since for a given level of QALYs the corresponding point on the increasing frontier requires a fewer number of mammograms than a point on the decreasing part of the frontier. This differs from the monotone increasing shape of expected number of mammograms versus mortality rates in Maillart et al.[28], because in terms of mortality there is no downside to prescribing more mammograms. However, when we consider QALYs, harms regarding the negative effect of too frequently prescribed mammograms are captured.

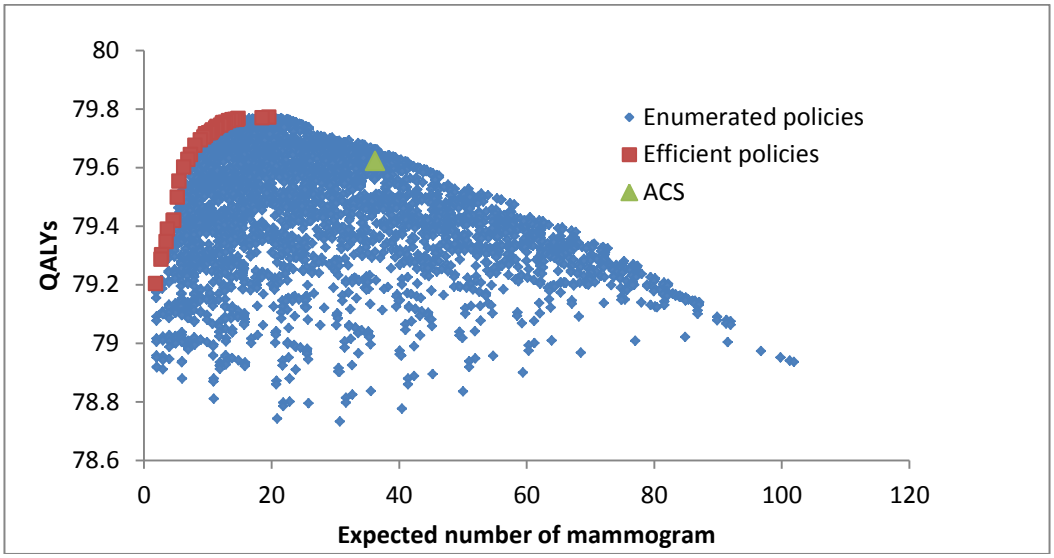

Figure 2.2 Scatter plot of screening policies without overdiagnosis.

Figure 2.3 shows the efficient portions of the Pareto frontier for the cases when overdiagnosis is present at levels from 0% to 50%. When the overdiagnosis rate is very high, it implies that fewer people have invasive cancer, resulting in higher QALYs. This

does not mean, however, that higher overdiagnosis rate has a positive effect on QALYs. Instead it reflects the modeling error arising from the ignorance of overdiagnosis.

A summary of the optimal screening policies from the Markov model are shown in Table 2.4. Full results are provided in Table 2.6. As the overdiagnosis rate increases, screenings will more likely detect overdiagnosis cases, which subsequently results in unnecessary treatment. When the overdiagnosis rate is high, therefore, patients should have less frequent screening to avoid overdiagnosis detection.

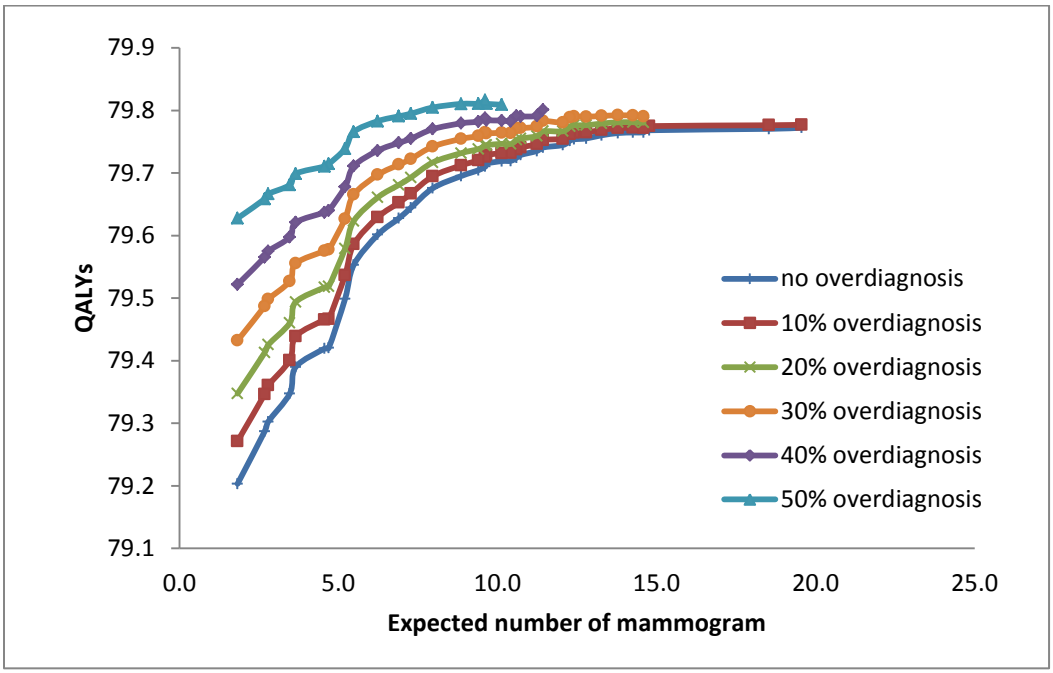

Figure 2.3 Comparisons under different overdiagnosis rates

The following example from the results illustrates the impact of overdiagnosis. If there is no overdiagnosis present, the efficient policy [35,3,50,2,85] would be adopted with a corresponding expected lifetime number of mammograms of 19.56 and QALY value of 79.7722. If, however, the actual overdiagnosis rate was unknowingly at 25%, policy [35,3,NA,NA,85] from Table 2.6 would yield even higher QALYs but only with

an expected number of mammograms of 14.77. Therefore, the cost of ignoring overdiagnosis in this case would be (19.56-14.77)/19.56= 24.5% additional mammograms. From another perspective, if the actual overdiagnosis rate is 25%, policy [35,3,50,2,85] results in 19.56 expected number of mammograms and 79.7864 QALYs according to the Table 2.6, and policy [35,3,NA,NA,85] with an expected number of mammograms of 14.77 and QALYs of 79.7880 outperforms [35,3,50,2,85] on both metrics. Hence, the harm of ignoring overdiagnosis is two-fold, additional screening effort with a lower benefit as measure by QALYs. As the level of overdiagnosis increases, the harms of ignoring overdiagnosis increases as well.

Rate of Overdiagnosis	Starting Age	Screening Interval	Switching Age	Screening Interval	Ending Age	Expected Number Mammograms over Lifetime
0%	35	3	40	2	90	22.4
10%	35	3	40	2	90	22.4
20%	35	3	50	2	85	19.6
30%	35	3	50	2	85	19.6
40%	35	3	(none)	3	90	15.4
50%	35	4	50	3	90	14.4

Table 2.4 Optimal policies for various levels of overdiagnosis.

2.5 Sensitivity analysis

Since we have conducted a sensitivity analysis on the overdiagnosis rate, we will change other input parameters by keeping the overdiagnosis rate constant at 0. We vary non-population based input data by ±5% and see how much the efficient frontier has deviates. We define percentage of frontier deviation as follows:

$$\frac{\text{number of no longer efficient policies with 5\% increase and -5\% decrease}}{2 \times \text{original number of efficient policies}}$$

Input data changed	Frontier deviation (%)
Sensitivity and Specificity	8.62%
Posttreatment survival	5.17%
Cost (including c^-, c^+, c)	6.90%

Table 2.5 Sensitivity analysis on input parameters

As seen from the results, the average deviations are very small. Thus the model is robust with respect to the input data.

2.6 Conclusions

Although the magnitude of overdiagnosis in breast cancer screening is not agreed upon in the literature, it is generally reported to occur at a significant level. In this study it was found that overdiagnosis had a negative impact on the effectiveness of screening policies. In particular, efficient screening intervals determined from a Markov model are increasing in the overdiagnosis rate. Further, ignoring overdiagnosis levels will lead to policies that have both a higher number of expected mammograms over a patient's life and a decrease in QALYs.

Although the 2009 USPTFS recommendations[18]of less frequent screening than previous recommendations created a controversy when published,[29,30] a recent study has shown that it has not lead to fewer mammograms in the US.[30] The findings here support

the USPTFS recommended biannual screening intervals for even moderate levels of overdiagnosis. If the rate of overdiagnosis exceeds 25%, then an even greater interval of 3 years may be beneficial.

Advances in technology that would allow the identification of early stage cancers that will not progress to later stages would of course increase the effectiveness of more frequent screening. In the absence of that technology, however, the benefit of early detection from shorter screening intervals must be weighed against the harms associated with overtreatment and more frequent screening.

2.7 Limitations

The use of overdiagnosis rate in the paper may raise some concerns, considering that what we use is a proxy. The overdiagnosis rate is difficult to measure directly since it primarily appears as an output of trials. Our approach is to reverse engineer the value and used it as an input to study screening policies. There will be some inevitable errors, but this approach does help to inform how to adjust screening policies in accordance with hidden overdiagnosis rates.

Different ages may have different overdiagnosis rates. However, we do not have accurate information on changes with age. We therefore assume overdiagnosis rate is constant across all ages for each case and then conduct a wide range of sensitivity analysis on the rates. If more accurate information about how overdiagnosis rate varies with age becomes available, the model may easily be modified to incorporate it.

There are other limitations that should be mentioned. First, our model of breast cancer progression is limited to a few states. We could include additional stages to represent the progression of breast cancer accurately if transition data were available. Second, we did not consider lead time issues. Finally, our use of Schairer data[34], also used in Maillart et al.[28] and Ayer et al.[32], may underestimate mortality risk for both younger and older ages. Therefore, the model may be biased against policies that start early and end later.

Our current research on breast cancer screening assumes that over-diagnosis rate is age independent. However, there is evidence indicating that younger women are more likely to be overdiagnosed than older women. This issue will be investigated in future studies and the model will be reanalyzed with the incorporation of age dependent overdiagnosis rates. Further, our recommended screening policy is population based, thereby failing to consider prior screening history and personal risk characteristics such as age, and family history. As a next step we will extend the model in Ayer et al. [32], where a personalized mammography screening policy is proposed, to capture the presence of over-diagnosis.

Starting age	First interval	Switching age	Second interval	Ending age	Num of mammograms	QALYs under different overdiagnosis rate										
						0	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
35	3	50	2	85	19.5637	79.7722	79.7747	79.7770	79.7798	79.7828	79.7864	79.7898	79.7939	79.7978	79.8025	79.8083
35	3	50	2	80	18.5375	79.7705	79.7735	79.7762	79.7796	79.7831	79.7873	79.7914	79.7962	79.8007	79.8062	79.8129
35	3	NA	NA	85	14.7745	79.7682	79.7717	79.7750	79.7789	79.7830	79.7880	79.7927	79.7984	79.8038	79.8104	79.8183
35	4	45	3	85	14.5903	79.7649	79.7686	79.7720	79.7760	79.7804	79.7856	79.7905	79.7965	79.8021	79.8090	79.8173
35	3	NA	NA	80	14.2624	79.7647	79.7686	79.7722	79.7765	79.7811	79.7866	79.7918	79.7980	79.8040	79.8112	79.8198
35	4	50	3	85	13.7881	79.7638	79.7679	79.7717	79.7762	79.7810	79.7867	79.7921	79.7986	79.8048	79.8123	79.8212
35	4	50	3	80	13.2760	79.7602	79.7647	79.7688	79.7738	79.7790	79.7853	79.7912	79.7983	79.8050	79.8130	79.8228
40	4	50	3	85	12.7892	79.7552	79.7603	79.7649	79.7705	79.7763	79.7833	79.7898	79.7975	79.8048	79.8136	79.8240
40	3	NA	NA	80	12.4105	79.7538	79.7592	79.7642	79.7701	79.7763	79.7836	79.7905	79.7987	79.8065	79.8157	79.8267
40	4	50	3	80	12.2771	79.7516	79.7571	79.7621	79.7681	79.7744	79.7819	79.7889	79.7972	79.8050	79.8144	79.8255
40	3	60	4	85	12.0540	79.7446	79.7499	79.7547	79.7605	79.7666	79.7738	79.7806	79.7887	79.7964	79.8056	79.8166
40	4	55	3	80	11.4331	79.7408	79.7470	79.7527	79.7595	79.7666	79.7750	79.7829	79.7923	79.8011	79.8115	79.8240
40	3	55	4	85	11.2396	79.7352	79.7409	79.7461	79.7524	79.7590	79.7668	79.7742	79.7830	79.7914	79.8013	79.8133
40	3	55	4	80	10.7275	79.7288	79.7351	79.7408	79.7477	79.7550	79.7636	79.7716	79.7813	79.7903	79.8012	79.8141
40	4	60	3	80	10.5979	79.7244	79.7314	79.7377	79.7452	79.7531	79.7625	79.7713	79.7817	79.7915	79.8032	79.8171
40	4	NA	NA	90	10.4211	79.7194	79.7259	79.7318	79.7389	79.7463	79.7552	79.7635	79.7733	79.7827	79.7938	79.8072
40	4	NA	NA	85	10.1340	79.7189	79.7256	79.7317	79.7389	79.7466	79.7556	79.7642	79.7743	79.7839	79.7953	79.8090
40	4	NA	NA	80	9.6528	79.7151	79.7222	79.7287	79.7365	79.7446	79.7543	79.7634	79.7742	79.7843	79.7964	79.8109
45	3	60	4	80	9.6167	79.7103	79.7185	79.7260	79.7348	79.7440	79.7548	79.7648	79.7767	79.7877	79.8008	79.8162
45	3	50	4	85	9.3985	79.7045	79.7127	79.7201	79.7290	79.7382	79.7490	79.7591	79.7710	79.7821	79.7953	79.8109
45	3	50	4	80	8.8534	79.6950	79.7039	79.7120	79.7217	79.7317	79.7434	79.7544	79.7673	79.7793	79.7936	79.8104
45	4	NA	NA	80	7.9641	79.6757	79.6857	79.6948	79.7056	79.7168	79.7300	79.7423	79.7567	79.7701	79.7860	79.8047
45	4	NA	NA	75	7.2754	79.6445	79.6563	79.6670	79.6796	79.6928	79.7082	79.7224	79.7392	79.7548	79.7732	79.7947
50	4	NA	NA	80	6.8994	79.6276	79.6408	79.6526	79.6666	79.6810	79.6979	79.7134	79.7315	79.7482	79.7678	79.7906
50	4	NA	NA	75	6.2353	79.6014	79.6161	79.6293	79.6449	79.6610	79.6797	79.6970	79.7172	79.7358	79.7576	79.7829
50	4	NA	NA	70	5.4777	79.5532	79.5705	79.5861	79.6044	79.6233	79.6453	79.6656	79.6892	79.7110	79.7363	79.7658
55	4	NA	NA	75	5.2032	79.4987	79.5187	79.5365	79.5575	79.5791	79.6040	79.6269	79.6534	79.6777	79.7060	79.7386
50	4	60	3	65	4.6931	79.4206	79.4450	79.4668	79.4925	79.5189	79.5495	79.5775	79.6100	79.6399	79.6745	79.7144
55	3	60	4	70	4.5649	79.4195	79.4438	79.4656	79.4911	79.5173	79.5477	79.5754	79.6076	79.6370	79.6712	79.7104
55	4	NA	NA	70	3.6523	79.3900	79.4158	79.4390	79.4660	79.4939	79.5261	79.5555	79.5897	79.6209	79.6572	79.6989
60	4	NA	NA	75	3.4670	79.3476	79.3757	79.4008	79.4302	79.4604	79.4952	79.5270	79.5637	79.5973	79.6362	79.6809
55	4	NA	NA	65	2.7886	79.3027	79.3332	79.3606	79.3926	79.4255	79.4635	79.4982	79.5384	79.5751	79.6176	79.6665
60	4	NA	NA	70	2.6748	79.2870	79.3184	79.3464	79.3792	79.4128	79.4516	79.4870	79.5280	79.5654	79.6087	79.6584
60	4	NA	NA	65	1.8231	79.2033	79.2392	79.2714	79.3089	79.3474	79.3918	79.4323	79.4791	79.5218	79.5711	79.6276

Table 2.6 Enumeration of frontier screening policies under different overdiagnosis rates.

Chapter 3 Cost-effectiveness of Sofosbuvir-based Treatments for Chronic Hepatitis C in US

3.1 Introduction

Chronic hepatitis C (CHC) is the leading cause of chronic liver disease and the primary reason for liver transplantation.[51,52] Approximately 170 million people worldwide are infected with hepatitis C virus (HCV), including 4 million people in the US.[53,54] CHC can go undetected for years, and once the symptoms do appear, liver damage has begun.[55] Approximately 42% of CHC patients will develop cirrhosis in their lifetime. [56] Further, 23% of these patients, if untreated, will eventually develop hepatocellular carcinoma, the primary cause of liver disease induced mortality. [57] In advanced stages of cirrhosis, liver transplantation is typically the only treatment option. [58]

In the last few years, the standard of care for untreated CHC patients changed from dual therapy with peginterfeon and ribavirin to triple treatment with peginterferon, ribavirin plus protease inhibitors (PI) e.g. telaprevir or boceprevir.[80] Although fairly effective compared to the old dual therapy, this triple therapy cannot achieve more than 75% sustained virologic response (SVR) [81], which is defined as HCV RNA less than lower limit of quantification (LLOQ) at 12 weeks after the end of treatment. Once SVR is achieved, relapse is very unlikely. However, injected interferon can lead to severe side effects such as fatigue, depression, and emotional liability.[52]

In Dec 2013, sofosbuvir (brand name Sovaldi) as a new component of interferon-free oral regimen has been approved by the U.S. Food and Drug Administration (FDA) for

treating CHC. The drug eliminates the need for some patients to take interferon, specifically patients with genotypes 2 and 3.[62] These patients can use sofosbuvir alone with ribavirin, whereas patients with genotype 1 are recommended to take sofosbuvir in combination with peginterferon and ribavirin[62]. After that, there appeared a number of potent inhibitors which were approved as all-oral regimen to treat genotype 1 (Table 3.1). In Oct 2014, the combination of ledpasvir-sofosbuvir (Harvoni) was approved by the FDA for the treatment of genotype 1 CHC patients with or without cirrhosis.[82] One month later, the use of simeprevir (brand name Olysio) in combination with sofosbuvir was also approved for genotype 1 patients.[83] Another month later, Viekira Pak comprised of four medications (ombitasvir, paritaprevir, ritonavir and dasabuvir) was approved for genotype 1 patients as well.[84] These new treatments are characterized by significant increases in SVR.[75] The traditional regimen of peginterferon plus ribavirin is effective in 50% to 70% of patients with CHC. These new regimens as combinations of inhibitors increased the effective rate to 80% to 95%.[64,65,66,82,83,84] However, as a popular component of new treatments, current market pricing of a 12-week course of sofosbuvir alone costs roughly $84,000.[67,68] We determine the cost-effectiveness of sofosbuvir-involved treatments in comparison with interferon-based treatments. To date, such analysis has not been reported, except for a recent study that found sofosbuvir-based treatments to be cost-effective for incarcerated persons.[69]

Genotype	Treatment	Duration (weeks)
1	Harvoni	12
	Olysio + Sovaldi with or without Ribavirin	12 (no cirrhosis), 24 (cirrhosis)
	Viekira Pak + Ribavirin	12 (no cirrhosis), 24 (cirrhosis)
	Sovaldi + Peginterferon + Ribavirin	12
2	Sovaldi + Ribavirin	12
3	Sovaldi + Ribavirin	24

Table 3.1 FDA recommendation.[67]

3.2 Methods

3.2.1 Model Description

We apply a Markov simulation model of CHC disease progression to evaluate the cost-effectiveness of different treatment strategies for CHC. The model calculates the expected lifetime medical costs and quality adjusted life years (QALYs) of hypothetical cohorts of identical patients receiving certain treatments. Treatments are compared based on the ratio of the additional cost of the more costly treatment divided by the additional effectiveness of the treatment. Reference patient cohorts are defined according to the average characteristics, gender and age, obtained from the trials used in this study (52-year old, 64% male, treatment-native who have CHC with or without cirrhosis).

At the beginning of a period, each hypothetical patient receives a designated treatment. If the patient shows detectable HCV RNA by PCR test throughout the therapy, he/she is classified as a non-responder. If a patient is HCV negative during therapy and also negative in the test 12 weeks after treatment, we assume SVR is achieved.

Otherwise, a relapse occurs. Whether a non-responder or relapser will receive follow-up treatment depends on the specific treatment plan. After receiving treatment, each patient enters a Markov process based on the viral response result, and subsequent long-term prognosis of each treatment group is estimated using simulation and the cohort is tracked as a patient moves through different health states until death. Transitions are made annually based on natural disease progression and each year patients may remain in the same state or progress to subsequent stages of liver diseases by given probabilities until they die of liver disease or natural causes.

Specifically, for each patient, we determine a value for the expected natural life span by using age-specific mortality tables, such that the patient's lifetime represents that of a random patient drawn from the population. The durations of disease occur in each patient are lined up sequentially until death occurs. Thus, for each individual cohort member, we assign the age at death, cause of death, and time spent in each disease state, then statistics such as QALYs and lifetime medical costs can be calculated. In accordance with literature[56,70,71,72,73] , the costs and benefits are discounted at an annual rate of 3%. All costs are adjusted to 2014 U.S. dollars.

3.2.2 Genotypes

Hepatitis C is divided into six distinct genotypes Genotype 1 to 6 throughout the world with multiple subtypes in each genotype class. A genotype is a classification of a virus based on the generic materials in the RNA (Ribonucleic acid) strands of the virus. Generally, patients are only infected with one genotype. Genotype 1 is the most common

type of Hepatitis C genotype in the United States and the most difficult to treat. Individuals with genotypes 2 and 3 are almost three times more likely than individuals with genotype 1 to respond to therapy. Furthermore, when using combination therapy, the recommended duration of treatment depends on the genotype. Thus for physicians, knowing the genotype of Hepatitis C is helpful in making a therapeutic recommendation.

3.2.3 Efficacy Rates of Treatments

Efficacy data associated with sofosbuvir-based new treatments are extracted from five clinical studies.[64,65,66] These studies include a total of 1724 HCV mono-infected patients with genotype 1 to 6 CHC. It should be noted that these five trials targeted different patient cohorts. The primary ending point was SVR at 12 weeks after the end of treatment. In the NEUTRINO study (327 patients),[65] a 12-week treatment was evaluated with Sovaldi + peginterferon-alpha + ribavirin in treatment-native subjects with genotype 1,4,5,6. In the study, 90% of patients had a SVR with 89% for patients with genotype 1. The SVR rate was 92% among patients without cirrhosis and 80% among those with cirrhosis. Aiming at treatment-naive subjects with genotype 2 and 3, the FISSON study (499 patients) compared 12-week treatment with Sovaldi and ribavirin to a 24-week treatment with peginterferon-alpha plus ribavirin.[65] SVR categorized by genotype and cirrhosis is shown in Table 3.2. The FUSION study (201 patients) conducted experiments on patients previously treated with interferon with genotype 2 or 3 who either relapsed or failed to respond, and 12 or 16 weeks treatment with Sovaldi and ribavirin was performed.[64] The VALENCE trial (419 patients) showed that for treatment-

naive genotype 3 patients, Sovaldi plus ribavirin for 24 weeks treatment obtained a 93% SVR with no cirrhosis and a 92% SVR with cirrhosis.[66] For Harvoni treatment, phase 3 studies (ION-1,ION-2,ION-3, 1952 patients in total) have consistently shown SVR rates greater than 90% with a 12-week course in patients of genotype 1 CHC with or without cirrhosis.[82] According to COSMOS study (167 patients), SVR rates were 95% for non-cirrhotic patients with 12-week treatment of Olysio + Sovaldi, 100% for cirrhotic patients with 24-week treatment.[83] Viekira Pak+ ribavirin regimens were characterized with 95% SVR for non-cirrhotic patients with 12-week treatment (SAPPHIRE-I study, 631 patients), also 95% for cirrhotic patients with 24-week treatment (TURQOUISE-II study, 380 patients).[84] In all these trials, treatments were not guided by subjects' HCV RNA levels implying that no response guided algorithm was used.

As benchmarks, we consider pegylated interferon, ribavirin plus telaprevir therapy as the standard care for genotype 1 and pegylated interferon plus ribavirin as the standard care for genotypes 2 and 3. They are commonly accepted treatments and acknowledged to be cost-effective in previous literature. Telaprevir is given with peginterferon and ribavirin for the first 12 weeks of therapy, followed by an additional 12 or 36 weeks of peginterferon and ribavirin depending on the response during therapy. If HCV RNA levels are undetectable at week 12 of treatment, an additional 12 weeks of peginterferon and ribavirin should be received, otherwise an additional 24 week of peginterferon and ribavirin are expected. As sofosbuvir also works for relapsers and non-responders with genotypes 2 and 3, we design a follow-up treatment of sofosbuvir plus ribavirin for patients who experience prior interferon treatment failure. We assume that all patients who participate in sofosbuvir involved therapy either as initial or follow-up treatment

complete the whole course of treatment. Subsequent prognosis of patients who relapse after the whole treatment is assumed to be identical to those who never have treatment.

In this study, we discuss treatment strategies based on genotypes. For genotype 1, the following treatment strategies are compared: (1) peginterferon + ribavirin + telaprevir for 12 weeks, followed by an additional 12 or 24 weeks treatment of peginterferon + ribavirin dependent on HCV RNA level at week 12. (2) Harvoni treatment, 12 weeks; (3) Olysio + Sovaldi, 12 weeks for patients without cirrhosis, 24 weeks for patients with cirrhosis; (4) Viekira Pak + ribavirin, 12 weeks for patients without cirrhosis, 24 weeks for patients with cirrhosis; (5) sofosbuvir + peginterferon + ribavirin, 12 weeks for patients with or without cirrhosis. For genotype 2 and 3, treatment strategies include (1) peginterferon + ribavirin, 24 weeks for treatment-naive patients; (2) sofosbuvir + ribavirin, 12 weeks for patients with genotype 2, 24 weeks for genotype 3; (3) peginterferon + ribavirin as initial treatment, 24 weeks for patients with genotype 2/3, follow-up treatment with sofosbuvir + ribavirin for 12/16 weeks are performed on non-responders and relapsers.

Genotype	Treatment	Duration (weeks)	SVR(%)	Trials
1	Telaprevir+Peginterferon-alpha+Ribavirin	12 + 12 or 12 + 36	75	ADVANCE
	Harvoni	12	98	ION
	Olysio + Sovaldi	12	95	COSMOS
	Viekira Pak + Ribavirin	12	96	SAPPHIRE
	Sofosbuvir +Peginterferon+Ribavirin	12	92	NEUTRINO
2	Peginterferon+Ribavirin (Treatment-naive)	24	81	FISSION
	Sofosbuvir +Ribavirin (Treatment-naive)	12	97	FISSION
	Sofosbuvir +Ribavirin (Non-responders & Relapsers)	12	90	FUSION
		16	92	FUSION
3	Peginterferon+Ribavirin (Treatment-naive)	24	81	FISSION
	Sofosbuvir +Ribavirin (Treatment-naive)	24	93	VALENCE
	Sofosbuvir +Ribavirin (Non-responders & Relapsers)	12	37	FUSION
		16	63	FUSION

Table 3.2.1 Response rate for patients without cirrhosis.

Genotype	Treatment	Duration (weeks)	SVR(%)	Trials
1	Telaprevir+Peginterferon-alpha+Ribavirin	12 + 12 or 12 + 36	75	ADVANCE
	Harvoni	12	98	ION
	Olysio + Sovaldi	24	100	COSMOS
	Viekira Pak + Ribavirin	24	95	TURQOUISE
	Sofosbuvir +Peginterferon+Ribavirin (Treatment-naive)	12	80	NEUTRINO
2	Peginterferon+Ribavirin (Treatment-naive)	24	62	FISSION
	Sofosbuvir +Ribavirin (Treatment-naive)	12	83	FISSION
	Sofosbuvir +Ribavirin (Non-responders & Relapsers)	12	60	FUSION
		16	78	FUSION
3	Peginterferon+Ribavirin (Treatment-naive)	24	30	FISSION
	Sofosbuvir +Ribavirin (Treatment-naive)	24	92	FISSION
	Sofosbuvir +Ribavirin (Non-responders & Relapsers)	12	19	FUSION
		16	61	FUSION

Table 3.2.2 Response rate for patients with cirrhosis.

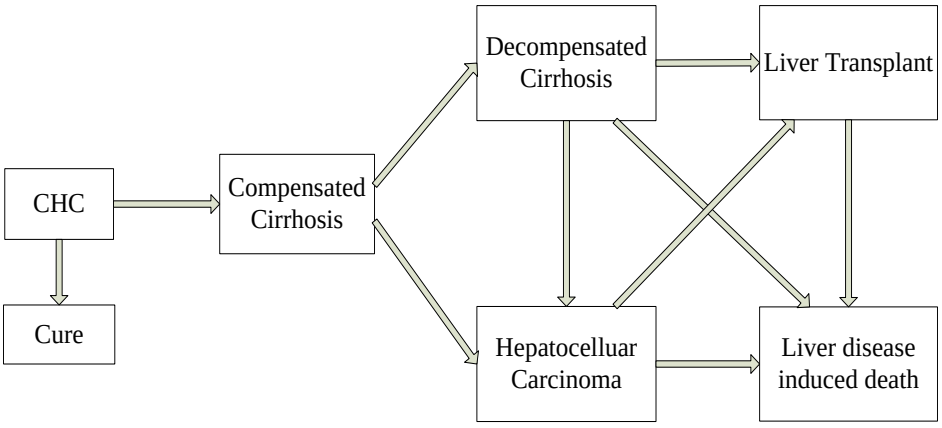

Figure 3.1 Natural history of HCV

The Markov simulation model includes the following CHC associated health states: treatment induced cure (healthy state), CHC, compensated cirrhosis, decompensated cirrhosis, hepatocellular carcinoma (HCC), liver transplant state, and liver disease induced death.[55,56] Transition relations are depicted in Figure 3.1.[74] In addition, patients at each health state are subject to the same age-dependent other cause induced death rate. Transition occurs annually and depends on health state-specific transition probabilities. Due to a lack of well-designed studies of patients with CHC, transition probabilities are estimated from the most widely quoted published data. Age-dependent death rates are obtained from the 2008 United States life table.[76]

Transition Probability	Baseline	Range	Reference
Chronic Hepatitis C			
to Compensated Cirrhosis	7.30%	1.0%-23.2%	[56,73]
Compensated Cirrhosis			
to Decompensated Cirrhosis	3.90%	2.0%-8.3%	[70,73,75]
to HCC	3.70%	1.0%-4.4%	[73,77,78]
Decompensated Cirrhosis			
to HCC	3.70%	1.0%-4.4%	[73,75,77]
to Liver Transplant	3%	1.0%-6.2%	[70,73]
to Liver-induced Death	12.90%	6.5%-19.3%	[70,72,73,75]
HCC			
to Liver Transplant	3%	1.0-6.2%	[70,73]
to Liver-induced Death	42.70%	33%-86%	[73,75]
Liver Transplant			
to Liver-induced Death, first year	13.70%	6%-42%	[73,79]
to Liver-induced Death, successive year	5.20%	2.4%-11%	[73,79]

Table 3.3 Annual transition probabilities.

3.2.4 Health-state Related Quality Adjusted Life Years

Quality of life specific to different health states are adjusted on an annual scale from 1 (perfect health) to 0 (death). Estimates of utilities were based on actual patients' utilities using the health utility index.[85] In order to estimate treatment-specific QALYs, the time spent in each health state was multiplied by each utility value and then summed over the life expectancy. As interferon based therapy has significant side effects, a 9% reduction in utility is assumed for interferon-based therapy.[70,73] Since most side effects are significantly more common in interferon containing regimens as compared to interferon-free ones, we assume that adding sofosbuvir in a regimen does not change QALY values.

QALYs	Baseline	Range	Reference
Uninfected	1	1	[85]
Chronic Hepatitis C	0.82	0.6-0.9	[85]
Compensated Cirrhosis	0.78	0.5-0.9	[85]
Decompensated Cirrhosis	0.65	0.3-0.88	[85]
HCC	0.25	0.1-0.5	[85]
Liver Transplant (1st year)	0.5	0.11-0.7	[85]
Liver Transplant (successive year)	0.7	0.24-0.87	[85]

Table 3.4 Health-state specific QALYs.

3.2.5 Medical Costs

Medical costs are summarized in Tables 3.5 and 3.6. Drug costs are estimated using approximate 2014 average wholesale acquisition price.[68,82,83,84] Other therapy related costs including screening, diagnostic and laboratory testing, drugs, monitoring costs during therapy and follow-up periods are estimated from the literature. Annual costs associated with each health state have been previously discussed and are inflated to 2014 US dollars using the medical care component of the Consumer Price Index.[86]

	Annual costs of care (2014 US$)	Reference
Chronic Hepatitis C	$572.69	[73]
Compensated Cirrhosis	$762.99	[73]
Decompensated Cirrhosis	$39,675.48	[73]
HCC	$25,862.67	[73]
Liver Transplant (1st year)	$483,057.01	[73]
Liver Transplant (successive year)	$46,515.46	[73]

Table 3.5 Annual costs of care.

Treatment	Cost per week (2014 US$)
Harvoni	$7,875.00
Olysio + Sovaldi	$12,500.00
Viekira Pak + Ribavirin	$7,000.00
Ribavirin	$250
Peginterferon+Ribavirin	$750
Sofosbuvir	$7,000

Table 3.6 Cost of treatments.

3.2.6 Sensitivity Analysis

In order to evaluate the robustness of the model, sensitivity analysis is performed for all parameters. Specifically, a 95% confidence interval is used for each entry of utility weights and natural history transition probabilities. Costs are halved and doubled. In terms of response rate, the model is reanalyzed for ±10% change of the value for each efficacy. A variable is considered to be potentially influential if it leads to the change of effectiveness for a treatment. In addition to one-way sensitivity analyses for all variables, we also conduct probabilistic sensitivity analyses. It is based on Monte Carlo simulation with 1000 runs, with parameters varied randomly according to associated distributions. This approach examines the effect of joint uncertainty in the model's variables. We assume that transition probabilities and utilities follow uniform distribution with range specified in Table 3.3&3.4, treatment efficacies follow Beta distribution and costs follow Gamma distribution all with standard deviation equal to 10% of the baseline value.

3.3 Results and Discussion

3.3.1 Base Case Analysis

Treatments aimed at the same group of patients (genotype, existence of cirrhosis) are compared. We consider results categorized by whether cirrhosis exists. The results for patients without cirrhosis are shown in Table 3.7.1, and the results for patients with cirrhosis are shown in Table 3.7.2. Incremental cost-effectiveness ratio (ICER) is defined as the ratio of the change in costs to incremental benefits of a medical intervention or treatment, specifically in our study it represents the additional money spent to gain one additional QALY. Note that the treatments are sorted according to Cost in ascending order. In both tables, ICER is calculated as additional Cost divided by additional Effectiveness between each treatment and the benchmark treatment. Specifically, all treatments are first compared to standard of care treatment (e.g. standard of care treatment for genotype 1 is Telaprevir + Peginterferon + Ribavirin), then compared to the adjacent efficient treatment, e.g. in Table 3.7.1, for genotype 2, in the second column of ICER, ICER between two-phase treatment with 24+12 versus Peginterferon + Riavirin is 4,233.09, ICER between two-phase treatment with 24+16 versus two-phase treatment with 24+12 is 44,457.83, and ICER between Sofobuvir + Ribavirin versus two-phase treatment with 24 + 16 is 1,805,952.38.

Genotype	Treatment	Duration (weeks)	Cost ($)	Effectiveness (QALYs)	ICER compared to benchmark ($/QALY)	Adjacent ICER ($/QALY)
1	Viekira Pak + Ribavirin	12	97,380	19.9659	* efficient	
	Harvoni	12	106,830	19.9618	* efficient	* inefficient
	Telaprevir + Peginterferon + Ribavirin for first 12 weeks, followed by additional 12 or 36 weeks of Peginterferon + Ribavirin	12 + 12 or 12 + 36	108,820	18.3364	* benchmark	* inefficient
	Sofosbuvir + Peginterferon + Ribavirin (Treatment-naive)	12	111,790	19.568	2,411.50 (* efficient)	* inefficient
	Olysio + Sovaldi	12	165,220	19.9356	35,267.63 (* efficient)	* inefficient
2	Peginterferon + Ribavirin (Treatment-naive)	24	45,560	18.7853	* benchmark	
	Initial: Peginterferon + Ribavirin; Follow-up: Sofosbuvir + Ribavirin for non-responders and relapsers.	24 + 12	50,340	19.9145	4,233.09 (* efficient)	4,233.09 (* efficient)
		24 + 16	54,030	19.9975	6,987.30 (* efficient)	44,457.83 (* efficient)
	Sofosbuvir + Ribavirin (Treatment-naive)	12	99,540	20.0227	43,623.73 (* efficient)	1,805,952.38 (* inefficient)
3	Peginterferon + Ribavirin (Treatment-naive)	24	52,810	18.2828	* benchmark	
	Initial: Peginterferon + Ribavirin; Follow-up: Sofosbuvir + Ribavirin for non-responders and relapsers.	24 +12	69,390	18.9351	25,417.75 (* efficient)	25,417.75 (* efficient)
		24 + 16	70,220	19.4892	14,431.37 (* efficient)	1,497.92 (* efficient)
	Sofosbuvir + Ribavirin (Treatment-naive)	24	187,880	19.8033	88,832.62 (* inefficient)	374,594.08 (* inefficient)

Table 3.7.1 Base case results for patients without cirrhosis.

Genotype	Treatment	Duration (weeks)	Cost ($)	Effectiveness (QALYs)	ICER compared to benchmark ($/QALY)	Adjacent ICER ($/QALY)
1	Harvoni	12	108,000	19.9787	* efficient	
	Telaprevir + Peginterferon + Ribavirin for first 12 weeks, followed by additional 12 or 36 weeks of Peginterferon + Ribavirin	12 + 12 or 12 + 36	122,420	17.2075	* benchmark	* inefficient
	Sofosbuvir + Peginterferon + Ribavirin (Treatment-naive)	12	130,750	17.8475	13,015.63 (* efficient)	* inefficient
	Viekira Pak + Ribavirin	24	186,820	19.7603	25,227.20 (* efficient)	* inefficient
	Olysio + Sovaldi	24	313,310	20.136	65,183.54 (* inefficient)	1,305,212.97 (* inefficient)
2	Peginterferon + Ribavirin (Treatment-naive)	24	81,070	15.874	* benchmark	
	Initial: Peginterferon + Ribavirin; Follow-up: Sofosbuvir + Ribavirin for non-responders and relapsers.	24 + 12	81,920	18.4461	330.47 (* efficient)	330.47 (* efficient)
		24 + 16	82,240	19.3293	338.61 (* efficient)	362.32 (* efficient)
	Sofosbuvir + Ribavirin (Treatment-naive)	12	119,940	18.4217	15,256.90 (* efficient)	* inefficient
3	Peginterferon + Ribavirin (Treatment-naive)	24	121,170	12.2543	* benchmark	
	Initial: Peginterferon + Ribavirin; Follow-up: Sofosbuvir + Ribavirin for non-responders and relapsers.	24 + 16	147,520	17.1869	5,342.01 (* efficient)	5,342.01 (* efficient)
		24 +12	162,490	13.8563	25,792.76 (* efficient)	* inefficient
	Sofosbuvir + Ribavirin (Treatment-naive)	24	191,280	19.3615	9,864.64 (* efficient)	20,123.24 (* efficient)

Table 3.7.2 Base case results for patients with cirrhosis.

For genotype 1 patients without cirrhosis, compared to the acknowledged efficient benchmark treatment (peginterferon + ribavirin + telaprevir), all new treatments are cost-

effective with ICER less than the threshold of $50,000/QALY. In particular, treatments Harvoni and Viekira Pak both achieve higher QALYs with reduced costs, which make the benchmark treatment no longer efficient, whereas treatments olysio + sovaldi and sofosbuvir + peginterferon + ribavirin both achieve higher QALYs but with increased costs compared to benchmark treatment. However, if compared to Viekira + Pak, Harvoni, olysio + sovaldi and sofosbucvir + peginterferon + ribavirin are no longer efficient characterized with higher costs and lower QALYs. Thus we conclude, all the four new regimens are alternatives of the current standard care of treatment (peginterferon + ribavirin + telaprevir), but of them all Viekira Pak is the most cost-effective for genotype 1 patients without cirrhosis, whereas the other three sofosbuvir based treatments are featured with higher costs and lower QALYs. For genotype 2 treatments, compared to standard care of treatment (peginterferon + ribavirin), all three sofosbuvir-based treatments are cost-effective. However, the comparative ICER of two-phase treatment with 16 weeks follow-up vs 12 week single treatment of sofosbuvir + ribavirin is $1,805,952.38/QALY, which is far beyond the threshold. It indicates that compared to two-phase treatment with 16 weeks follow-up, patients have to pay $1,805,952.38 for one additional QALY increase by adopting the 12 week single treatment of sofosbuvir + ribavirin, which is not cost-effective. Thus, for genotype 2, the two-phase treatments with peginterferon + ribavirin as initial and 12 & 16 week of sofosbuvir as follow-up are cost-effective whereas single treatment with sofosbuvir + ribavirin is not. For genotype 3, similarly, two-phase sofosbuvir-based treatments are cost-effective compared to standard care of treatment, and ICERs between adjacent treatments are also below the threshold. Whereas the 24 week single treatment with

sofosbuvir + ribavirin is not cost-effective compared to standard care of treatment with ICER $88,832.62/QALY, also not cost-effective compared to two-phase treatment with 16 weeks follow-up with an ICER of $374,594.08/QALY. Overall, single phase treatment with sofobuvir for both genotype 2 and 3 patients dominates two-phase treatments using peginterferon + ribavirin regimen as initial and sofosbuvir based new treatment as follow-up. The reason is attributable to the fact that traditional regimen (peginterferon + ribavirin) has rather good SVR for genotype 2 and 3, and the new treatment's additional SVR increase does not justify its much higher cost.

When patients with cirrhosis are considered, for genotype 1 all new treatments are efficient compared to standard care of treatment, whereas adjacent ICERs show that sofosbuvir + peginterferon + ribavrin and Viekira Pak are both inefficient compared to Harvoni with higher costs and lower QALYs. Although olysio + sovaldi has higher cost as well as higher QALYs compared to Harvoni, but ICER is $1,305,212.97 /QALY far beyond threshold, thus not efficient as well. Therefore, for genotype 2 patients with cirrhosis Harvoni is the most cost-effective treatment. For genotype 2, similar to the non-cirrhotic cases, both two-phase treatments with 12 week and 16 week as follow-up are cost-effective whereas 12 week single treatment with sofosbuvir + ribavirin is not cost-effective. Regarding genotype 3 treatments, 24 week single treatment with sofosbuvir + ribavirin is cost-effective, so is the two-phase treatment with 16 week follow-up. In general, sofosbuvir is not recommended as initial treatment for patients with genotype 2 but is recommended as initial treatment for genotype 3.

3.3.2 Sensitivity Analysis

One-way sensitivity analyses are performed on prices, transition probabilities, SVR rates, utility weights and cost of care. Only when the sofosbuvir price is reduced by at least 30%, olysio + sovaldi and sofosbuvir + peginterferon + ribavirin can achieve cost-effectiveness compared to Harvoni or Viekira Pak for genotype 1 patients. However, even when the sofosbuvir price is halved, using sofosbuvir as initial treatment is still not cost-effective compared to two-phase treatments that use sofosbuvir as follow-up treatment for genotypes 2 and 3. Further, increasing SVR for sofosbuvir-based treatments by 10% does make them cost-effective compared to Harvoni and Viekira Pak for genotype 1 patients, and reducing SVR for Havoni and Viekira Pak by 10% also makes sofosbuvir-based treatments cost-effective. Changing SVR rates of sofosbuvir-based treatments for genotypes 2 and 3 does not change effectiveness of the treatments. Two-phase treatments of 24 + 12 weeks are always cost-effective for both genotypes 2 and 3, compared to which, change in the values of cost of care or utility weights can push 24 + 16 weeks treatments' ICER beyond or below $50000/QALY benchmarked with the base case. However, it does not change the conclusion that single phase treatment of sofosbuvir + ribavirin has always been not cost-effective for both genotypes 2 and 3. Thus, for genotype 1 patients, reduction in sofosbuvir price or increase in sofosbuvir-based treatment SVR rates can improve the cost-effectiveness of sofosbuvir-based treatments compared to alternatives Harvoni and Viekira Pak. For genotypes 2 and 3, sofosbuvir-based new treatments serve better as follow-up treatments rather than initial treatments. Probabilistic analysis results are shown via the cost-effectiveness

acceptability curves (Figure 3.2), which are interpreted as the probability that the data are consistent within a true cost-effectiveness ratio falling below that value between two treatments. The comparison treatment pairs in our study include olysio + sovaldi versus Harvoni, Viekira Pak versus Harvoni, sofosbuvir + peginterferon + ribavirin versus Harvoni. From Figure 3.2.1, we can observe that ICER of olysio + sovaldi versus Harvoni falls between 0 and $50,000/QALY with approximate probability less than 5%, in other words with probability 95% it will fall either below 0 or above $50,000/QALY, implying that compared to Harvoni, olysio + sovaldi either achieves lower QALYs with higher costs or achieves higher QALYs with higher cost but with ICER exceeding the threshold $50,000/QALY. Thus with 95% confidence interval we conclude that olysio + sovaldi is not cost-effective. But for Viekira Pak and sofosbuvir + peginterferon + ribavirin, compared to Harvoni, they both have around 0.5 probability to be effective and 0.5 probability to be not effective according to Figure 3.2.2&3.2.3. For genotype 2, ICER of two-phase treatment with peginterferon + ribavirin as initial and 12 week sofosbuvir as follow-up versus standard of treatment is below $10,000/QALY with 95% probability.

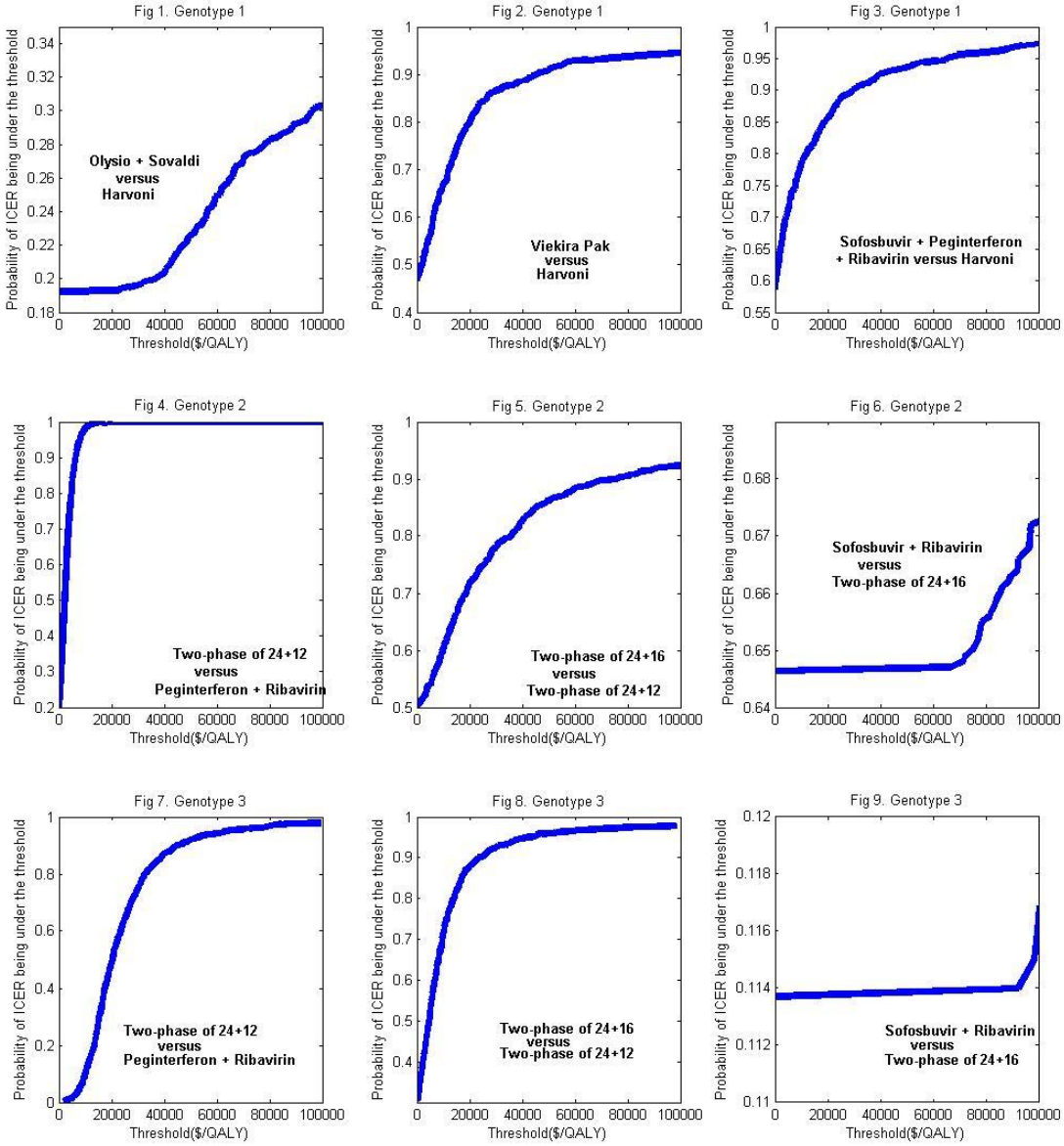

Figure 3.2 Acceptability curves

3.4 Conclusion

We analyzed the cost-effectiveness of sofosbuvir-based new treatments for genotypes 1, 2 and 3. Data regarding the natural history of hepatitis C, utility weights, and various costs and transition probabilities were obtained from the literature, and sustained

virologic response data associated with new treatments were extracted from clinical studies. Treatment strategies compared in this study were designed based on the data available while complying with drug dosage and administrative recommendations. It is important to note that results obtained in this study should be interpreted within the model assumptions. Our analyses conclude that Viekira Pak is cost-effective for genotype 1 patients without cirrhosis, while Harvoni is cost-effective for genotype 1 patients with cirrhosis. Sofosbuvir-based treatments for genotype 1 in general are not cost-effective due to its substantial high costs. For genotype 2, 3, generally speaking, sofosbuvir + ribavirin as initial treatment comes with a large increase in cost and small increase in effectiveness. Therefore, it is not recommended as initial treatment for patients with genotype 2 and 3, except for genotype 3 with cirrhosis, in which case, 24 week sofosbuvir + ribavirin treatment is cost-effective as it leads to much higher SVR compared to alternative treatments. In general, sofosbuvir + ribavirin are cost-effective as second-phase treatments following peginterferon + ribavirin initial treatment. To assure the robustness of our conclusions, we performed sensitivity analyses with wide ranges for all model parameters, and they did not significantly impact our final conclusions.

Chapter 4 ICU Stepdown Policies

4.1 Introduction

ICUs are medical units providing care for the sickest and most unstable patients in a hospital. They are typically the most richly staffed and always highly congested. Over the last 20 years, approximately 25% in the number of hospital beds nationwide has been reduced according to statistics from American Hospital Association[40]. This makes limited resources even more scarce. In fact, 90% of ICUs will not have the capacity to provide beds when needed[89]. Reports in the news media also indicate a nationwide increase in the number of hospitals turning away ambulances due to a lack of inpatient beds, as well as an increase in the frequency and duration of such diversions (New York Times 2002 [99]). After a patient experiences some trauma or completing surgery, he or she is admitted into ICU for further monitor and recovery. Although it is possible to hold patients in other areas temporarily, e.g., emergency department pending bed available (which is called boarding), it is rather undesirable to do so. Delays in providing intensive care can result in serious consequences and an increase in time spent by patients in emergency rooms and hallways waiting for a bed is strongly correlated with increasing mortality rate.[87] Therefore, it has become more and more important to develop efficient and cost-effective solution to resolve ICU congestion issues from medical perspective. In addition, despite the fact that ICU beds occupying only 5-10% of inpatient beds, they consume 20-35% of total health care costs[90] . Approximately $82 billion[102] is spent annually on ICU bed management, which makes it critical to improve the function of the

ICU. In this study, we propose early stepdown policies for the ICU to resolve congestion issues. We assume that ICU patients in a less severe condition can be stepped down and designated to lower level care units so that beds can be vacated to accommodate new arrivals. In particular, we confine our discussions to threshold policies where ICU patients are stepped down if and only if the total number of patients in the ICU reaches a certain threshold. We analyze scenarios when lower level care units (e.g. GCU) have infinite and limited capacity and investigate its impact on the optimal policies. The service system is modeled as birth-death process and steady state distribution can be obtained by solving balanced equations, with which specific questions regarding when and how to early step down patients from the ICU to lower level care units under different scenarios are answered.

4.2 Literature Review

There has been a significant amount of research that has tried to address facility congestion problems from bed capacity planning perspective [90,91,93]. However, prospective scheduling for ICUs is difficult due to the stochastic nature of arrival and service processes. Conventionally, it is common practice that the bed requirement are derived as an average of number of daily admissions times the average length of stay divided by the average bed occupancy rate [90]. This estimate ignores the stochastic nature of the issue and gives a rather rough estimate, typically underestimating true demand for beds due to the averaging.

Queuing theory has been a particularly useful modeling framework for the question of capacity, staffing and other tactical decision in the health care area. In Huang et al.[90], stochasticity is captured by assuming that patient admissions follow a Poisson distribution and the patients' length of stay follows a general distribution. Shonick and Jackson [94] present a stochastic model for the behavior of the daily census in general-acute hospitals, where only emergency patients are admitted when the number of occupied beds reaches a threshold. Among the health care literature on bed capacity planning, most papers deal with the determination of bed requirements for specific patient care units. For example, Sissours and Moore [96] determine bed needs for cardiac care units. Thompson and Fette [97] study the requirements for maternity facilities, and Kao and Tung [91] present an approach for periodic bed reallocations due to changing demand patterns to minimize expected overflow. Although it is very important to plan ahead for the number of beds in each unit, we cannot get around the uncertainty of patient arrivals. Thus in the face of given bed capacity, a strategy is needed for cases when units are experiencing overcrowding.

On the other hand, considering that ICUs are characterized with high risk and constrained capacity, expanding or reducing the capacity is not always a viable option. Some hospitals resort to solutions of discharging patients early or rationing admissions in the first place. Diwas and Terwiesch [92] study the ICU of a cardiac surgery service and show that when the ICUs are congested, the treatment of patients tends to speed up. There are also hospitals which discharge a patient currently residing in the ICU to accommodate a newly admitted patient because new arrivals are typically very high priority patients [88]. Patients that are discharged early may require readmission, however,

due to the high risk of physiological deterioration and can therefore impose an additional load on ICU resources. Chan et al. [88] study the impact of several different ICU discharge strategies on patient mortality and total readmission load where patients are prioritized and discharged based on measures of criticality. Shmueli et al. [95] consider different admission policies including first come first serve (FCFS) and bed specific hurdle (BSH) in order to maximize the expected incremental number of lives saved from operating ICU. Kim et al. [93] extended this research by using additional controls in the estimation with detailed data.

The research that is the closest to ours is Zhu et al.[98] who use stepdown units as intermediate level of care between ICUs and GCUs. Namely, patients with higher severity but that do not require intensive monitoring are stepped down. In this way, pressure is removed from ICUs, which results in higher efficiency. Rather than focusing on how to allocate resources between different units, our research studies when the beds in ICUs and GCUs are fixed, how to optimize stepdowns of patients from ICUs to GCUs such that hospital resources are fully utilized. In this way, patient inflow pressure is relieved from ICUs. We model units loading as multi-source multi-server queues, and we restrict our search to threshold policies, i.e. when the number of patients reaches a threshold in ICUs, early stepdown is conducted. Under given costs related to rejection of ICU and GCU patients, an optimal threshold can be obtained by solving steady state birth-death balance equations. We also study various stepdown policies based on different priority rules.

4.3 Models and Results

The ICU is the designated location for the care of the sickest and most unstable patients in a hospital. Critically ill patients, who may be admitted to a hospital due to multiple illnesses, including trauma, need urgent admission to the ICU. While it is possible to hold these patients in other areas pending bed availability e.g. the emergency department, this is quite undesirable, since delays in providing intensive care are associated with worse outcomes. Consequently, in such situations, clinicians may elect to discharge or step down a patient currently in the ICU to make room for a more acute patient. We refer to this as a demand-driven stepdown. The patient selected for such stepdown must be sufficiently stable to be transferred to a less richly staffed setting (such as the General Care Unit (GCU) or Medical Surgical Unit (MSU)). There are various ways to measure patient criticality such as the Acute Physiology, Age, Chronic Health Evaluation (APACHE) score.[101] We assume there are two stages for patients in the ICU, stage 1 represents critical condition stage and patients in stage 1 must be treated in the ICU, stage 2 represents recovery stage, during which patients can either be treated in the ICU or transferred to a lower care level units as early stepdown.

Figure 4.1 shows the network structure. We assume ICU patients enter the system following a Poisson process with rate λ_I. The number of beds in the ICU is fixed at N_I. Queue capacities for the ICU is Q_I, which is considered as boarding capacity. Once the queue is full, new arrivals are turned away. After a patient finishes the course in ICU, he/she will be discharged or stepped down to recovery units. Newly arriving ICU patients enter stage 1, and treated at service rate μ_1. On completion of stage 1 service, patients

enter stage 2, and are treated with service rate μ_2 or transferred to lower care level units as early stepdown driven by demand. It is considered better for patients to finish all treatments in the ICU, and only when the ICU encounters congestion will patients at stage 2 be considered to be transferred to lower level care units. A penalty cost is applied so that room can be made for newly arrival ICU patients.

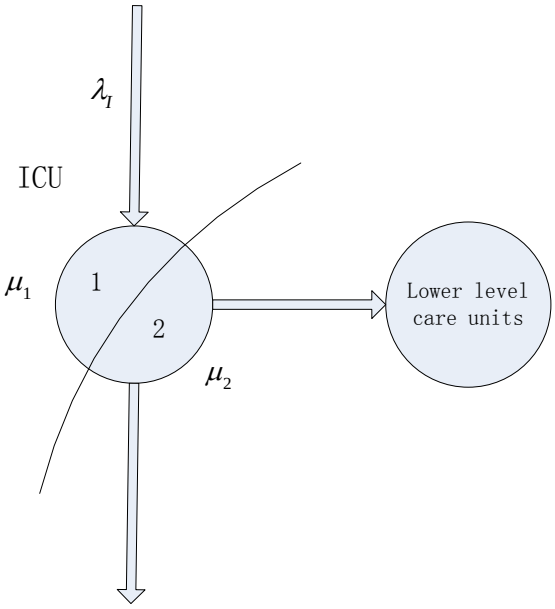

Figure 4.1 Network structure of ICU

4.3.1 ICU Stepdowns

To begin, we assume that lower level care units have infinity capacity, meaning that whenever the ICU requests a stepdown, there will always be room in lower level care units to accommodate transferred patient. Then the question becomes at what point the ICU should early step down a patient. As early stepdown comes with penalty costs,

transfer should only be considered when the ICU faces crowding. In this study, we limit our discussion to threshold policies, meaning that when the total number of patients in the ICU exceeds a threshold M, a stage 2 patient can be early stepped down. M is in the range from 1 to $N_I + Q_I$.

We write $P_{(i,j,v)}(t)$ as the probability that there are i stage 1 patients, j stage 2 patients in ICU, v patients in ICU queue at time t, and let the equilibrium (steady state) probability be $P_{(i,j,v)}$. Using assumptions of the birth-death process formulations, we obtain the following transient equations:

when $N_I < M \le N_I + Q_I$

$$P_{(i,j,0)}(t+\Delta t) = P_{(i-1,j,0)}(t)\lambda_I 1_{\{i\ge 1\}}\Delta t + P_{(i+1,j-1,0)}(t)(i+1)\mu_1 1_{\{j\ge 1\}}\Delta t + P_{(i,j+1,0)}(t)(j+1)\mu_2\Delta t$$
$$+P_{(i,j,0)}(t)[1-(\lambda_I + i\mu_1 + j\mu_2)\Delta t] \qquad i+j \le N_I - 1$$

$$P_{(i,j,v)}(t+\Delta t) = P_{(i-1,j+1,v)}(t)\lambda_I 1_{\{i\ge 1 \& v=M-N_I-1\}}\Delta t + P_{(i-1,j,0)}(t)\lambda_I 1_{\{i\ge 1 \& v=0\}}\Delta t + P_{(i,j,v-1)}(t)\lambda_I 1_{\{v\ge 1\}}\Delta t$$
$$+P_{(N_I,0,v+1)}(t)N_I\mu_1 1_{\{j=0 \& v=M-N_I-1\}}\Delta t + P_{(i+1,j-1,v)}(t)(i+1)\mu_1 1_{\{j\ge 1\}}\Delta t + P_{(i-1,j+1,v+1)}(t)(j+1)\mu_2 1_{\{i\ge 1 \& v\le M-2-N_I\}}\Delta t$$
$$+P_{(i,j,v)}(t)[1-(\lambda_I 1_{\{v\le Q_I-1\}} + i\mu_1 + j\mu_2)\Delta t] \qquad i+j = N_I, v = 0,1,....,M-N_I-1$$

$$P_{(N_I,0,v)}(t+\Delta t) = P_{(N_I,0,v-1)}(t)\lambda_I 1_{\{v\ge 1\}}\Delta t + P_{(N_I,0,v+1)}(t)N_I\mu_1 1_{\{v\le Q_I-1\}}\Delta t + P_{(N_I,0,v)}(t)[1-(\lambda_I 1_{\{v\le Q_I-1\}} + N_I\mu_1)\Delta t]$$
$$i = N_I,\ j = 0, v = M-N_I,....,Q_I$$

when $M \le N_I$

$$P_{(i,j,0)}(t+\Delta t) = P_{(i-1,j+1,0)}(t)\lambda_I 1_{\{i+j=M-1 \& i\ge 1\}}\Delta t + P_{(i+1,0,0)}(t)(i+1)\mu_1 1_{\{i=M-1\}}\Delta t + P_{(i-1,j,0)}(t)\lambda_I 1_{\{i\ge 1\}}\Delta t$$
$$+P_{(i+1,j-1,0)}(t)(i+1)\mu_1 1_{\{j\ge 1\}}\Delta t + P_{(i,j+1,0)}(t)(j+1)\mu_2 1_{\{i+j\le M-2\}}\Delta t + P_{(i,j,0)}(t)[1-(\lambda_I + i\mu_1 + j\mu_2)\Delta t]$$
$$i+j \le M-1$$

$$P_{(i,0,0)}(t+\Delta t) = P_{(i+1,0,0)}(t)(i+1)\mu_1\Delta t + P_{(i-1,0,0)}(t)\lambda_I 1_{\{i\ge 1\}}\Delta t + P_{(i,0,0)}(t)[1-(\lambda_I + i\mu_1)\Delta t]$$
$$M \le i \le N_I - 1,\ j = 0$$

$$P_{(N_I,0,v)}(t+\Delta t) = P_{(N_I,0,v+1)}(t)N_I\mu_1 1_{\{v\le Q_I-1\}}\Delta t + P_{(N_I-1,0,0)}(t)\lambda_I 1_{\{v=0\}}\Delta t + P_{(N_I,0,v-1)}(t)\lambda_I 1_{\{v\ge 1\}}\Delta t$$
$$+P_{(N_I,0,v)}[1-(1_{\{v\le Q_I-1\}}\lambda_I + N_I\mu_1)\Delta t] \qquad i = N_I,\ j = 0, 0 \le v \le Q_I$$

The resulting steady-state difference equations are:

when $N_I < M \le N_I + Q_I$

$$P_{(i,j,0)}(\lambda_I + i\mu_1 + j\mu_2) = P_{(i-1,j,0)}\lambda_I 1_{\{i\ge 1\}} + P_{(i+1,j-1,0)}(i+1)\mu_1 1_{\{j\ge 1\}} + P_{(i,j+1,0)}(j+1)\mu_2 \quad i+j \le N_I - 1$$

$$P_{(i,j,v)}(\lambda_I 1_{\{v\le Q_I-1\}} + i\mu_1 + j\mu_2) = P_{(i-1,j+1,v)}\lambda_I 1_{\{i\ge 1 \& v=M-N_I-1\}} + P_{(i-1,j,0)}\lambda_I 1_{\{i\ge 1 \& v=0\}} + P_{(i,j,v-1)}\lambda_I 1_{\{v\ge 1\}}$$
$$+P_{(N_I,0,v+1)}N_I\mu_1 1_{\{j=0 \& v=M-N_I-1\}} + P_{(i+1,j-1,v)}(i+1)\mu_1 1_{\{j\ge 1\}} + P_{(i-1,j+1,v+1)}(j+1)\mu_2 1_{\{i\ge 1 \& v\le M-2-N_I\}}$$
$$i+j = N_I, v = 0,1,....,M-N_I-1$$

$$P_{(N_I,0,v)}(\lambda_I 1_{\{v\le Q_I-1\}} + N_I\mu_1) = P_{(N_I,0,v-1)}\lambda_I 1_{\{v\ge 1\}} + P_{(N_I,0,v+1)}N_I\mu_1 1_{\{v\le Q_I-1\}} \quad i = N_I, j = 0, v = M-N_I,....,Q_I$$

when $M \le N_I$

$$P_{(i,j,0)}(\lambda_I + i\mu_1 + j\mu_2) = P_{(i-1,j+1,0)}1_{\{i+j=M-1 \& i\ge 1\}}\lambda_I + P_{(i+1,0,0)}1_{\{i=M-1\}}(i+1)\mu_1 + P_{(i-1,j,0)}1_{\{i\ge 1\}}\lambda_I$$
$$+P_{(i+1,j-1,0)}1_{\{j\ge 1\}}(i+1)\mu_1 + P_{(i,j+1,0)}1_{\{i+j\le M-2\}}(j+1)\mu_2 \qquad i+j \le M-1$$

$$P_{(i,0,0)}(\lambda_I + i\mu_1) = P_{(i+1,0,0)}(i+1)\mu_1 + P_{(i-1,0,0)}1_{\{i\ge 1\}}\lambda_I \qquad M \le i \le N_I - 1, j = 0$$

$$P_{(N_I,0,v)}(1_{\{v\le Q_I-1\}}\lambda_I + N_I\mu_1) = P_{(N_I,0,v+1)}1_{\{v\le Q_I-1\}}N_I\mu_1 + P_{(N_I-1,0,0)}1_{\{v=0\}}\lambda_I + P_{(N_I,0,v-1)}1_{\{v\ge 1\}}\lambda_I$$
$$i = N_I, j = 0, 0 \le v \le Q_I$$

The objective function is to minimize the total costs which consist of early stepdown penalty costs, patient waiting costs as well as patient rejection costs (costs for turning away patients). As more patients are stepped down early, the ICU will accommodate new patients, taking less time for patients in the queue to wait and get admitted. Similarly if fewer patients are stepped down, patients will spend more time waiting in the queue, and more patients will be rejected due to full capacity of the ICU. A balance between these three costs can be achieved by adjusting the threshold level. We use length of queue LOQ in the system to measure the equilibrium average number of patients in the queue, rate of patient rejection REJ to measure how many patients are rejected due to ICU full capacity

within one time unit, and rate of early stepdown TRA to measure how many patients are early stepped down within one time unit. They are calculated as follows:

$$LOQ = \sum_{i+j=N_I, v=1,2,\ldots,Q_I} P_{(i,j,v)} v,$$

$$REJ = \sum_{i+j=N_I} P_{(i,j,Q_I)} \lambda_I,$$

$$TRA = \sum_{i+j=N_I} P_{(i,j,M-N_I-1)} \lambda_I + \sum_{v=M-N_I,\ldots,Q_I} P_{(N_I,0,v)} N_I \mu_1$$

Assume that the cost for one patient to wait one time unit in the queue is C_W, the cost for rejecting one patient is C_R, the cost for stepping down one patient early is C_T. The total cost within time period Δt we aim to minimize is written as $C_W LOQ\Delta t + C_R REJ\Delta t + C_T TRA\Delta t$. As steady-state difference equations are linear, exact solutions can be obtained by solving the linear system. A real sized numerical example is conducted with parameters given in the following table:

$N_I = 50$	$Q_I = 6$	$\lambda_I = 20$	$\mu_1 = 0.5$	$\mu_2 = 0.5$

Table 4.1 Numerical parameters when lower level care units have infinite capacity where it is assumed that there are 50 beds available in the ICU, boarding capacity is 6, the average rate of patients arrival is 20 per day, the rate of finishing stage 1 treatment is 0.5 patient per day, and the rate of finishing stage 2 treatment is 0.5 patient per day. Threshold M ranges from 1 to 56.

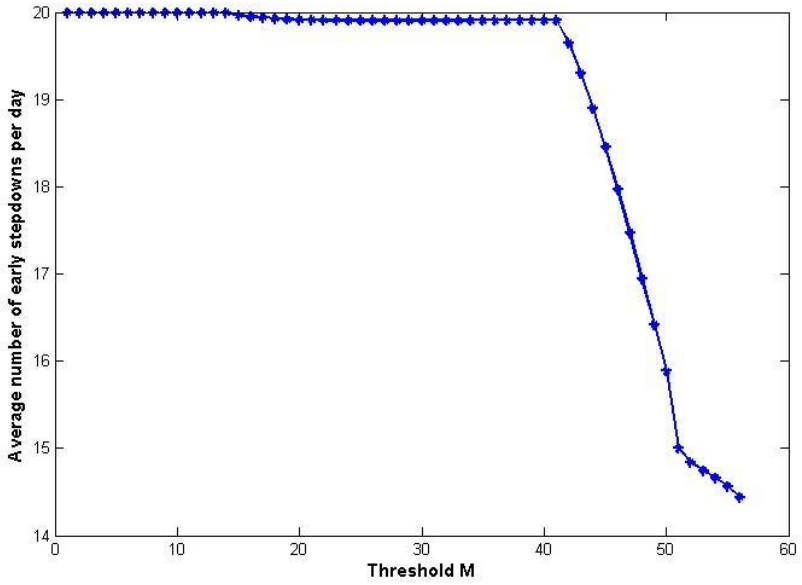

Figure 4.2 Average number of early stepdowns per day for different thresholds

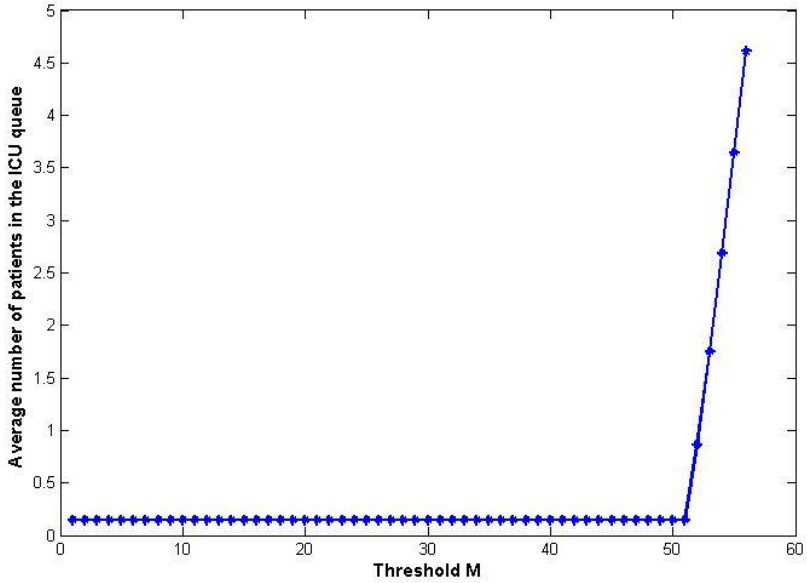

Figure 4.3 Average number of patients in the ICU queue for different thresholds

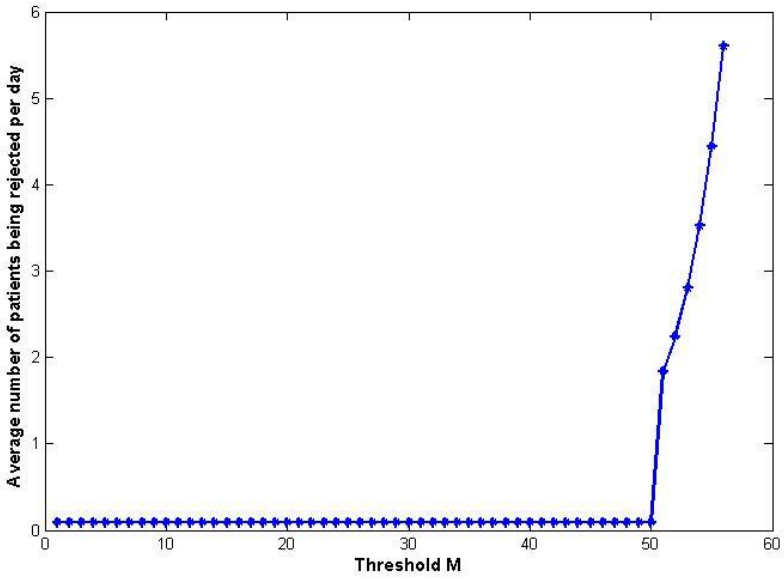

Figure 4.4 Average number of patients being rejected per day for different thresholds

As seen in Figures 4.3 and 4.4, when the threshold is below 50 the average number of patients in the ICU queue stays constant, and so does the average number of patients being rejected per day. However, the average number of early stepdowns per day is decreasing. The results imply that although more patients are stepped down early, the congestion situation is not improved according to the constant queue length and rejection rate. The reason is due to the infinite capacity in the lower level care units. The ICU can step down patients any time it needs to do so, as long as new arrival can be accommodated, there is no need to reserve room in advance. Thus there is no benefit to step down patients early before the ICU gets full.

When the threshold is greater than the ICU capacity, as it increases, the number of early stepdowns decreases whereas the average number of patients in the queue increases, and the average number of patients being rejected due to the ICU full capacity increases as well. The reason is that when the threshold is set higher, it is relatively harder to reach

this threshold compared to when threshold is low, thus fewer transfers will be made and patients have to wait longer in the queue before getting admitted. In particular, the detailed steady state probability distribution for the ICU queue is given in Figure 4.5. When threshold M is set to 51, the steady state probability of having no patient waiting the queue is greater than 0.9. When threshold M increases by 1, the probability of having one patient waiting is approximately 0.7. According to Figures 4.2 and 4.3., when threshold M increases by 1, the average number of patients in the ICU queue increases by approximate 1 as well whereas the average number of patients transferred decreases by around 0.15. The total costs consist of three components, which are the cost of early stepping down patients, the cost of having patients waiting in the queue, and the cost of rejecting patients due to full capacity.

The optimal threshold which achieves the minimum total cost depends on the relative cost ratios. Without loss of generality, we standardize the cost of having one patient waiting in the queue for one time unit (one day in this numerical example) to be 1, then vary the cost of early stepping down one patient and rejecting one patient from 0.1 to 10. The optimal threshold is shown as in Figures 4.6.1 and 4.6.2 (note thatFigure 4.6.2 is the projection of Figure 4.6.1). We can see that the optimal threshold alters as the costs of early stepping down and rejecting patients change. When the cost of early stepdown is in range 0.1 to approximate 4.5, the optimal threshold remains 51 regardless the changes in cost of rejecting patients. In addition, the optimal threshold grows to 52 as the cost of early stepdown further increases. Only when the cost of early stepdown is relatively large (approaching 10 in this example) and the cost of rejecting one patient stays small, a minimum cost is achieved when the threshold is set to 56. Thus we conclude that the

optimal threshold which achieves the minimum cost depends on the cost ratios. In particular, it depends on the cost of stepping down one patient early versus the cost of having one patient waiting in the queue for one time unit, and the cost of rejecting one patient from admission versus the cost of having one patient waiting in the queue. When the cost of early stepdown is relatively large and the cost of rejecting one patient is relatively small, a large threshold is optimal.

We name a policy a "*bumpout* policy" when the ICU inpatient beds are fully occupied, and a patient residing in the ICU is stepped down due to new patient arrival. The policy works well in practice since the goal of stepping down patients early doesn't block the inflow of new arrival ICU patients. It is implemented by many hospitals with or without their awareness. In fact, the *bumpout* policy is only a special case of the threshold policy where $M = N_I + 1$. With previous discussions, we conclude that the widely adopted *bumpout* policy does not always yield minimum total cost, even when comparison is limited to threshold policies. Instead, it depends on the costs concerned and corresponding cost ratios, a larger threshold might give better solutions when lower level care units have infinite capacity.

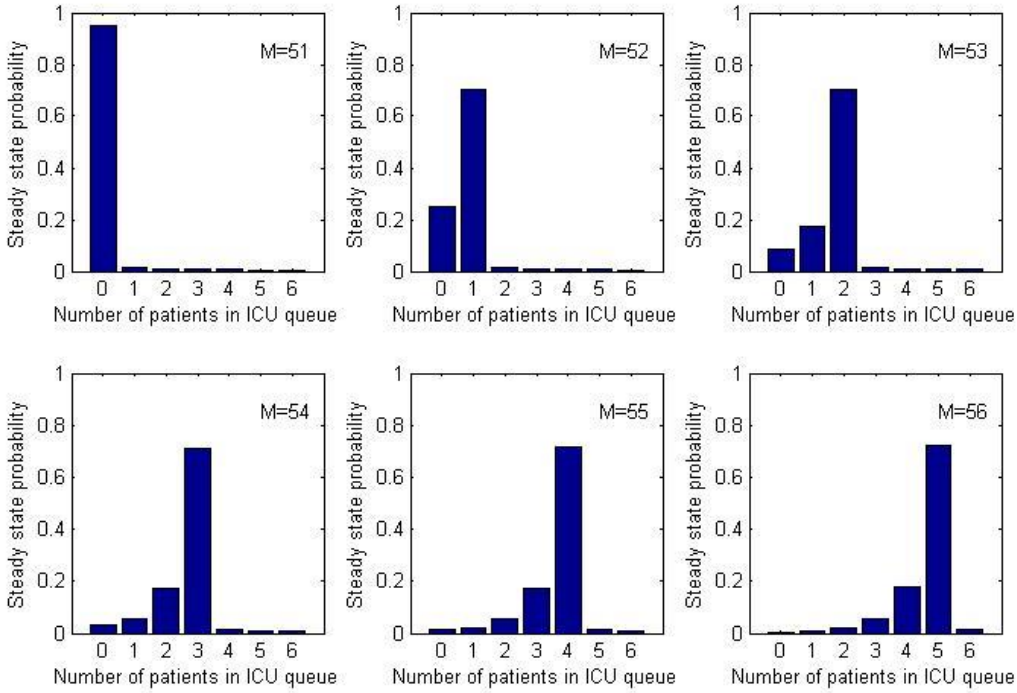

Figure 4.5 Probability distribution of number of patients in ICU queue for different thresholds

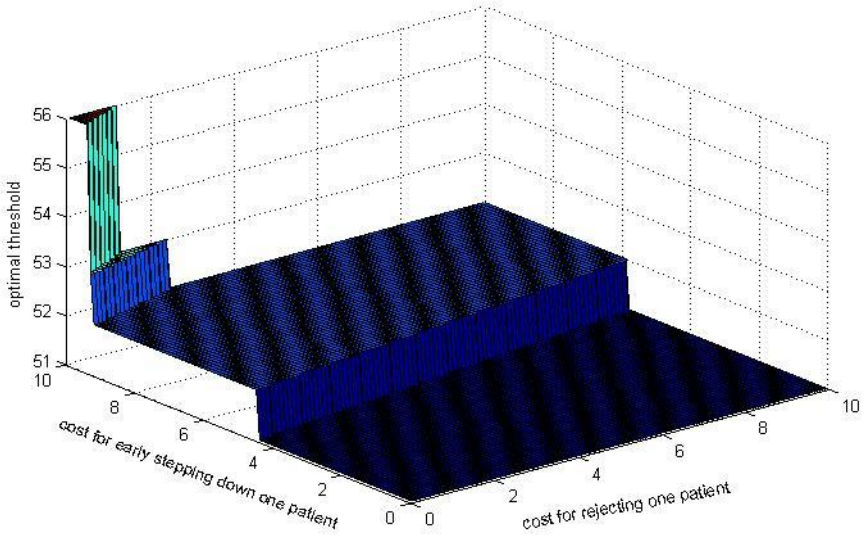

Figure 4.6.1 Optimal threshold

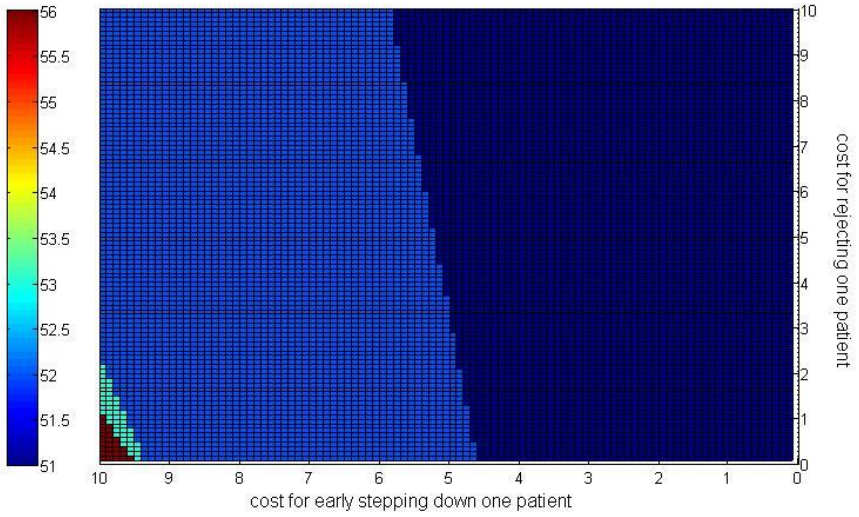

Figure 4.6.2 Optimal threshold projection

4.3.2 ICU Stepdowns with Readmissions

Patients who are subject to a demand-driven early stepdown might also potentially risk physiologic deterioration, which might ultimately lead to readmission. Not only the patients who are readmitted have a higher mortality rate than first-time patients, it also imposes an additional load on the capacity limited ICU resources. Next we will investigate the impact of readmission on the optimal threshold. We assume patients who have been stepped down early will be readmitted with probability. h The average interval time follows exponential distribution with rate λ_R. In this scenario, too many early stepdowns may will help with temporary capacity issues. However in the long term it might aggravate the congestion problem. Considering the fact that that readmitted patients have a higher mortality rate [88], we also impose a cost associated with readmission. Let C_A be the cost for one patient to be readmitted, and RAD be the number

of patients being readmitted within one time unit. Under the new assumptions, we conduct the numerical example again with $h = 0.2, \lambda_R = 0.1$, meaning that for an early stepdown patient, with probability 0.1 he/she will be readmitted with rate 0.1 patient per day.

$N_I = 50$	$Q_I = 6$	$\lambda_I = 20$	$\mu_1 = 0.5$	$\mu_2 = 0.5$	$h = 0.2$	$\lambda_R = 0.1$

Table 4.2 Numerical parameters when ICU readmission is considered

As shown in Figure 4.7, when the threshold increases, the average number of patients in the ICU queue and average number of patients being rejected per day increases, whereas the average number of early stepdowns per day as well as average number of readmissions per day both decrease. It comes as no surprise that when the threshold is high, fewer transfers will be made, and therefore fewer readmissions will occur. To compute the optimal threshold, we standardize the cost of having one patient waiting for one day in the queue to be 1, as well as the cost of rejecting one patient to be 1. We vary the costs of stepping down one patient early and readmitting one patient from 0.1 to 10. As shown in Figure 4.8.1, when the cost for early stepping down one patient is relatively small, the optimal threshold stays small; when the cost for early stepping down one patient is large, the optimal threshold is large at 56. The cost of readmitting one patient does not seem to have substantial impact on the alteration of optimal threshold. However when we set $h = 0.5$, as shown in Figure 4.8.2, the optimal threshold is small only when the costs for early stepping down one patient and readmitting one patient are both small, as long as one cost of the two is large, the optimal threshold is large. The reason is that only when there is a substantial proportion of patients getting readmitted among those stepped down early, the optimal threshold will then alter accordingly. Overall, the greater the

probability that an early stepdown patient gets readmitted, and the larger the cost of readmission, the greater is the resulting optimal threshold.

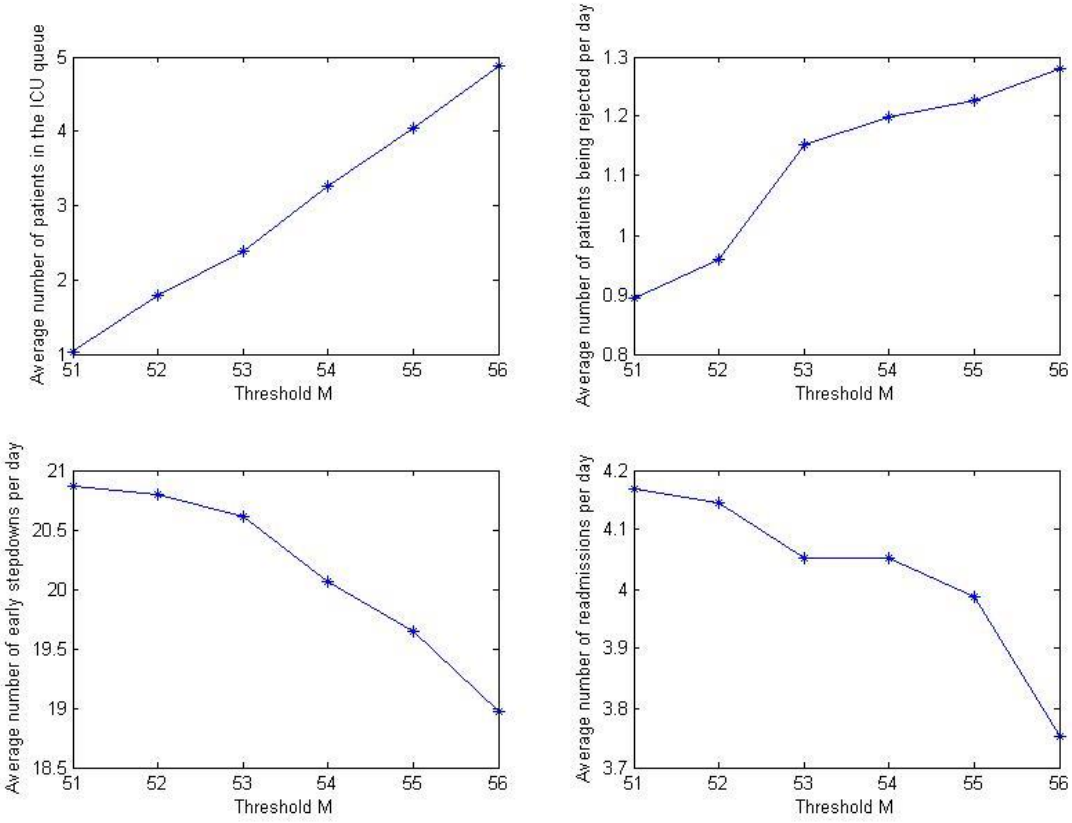

Figure 4.7 Costs versus threshold M considering readmission

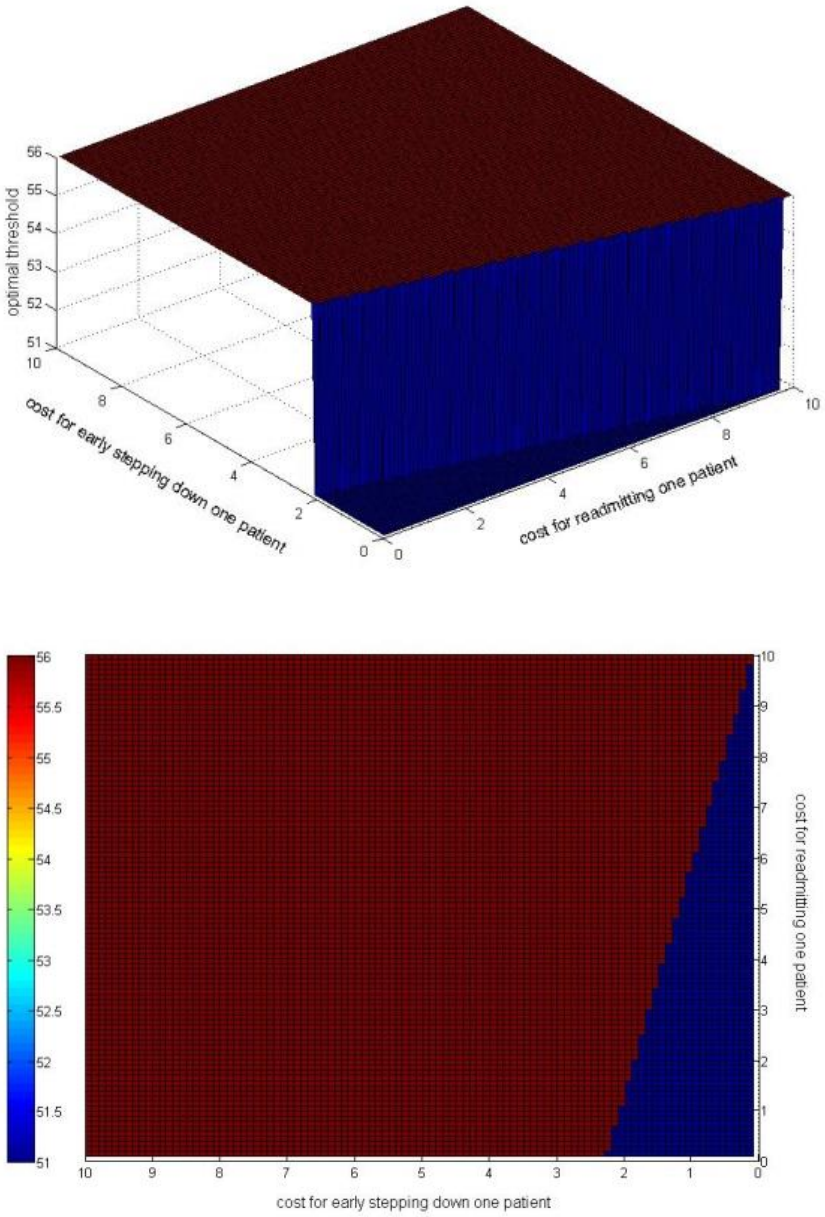

Figure 4.8.1 Optimal threshold when $h = 0.2$

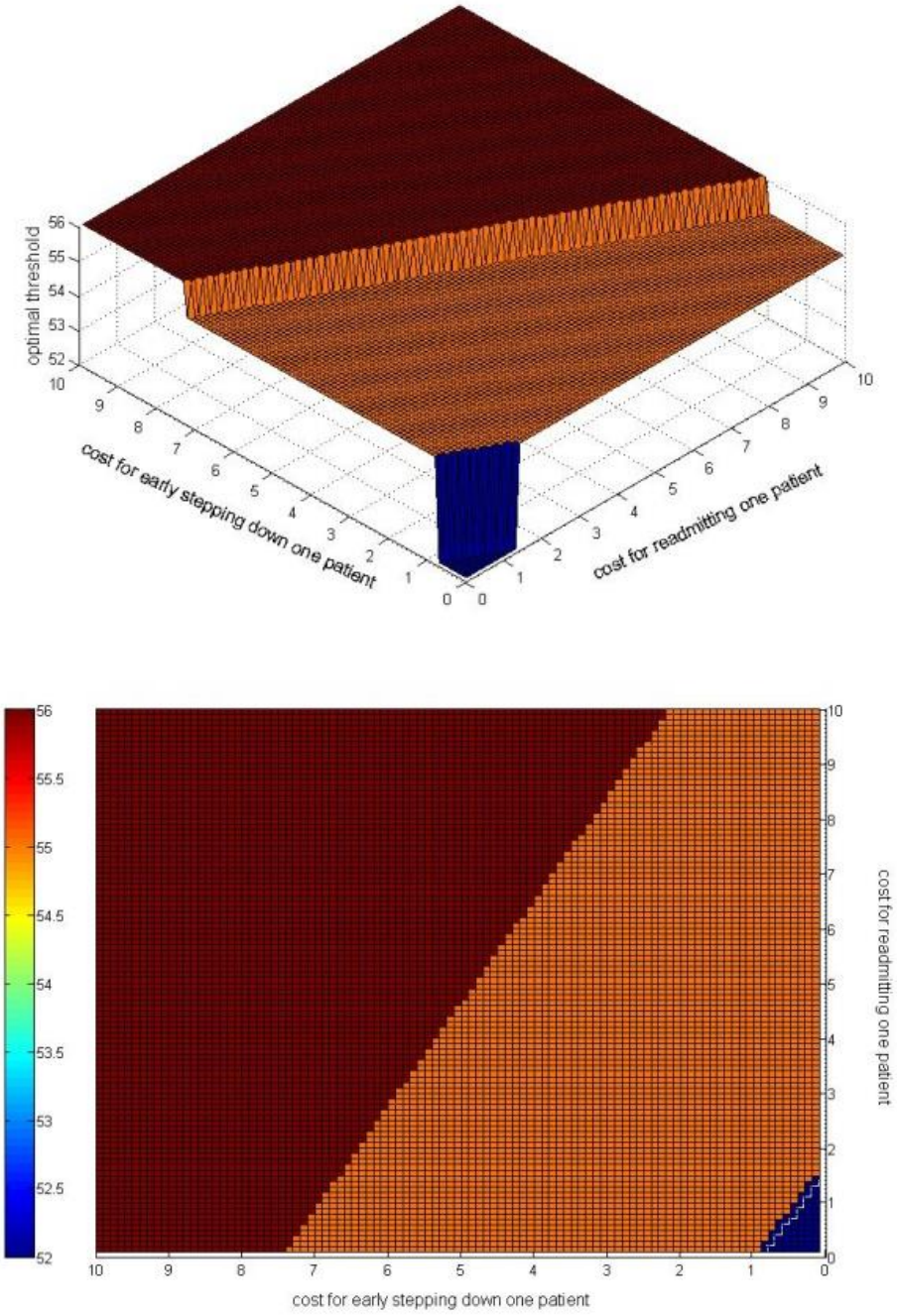

Figure 4.8.2 Optimal threshold when $h = 0.5$

4.3.3 ICU and GCU Coordination

In this section, we will discuss assumptions and models when two units ICU and GCU are involved. Rather than assuming that the ICU patients step down to the lower level care units which have infinite capacity, we assume that the ICU patients step down to GCU whose capacity is limited. Figure 4.9 shows the basic network structure of the ICU and GCU. We assume ICU patients enter the ICU admission system and GCU patients enter the GCU admission system. The numbers of beds in ICUs and GCUs are fixed to N_I and N_G respectively. Queue capacities for ICUs and GCUs are Q_I and Q_G. Once the queue is full, new arrivals are rejected and turned away.

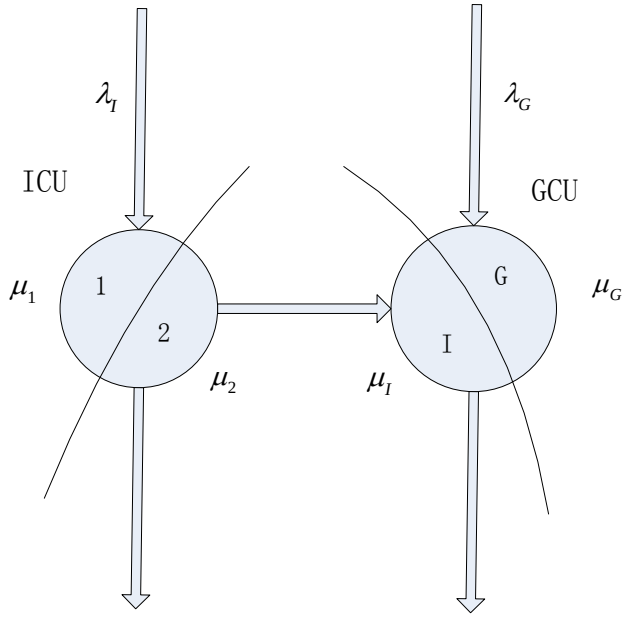

Figure 4.9 Network structure of the ICU and GCU

Following the same assumptions as in the previous section, the arrival of ICU patients follows a Poisson process with rate λ_I. There are two stages for patients in the ICU, stage 1 with service rate μ_1 and stage 2 with service rate μ_2. The arrival of GCU

patients follows a Poisson process with rate λ_G and the service rate is μ_G. For the ICU patient who is stepped down early to the GCU, the service rate is μ_I. It is assumed that $\mu_2 \geq \mu_I$ since the ICU generally provides much more invasively monitored care than the GCU. Therefore, it typically takes less time for patients to recover in the ICU than GCU.

Early stepdown happens if and only if there is stage 2 patient in the ICU and the total number of patients in the ICU system (including inpatients and patients in the queue) reaches the threshold M $(1 \leq M \leq N_I + Q_I)$. Meanwhile if there is a vacant bed in the GCU, a stage 2 patient will be transferred to the GCU. Considering that the GCU has limited capacity, implying that the ICU may not be able to make a stepdown anytime it is requested, it might be necessary to step down patients early even when the ICU is not full, thus $1 \leq M \leq N_I + Q_I$.We assume that the transfer is instantaneous with no preemption. Due to the memoryless property of the exponential distribution, stage 2 patient transferred to GCU will start a new stay with rate μ_I no matter how long they have stayed in the ICU. It is assumed that ICU patients always have a higher priority than GCU patients, meaning that if there is a bed in GCU vacated, an ICU stage 2 patient will be stepped down before a GCU patient in the queue is admitted. Since two care units are involved, the dilemma lies in how to arrange the early stepdown such that not only the ICU can resolve its congestion issues but meanwhile not affect operations of the GCU. Without losing the insight we neglect the early stepdown costs. The objective function is to minimize the costs that only consist of the penalty for having ICU and GCU patients waiting in the queue and rejecting ICU and GCU patients due to full capacities.

Let $C_{(p,q,u)(i,j,v)}$ denote the steady state probability that there are p GCU patients in the GCU, q transferred stage 2 ICU patients, u patients in the GCU queue, and there are i stage 1 ICU patients, j stage 2 ICU patients, v patients in the ICU queue. Similar to the case when only the ICU is considered, steady-state balanced equations can be obtained using birth-death processes. Specific formulations can be found in Appendix A. Since the dimension of the linear system grows in polynomial order with parameters, the balanced model only works for relatively small-scale systems. Two metrics are measured for each unit, rate of patient rejection REJ and the steady state average length of queue LOQ. They are computed as:

$$REJ_{ICU} = \sum_{p,q,u,i+j=N_I} C_{(p,q,u)(i,j,Q_I)}, \quad REJ_{GCU} = \sum_{i,j,v,p+q=N_G} C_{(p,q,Q_G)(i,j,v)}$$

$$LOQ_{ICU} = \sum_{v=0}^{Q_I} \sum_{p,q,u,i+j=N_I} vC_{(p,q,u)(i,j,v)}, \quad LOQ_{GCU} = \sum_{u=0}^{Q_G} \sum_{i,j,v,p+q=N_G} uC_{(p,q,u)(i,j,v)}$$

As we are interested in ICU-GCU coordination when the ICU is congested, we keep condition the $\frac{\lambda_I}{\mu_1} + \frac{\lambda_I}{\mu_2} > N_I$ satisfied when conducting numerical examples. The parameters for the numerical example are given in Table 4.3.

ICU	$N = 8$	$Q = 3$	$\lambda = 120$	$\mu_1 = 30$	$\mu_2 = 6$
GCU	$N = 11$	$Q = 3$	$\lambda = 60$	$\mu_G = 7$	$\mu_I = 6$

Table 4.3 Numerical parameters when GCU has finite capacity

When the two units work independently (i.e. the ICU does not step down patients early), results are given as Table 4.4. We can see that in this scenario, the GCU can only handle its loads with small rejection rate of 0.0410, whereas the ICU has serious

congestion problems with a rejection rate 0.6671, meaning that if there is a new arrival in the ICU, with probability 0.667 that the patient will be rejected and turned away.

REJ_{ICU}	LOQ_{ICU}	REJ_{GCU}	LOQ_{GCU}
0.6671	2.5366	0.0410	0.2959

Table 4.4 Results when the ICU does not early step down patients

M	d	REJ_{GCU}	LOQ_{GCU}	REJ_{ICU}	LOQ_{ICU}
1	4248	0.969846276	2.955755	0.244956595	1.304758
2	4314	0.969643695	2.955442	0.244978143	1.304879
3	4446	0.968163231	2.953066	0.245171372	1.30597
4	4644	0.962928666	2.944539	0.245974568	1.310509
5	4908	0.950613346	2.924264	0.248148039	1.322813
6	5238	0.927957156	2.886468	0.252705214	1.348691
7	5634	0.891657197	2.824855	0.261019134	1.396111
8	6096	0.837060025	2.730237	0.275312554	1.478124
9	6624	0.755606493	2.585113	0.299808431	1.61963
10	7152	0.62988172	2.348946	0.344061979	1.896043
11	7680	0.41922899	1.880257	0.421754017	2.222736

Table 4.5 Results when the ICU early step down patients to the GCU

When the ICU early steps down patients to the GCU, Table 4.5 shows the results for all enumerations of thresholds M from 1 to 11. Compared to the no early stepdown case, the ICU rejection rate is reduced to the range 0.24-0.42 from 0.6671 depending on the specific threshold. This reduction comes with an increased GCU rejection rate and longer average waiting times in the GCU queue. We see that, as the threshold increases, the rejection rate of the GCU decreases whereas rejection rate of the ICU increases. This is intuitive since as the threshold increases, the early stepdown conditions become more restrictive, resulting in fewer early stepdowns.

As mentioned in previous section, a *bumpout* policy is widely implemented in practice, but it does not always yield optimal solution even when the lower level care units have infinite capacity (though stepping down early before the ICU reaches full capacity is unnecessary in this scenario). When lower level care units are constrained to have limited capacity, stepping down patients even when the ICU still has empty beds will yield a better solution, since the ICU may not be able to make an immediate transfer whenever it needs to. If the GCU does not have empty beds, the transfer will have to wait until a GCU patient leaves and this could eventually lead to the congestion of the system. In this case, it makes a big difference to use a small M. In the numerical example, $M=9$ represents a bump out policy, rejection rate for the ICU is 0.2998, whereas when $M=1$ (i.e. always transfer stage 2 patient to GCU if possible), rejection rate for the ICU is 0.2449. Thus the ICU rejection probability is reduced by (0.2998-0.2449)/0.2998=18.3% by setting $M=1$.

	c=1	c=0.9	c=0.8	c=0.7	c=0.6	c=0.5	c=0.4	c=0.3	c=0.2	c=0.1	c=0
M=1	1.214803	1.117818	1.020834	0.92384899	0.82686436	0.729879733	0.632895105	0.53591048	0.43892585	0.3419412	0.2449566
M=2	1.214622	1.117657	1.020693	0.92372873	0.82676436	0.729799991	0.632835621	0.53587125	0.43890688	0.3419425	0.24497814
M=3	1.213335	1.116518	1.019702	0.92288563	0.82606931	0.729252988	0.632436664	0.53562034	0.43880402	0.3419877	0.24517137
M=4	1.208903	1.11261	1.016318	0.92002463	0.82373177	0.727438901	0.631146034	0.53485317	0.4385603	0.3422674	0.24597457
M=5	1.198761	1.1037	1.008639	0.91357738	0.81851605	0.723454712	0.628393377	0.53333204	0.43827071	0.3432094	0.24814804
M=6	1.180662	1.087867	0.995071	0.90227522	0.80947951	0.716683792	0.623888076	0.53109236	0.43829665	0.3455009	0.25270521
M=7	1.152676	1.063511	0.974345	0.88517917	0.79601345	0.706847733	0.617682013	0.52851629	0.43935057	0.3501849	0.26101913
M=8	1.112373	1.028667	0.944961	0.86125457	0.77754857	0.693842567	0.610136564	0.52643056	0.44272456	0.3590186	0.27531255
M=9	1.055415	0.979854	0.904294	0.82873298	0.75317233	0.677611678	0.602051028	0.52649038	0.45092973	0.3753691	0.29980843
M=10	0.973944	0.910956	0.847967	0.78497918	0.72199101	0.659002839	0.596014667	0.5330265	0.47003832	0.4070502	0.34406198
M=11	0.840983	0.79906	0.757137	0.71521431	0.67329141	0.631368512	0.589445613	0.54752271	0.50559982	0.4636769	0.42175402

Table 4.6 Costs corresponding to different thresholds and cost ratios

The optimal threshold depends on the cost ratios of rejecting the ICU versus GCU patients, and having the ICU versus GCU patients waiting. Since the rejection rate REJ and average queue length LOQ behave consistently when the threshold is varied, we only use REJ to calculate total costs. If only the ICU is concerned, the optimal threshold will

be set to 1, which minimizes the ICU costs. If both the ICU and GCU are involved, the optimal threshold will depend on the cost ratios. Assume that the cost for rejecting one ICU patient is normalized to 1, and the cost for rejecting one GCU patient is varied from 1 to 0. Table 4.6 shows total costs for each threshold when the cost is enumerated. The value in red in each column is the minimal cost for a given GCU rejection cost among enumerations of the threshold, and the corresponding M is the optimal threshold. It is noted that when the cost for rejecting one GCU patient decreases, the optimal threshold also decreases, the rationale lies in that more stepdowns from the ICU to GCU are preferred when rejecting ICU patients comes at a bigger cost then rejecting GCU patients.

4.3.4 Alternative Stepdown Policies

The GCU helps the ICU to reduce traffic loads by receiving early stepdowns from the ICU. There are two alternative ways to discipline the stepdown regarding whether GCU places its own patients' priority above the ICU stepdown patients. If the GCU gives higher priority to its own patients, the ICU patients will only be stepped down if the GCU finishes serving its own patients (i.e. no GCU patient is waiting in the queue). We call this the GCU priority policy. If the GCU gives higher priority to ICU patients, it will serve ICU patients first and then its own patients waiting in the queue, We name this the ICU priority policy. We compare the two alternative stepdown policies to the threshold policy. In particular, stability properties are investigated by increasing the patient load ρ from 0 to 1.

ICU	$N=50$	$Q=\infty$	$\lambda=30\rho$	$\mu_1=1.2$	$\mu_2=1$
GCU	$N=80$	$Q=\infty$	$\lambda=50\rho$	$\mu_G=1$	$\mu_I=0.8$

Table 4.7 Numerical parameters for alternative stepdown policies

According to Figure 4.10, when the patient load increases, the proportion of ICU patients served by the GCU decreases under both policies. The reason is that when patient load increases, fewer beds are available for ICU stepdown patients. As the patient load increases, the difference between ICU priority policy and GCU priority policy grows larger in terms of the proportion of ICU patients being served by the GCU. When the patient load approaches 0.9, the GCU queue grows extremely rapidly and hence the GCU is no longer stable under ICU priority policy shown in Figure 4.11.1. Under the GCU priority policy, both the ICU and GCU are stable, handling the patient load well as shown in Figure 4.11.2.

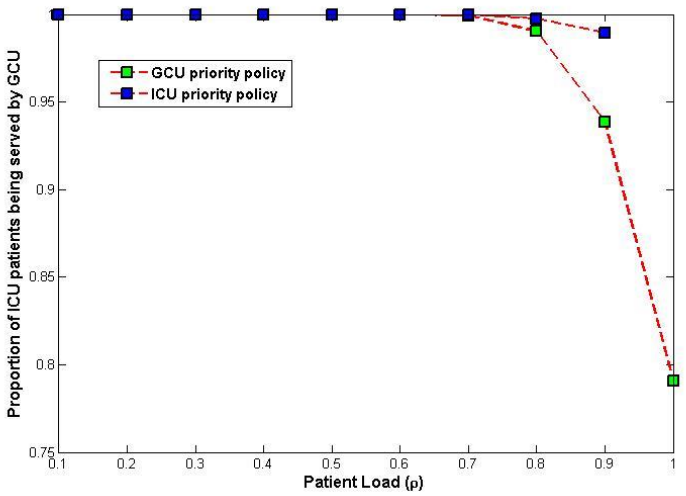

Figure 4.10 Proportion of ICU patients served by GCU versus patient load

Figure 4.12.1 and 4.12.2 shed light on the utilizations of ICU and GCU under these two policies. We can see that for the ICU priority policy, as the patient load increases, the

GCU spends more and more time serving ICU patients. When patient load reaches 0.9, the GCU is busy serving both ICU and GCU patients, while ICU still has around 50% idle capacity. It is clear that the GCU is over-utilized, and it is serving patients that should remain in the ICU, which explains the cause for GCU instability -- it offers so much help to ICU patients that its own patients are sacrificed..

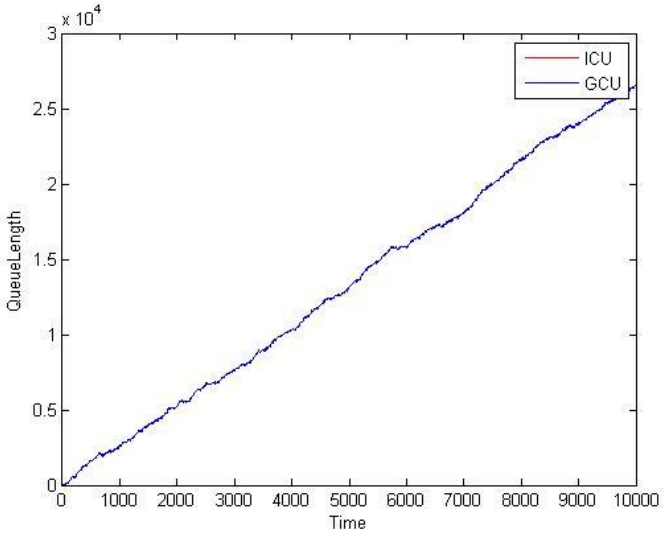

Figure 4.11.1 ICU priority policy when $\rho = 0.95$

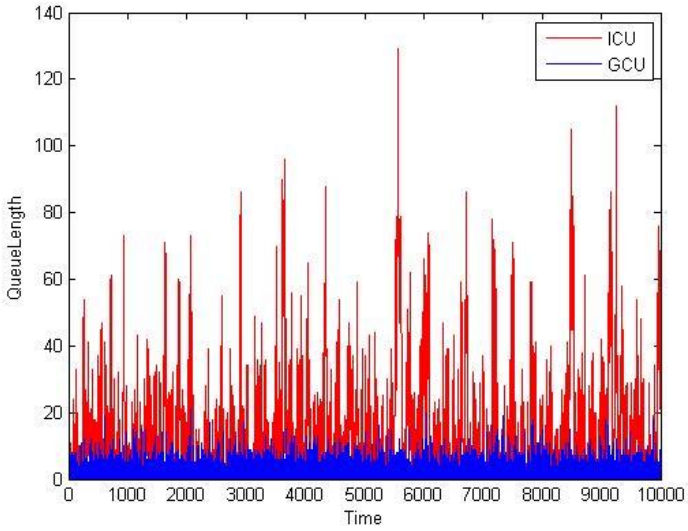

Figure 4.11.2 GCU priority policy when $\rho = 1$

For the GCU priority policy shown in Figure 4.12.2, as traffic load increases, both the utilizations of the ICU and the GCU increase. When patient load approaches 1, neither the ICU or GCU has much idle capacity, with the ICU serving ICU patients and the GCU serving both GCU and ICU patients. In contrast to ICU priority policy, time spent on serving ICU patients by the GCU increases first and then decreases under GCU priority policy. It implies that when patient load is heavy, the GCU devotes more time to serving its own patients and the ICU is forced to reduce the number of early stepdowns. This leads to full utilization of both units for GCU priority policy.

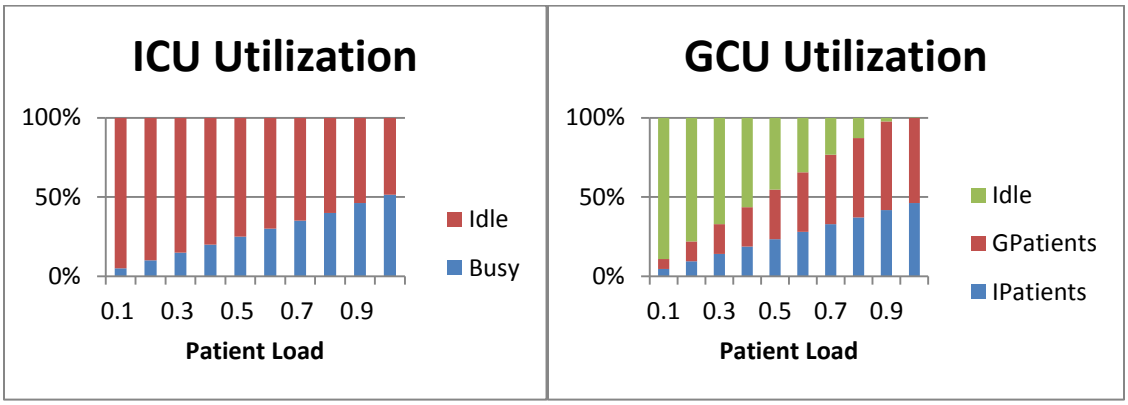

Figure 4.12.1 Unit utilizations for ICU priority policy

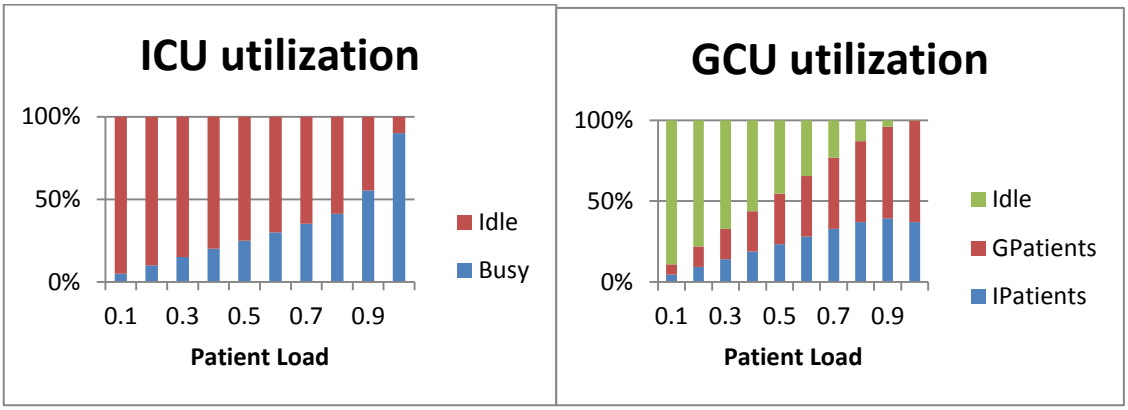

Figure 4.12.2 Unit utilizations for GCU priority policy

Under heavy patient load, only limited help can be offered to the ICU following a GCU priority policy. Threshold policies as a variation of the ICU priority policy comes with more flexibility. It requires that the GCU only helps the ICU only when the number of patients in the ICU reaches a certain threshold. In the numerical example, when the threshold is set to 50, 70, and 90, we see that both the ICU and GCU are stable according to Figure 4.13.1, 4.13.2, and 4.13.3. As expected, the higher the threshold is, the longer average queue length ICU has and the shorter average queue length GCU has. To decide

the optimal threshold, the relative costs of rejecting patients in the ICU and GCU are required. But we can see that the threshold policy can remedy the instability of ICU priority policy and offer more flexibility on the level of the ICU being helped than GCU priority policy.

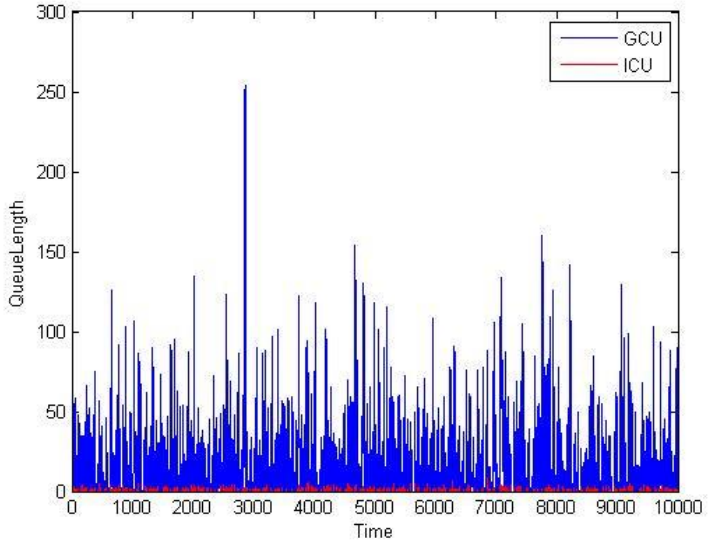

Figure 4.13.1 Threshold=50, $\rho = 0.95$

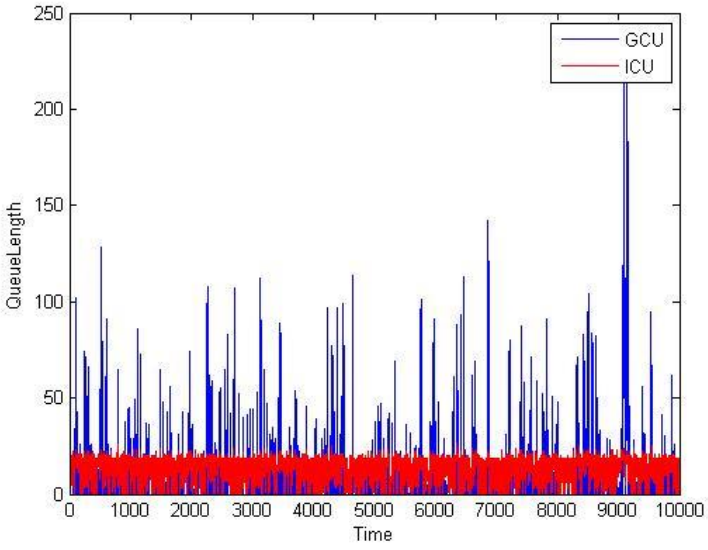

Figure 4.13.2 Threshold=70, $\rho = 0.95$

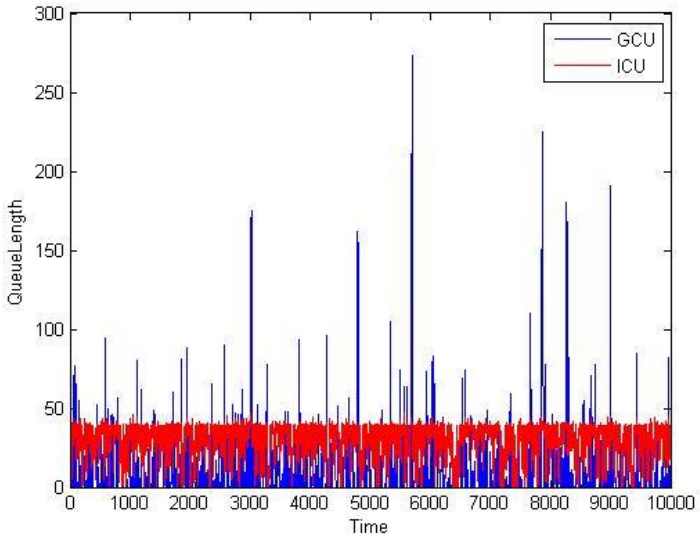

Figure 4.13.3 Threshold=90, $\rho = 0.95$

4.4 Conclusion

ICUs are among the most richly staffed and highly congested facilities providing care for the sickest and most unstable patients in a hospital. In this study, we propose early stepdown policies for the ICU to resolve congestion issues. We assume that ICU patients in a less severe condition can be stepped down early and designated to lower level care units so that beds can be vacated to accommodate new arrivals. In particular, we confine our discussions to threshold policies where ICU patients are stepped down early if and only if the total number of patients in the ICU reaches a certain threshold. We analyze scenarios when lower level care units (e.g. GCU) have infinite and limited capacity and investigate its impact on the optimal policies. We formulate the problem as birth-death process and steady state probability distribution can be obtained by solving balanced models.

Specifically, when lower level care units are assumed to have infinite capacity meaning that whenever the ICU requests an early stepdown there will be room to accommodate the transfer, it is not optimal to step down patients early before the ICU gets fully occupied. When the threshold is set greater than the ICU capacity, the number of patients being early stepped down decreases and the number of patients waiting in the queue increases as the threshold increases. The optimal threshold depends on the ratios of costs being concerned. If the cost of early stepping down patients is relatively large compared to keeping patients in the queue and rejecting patient, a large threshold will render minimum costs. Furthermore, considering that patients being early stepped down potentially risk physiologic deterioration, readmission is assumed and added to the model framework. Numerical examples illustrate that when there is a substantial proportion of patients getting readmitted among those early stepped down, the optimal threshold will then alter accordingly. Overall, the greater probability an early stepdown patient gets readmitted, the larger cost of readmission is, the greater optimal threshold is. In addition, we also show that the widely adopted *bumpout* policy (i.e. a patient residing in the ICU is stepped down due to new patient arrival when the ICU inpatient beds are fully occupied) is only a special case of the threshold policies. Our discussions imply that the *bumpout* policy does not always yield minimum total cost; alternative thresholds might give better solutions depending on the costs concerned and corresponding cost ratios, though it should be noted that stepping down patients before the ICU gets full is not optimal in this scenario. However, when lower level care units are constrained to have limited capacity, stepping down patients even when the ICU still has empty bed might yield better solution, as a result of the fact that the ICU may not be able to make an immediate

transfer whenever it needs to. If only the ICU costs are concerned, the threshold should be set as small as possible.

We also investigate stability properties of different stepdown policies characterized with different priority rules. In particular if each unit has highest priority for its own patients, system stability will be sustained, but it also limits the extent of help ICU can be offered. Otherwise, some units will become unstable as a result of over helping ICU and neglecting their own patients. That's where the threshold policy comes into play, which allows for early stepdowns only if the total number of patients in the ICU reaches a threshold. It can remedy the instability by forcing lower level care units not to help ICU too much, but also providing reasonable help when ICU is highly congested. The optimal threshold depends on the relative costs of rejecting different units' patients. If only ICU is concerned, the optimal threshold will be set as low as possible, which minimizes ICU costs, i.e. total costs. If different cost ratios are given to each unit, then the optimal threshold will be the one reaching a balance between these costs.

Our work provides a framework within which different coordination policies among units can be investigated to address unit congestion issues. The work also suggests potential direction for future theoretical research, for instance, a fluid and diffusion analysis can be used to address the problem and asymptotic optimal policies might be found within the framework.

Chapter 5 Conclusions

Cost-effectiveness analysis as a means to assess the tradeoff between consumed resources and achieved health outcomes has become an essential component in determining the allocation of scarce health resources. It can help inform policy makers for better ways to allocate limited resources. Some form of cost-effectiveness is now required for health interventions to be covered by many insurers. Health technology assessment agencies such as the National Institute for Health and Clinical Excellence (NICE) place considerable weight on the relative cost-effectiveness of therapies when making their judgments. It is one of a number of techniques of economic evaluation, where the choice of technique depends on the nature of the benefits specified. Cost-effectiveness analysis has been defined by the NICE as an economic study design in which consequences of different interventions are measured using a single outcome. Alternative interventions are then compared in terms of cost per unit of effectiveness. The results provide information that help decision makers determine which policy best serves their programmatic and financial needs. In this dissertation, three examples which fall under the cost-effectiveness framework are studied and addressed in detail.

The first study was to determine an optimal breast cancer screening policy in the presence of overdiagnosis. This study examines the impact of overdiagnosis on quality adjusted life years (QALYs) and the expected number of mammograms over a lifetime under various screening policies. An extension of a current Markov model of breast cancer screening was developed that includes an overdiagnosis state. Sensitivity analysis over a range of overdiagnosis values of 0% to 50% was conducted. Numerical results

show that overdiagnosis has a negative impact on the effectiveness of screening policies. In particular, the recommended screening interval increases in the rate of overdiagnosis. Further, ignoring overdiagnosis levels will lead to policies that have both a higher number of expected mammograms over a patient's life and a decrease in QALYs, which is not cost-effective. The 2009 USPTFS[18] recommended less frequent screening than previous recommendations. The findings here support the USPTFS recommended biannual screening intervals for even moderate levels of overdiagnosis. If the rate of overdiagnosis exceeds 25%, then an even greater interval of 3 years may be beneficial. Before technology would allow the identification of early stage cancers that will not progress to later stages, the benefit of early detection from shorter screening intervals must be weighed against the harms associated with frequent screening and overtreatment. Overdiagnosis should always be a consideration in future recommendations of screening intervals.

It is difficult to measure the overdiagnosis rate directly since it primarily appears as an output of trials. Our approach was to "reverse engineer" the value and use it as an input to study screening policies. We did not model how different ages should have different overdiagnosis rates due to a lack of accurate information, and assumed the overdiagnosis rate is constant across all ages for each case and then conducted a wide range of sensitivity analysis on the rates. If the accurate information about how overdiagnosis rate varies with age becomes available, the model can easily be modified to incorporate it.

The second example is a cost-effectiveness analysis of treatments for hepatitis C disease. The standard care of treatments interferon plus ribavirin (plus protease inhibitor

for genotype 1) are effective in 50% to 70% of patients with CHC of all genotypes. Recently several new treatments, e.g. Harvoni, Olysio + Sovaldi, Viekira Pak (for genotype 1) and Sofosbuvir-based regimens (for all genotypes) characterized with potent inhibitors have been approved by FDA providing more options for CHC patients. Trials have shown that the new treatments increased the average rate to 80% to 95%, though with a substantial increase in the costs. In particular, current market pricing of a 12-week course of sofosbuvir is reaching approximately $84,000. In this study, we apply a Markov simulation model of CHC disease progression to evaluate the cost-effectiveness of new treatment strategies in comparison with standard care of treatments. The model calculates the expected lifetime medical costs and quality adjusted life years (QALYs) of hypothetical cohorts receiving certain treatments. Treatments for genotype 1,2,3, are first compared to corresponding standard care of treatments, then compared with each other within the same genotypes. Results show that Viekira Pak is cost-effective for genotype 1 patients without cirrhosis, whereas Harvoni is cost-effective for genotype 1 patients with cirrhosis. Sofosbuvir-based treatments for genotype 1 in general are not cost-effective due to its substantial high costs. Two-phase treatments with initial standard care of treatment and sofosbuvir-based regimens as follow-up are cost-effective for genotype 3 patients and for genotype 2 patients with cirrhosis. The results are shown to be robust to a broad range of parameter values through a sensitivity analysis. However, there is limited data on sofosbuvir-involved treatment, and the results obtained in this study must be interpreted within the model assumptions.

The third example is aimed to resolve congestion issues of intensive care units (ICUs). Due to limited resources, it is critical to effectively use medical units and

maximize throughput from a cost-effective perspective. In this research, we assume that ICU patients in a less severe condition can be early stepped down and designated to lower level care units so that beds can be vacated to accommodate new arrivals. In particular, we confine our discussions to threshold policies where ICU patients are early stepped down if and only if the total number of patients in the ICU reaches a certain threshold. We analyzed scenarios when lower level care units (e.g. GCU) have infinite and limited capacity and investigate its impact on the optimal policies. When lower level care units are assumed to have infinite capacity, numerical examples showed that it is not optimal to early step down patients before the ICU gets fully occupied. When the threshold is set greater than the ICU capacity, the number of patients being stepped down early decreases and the number of patients waiting in the queue increases as the threshold increases. The optimal threshold depends on the ratios of costs being concerned. If the cost of early stepping down patients is relatively large compared to keeping patients in the queue and rejecting patient, a large threshold will render minimum costs. Furthermore, considering that patients being early stepped down potentially risk physiologic deterioration, readmission is assumed and added to the model framework. Numerical examples show that when there is a substantial proportion of patients getting readmitted among those early stepped down, the optimal threshold will then alter accordingly. We also point out that the widely adopted bumpout policy (i.e. a patient residing in the ICU is stepped down due to new patient arrivals when the ICU inpatient beds are fully occupied) is only a special case of the threshold policies. We show that the bumpout policy does not always yield minimum total cost. Alternative thresholds can give better solutions depending on the costs concerned and corresponding cost ratios,

though it should be noted that stepping down patients before the ICU gets full is not optimal in this scenario. However, when lower level care units are constrained to have limited capacity, stepping down patients even when the ICU still has empty bed might yield better solution, as a result of the fact that the ICU may not be able to make an immediate transfer whenever it needs to. We next investigated stability properties of different stepdown policies characterized with different priority rules. In particular if each unit has highest priority for its own patients, system stability will be sustained, but it also limits the extent of help ICU can be offered. A threshold policy had the highest performance because it can not only remedy the instability by forcing lower level care units not able to help the ICU, but it also provids reasonable help when ICU is highly congested.

Our work provides a framework within which different coordination policies among units can be studied to address unit congestion issues. The work also suggests potential direction for future theoretical research, for instance, a fluid and diffusion analysis can be used to address the problem and asymptotic optimal policies might be found within the framework.

Cost-effectiveness analysis is far from being a precise science, and there is often considerable uncertainty associated with the findings and wide variation around the estimate generated. It is therefore imperative that the assessment of cost-effectiveness should be subjected to a sensitivity analysis to enable decision-makers to be fully aware of the range of possible eventualities. In the breast cancer research, sensitivity analysis is conducted on the overdiagnosis rate in the range of 0% to 50% and suggestions of screening intervals corresponding to overdiagnosis rate are given. In HCV research,

sensitivity analysis is performed for parameters including utility weights, transition probabilities, costs and efficacy rates. In addition to one-way sensitivity analyses for all variables, we also conduct probabilistic sensitivity analyses to examine the effect of joint uncertainty in the model's variables. The results are shown to be robust.

While cost-effectiveness analysis is a useful technique for assisting in the decision-making process, there are important issues to consider. First of all, cost-effectiveness can indicate which one of a number of alternative interventions represents the best value for money, but it is not as useful when comparisons need to be made across different areas of health care, since the outcome measures used may be very different. The quality of cost-effectiveness analysis is highly dependent on the quality of effectiveness data used, and all cost-effectiveness analysis should include a detailed sensitivity analysis to test the extent to which changes in the parameters used in the analysis may affect the results obtained. Also cost-effectiveness is only one of a number of criteria that should be employed in determining whether interventions are made available. Issues of equity, needs, priorities and so on should also form part of the decision-making process.

References

1. Weinstein, M.C. and W.B. Stason. (1977). Foundations of cost-effectiveness analysis for health and medical practices. *N Eng J Med*;296:716-21.

2. Arrow, K.J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*;53:941-73.

3. Garber, A.M. and C.E. Phelps. (1995). Economic foundations of cost-effectiveness analysis. *National Bureau of Economic Research*.

4. Kamlet, M.S. (1992). The comparative benefits modeling project: A framework for cost-utility analysis of government health care programs. Washington, DC: U.S. Department of Health and Human Services, Public Health Service.

5. Gold, M.R., J.E. Siegel, L.B. Russell and M.C. Weinstein. (1996). Cost-effectiveness in Health and Medicine (*Oxford University Press*, New York).

6. Karlsson G., and M. Johannesson. (1996). The decision rules of cost-effectiveness analysis. *PharmacoEconomics*;9:113-120.

7. Mullahy, J., and W.G. Manning. (1994). Statistical issues in cost-effectiveness analysis, Valuing Health Care: Costs, Benefits and Effectiveness of Pharmaceuticals and Other Medical Technologies (*Cambridge University Press*, New York).

8. O'Brien, B.J., M.F. Drummond, R.J. Labelle and A. Willan. (1994). In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Medical Care*;32:150-163.

9. Briggs, A., M. Sculpher and M. Buxton. (1994). Uncertainty in the economic evaluation of health care technologies: the role of sensitivity analysis, *Health Economics*;3:95-104.

10. Eddy, D.M. (1990). Screening for cervical cancer. *Ann Intern Med*;113:214-26.

11. U.S. Cancer Statistics Working Group. (2013). United States Cancer Statistics: 1999–2009 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; Available at: http://www.cdc.gov/uscs.

12. Humphrey LL, Helfand M, Chan BKS, Woolf SH. (2002). Breast cancer screening: a summary of the evidence for the US Preventive Services Task Force. *Ann Intern Med*;137(5):347-

60.

13. Nelson HD, Tyne K, Naik A, Bougatsos C, Chan BK, Humphrey L. (2009). Screening for breast cancer: an update for the U.S. Preventive Services Task Force. *Ann Intern Med*;151(10):727-37.

14. Moss SM, Cuckle H, Evans A, Johns L, Waller M, Bobrow L. (2006). Effect of mammographic screening from age 40 years on breast cancer mortalityat 10 years' follow up: a randomized controlled trial. *Lancet*;368(9):2053-60.

15. Smith RA, Duffy SW, Gabe R, Tabar L, Yen AMF, Chen THH. (2004). The randomized trials of breast cancer screening: what have we learned? *Radiol Clin N Am*;42:793-806.

16. Wu D, Perez A. (2011). A limited review of overdiagnosis methods and long-term effects in breast cancer screening. *Oncol Rev*;5(3):143-7.

17. Smith RA, Saslow D, Sawyer KA, Burke W, Costanza ME, Evans WP, Foster RS, Hendrick E, Eyre HJ, Sener S. (2003). American Cancer Society guidelines for breast cancer screening: update 2003. *Cancer*;53(3):141-69.

18. Calogne N, Petitti DB, DeWitt TG, Dietrich AJ, Gregory KD, et al. (2009). Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med*;151(10):716-26.

19. Welch HG, Black WC. (2010). Over-diagnosis in cancer. *J Natl Cancer Inst*;102(9):605-13.

20. Zahl P-H, Maehlen J, Welch HG. (2008). The natural history of invasive breast cancers detected by screening mammography. *Arch Intern Med*;168(21):2311-16.

21. Bleyer A, Welch HG. (2012). Effect of three decades of screening mammography on breast-cancer incidence. *N Engl J Med*;367:1998-2005.

22. Duffy SW, Agbaje O, Tabar L, Vitak B, Bjurstam N, Bjorneld L, Myles JP, Warwick J. (2005). Estimates of over-diagnosis from two trials of mammographic screening for breast cancer. *Breast Cancer Res*;7(6):258-265.

23. Zackrisson S, Andersson I, Janzon L, Manher J, Garne JP. (2006). Rate of over-diagnosis of breast cancer 15 years after end of Malmo mammographic screening trials: follow-up study. *Br Med J*;322:689.

24. Paci E, Warwick J, Falini P, Duffy SW. (2004). Over-diagnosis in screening: is the increase in breast cancer incidence rates a cause for concern? *J Med Screen*;11:23-7.

25. Jonsson H, Johansson R, Lenner P. (2005). Increased incidence of invasive breast cancer after the introduction of service screening with mammography in Sweden. *Int J Cancer*; 117:842-7

26. Olsen AH, Jensen A, Njor SH, Villadsen E, Schwartz W, Vejborg I, Lynge E. (2003). Breast cancer incidence after the start of mammography screening in Denmark. *Br J Cancer*;88:362-5.

27. Zahl PH, Strand BH, Maehlen J. (2004). Incidence of breast cancer in Norway and Sweden during introduction of nationwide screening: prospective cohort study. *Br Med J*;328:921-4.

28. Maillart LM, Ivy JS, Ransom S, Diehl K. (2008). Assessing dynamic breast cancer screening policies. *Operations Res*;56(6):1411-27.

29. Colbert JA, Adler JN. (2013). Mammography Screening — Polling Results. *New England Journal of Medicine*;368:e12.

30. Pace LE, He Y, Keating NL. (2013). Trends in mammography screening rates after publication of the 2009 US Preventive Services Task Force recommendations. *Cancer*; doi: 10.1002/cncr.28105.

31. Arias E. United States life table 2004. *Natl Vital Stat Rep* 2007;54(14) 1-39.

32. Ayer T, Alagoz O, Stout NK. (2012). A POMDP approach to personalized mammography screening decisions. *Operations Res*;60(5):1110-9.

33. Sonnenberg FA, Beck JR. (1993). Markov models in medical decision making: A practical guide. *Med Decis Making*;13(4):322-38

34. Schairer C, Mink PJ. Carroll L, Devesa SS. (2004). Probabilities of death from breast cancer and other causes among female breast cancer patients. *JNCI J Natl Cancer Inst*;96(17):1311–1321

35. Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F et al. (1988). Mammographic screening and mortality from breast cancer: the Malmo mammographic screening trial. *BMJ*;297:943-8.

36. Peeters PH, Verbeek AL, Straatman H, Holland R, Hendriks JH, Mravunac M, et al. (1989). Evaluation of over-diagnosis of breast cancer in screening with mammography: results of the Nijmegen programme. *Int J Epidemiol*; 18:295-9

37. Gotzsche PC. (2004). On the benefits and harms of screening for breast cancer. *Int J*

Epidemiol;33:56-64

38. Gotzsche PC and Nielsen M. (2011). Screening for breast cancer with mammography. *Cochrane Database Syst Rev* (1):CD001877.

39. Centers for Disease Control and Prevention (CDC). (1995). The national breast and cervical cancer early detection program at a glance. *U.S. Department of Health and Human Services*. Centers for Disease Control and Prevention, Atlanta.

40. American Hospital Association. (2000). Hospital Statistics 2000. Chicago,III. American Hospital Association.

41. Jorgensen KJ, Gotzsche PC. (2009). Over-diagnosis in publicly organized mammography screening programmes: systematic review of incidence trend. *BMJ*;339:b2587.

42. Miller AB, Wall C, Baines CJ, SUn P, To T, Narod SA. (2014). Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ*;348:g366.

43. Lydia E. Pace, Nancy L. Keating. (2014). A systematic assessment of benefits and risks to guide breast cancer screening decisions. *The Journal of the American Medical Association*;Vol 311, No.13.

44. Miller AB, To T, Baines CJ, Wall C. (2002). The Canadian National Breast Screening Study-1: breast cancer mortality after 11 to 16 years of follow-up: a randomized screening trial of mammography in women age 40 to 49 years. *Ann Intern Med*;137(5, part 1):305-312.

45. Miller AB, To T, Baines CJ, Wall C. (2000). Canadian National Breast Screening Study-2: 13-year results of a randomized trial in women aged 50-59 years. *J Natl Cancer Inst*;92(18):1490-1499.

46. Gotzsche PC. (2006). Ramifications of screening for breast cancer: overdiagnosis in the Malmo trial was considerably underestimated. *BMJ*;332(7543):727.

47. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. (2013). The benefits and harms of breast cancer screening: an independent review. *Br J Cancer*;108(11):2205-2240.

48. Lyman GH, Giulaian AE, Somerfield MR, Benson AB, Bodurka DC, Burstein HJ, et al. (2005). American Society of Clinical Oncology guideline recommendations for sentinel lymph node biopsy in early-stage breast cancer. *J Clin Oncol*;23(30):7703-20.

49. Anthony BM, Claus W, Cornelia JB, Ping S, Teresa T, Steven AN. (2014). Twenty five

year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomized screening trial. *BMJ*;348:g366.

50. James R, Maria C. (2011). Possible net harms of breast cancer screening: updated modelling of Forrest report. *BMJ*;343:d7627.

51. Verna EC, Brown RS Jr. (2006). Hepatitis C virus and liver transplantation. *Clin Liver Dis*; 10:919-940.

52. Strader DB, Wright T, Thomas DL, et al. (2004). Diagnosis, management, and treatment of hepatitis C. *Hepatology*;39:1147-71.

53. Awad T, Thorlund K, Hauser G, Stimac D, Mabrouk M, Gluud C. (2010). Peginterferon alpha-2a is associated with higher sustained virological response than peginterferon alfa-2b in chronic hepatitis C: systematic review of randomized trials. *Hepatology*;51:1176–84.

54. Averhoff FM, Glass N, Holtzman D. (2012). Global burden of hepatitis C: considerations for healthcare providers in the United States. *Clin Infect Dis*;55 Suppl 1:S10–5.

55. Chen S.L, Morgan T.R. (2006). The natural history of hepatitis C virus (HCV) infection. *Int J Med Sci*; 3:47-52.

56. Sullivan S.D, Jensen M, Bernstein D.E, Hassanein T.I, Foster G.R, Lee S.S, Cheinquer H, Craxi A, Cooksley G., Klaskala W, Perttit K, Patel K.K, Green J. (2004). Cost-effectiveness of combination peginterferon alpha-2a and ribavirin compared with interferon alpha-2b and ribavirin in patients with chronic hepatitis C. *Am J Gastroenterol*;99:1490-96.

57. Sangiovanni A, Prati GM, Fasani P, et al. (2006). The natural history of compensated cirrhosis due to hepatitis C virus: a 17-year cohort study of 214 persons. *Hepatolog.y*;43:1303-10.

58. Ghany MG, Strader DB, Thomas DL,et al. (2009). Diagnosis, management, and treatment of hepatitis C: an update. *Hepatology*;49:1335-74.

59. Heathcote EJ, Shiffman ML, Cooksley WG, et al. (2000). Peginterferon alpha-2a in patients with chronic hepatitis C and cirrhosis. *N Engl J Med*; 343:1673-80.

60. Fattovich G, Guistina G, Degas E. (1997). Morbidity and mortality in compensated cirrhosis type C: a retrospective follow up study of 384 patients. *Gastroenterology*;112:463-72.

61. Siebert U, Sroczynski G, Rossol S, Wasem J. (2003). Cost effectiveness of peginterferon alpha-2b plus ribavirin versus interferon alpha-2b plus ribavirin for initial treatment of chronic hepatitis C. *Gut*.;52:425-32.

62. Tucker M. FDA approves 'game changer' hepatitis C drug Sofosbuvir''. Medscape.

December 6, 2013. http://www.medscape.com/viewarticle/817371 (accessed 1 Dec 2014).

63. Singer M, Younossi Z. (2001). Cost effectiveness of screening for hepatitis C virus in asymptomatic, average –risk adults. *Am J Med*; 111:614-21.

64. Jacobson IM, Gordon SC, Kowdley KV, Yoshida EM, Rodriguez-Torres M, Sulkowski MS, Shiffman ML, Lawitz E, Everson G, Bennett M, Schiff E, Al-Assi MT, Subramanian GM, An D, Lin M, McNally J, Brainard D, Symonds WT, McHutchison JG, Patel K, Feld J, Pianko S, Nelson DR. (2013). Sofosbuvir for hepatitis C genotype 2 or 3 in patients without treatment options. *N Engl J Med*;368:1867-77

65. Lawitz E, Mangia A, Wyles D, Rodriguez-Torres M, Hassanein T, Gordon SC, Schultz M, Davis MN, Kayali Z, Reddy KR, Jacobson IM, Kowdley KV, Nyberg L, Subramanian GM, Hyland RH, Arterburn S, Jiang D, McNally J, Brainard D, Symonds WT, McHutchison JG, Sheikh AM, Younossi Z, Gane EJ. (2013). Sofosbuvir for previously untreated chronic hepatitis C infection. *N Engl J Med*;368:1878-87.

66. Sovaldi drug description.
http://www.rxlist.com/sovaldi-drug.htm. (accessed 1 Mar 2015).

67. U.S. Food and Drug Administration Approves Gilead's Sovaldi™ (Sofosbuvir) for the Treatment of Chronic Hepatitis C. Gilead [eLetter] 6 December 2013.
http://www.gilead.com/news/press-releases/2013/12/us-food-and -drug-administration-approves-gileads-sovaldi-sofosbuvir-for-the-treatment-of-chronic-hepatitis-c#sthash.T9uTbSWK.dpuf". (accessed 1 Dec 2014).

68. Peginterferon alfa-2a (Pegasys). Hepatitis C Online.
http://www.hepatitisc. uw.edu/page/treatment/drugs/peginterferon-alfa-drug. (accessed 1 Dec 2014).

69. Liu S, Watcha D, Holodniy M, Goldhaber-Fiebert, J. (2014). Sofosbuvir-based treatment regimens for chronic, genotype 1 hepatitis C virus infection in U.S. incarcerated populations. *Ann Intern Med*;161:546-53.

70. Bennett WG, Inoue Y, Beck JR, Wong JB, Pauker SG, Davis GL. (1997). Estimates of the cost-effective of a single course of interferon-alpha2b in patients with histologically mild chronic hepatitis C. *Ann Intern Med*;127:855-65.

71. Davis GL, Alter MJ, El-Serag H, Poynard R, Jennings LW. (2010). Aging of the hepatitis C virus infected persons in the United States: a multiple cohort model of HCV prevalence and

disease progression. *Gastroenterology*;138:513-21.

72. Kim WR, Poterucha JJ, Hermans JE, Therneau TM, Dickson ER, Evans RW, Gross JB. (1997). Cost-effectiveness of 6 and 48 weeks of interferon-alpha therapy for chronic hepatitis C. *Ann Intern Med*; 127:866-74.

73. Younossi Z, Singer M, Mchutchison J, Shermock K. (1999). Cost effectiveness of interferon alpha2b combined with ribavirin for the treatment of chronic hepatitis C. *Hepatology*;30:1318-24.

74. Salomon JA, Weinstein MC, Hammitt JK. (2003). Cost-effectiveness of treatment for chronic hepatitis C infection in an evolving patient population. *JAMA*;290:228-37.

75. Fattovich G, Guistina G, Degas E. (1997). Morbidity and mortality in compensated cirrhosis type C: a retrospective follow up study of 384 patients. *Gastroenterology*;112:463-72.

76. E Arias. United States Life Tables, 2008. *National Vital Statistics Reports*. 2012;61. http://www.cdc.gov/nchs/data/nvsr/nvsr59/nvsr59_10.pdf. (accesses 1 Dec 2014).

77. Degos F, Christidis C, Ganne-Carrie N, Farmachidi J.P, Degott C, Guettier C, Trinchet J.C, Beaugrand M, Chevret S. (2000). Hepatitis C virus related cirrhosis: time to occurrence of hepatocellular carcinoma and death. *Gut*;47:131-36.

78. Fattovich G, Stroffolini T, Zagni I, Donato F. (2004). Hepatocellular carcinoma in cirrhosis: incidence and risk factors. *Gastroenterology*;127:S35-50.

79. Forman L. M, Lewis J.D, Berlin J.A, Feldman H.I, Lucey M.R. (2002). The association between hepatitis C infection and survival after orthotopic liver transplantation. *Gastroenterology*;122:889-96.

80. Salvatore Petta, Giuseppe Cabibbo et al. (2014). Cost-Effectiveness of Sofosbuvir-Based Triple Therapy for Untreated Patients with Genotype 1 Chronic Hepatitis C. *Hepatology*.

81. Jacobson IM, McHutchison JG, Dusheiko G et al. (2011). Telaprevir for previously untreated chronic hepatitis C virus infection. *N Engl J Med*; 364(25):2405-2416.

82. Ledipasvir-Sofosbuvir (Harvoni). Hepatitis C Online. http://www.hepatitisc.uw.edu/page/treatment/drugs/ledipasvir-sofosbuvir (accessed 1 Mar 2015).

83. Simeprevir (Olysio). Hepatitis C Online. http://www.hepatitisc.uw.edu/page/treatment/drugs/simeprevir-drug (accessed 1 Mar 2015).

84. Ombitasvir-Paritaprevir-Ritonavir and Dasabuvir (Viekira Pak). Hepatitis C Online.

http://www.hepatitisc.uw.edu/page/treatment/drugs/3d (accessed 1 Mar 2015).

85. Younossi Z, McCormick M, Boparai N, Price L, Fqrquhar L, Guyatt G. Assessment of utilities and health-related quality of life in patients with chronic liver disease. Gatroenterology 1999; 116:A1292.

86. Bureau of Labor Statistics

http://data.bls.gov/timeseries/CUUR0000SAM?output_view=pct_12mths (accessed 1 Mar 2015).

87. Chalfin D.B., Trzeciak S., Likourezos A., Baumann B.M., Dellinger R.P. (2007). Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine*; 35:1477-1483.

88. Chan Carri W., Farias Vivek F., Bambos Nicholas, Escobar Gabriel J. (2011). Optimizing ICU discharge decisions with patient readmissions. *Operations Research.*

89. Green, L. V. (2003). How many hospital beds? *Inquiry*;39 400-412.

90. Huang Xiaoming. (1995). A planning model for requirement of emergency beds. *IMA Journal of Mathematics Applied in Medicine and Biology*;12,345-353

91. Kao Edward P.C.and Tung Grace G. (1981). Bed allocation in a public health care delivery system. *Management Science*; Vol.27, No. 5.

92. Diwas, K.C., Terwiesch C..(2012). An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management*; 14(1) 50-65.

93. Kim, S-H, Chan C.W., Olivares M., Escobar G. (2012). ICU Admission Control: An Empirical Study of Capacity Allocation and its Implication on Patient Outcomes. Columbia Business School Research Paper.

94. Shonick William and Jackson James R. (1973). An Improved Stochastic Model for Occupancy-Related Random Variables in General-Acute Hostpitals. *Operations Research*;Vol.21 No.4, pp 952-65.

95. Shmueli Amir, Sprung Charles L., Kaplan Edward H. (2003). Optimizing admissions to an intensive care unit. Health Care *Management Science*;6,131-136.

96. Sissouras A. A. and Moores B. (1976). The Optimum Number of Beds in a Coronary Care Unit, OMEGA;Vol. 4, pp.59-65

97. Thompson J.D. and Fetter R.B. (1964). Research Helps Calculate OB Bed Needs. Mod. Hosp. Vol.102, pp.98-101

98. Zhu Bo, Armony Mor, Chan Carr W. (2013). Critical care in hospitals: When to introduce a Step Down Unit?

99. New York Times. (2002). 1 in 3 hospitals say they divert ambulances. (April 9).

100. American Joint Committee on Cancer https://cancerstaging.org/references-tools/quickreferences/Documents/BreastMedium.pdf. (accessed 25 April 2015).

101. Knaus W, Wagner D, Draper E, et al. (1991). The Apache Iii Prognostic System. Risk Prediction Of Hospital Mortality For Critically Ill Hospitalized Adults. *Chest* 1991;100(6):1619-1636.

102. Halpern N.A., Pastores S.M. (2010). Critical care medicine in the United States 2000-2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Crit Care Med*;38 65-71.

103. Barro, J.R., Cutler D.M. (1997). Consolidation in theMedical Care Marketplace: A Case Study from Massachusetts. NBER Working Paper 5957. Cambridge, National Bureau of Economic Research.

104. Rivera A., Dasta J.F., Varon J. (2009). Critical Care Economics. *Critical Care & Shock;* 12 124-129.

105. Allon G., Deo S., Lin W. (2013). The impact of hospital size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research;* 61 554-562.

Appendix

ICU Stepdowns Supplements

The definition domain D is given as follows. The balanced model only works when $Q_I < \infty, Q_G < \infty$, otherwise system has infinity states. Using a birth-death process, steady-state balanced equations can be obtained.

$$\begin{aligned} D = \{ & (p,q,u)(i,j,v): \\ & 0 \le p+q < N_G, u = 0; p+q = N_G, u = 0,1,...,Q_G \\ & 0 \le i+j < N_I, v = 0; i+j = N_I, v = 0,1,...,Q_I \\ & if\ i+j+v \ge M,\ j = 0\ or\ j > 0, p+q = N_G \} \end{aligned}$$

If threshold M satisfies $1 \le M \le N_I$, the total number of equations is

$$\frac{(1+N_G)N_G}{2}[\frac{(1+M)M}{2} + (N_I - M + 1) + Q_I] + (N_G+1)(Q_G+1)[\frac{(1+N_I)N_I}{2} + (N_I+1)(Q_I+1)]$$

If $N_I + 1 \le M \le N_I + Q_I$, the total number of equations is:

$$\frac{(1+N_G)N_G}{2}[\frac{(1+N_I)N_I}{2} + Q_I + 1] + (N_G+1)(Q_G+1)[\frac{(1+N_I)N_I}{2} + (N_I+1)(Q_I+1)]$$

When $1 \le M \le N_I$ and state 2 patient has higher priority

$$\begin{aligned} & C_{(p,q,0)(i,j,0)}(\lambda_G + p\mu_G + q\mu_I + \lambda_I + i\mu_1 + j\mu_2) = C_{(p-1,q,0)(i,j,0)}1_{\{p\ge 1\}}\lambda_G + C_{(p,q-1,0)(i-1,j+1,0)}1_{\{q\ge 1 \& i+j=M-1 \& i\ge 1\}}\lambda_I \\ & + C_{(p,q-1,0)(i+1,0,0)}1_{\{q\ge 1 \& i=M-1\}}(i+1)\mu_1(1-h) + C_{(p+1,q,0)(i,j,0)}(p+1)\mu_G + C_{(p,q+1,0)(i,j,0)}(q+1)\mu_I + C_{(p,q,0)(i-1,j,0)}1_{\{i\ge 1\}}\lambda_I \\ & + C_{(p,q,0)(i+1,j-1,0)}1_{\{j\ge 1\}}(i+1)\mu_1(1-h) + C_{(p,q,0)(i+1,j,0)}1_{\{i+j\le M-2 \| j=0\}}(i+1)\mu_1 h + C_{(p,q,0)(i,j+1,0)}1_{\{i+j\le M-2\}}(j+1)\mu_2 \\ & \qquad\qquad p+q \le N_G - 1, u = 0, i+j \le M-1 \end{aligned}$$

$$\begin{aligned} & C_{(p,q,0)(i,0,0)}(\lambda_G + p\mu_G + q\mu_I + \lambda_I + i\mu_1) = C_{(p-1,q,0)(i,0,0)}1_{\{p\ge 1\}}\lambda_G + C_{(p,q-1,0)(i+1,0,0)}1_{\{q\ge 1\}}(i+1)\mu_1(1-h) \\ & + C_{(p+1,q,0)(i,0,0)}(p+1)\mu_G + C_{(p,q+1,0)(i,0,0)}(q+1)\mu_I + C_{(p,q,0)(i+1,0,0)}(i+1)\mu_1 h + C_{(p,q,0)(i-1,0,0)}1_{\{i\ge 1\}}\lambda_I \\ & \qquad\qquad p+q \le N_G - 1, u = 0, M \le i \le N_I - 1, j = 0 \end{aligned}$$

$$\begin{aligned} & C_{(p,q,0)(N_I,0,v)}(\lambda_G + p\mu_G + q\mu_I + 1_{\{v\le Q_I - 1\}}\lambda_I + N_I\mu_1) = C_{(p-1,0)(N_I,0,v)}1_{\{p\ge 1\}}\lambda_G + C_{(p,q-1,0)(N_I,0,v+1)}1_{\{q\ge 1 \& v\le Q_I - 1\}}N_I\mu_1(1-h) \\ & + C_{(p+1,q,0)(N_I,0,v)}(p+1)\mu_G + C_{(p,q+1,0)(N_I,0,v)}(q+1)\mu_I + C_{(p,q,0)(N_I-1,0,0)}1_{\{v=0\}}\lambda_I + C_{(p,q,0)(N_I,0,v-1)}1_{\{v\ge 1\}}\lambda_I \\ & + C_{(p,q,0)(N_I,0,v+1)}1_{\{v\le Q_I - 1\}}N_I\mu_1 h \\ & \qquad\qquad p+q \le N_G - 1, u = 0, i = N_I, j = 0, 0 \le v \le Q_I \end{aligned}$$

$$C_{(p,q,0)(i,j,0)}(1_{\{Q_G\geq 1\}}\lambda_G + p\mu_G + q\mu_I + \lambda_I + i\mu_1 + j\mu_2) = C_{(p-1,q,0)(i,j,0)}1_{\{p\geq 1\&(i+j\leq M-1\|j=0)\}}\lambda_G$$
$$+C_{(p,q-1,0)(i+1,0,0)}1_{\{q\geq 1\&i\geq M-1\&j=0\}}(i+1)\mu_1(1-h) + C_{(p,q-1,0)(i-1,j+1,0)}1_{\{q\geq 1\&i+j=M-1\&i\geq 1\}}\lambda_I + C_{(p,q,1)(i,j,0)}1_{\{i+j\leq M-1\|j=0\}}p\mu_G$$
$$+C_{(p-1,q+1,1)(i,j,0)}1_{\{p\geq 1\&(i+j\leq M-1\|j=0)\}}(q+1)\mu_I + C_{(p,q,0)(i-1,j,0)}1_{\{i\geq 1\}}\lambda_I + C_{(p,q,0)(i+1,j-1,0)}1_{\{j\geq 1\}}(i+1)\mu_1(1-h)$$
$$+C_{(p,q,0)(i+1,j,0)}(i+1)\mu_1 h + C_{(p,q,0)(i,j+1,0)}\{(j+1)\mu_2 + 1_{\{i+j\geq M-1\}}q\mu_I\} + C_{(p+1,q-1,0)(i,j+1,0)}1_{\{i+j\geq M-1\&q\geq 1\}}(p+1)\mu_G$$
$$p+q=N_G, u=0, i+j\leq N_I-1$$

$$C_{(p,q,0)(i,j,v)}(1_{\{Q_G\geq 1\}}\lambda_G + p\mu_G + q\mu_I + 1_{\{v\leq Q_I-1\}}\lambda_I + i\mu_1 + j\mu_2) = C_{(p-1,q,0)(N_I,0,v)}1_{\{j=0\}}\lambda_G$$
$$+C_{(p,q-1,0)(N_I,0,v+1)}1_{\{j=0\&v\leq Q_I-1\}}N_I\mu_1(1-h) + C_{(p,q,1)(N_I,0,v)}1_{\{Q_G\geq 1\&j=0\}}p\mu_G + C_{(p-1,q+1,1)(N_I,0,v)}1_{\{Q_G\geq 1\&j=0\}}(q+1)\mu_I$$
$$+C_{(p,q,0)(i-1,j,0)}1_{\{i\geq 1\&v=0\}}\lambda_I + C_{(p,q,0)(i,j,v-1)}1_{\{v\geq 1\}}\lambda_I + C_{(p,q,0)(i+1,j-1,v)}1_{\{j\geq 1\}}(i+1)\mu_1(1-h) + C_{(p,q,0)(i,j,v+1)}1_{\{v\leq Q_I-1\}}i\mu_1 h$$
$$+C_{(p,q,0)(i-1,j+1,v+1)}1_{\{v\leq Q_I-1\&i\geq 1\}}\{(j+1)\mu_2 + q\mu_I\} + C_{(p+1,q-1,0)(i-1,j+1,v+1)}1_{\{v\leq Q_I-1\&i\geq 1\}}(p+1)\mu_G$$
$$p+q=N_G, u=0, i+j=N_I, 0\leq v\leq Q_I$$

$$C_{(p,q,u)(i,j,0)}(1_{\{u\leq Q_G-1\}}\lambda_G + p\mu_G + q\mu_I + \lambda_I + i\mu_1 + j\mu_2) = C_{(p,q,u-1)(i,j,0)}\lambda_G + C_{(p,q,u+1)(i,j,0)}1_{\{u\leq Q_G-1\&(i+j\leq M-1\|j=0)\}}p\mu_G$$
$$+C_{(p-1,q+1,u+1)(i,j,0)}1_{\{p\geq 1\&u\leq Q_G-1\&(i+j\leq M-1\|j=0)\}}(q+1)\mu_I + C_{(p,q,u)(i-1,j,0)}1_{\{i\geq 1\}}\lambda_I + C_{(p,q,u)(i+1,j-1,0)}1_{\{j\geq 1\}}(i+1)\mu_1(1-h)$$
$$+C_{(p,q,u)(i+1,j,0)}(i+1)\mu_1 h + C_{(p,q,u)(i,j+1,0)}\{(j+1)\mu_2 + 1_{\{i+j\geq M-1\}}q\mu_I\} + C_{(p+1,q-1,u)(i,j+1,0)}1_{\{q\geq 1\&i+j\geq M-1\}}(p+1)\mu_G$$
$$p+q=N_G, 1\leq u\leq Q_G, i+j\leq N_I-1$$

$$C_{(p,q,u)(i,j,v)}(1_{\{u\leq Q_G-1\}}\lambda_G + p\mu_G + q\mu_I + 1_{\{v\leq Q_I-1\}}\lambda_I + i\mu_1 + j\mu_2) = C_{(p,q,u-1)(i,j,v)}\lambda_G + C_{(p,q,u+1)(i,j,v)}1_{\{u\leq Q_G-1\&j=0\}}p\mu_G$$
$$+C_{(p-1,q+1,u+1)(i,j,v)}1_{\{p\geq 1\&u\leq Q_G-1\&j=0\}}(q+1)\mu_I + C_{(p,q,u)(i-1,j,0)}1_{\{i\geq 1\&v=0\}}\lambda_I + C_{(p,q,u)(i,j,v-1)}1_{\{v\geq 1\}}\lambda_I$$
$$+C_{(p,q,u)(i+1,j-1,v)}1_{\{j\geq 1\}}(i+1)\mu_1(1-h) + C_{(p,q,u)(i,j,v+1)}1_{\{v\leq Q_I-1\}}i\mu_1 h + C_{(p,q,u)(i-1,j+1,v+1)}1_{\{i\geq 1\&v\leq Q_I-1\}}\{(j+1)\mu_2 + q\mu_I\}$$
$$+C_{(p+1,q-1,u)(i-1,j+1,v+1)}1_{\{q\geq 1\&i\geq 1\&v\leq Q_I-1\}}(p+1)\mu_G$$
$$p+q=N_G, 1\leq u\leq Q_G, i+j=N_I, 0\leq v\leq Q_I$$

When $1\leq M\leq N_I$ and GCU patient has higher priority

$$C_{(p,q,0)(i,j,0)}(\lambda_G + p\mu_G + q\mu_I + \lambda_I + i\mu_1 + j\mu_2) = C_{(p-1,q,0)(i,j,0)}1_{\{p\geq 1\}}\lambda_G + C_{(p,q-1,0)(i-1,j+1,0)}1_{\{q\geq 1\&i+j=M-1\&i\geq 1\}}\lambda_I$$
$$+C_{(p,q-1,0)(i+1,0,0)}1_{\{q\geq 1\&i=M-1\}}(i+1)\mu_1(1-h) + C_{(p+1,q,0)(i,j,0)}(p+1)\mu_G + C_{(p,q+1,0)(i,j,0)}(q+1)\mu_I + C_{(p,q,0)(i-1,j,0)}1_{\{i\geq 1\}}\lambda_I$$
$$+C_{(p,q,0)(i+1,j-1,0)}1_{\{j\geq 1\}}(i+1)\mu_1(1-h) + C_{(p,q,0)(i+1,j,0)}1_{\{i+j\leq M-2\|j=0\}}(i+1)\mu_1 h + C_{(p,q,0)(i,j+1,0)}1_{\{i+j\leq M-2\}}(j+1)\mu_2$$
$$p+q\leq N_G-1, u=0, i+j\leq M-1$$

$$C_{(p,q,0)(i,0,0)}(\lambda_G + p\mu_G + q\mu_I + \lambda_I + i\mu_1) = C_{(p-1,q,0)(i,0,0)}1_{\{p\geq 1\}}\lambda_G + C_{(p,q-1,0)(i+1,0,0)}1_{\{q\geq 1\}}(i+1)\mu_1(1-h)$$
$$+C_{(p+1,q,0)(i,0,0)}(p+1)\mu_G + C_{(p,q+1,0)(i,0,0)}(q+1)\mu_I + C_{(p,q,0)(i-1,0,0)}1_{\{i\geq 1\}}\lambda_I + C_{(p,q,0)(i+1,0,0)}(i+1)\mu_1 h$$
$$p+q\leq N_G-1, u=0, M\leq i\leq N_I-1, j=0$$

$$C_{(p,q,0)(N_I,0,v)}(\lambda_G + p\mu_G + q\mu_I + 1_{\{v\leq Q_I-1\}}\lambda_I + N_I\mu_1) = C_{(p-1,0)(N_I,0,v)}1_{\{p\geq 1\}}\lambda_G + C_{(p,q-1,0)(N_I,0,v+1)}1_{\{q\geq 1\&v\leq Q_I-1\}}N_I\mu_1(1-h)$$
$$+C_{(p+1,q,0)(N_I,0,v)}(p+1)\mu_G + C_{(p,q+1,0)(N_I,0,v)}(q+1)\mu_I + C_{(p,q,0)(N_I-1,0,0)}1_{\{v=0\}}\lambda_I + C_{(p,q,0)(N_I,0,v-1)}1_{\{v\geq 1\}}\lambda_I$$
$$+C_{(p,q,0)(N_I,0,v+1)}1_{\{v\leq Q_I-1\}}N_I\mu_1 h$$
$$p+q\leq N_G-1, u=0, i=N_I, j=0, 0\leq v\leq Q_I$$

$$C_{(p,q,0)(i,j,0)}(1_{\{Q_G\geq 1\}}\lambda_G + p\mu_G + q\mu_I + \lambda_I + i\mu_1 + j\mu_2) = C_{(p-1,q,0)(i,j,0)}1_{\{p\geq 1\&(i+j\leq M-1\|j=0)\}}\lambda_G$$
$$+C_{(p,q-1,0)(i+1,0,0)}1_{\{q\geq 1\&i\geq M-1\&j=0\}}(i+1)\mu_1(1-h) + C_{(p,q-1,0)(i-1,j+1,0)}1_{\{q\geq 1\&i+j=M-1\&i\geq 1\}}\lambda_I + C_{(p,q,1)(i,j,0)}p\mu_G$$
$$+C_{(p-1,q+1,1)(i,j,0)}1_{\{p\geq 1\}}(q+1)\mu_I + C_{(p,q,0)(i-1,j,0)}1_{\{i\geq 1\}}\lambda_I + C_{(p,q,0)(i+1,j-1,0)}1_{\{j\geq 1\}}(i+1)\mu_1(1-h)$$
$$+C_{(p,q,0)(i+1,j,0)}(i+1)\mu_1 h + C_{(p,q,0)(i,j+1,0)}\{(j+1)\mu_2 + 1_{\{i+j\geq M-1\}}q\mu_I\} + C_{(p+1,q-1,0)(i,j+1,0)}1_{\{i+j\geq M-1\&q\geq 1\}}(p+1)\mu_G$$
$$p+q=N_G, u=0, i+j\leq N_I-1$$

$$C_{(p,q,0)(i,j,v)}(1_{\{Q_G\geq 1\}}\lambda_G + p\mu_G + q\mu_I + 1_{\{v\leq Q_I-1\}}\lambda_I + i\mu_1 + j\mu_2) = C_{(p-1,q,0)(N_I,0,v)}1_{\{p\geq 1\&j=0\}}\lambda_G$$
$$+C_{(p,q-1,0)(N_I,0,v+1)}1_{\{q\geq 1\&j=0\&v\leq Q_I-1\}}N_I\mu_1(1-h) + C_{(p,q,1)(i,j,v)}1_{\{Q_G\geq 1\}}p\mu_G + C_{(p-1,q+1,1)(i,j,v)}1_{\{Q_G\geq 1\&p\geq 1\}}(q+1)\mu_I$$
$$+C_{(p,q,0)(i-1,j,0)}1_{\{i\geq 1\&v=0\}}\lambda_I + C_{(p,q,0)(i,j,v-1)}1_{\{v\geq 1\}}\lambda_I + C_{(p,q,0)(i+1,j-1,v)}1_{\{j\geq 1\}}(i+1)\mu_1(1-h) + C_{(p,q,0)(i,j,v+1)}1_{\{v\leq Q_I-1\}}i\mu_1 h$$
$$+C_{(p,q,0)(i-1,j+1,v+1)}1_{\{v\leq Q_I-1\&i\geq 1\}}\{(j+1)\mu_2 + q\mu_I\} + C_{(p+1,q-1,0)(i-1,j+1,v+1)}1_{\{q\geq 1\&v\leq Q_I-1\&i\geq 1\}}(p+1)\mu_G$$
$$p+q=N_G, u=0, i+j=N_I, 0\leq v\leq Q_I$$

$$C_{(p,q,u)(i,j,0)}(1_{\{u\leq Q_G-1\}}\lambda_G + p\mu_G + q\mu_I + \lambda_I + i\mu_1 + j\mu_2) = C_{(p,q,u-1)(i,j,0)}\lambda_G + C_{(p,q,u+1)(i,j,0)}1_{\{u\leq Q_G-1\}}p\mu_G$$
$$+C_{(p-1,q+1,u+1)(i,j,0)}1_{\{p\geq 1\&u\leq Q_G-1\}}(q+1)\mu_I + C_{(p,q,u)(i-1,j,0)}1_{\{i\geq 1\}}\lambda_I + C_{(p,q,u)(i+1,j-1,0)}1_{\{j\geq 1\}}(i+1)\mu_1(1-h)$$
$$+C_{(p,q,u)(i+1,j,0)}(i+1)\mu_1 h + C_{(p,q,u)(i,j+1,0)}(j+1)\mu_2$$
$$p+q=N_G, 1\leq u\leq Q_G, i+j\leq N_I-1$$

$$C_{(p,q,u)(i,j,v)}(1_{\{u\leq Q_G-1\}}\lambda_G + p\mu_G + q\mu_I + 1_{\{v\leq Q_I-1\}}\lambda_I + i\mu_1 + j\mu_2) = C_{(p,q,u-1)(i,j,v)}\lambda_G + C_{(p,q,u+1)(i,j,v)}1_{\{u\leq Q_G-1\}}p\mu_G$$
$$+C_{(p-1,q+1,u+1)(i,j,v)}1_{\{p\geq 1\&u\leq Q_G-1\}}(q+1)\mu_I + C_{(p,q,u)(i-1,j,0)}1_{\{i\geq 1\&v=0\}}\lambda_I + C_{(p,q,u)(i,j,v-1)}1_{\{v\geq 1\}}\lambda_I$$
$$+C_{(p,q,u)(i+1,j-1,v)}1_{\{j\geq 1\}}(i+1)\mu_1(1-h) + C_{(p,q,u)(i,j,v+1)}1_{\{v\leq Q_I-1\}}i\mu_1 h + C_{(p,q,u)(i-1,j+1,v+1)}1_{\{i\geq 1\&v\leq Q_I-1\}}(j+1)\mu_2$$
$$p+q=N_G, 1\leq u\leq Q_G, i+j=N_I, 0\leq v\leq Q_I$$

When $N_I+1\leq M\leq N_I+Q_I$ and state 2 patient has higher priority

$$C_{(p,q,0)(i,j,0)}(\lambda_G + p\mu_G + q\mu_I + \lambda_I + i\mu_1 + j\mu_2) = C_{(p-1,q,0)(i,j,0)}1_{\{p\geq 1\}}\lambda_G + C_{(p+1,q,0)(i,j,0)}(p+1)\mu_G$$
$$+C_{(p,q+1,0)(i,j,0)}(q+1)\mu_I + C_{(p,q,0)(i-1,j,0)}1_{\{i\geq 1\}}\lambda_I + C_{(p,q,0)(i+1,j-1,0)}1_{\{j\geq 1\}}(i+1)\mu_1(1-h) + C_{(p,q,0)(i+1,j,0)}(i+1)\mu_1 h$$
$$+C_{(p,q,0)(i,j+1,0)}(j+1)\mu_2$$
$$p+q\leq N_G-1, u=0, i+j\leq N_I-1$$

$$C_{(p,q,0)(i,j,v)}(\lambda_G + p\mu_G + q\mu_I + 1_{\{v\leq Q_I-1\}}\lambda_I + i\mu_1 + j\mu_2) = C_{(p-1,q,0)(i,j,v)}1_{\{p\geq 1\}}\lambda_G + C_{(p,q-1,0)(i-1,j+1,v)}1_{\{q\geq 1\&v=M-N_I-1\&i\geq 1\}}\lambda_I$$
$$+C_{(p,q-1,0)(N_I,0,v+1)}1_{\{q\geq 1\&j=0\}}N_I\mu_1(1-h) + C_{(p+1,q,0)(i,j,v)}(p+1)\mu_G + C_{(p,q+1,0)(i,j,v)}(q+1)\mu_I + C_{(p,q,0)(i-1,j,0)}1_{\{i\geq 1\&v=0\}}\lambda_I$$
$$+C_{(p,q,0)(i,j,v-1)}1_{\{v\geq 1\}}\lambda_I + C_{(p,q,0)(i+1,j-1,v)}1_{\{j\geq 1\}}(i+1)\mu_1(1-h) + C_{(p,q,0)(i,j,v+1)}1_{\{v\leq M-2-N_I\|j=0\}}i\mu_1 h$$
$$+C_{(p,q,0)(i-1,j+1,v+1)}1_{\{i\geq 1\&v\leq M-2-N_I\}}(j+1)\mu_2$$
$$p+q\leq N_G-1, u=0, i+j=N_I, v=0,1,...,M-N_I-1$$

$$C_{(p,q,0)(N_I,0,v)}(\lambda_G + p\mu_G + q\mu_I + 1_{\{v\leq Q_I-1\}}\lambda_I + N_I\mu_1) = C_{(p-1,0)(N_I,0,v)}1_{\{p\geq 1\}}\lambda_G + C_{(p,q-1,0)(N_I,0,v+1)}1_{\{q\geq 1\&v\leq Q_I-1\}}N_I\mu_1(1-h)$$
$$+C_{(p+1,q,0)(N_I,0,v)}(p+1)\mu_G + C_{(p,q+1,0)(N_I,0,v)}(q+1)\mu_I + C_{(p,q,0)(N_I,0,v-1)}1_{\{v\geq 1\}}\lambda_I + C_{(p,q,0)(N_I,0,v+1)}1_{\{v\leq Q_I-1\}}N_I\mu_1 h$$
$$p+q\leq N_G-1, u=0, i=N_I, j=0, v=M-N_I,....,Q_I$$

$$C_{(p,q,u)(i,j,0)}(1_{\{u\le Q_G-1\}}\lambda_G+p\mu_G+q\mu_I+\lambda_I+i\mu_1+j\mu_2)=C_{(p-1,q,0)(i,j,0)}1_{\{p\ge 1\&u=0\}}\lambda_G+C_{(p,q,u-1)(i,j,0)}1_{\{u\ge 1\}}\lambda_G$$
$$+C_{(p,q,u+1)(i,j,0)}1_{\{u\le Q_G-1\}}p\mu_G+C_{(p-1,q+1,u+1)(i,j,0)}1_{\{p\ge 1\&u\le Q_G-1\}}(q+1)\mu_I+C_{(p,q,u)(i-1,j,0)}1_{\{i\ge 1\}}\lambda_I$$
$$+C_{(p,q,u)(i+1,j-1,0)}1_{\{j\ge 1\}}(i+1)\mu_1(1-h)+C_{(p,q,u)(i+1,j,0)}(i+1)\mu_1 h+C_{(p,q,u)(i,j+1,0)}(j+1)\mu_2$$
$$p+q=N_G,0\le u\le Q_G,i+j\le N_I-1$$

$$C_{(p,q,0)(i,j,v)}(1_{\{Q_G\ge 1\}}\lambda_G+p\mu_G+q\mu_I+1_{\{v\le Q_I-1\}}\lambda_I+i\mu_1+j\mu_2)=C_{(p-1,q,0)(i,j,v)}1_{\{p\ge 1\&(v\le M-1-N_I\|j=0)\}}\lambda_G$$
$$+C_{(p,q-1,0)(i-1,j+1,v)}1_{\{q\ge 1\&v=M-N_I-1\&i\ge 1\}}\lambda_I+C_{(p,q-1,0)(N_I,0,v+1)}1_{\{q\ge 1\&\,j=0\&M-N_I-1\le v\le Q_I-1\}}N_I\mu_1(1-h)$$
$$+C_{(p,q,1)(i,j,v)}1_{\{Q_G\ge 1\&(v\le M-1-N_I\|j=0)\}}p\mu_G+C_{(p-1,q+1,1)(i,j,v)}1_{\{p\ge 1\&Q_G\ge 1\&(v\le M-1-N_I\|j=0)\}}(q+1)\mu_I+C_{(p,q,0)(i-1,j,0)}1_{\{i\ge 1\&v=0\}}\lambda_I$$
$$+C_{(p,q,0)(i,j,v-1)}1_{\{v\ge 1\}}\lambda_I+C_{(p,q,0)(i+1,j-1,v)}1_{\{j\ge 1\}}(i+1)\mu_1(1-h)+C_{(p,q,0)(i,j,v+1)}1_{\{v\le Q_I-1\}}i\mu_1 h$$
$$+C_{(p,q,0)(i-1,j+1,v+1)}1_{\{v\le Q_I-1\&i\ge 1\}}\{(j+1)\mu_2+1_{\{v\ge M-1-N_I\}}q\mu_I\}+C_{(p+1,q-1,0)(i-1,j+1,v+1)}1_{\{M-1-N_I\le v\le Q_I-1\&i\ge 1\&q\ge 1\}}(p+1)\mu_G$$
$$p+q=N_G,u=0,i+j=N_I,0\le v\le Q_I$$

$$C_{(p,q,u)(i,j,v)}(1_{\{u\le Q_G-1\}}\lambda_G+p\mu_G+q\mu_I+1_{\{v\le Q_I-1\}}\lambda_I+i\mu_1+j\mu_2)=C_{(p,q,u-1)(i,j,v)}\lambda_G$$
$$+C_{(p,q,u+1)(i,j,v)}1_{\{u\le Q_G-1\&(v\le M-1-N_I\|j=0)\}}p\mu_G+C_{(p-1,q+1,u+1)(i,j,v)}1_{\{p\ge 1\&u\le Q_G-1\&(v\le M-1-N_I\|j=0)\}}(q+1)\mu_I$$
$$+C_{(p,q,u)(i-1,j,0)}1_{\{i\ge 1\&v=0\}}\lambda_I+C_{(p,q,u)(i,j,v-1)}1_{\{v\ge 1\}}\lambda_I+C_{(p,q,u)(i+1,j-1,v)}1_{\{j\ge 1\}}(i+1)\mu_1(1-h)$$
$$+C_{(p,q,u)(i,j,v+1)}1_{\{v\le Q_I-1\}}i\mu_1 h+C_{(p,q,u)(i-1,j+1,v+1)}1_{\{i\ge 1\&v\le Q_I-1\}}\{(j+1)\mu_2+1_{\{v\ge M-1-N_I\}}q\mu_I\}$$
$$+C_{(p+1,q-1,u)(i-1,j+1,v+1)}1_{\{q\ge 1\&i\ge 1\&M-1-N_I\le v\le Q_I-1\}}(p+1)\mu_G$$
$$p+q=N_G,1\le u\le Q_G,i+j=N_I,0\le v\le Q_I$$

When $N_I+1\le M\le N_I+Q_I$ and GCU patient has higher priority

$$C_{(p,q,0)(i,j,0)}(\lambda_G+p\mu_G+q\mu_I+\lambda_I+i\mu_1+j\mu_2)=C_{(p-1,q,0)(i,j,0)}1_{\{p\ge 1\}}\lambda_G+C_{(p+1,q,0)(i,j,0)}(p+1)\mu_G$$
$$+C_{(p,q+1,0)(i,j,0)}(q+1)\mu_I+C_{(p,q,0)(i-1,j,0)}1_{\{i\ge 1\}}\lambda_I+C_{(p,q,0)(i+1,j-1,0)}1_{\{j\ge 1\}}(i+1)\mu_1(1-h)+C_{(p,q,0)(i+1,j,0)}(i+1)\mu_1 h$$
$$+C_{(p,q,0)(i,j+1,0)}(j+1)\mu_2$$
$$p+q\le N_G-1,u=0,i+j\le N_I-1$$

$$C_{(p,q,0)(i,j,v)}(\lambda_G+p\mu_G+q\mu_I+1_{\{v\le Q_I-1\}}\lambda_I+i\mu_1+j\mu_2)=C_{(p-1,q,0)(i,j,v)}1_{\{p\ge 1\}}\lambda_G+C_{(p,q-1,0)(i-1,j+1,v)}1_{\{q\ge 1\&v=M-N_I-1\&i\ge 1\}}\lambda_I$$
$$+C_{(p,q-1,0)(N_I,0,v+1)}1_{\{q\ge 1\&\,j=0\}}N_I\mu_1(1-h)+C_{(p+1,q,0)(i,j,v)}(p+1)\mu_G+C_{(p,q+1,0)(i,j,v)}(q+1)\mu_I+C_{(p,q,0)(i-1,j,0)}1_{\{i\ge 1\&v=0\}}\lambda_I$$
$$+C_{(p,q,0)(i,j,v-1)}1_{\{v\ge 1\}}\lambda_I+C_{(p,q,0)(i+1,j-1,v)}1_{\{j\ge 1\}}(i+1)\mu_1(1-h)+C_{(p,q,0)(i,j,v+1)}1_{\{v\le M-2-N_I\|j=0\}}i\mu_1 h$$
$$+C_{(p,q,0)(i-1,j+1,v+1)}1_{\{i\ge 1\&v\le M-2-N_I\}}(j+1)\mu_2$$
$$p+q\le N_G-1,u=0,i+j=N_I,v=0,1,...,M-N_I-1$$

$$C_{(p,q,0)(N_I,0,v)}(\lambda_G+p\mu_G+q\mu_I+1_{\{v\le Q_I-1\}}\lambda_I+N_I\mu_1)=C_{(p-1,q,0)(N_I,0,v)}1_{\{p\ge 1\}}\lambda_G+C_{(p,q-1,0)(N_I,0,v+1)}1_{\{q\ge 1\&v\le Q_I-1\}}N_I\mu_1(1-h)$$
$$+C_{(p+1,q,0)(N_I,0,v)}(p+1)\mu_G+C_{(p,q+1,0)(N_I,0,v)}(q+1)\mu_I+C_{(p,q,0)(N_I,0,v-1)}1_{\{v\ge 1\}}\lambda_I+C_{(p,q,0)(N_I,0,v+1)}1_{\{v\le Q_I-1\}}N_I\mu_1 h$$
$$p+q\le N_G-1,u=0,i=N_I,j=0,v=M-N_I,....,Q_I$$

$$C_{(p,q,u)(i,j,0)}(1_{\{u\le Q_G-1\}}\lambda_G+p\mu_G+q\mu_I+\lambda_I+i\mu_1+j\mu_2)=C_{(p-1,q,0)(i,j,0)}1_{\{p\ge 1\&u=0\}}\lambda_G+C_{(p,q,u-1)(i,j,0)}1_{\{u\ge 1\}}\lambda_G$$
$$+C_{(p,q,u+1)(i,j,0)}1_{\{u\le Q_G-1\}}p\mu_G+C_{(p-1,q+1,u+1)(i,j,0)}1_{\{p\ge 1\&u\le Q_G-1\}}(q+1)\mu_I+C_{(p,q,u)(i-1,j,0)}1_{\{i\ge 1\}}\lambda_I$$
$$+C_{(p,q,u)(i+1,j-1,0)}1_{\{j\ge 1\}}(i+1)\mu_1(1-h)+C_{(p,q,u)(i+1,j,0)}(i+1)\mu_1 h+C_{(p,q,u)(i,j+1,0)}(j+1)\mu_2$$
$$p+q=N_G,0\le u\le Q_G,i+j\le N_I-1$$

$$C_{(p,q,0)(i,j,v)}(1_{\{Q_G\geq 1\}}\lambda_G + p\mu_G + q\mu_I + 1_{\{v\leq Q_I-1\}}\lambda_I + i\mu_1 + j\mu_2) = C_{(p-1,q,0)(i,j,v)}1_{\{p\geq 1\&(v\leq M-1-N_I \| j=0)\}}\lambda_G$$
$$+C_{(p,q-1,0)(i-1,j+1,v)}1_{\{q\geq 1\&v=M-N_I-1\&i\geq 1\}}\lambda_I + C_{(p,q-1,0)(N_I,0,v+1)}1_{\{q\geq 1\&j=0\&M-N_I-1\leq v\leq Q_I-1\}}N_I\mu_1(1-h) + C_{(p,q,1)(i,j,v)}1_{\{Q_G\geq 1\}}p\mu_G$$
$$+C_{(p-1,q+1,1)(i,j,v)}1_{\{p\geq 1\&Q_G\geq 1\}}(q+1)\mu_I + C_{(p,q,0)(i-1,j,0)}1_{\{i\geq 1\&v=0\}}\lambda_I + C_{(p,q,0)(i,j,v-1)}1_{\{v\geq 1\}}\lambda_I + C_{(p,q,0)(i,j,v+1)}1_{\{v\leq Q_I-1\}}i\mu_1 h$$
$$+C_{(p,q,0)(i+1,j-1,v)}1_{\{j\geq 1\}}(i+1)\mu_1(1-h) + C_{(p,q,0)(i-1,j+1,v+1)}1_{\{v\leq Q_I-1\&i\geq 1\}}\{(j+1)\mu_2 + 1_{\{v\geq M-1-N_I\}}q\mu_I\}$$
$$+C_{(p+1,q-1,0)(i-1,j+1,v+1)}1_{\{M-1-N_I\leq v\leq Q_I-1\&i\geq 1\&q\geq 1\}}(p+1)\mu_G$$
$$p+q=N_G, u=0, i+j=N_I, 0\leq v\leq Q_I$$

$$C_{(p,q,u)(i,j,v)}(1_{\{u\leq Q_G-1\}}\lambda_G + p\mu_G + q\mu_I + 1_{\{v\leq Q_I-1\}}\lambda_I + i\mu_1 + j\mu_2) = C_{(p,q,u-1)(i,j,v)}\lambda_G + C_{(p,q,u+1)(i,j,v)}1_{\{u\leq Q_G-1\}}p\mu_G$$
$$+C_{(p-1,q+1,u+1)(i,j,v)}1_{\{p\geq 1\&u\leq Q_G-1\}}(q+1)\mu_I + C_{(p,q,u)(i-1,j,0)}1_{\{i\geq 1\&v=0\}}\lambda_I + C_{(p,q,u)(i,j,v-1)}1_{\{v\geq 1\}}\lambda_I$$
$$+C_{(p,q,u)(i+1,j-1,v)}1_{\{j\geq 1\}}(i+1)\mu_1(1-h) + C_{(p,q,u)(i,j,v+1)}1_{\{v\leq Q_I-1\}}i\mu_1 h + C_{(p,q,u)(i-1,j+1,v+1)}1_{\{i\geq 1\&v\leq Q_I-1\}}(j+1)\mu_2$$
$$p+q=N_G, 1\leq u\leq Q_G, i+j=N_I, 0\leq v\leq Q_I$$

CURRICULUM VITAE

Sai Zhang

SKILLS

Language: C++, Python, Java, Perl, Javascript, Shell Script, MySQL, Matlab, R.
Environment: Linux, Windows, Subversion
Knowledge: Highly Skilled in Data Structure and Algorithm, Experienced with Object Oriented Design
Projects experience in Data Mining and Machine Learning
Knowledge on Multithreaded Programming and Distributed Systems

EDUCATION

The Pennsylvania State University, *University Park, PA* *8/2010-8/2015*
PhD in Operations Research, GPA ***3.95****/4.0*
Master in Statistics and Computational Science

Peking University, *School of Mathematical Science, Beijing, China* *9/2006-7/2010*
Bachelor of Science in Computational Mathematics**,** GPA ***3.78****/4.0*

EXPERIENCE

Quantitative Developer, *J.P.Morgan, New York* *2/2014-Present*

- Develop the firm-wide core quantitative derivatives pricing library in C++, Python.
- Support and maintain the library development infrastructure including continuous integration.
- Migrate testing framework to distributed system and split tasks by web query.
- Write up usage monitoring code and hook to main library with multithread.

Research Assistant, *The Pennsylvania State University, University Park* *8/2010-5/2015*

- Modeled breast cancer mammogram screening as Markov Decision Process and obtained optimal policies which could potentially increase average female life by 0.035 years.
- Modeled hepatitis C disease progression as Markov Chain and conducted cost-effectiveness analysis on newly invented treatments.

Project, Web Search Engine, *University Park* *9/2012-12/2012*

- Implemented search engine for crawling and building inverted index with Crawler4j and Lucene.
- Analyzed and cleaned data such as duplicate pages elimination and calculated TF-IDF value etc.
- Implemented ranking system incorporating Page Rank, TF-IDF scores and page titles etc.

Project, Oil Production Forecast Based on Well Design Architecture, *University Park* *2/2011-5/2011*

- Extracted features from wellbore design parameters using Principle Component Analysis (PCA).
- Trained production data with Random Forest and Adaboost.
- Conducted Cross-Validation and achieved 3.96% prediction error rate.

LEADERSHIP

Vice President of INFORMS Penn State Student Chapter *3/2012-8/2013*
Volunteer Leader of Beijng Bird's Nest Stadium during 2008 Olympic Games *7/2008-9/2008*