Three Essays on Data-Driven Models in Health Care Operations Management

by

Cheng Zhu

Desautels Faculty of Management,

McGill University, Montreal

August 2017

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

Abstract

Though 20th century has seen life expectancy largely lengthened worldwide, aging population, chronic diseases, worsening food supply with deficit nutrition and environmental problems add to the burden of healthcare systems all around the world. Data analytics, which has been seen as a significant power in other industries, is expected to contribute to the improvement of efficiency and effectiveness in healthcare. This thesis aims to identify and promote more effective and efficient strategic, operations and clinical policies in healthcare systems through descriptive, predictive and prescriptive analytics. To this end, this thesis focuses on three essays, i.e. three data-driven problems based on medium to large size of real life datasets, on: i) design of financial incentive systems for maternity care; ii) design of specialist response policies and modified triage coding to reduce waiting times in emergency departments (EDs), and iii) design of observation units for hearth failure patients.

The first essay focuses on strategic level and aims to design a two-level financial incentive mechanisms to reimburse physicians, in order to reduce unnecessary C-sections while retain it for those who need it, resulting in enhanced birth quality with alleviated economic burden for overall health care system. Contributing to clinical decision-making, we first cluster the patients according to their pregnancy complexities, and characterize a threshold between spontaneous birth and medically necessary planned C-section by analyzing 12.7 million annual birth records from National Bureau of Economics Research through statistical learning methods. Then we compare payment systems analytically vis-á-vis a variety of performance measures within two-level hierarchy, (i) mainstream payment models and (ii) compensation on the top of mainstream payment, and provide insights about the effectiveness of alter-

native payment models in the context of maternity care. Finally, we propose optimal payment for physicians to maximize the value for patients under the principal and agent framework, from the strategic perspective.

The second paper focuses on operational level and targets to reduce the length of stay in EDs by designing a systematic response policy for various specialists depending on ED clinical demands. This work is motivated by and verified with 40,000 ED visits to a local community hospital in Montreal. We first identify a class of patients who are more likely to require specialist consultation based on their clinical information available at the triage stage through statistical analysis. Then we analyze several alternative policies for specialists' response to consultation requests using queuing models with non-homogeneous Poisson arrival rates. Moreover, we examine an integrated ED decision-making by incorporating specialist consultation requests in the triage system. Finally, our proposed optimal specialist response policy and associated modified triage coding are verified through a comprehensive simulation model. We provide a feasible guideline of integrated patient streamlining to shorten length of stay and alleviate overcrowding in ED.

The third paper focuses on clinical level and propose a framework to design a dedicated observation unit for acute decomposition heart failure patients, in order to provide proper treatment and reduce unnecessary hospitalization and chance of post-discharge events. To this end, we, first, use multiple analytical models to figure out the proper number of bed for this observation unit based on historical patient arrival data from a local community hospital. Based on the confined range of analytical capacity, we use simulation models to analyze different discharge and admission policies. We propose an optimal discharge-admission criteria for this dedicated observation unit to realize cost-saving and quality enhancement of treating acute decomposition heart failure patients.

Abrégé

Bien que le 20ème siècle ait vu l'espérance de vie en grande partie allongée dans le monde entier, le vieillissement de la population, les maladies chroniques, l'aggravation de l'approvisionnement en nourriture avec une nutrition déficitaire et des problèmes environnementaux augmentent le fardeau des systèmes de santé partout dans le monde. L'analyse des données, qui a été considérée comme un pouvoir important dans d'autres industries, devrait contribuer à l'amélioration de l'efficacité et de l'efficience des soins de santé. Cette thèse contribue à la gestion des opérations de soins de santé à partir des perspectives de décision stratégique, opérationnelle et clinique. Spécialement, il résout trois différents problèmes liés aux données réelles de moyenne à grande taille , ainsi que la modélisation mathématique.

Le premier vise à améliorer la qualité des soins maternels sans augmenter les dépenses liées à la naissance en concevant des incitations financières optimales pour les soins maternels des médecins. Contribuant à la prise de décision clinique, nous avons d'abord déterminé le seuil optimal de complications de la grossesse entre la césarienne prévue et la naissance vaginale avec des méthodes d'apprentissage statistique et 4 millions de naissances annuelles du National Bureau of Economics Research (NBER). En suite, nous analysons les mécanismes de paiement existants et proposer un paiement optimal pour les médecins pour maximiser la valeur des patients dans le cadre principal et agent, du point de vue stratégique.

Le deuxième document vise à réduire la durée de séjour dans les départements émergents en concevant une politique d'intervention systématique pour divers spécialistes en fonction des demandes cliniques. Nous définissons d'abord une classe de patients qui sont plus susceptibles d'avoir besoin d'un conseil spécialisé sur la base

d'informations sur les triages avec 40 000 visites ED annuelles dans un hôpital communautaire local de Montréal. Ensuite, nous analysons l'heure d'arrivée optimale pour les spécialistes en fonction des taux d'arrivée dépendant du temps des patients avec la modélisation des files d'attente, et nous recommandons enfin les différentes politiques d'arrivée spécialisées pour différents spécialistes en fonction des volumes de la demande. Nous menons également une simulation complète pour comparer la politique de triage fondée sur les ressources et la politique de triage traditionnelle avec des politiques d'arrivée spécialisées spécialisées.

Le troisième document tente de concevoir une unité d'observation dédiée pour les patients atteints d'insuffisance cardiaque décomposée en phase aiguë (ADHF), afin de fournir un traitement adéquat et suffisant, et de réduire les événements post-décharge tout en allégeant la surdité, et économiser les ressources limitées dans les salles d'hospitalisation. Tout d'abord, nous décrivons la quantité appropriée de lit pour cette OU en fonction des demandes historiques de patients atteints d'une hospitalisation communautaire locale, en utilisant plusieurs modèles. Ensuite, nous concevons un critère optimal d'admission et de sortie pour l'unité d'organisation dans le cas où de nouveaux patients arrivent á une entière OU. Les critères peuvent être équilibrés si l'on admet de nouveaux patients et un patient précoce dépendant du patient en fonction des progrès de l'ADHF, qui provient de la littérature clinique.

Acknowledgements

First I would like to thank Dr. Beste Kucukyazici Verter for her close supervision, patient advice, professional guidelines and financial support throughout my doctoral research.

I would like to thank Research Center of St Mary's Hospital for its generous financial support. I am grateful to Fonds de recherche Socété et culture Québec for granting me PhD scholarship for the last four semesters of my doctoral study.

I would like to thank Eric Belzile of Research Center, St Mary's Hospital, Montreal for his dedicated help in data collection and preparation. These data is of great helo to the last two chapters of my thesis. I am grateful to Rick Mah, Chief of Emergency Department in St Mary's Hospital, Montreal for his helpful insights and professional advice that helped me get results of better quality. I would like to thank Dr. Yasmina Maize for her helpful and patient tutorial on simulation models.

I am also grateful to Prof. Morty Yalovsky, Prof. Vedat Verter, Dr. Brian Smiths and members of my committee - Prof. Mehmet Gumus and Prof. Raf Jans - for their patience and support in overcoming numerous obstacles I have been facing through my doctoral study.

Last but not the least, I would like to thank my father and friends for supporting me spiritually throughout writing this thesis and my life in general.

Preface and Statement of Co-Authorship

All the chapters are original scholar works and distinct contributions to the best of our knowledge. First four chapters are co-authored with my supervisor Dr. Beste Kucukyazici. I conduct the research on modeling and data analysis with discussion, advice and close supervision of Dr. Beste Kucukyazici. The last essay is co-authored with Dr. Beste Kucukyazici and Dr. Yasmina Maizi. I conduct the research with close supervision of Dr. Beste Kucukyazici. I collaborate with Dr. Yasmina Maizi while designing the simulation models.

Acronyms and Abbreviations

ADHF	Acute Decompensated Heart Failure
BNP	B-type Natriuretic Peptide
CS	Cesarean Section
ED	Emergency Department
FCFS	First Come First Serve
FFS	Fee-for-Service
FT	Fixed Time Policy
HF	Heart Failure
ICU	Intensive Care Unit
LOS	Length of Stay
NB	Natural Birth
OU	Observation Unit
P4P	Pay-for-Performance
PDE	Post-discharge Event
SB	Spontaneous Birth
SC	Specialist Consultation
R2R	Request to Realization
TL	Timeline Policy
TTFT	Time To the First Treatment in ED

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Thanks to the achievements on life science and technology advancement that lead to a large decrease of infant mortality and death from diseases, the 20th century has seen life expectancy largely lengthened worldwide. Global life expectancy grows to 71.4 years old in 2015, with an average annual growth rate of 5 years old from 2000 (World Health Organization, 2016). In Canada, life expectancy increases to almost 82 years old in 2011 compared with 70 years old in the 1950s (Decady and Greenberg, 2014). This brings new challenges for healthcare systems, which have already been a challenging issue worldwide, especially in high-income industrial countries. Indeed, World Health Organization (WHO) projected that almost 25 % of population would be over 65 years old in 2030 in OECD (Organisation for Economic Co-operation and Development) countries, of which the percentage is around 15 % in 2015; whereas in BRICS countries (Brazil, Russian Federation, India, China, South Africa) only 5 % of population would be over 65 years old by 2030 (World Bank, 2014; Pruss-Ustun et al., 2016). Increasing life expectancy does not necessarily increases quality of life, or health span, which is defined as the duration of healthy life without debilitating disease (EBioMedicine, 2015; Sagner et al., 2017).

Actually, aging population tends to have prevalent chronic conditions requiring more extensive and intensive healthcare service. The leading factors of morbidity and mortality, such as cardiovascular and pulmonary diseases, diabetes and certain types of cancers, have exposed the largest threaten to human life worldwide (Wagner

and Brath, 2012; Arena et al., 2015). Mental illnesses have also been one of the most important factors threatening health, productivity and wellbeing (Birnhaum et al., 2010; Kessler et al., 2006). Worsening food supply with deficit nutrition, due to early picking, pesticide and chemical abuse and deplete soil, has also been contributing to the burden of healthcare (Helweil, 2007; Davis, 2009). Moreover, pollution and other environmental problems expose threaten to health. In fact, currently environment factors account for 23 % global burden of disease (in DALY disability-adjusted life year), without significant decrease from 24 % in 2002 (Pruss-Ustun et al., 2016).

Healthcare resources are limited, as a result of constantly growing demand. Actually, it is challenging for almost all countries to raise sufficient funds to finance healthcare services for all citizens (World Health Organization 2010, OECD2014). Indeed, the proportion of rising healthcare costs, which now constitute over 10% of the GDP in most large OECD economies, continues to outpace growth of both inflation and national GDP (Canadian Institute for Health Information, 2012). However, increasing healthcare expenses do not lead to better quality of care. For instance, in the case of maternal care, as an expensive operation, Caesarean section can expose potential harms on both the mother and the newborn(s) (e.g. Knight et al., 2008; Goer et al., 2012). Efficiency and effectiveness of healthcare have remained the biggest concerns of governments, policy makers and societies all over the world (Peacock and Segal, 2000; Biorn et al., 2009; Health Canada, 2012).

Operations Management (OM) has been contributing to the enhancement of effectiveness and efficiency in healthcare system from mainly three perspectives (Brandeau et al., 2004):

Strategic level. It is a high-level policy making on planning, structure and economics of healthcare system.

Operational level. It focuses on the optimization of process, prioritization and system of healthcare delivery.

Clinical level. It refers to the decision-making regarding selection of technologies and procedures based on medical information and clinical research.

Interested readers can refer to Brandeau et al. (2004) and more recent Zaric (2013) for an introduction and overview of existing OM research on health care system in breath. Here we highlight some of the most recent and noticeable contributions for the illustrative purpose. From strategic perspective, the study of Levi et al. (2016) contributed to improving effectiveness of uniform subsidies regarding maximization of market consumption of malaria drugs. Adida et al. (2017) studied the advantages and disadvantages of different payment schemes in the general context of healthcare systems with mathematical models. Hua et al. (2016) evaluated the co-existence of private and public hospitals regarding government fiscal policies and quality of healthcare services. On the operational level, besides extensions and further works of traditional scheduling and streaming on patient flows (e.g. Kocaga et al., 2015; Defraeye and Van Nieuwenhuyse, 2016), recent works of Chan et al. (2016) and Dai and Shi (2017) studied the more realistic scenarios in the hospital setting of a queue system with time-varying periodic Poisson arrival process. While Chan et al. (2016) figured out the optimal frequency of discharge inspection in an inpatient ward, Dai and Shi (2017) found that to advance the discharge time can alleviate the over-crowding of peak arrivals. OM researchers have also made significant contribution to clinical decision-making. For instance, recent work of Ibrahim et al. (2016) designed a two-stage personalized treatment for anticoagulation therapy with (partially observable) Markov decision process, which offered clinical insights of great value regarding efficiency and effectiveness of the treatment.

1.1 Health Analytics

Data-driven studies, or analytics, an emerging area in OM, has been proven important and becomes more and more attractive in healthcare OM field; as the value of data analysis lays on real-life problem solving and real improvement in healthcare systems (Staheli, 2014). Healthcare analytics is at the core of healthcare transformation and with great potential contributions to clinical decision-making, cost savings and improvement of quality, efficiency and effectiveness, though research in this field

is still in a nascent stage (Raghupathi and Raghupathi, 2014). A recent survey conducted by Health Catalyst among members of the College of Healthcare Information Management Executives (CHIME) showed that healthcare analytics is the highest priority among IT relevant initiatives (Haughom et al., 2014). The promotion and more extensive application of Electronic Health Records (EHRs) will lead to effective information sharing, improved efficiency and efficient integration of healthcare information system (Kwapien, 2016). Besides, from the clinical perspective, data analytics can also be applied to decision-making on personalized treatment, drug design and medicine research (Marr, 2015). Health analytics involves three main categories (Health Analytics, 2015):

Descriptive analysis focuses on what has already happened. Although it seems straightforward to report descriptive statistics from existing data, it can be difficult to derive valuable insights to explain the rationales behind those statistics.

Predictive analysis tends to provide prediction of more likely consequences from symptoms or clinical procedures, based on historical patient information and controlled experiments. It shows a more promising way to help decision-making in healthcare, though it is still an emerging field in health analytics.

Prescriptive analysis works on the solutions to those problems that are likely to happen. This requires integration with other categories of analysis and more advanced tools, and is expected to be the real valuable future of health analytics.

1.2 Content of Thesis

This thesis focuses on three different data-driven problems, in order to demonstrate the contribution of data-driven research to all three crucial perspectives - strategic, operational and clinical decision-making. These studies feature in combining mathematical modelling with all three categories of health analytics - description, prediction and prescription. All the analytical models are verified with medium to large size of real life datasets.

The first essay focues on strategic level and aims to improve the quality of maternal care without increasing birth-relevant expenses by designing optimal financial incentives for physicians. From the perspective of clinical decision-making, this chapter first proposes a clustering approach for the patients according to their pregnancy complexities, and a method to characterize a threshold between spontaneous birth and medically necessary planned C-section with statistical learning methods based on over 12.7 million annual birth records from National Bureau of Economics Research (NBER). From the strategic perspective, this work then analyzes the advantages and drawbacks of existing payment mechanisms and potential bonus schemes through analytical models under the principal-agent framework. Sequentially, we propose an optimal payment for physicians to align their goals with healthcare payers to maximize the value for patients.

The second essay focuses on operational level and targets to reduce the length of stay (LOS) in emergency departments (ED) by designing a systematic response policy for various specialists who are not based in ED all the time. This work is motivated by the prolonged consultation delays in EDs and based on the dataset of forty thousand annual ED visits to a local community hospital in Montreal. To this end, we, first, investigate the optimal timing of a specialist's consultation session analytically in a queue with time-dependent customer arrivals. Then, we analyze and compare two potential response policies and determine the optimal ones for different specialists based on the patient volume and arrival patterns. We also explore the impact of a possible integration of ED decision-making by examining resource-based triage given the optimal specialist response policies through a comprehensive simulation model.

The last essay focuses on clinical level and attempts to design a dedicated observation unit (OU) for acute decomposition heart failure (ADHF) patients, in order to provide proper treatment, as well as reduce unnecessary hospitalization and chance of post-discharge events. Our ultimate goals are to alleviate overcrowding in ED, and provide effective use of scarce resources in inpatient wards without sacrificing quality of care or increasing relevant healthcare expenses. First, we use several featured analytical models to determine the capacity of this OU based on historical data of

patient flows and aggregated process of ADHF. Within the confined range of capacity, we use simulation models to examine alternative admission and discharge policies for this potential OU. Finally, we propose an iterative admission-discharge strategy for this potential OU in order to realize minimal total relevant costs and best possible quality of care.

1.3 Thesis Contributions

In this section we discuss the potential contributions of this thesis.

Our first essay on the design of financial incentives for maternity care makes the following contributions to the literature:

1. In this essay, we propose a reliable cut-off point between two typical procedures in the setting of maternity care (Section 3.3), through a detailed statistical analysis on the patients' complexity based on a large dataset of 12.7 million individual records. Compared to existing literature with a small dataset of hundreds of patients, our census data set is huge and reliable. Moreover, this compliments literature with a reliable and feasible method to predict whether a planned caesarean section is medically necessary or not according to the given clinical information before the onset of labor; whereas existing medical literature focuses on varied indicators of caesarean section during labor.

2. As a modeling approach, we propose a modified gatekeeper model in principle-agent modeling framework. In contrast to the traditional gatekeeping models where gatekeepers refer those clients beyond their capacities, our modified model considers the fact that consulting physicians have a typical dual role of both consulting and delivery, and thus they do not necessarily refer patients to their colleagues. With the framework of games and contract theory, we provide analytical analysis to this innovative model in the setting of maternity care.

3. In our modeling framework, quality of care and physicians' behaviors are explicitly incorporated into objective functions (Section 3.5). This fills in the gap in the OM literature given that in the existing literature the decision makers consider only

expenditures of healthcare services.

4. In our model, we incorporate physicians' efforts and patients' benefits explicitly in the physician's utility function, in contrast to the simplified utility functions in existing literature. Although in our approach the physician's utility function is more complex and intractable, we are able to provide analytical analysis to explain the rationales of physician's behaviours and decision-making.

5. In the setting of maternity care, we analytically model different mainstream payment mechanisms, namely fee-for-service, bundled and blended payments (Section 3.6). Moreover, our analytical results verify the existing empirical studies of those payment mechanisms. Therefore, our analytical model can be applied to those settings where new payment mechanisms need to be tested.

6. We propose several feasible outcome or process-oriented metrics for the pay-for-performance bonus. These performance measures are easy to observe and measure, resulting in a feasible incentive-based payment mechanism that can increase quality of maternity care without increasing the relevant expenses (Section 3.8).

Our second essay on the design of specialist response polices in ED makes the following contributions to the literature:

7. In the queueing models with time-varying arrival rates, we show the closed form of optimal response times for a specialist's one daily visit to ED, based on the pattern and volumes of the consultation demands (Section 4.3.2 and 4.5.3). This fills in the gap of queueing literature by providing the characteristics of average waiting time in a non-homogenous queue.

8. We provide valuable insights to healthcare practitioners and managers by proposing a feasible systematic guideline of determining the best specialist response rules (Section 4.3.4). This guideline can largely reduce the delay of specialist consultation in ED, and it is also easy to implement.

9. We contribute to the medical literature by designing a reliable statistical method of predicting whether the patient will require specialist consultation or not, based on limited clinical information at the triage stage (Section 4.5.2). This method can be of value for other studies, which will benefit from the accurate estimation of

likelihoods of specialist consultation demands.

Our third essay on the design of observation units makes the following contributions to the literature:

10. We design a systematic guideline of designing an ADHF dedicated OU (Chapter 5). This guideline provides valuable insights in determining the capacity as well as admission-discharge strategies for such an OU. Moreover, this guideline is verified to ensure an enhanced quality of care and reduced overall expenses of care, through comprehensive simulation models.

1.4 Thesis Organization

The rest of the thesis is organized as follows. The following chapter is a comprehensive literature review on financial incentives and payment schemes in Healthcare OM. The third chapter presents the first essay on the design of financial incentive systems for maternity care. The forth chapter focuses on the second essay: design of specialist response policies and modified triage coding to reduce waiting times in emergency departments. The fifth chapter presents the third essay on design of observation units for hearth failure patients. Finally, chapter six discusses conclusion and future research of this thesis. Detailed proofs to all the theorems of these chapters are in the Appendix.

Chapter 2

Literature Review on Design of Financial Incentives and Payment Schemes in Healthcare Systems

2.1 Introduction

Health systems aim at providing high quality care by the right providers at the right time and place to all citizens. Using resources efficiently is essential for such systems to remain sustainable. However, every county has encountered problems in financing healthcare services for all (World Health Report 2010). Indeed, the specter of rising healthcare costs, which now constitute over 10% of the GDP in most large OECD economies, continues to loom large over governments wanting to meet wide-ranging healthcare needs (OECD Health Statistics (Database) 2014). The rate of cost increase has outpaced both inflation and national GDP growth (CIHI 2012), making control of healthcare costs a priority of policymakers and academics alike. In response, financial incentives have been introduced by policymakers to steer healthcare providers toward intended and desirable outcomes that also curtail costs and increase efficiency. Popular financial incentive schemes around the globe typically aim to improve healthcare services in four key areas.

(i) ***Access***

Healthcare system should provide timely and proper diagnosis, treatment or other services to anyone who need them when necessary. The barriers of demographic and socioeconomic factors, such as geography, sex, race, and socioeconomic status must be overcome, in order to extend access to healthcare services including hospices, home care, primary care, and mental care (Biorn, Hagen et al. 2009).

(ii) ***Quality***

Quality of healthcare may include but not limit to accurate test results and diagnosis, effective treatment as well as other necessary services. To assess the effectiveness of treatment, the measurement of healthcare quality must be perfected. Reaching this goal is challenging. Furthermore, a strictly positive correlation between healthcare expenses and outcomes has not been achieved. Nevertheless, the ultimate goal of healthcare services has always been to provide safe and effective treatment (Peacock and Segal 2000).

(iii) ***Efficiency***

Healthcare systems aim to provide greater quantity of qualified services within the constraint of limited healthcare resources.The discrepancy between actual and optimal productivity can be calculated in different ways, for example by using technology to assess the ability to decrease inputs while keeping output constant (Biorn, Hagen et al. 2009).

(iv) ***Integration and cooperation***

Healthcare services, namely diagnosis, tests, treatment and recovery care, are impossible to segment with others. Similarly, healthcare providers, including physicians, nurses as well as clinics and hospitals, should cooperate towards achieving common goals in an integrated system. Several integrated healthcare delivery models have been successfully pioneered in recent decades, such as Kaiser Permanente, the Geisinger Health System. However, implementing an associated integral payment mechanism has remained an unsolved problem, which demands further investigation (Sutherland and Crump 2011). Due to the complexity of intended and unintended (but unavoidable) consequences, achieving a perfect remuneration mechanism is not easy.

The last few decades have seen continuing critical reviews and corresponding reforms, and governments currently use a wide range of methods to fund their healthcare services and design their financial incentives. These methods, geared to accomplish several health system objectives, range from global budgeting to payment mechanisms based on the volume and characteristics of patients. Several funding mechanisms are common to different countries, and each mechanism demonstrates both strengths and weaknesses. Certain mechanisms have proven sufficiently successful to be widely adopted. They include the Diagnosis Related Group (DRG) system, used for in-patient payment settings, that effectively shortens hospital bed-days and reduces inpatient costs. The fee-for-service (FFS) approach has always been popular, while activity-based funding, capitated managed care, shared savings, bundled payments, and pay-for-performance (P4P) have been more recently developed to overcome the low efficiencies and potential abuses resulting from FFS. In addition to these methods, many public and private healthcare insurers provide other financial incentives for specific goals. In the USA, the Center for Medicare & Medicaid Services

(CMS) introduced Electronic Health Records Incentive Programs, which pay bonus funds to participating healthcare professionals, hospitals, and critical access hospitals. The programs provide financial motivation to install and improve electronic health records technology (Center for Medicare & Medicaid Services 2014). According to the Quality Incentive Programs report by the American Academy of Physician Assistants (2008), the Leapfrog Group, comprising some large employers aiming to assess their healthcare purchases for employees, developed its Incentive and Reward Compendium to reward contracted providers for improving quality and efficiency. Moreover, the Government of Canada announced an investment in March 2007 of approximately $30 million over three years in the Patient Wait Times Guarantee (PWTG) Pilot Project Fund, aiming to establish guaranteed clinical treatment timeframes and offer incentives for care providers to shorten wait times (Health Canada 2012).

Healthcare service providers, namely physicians, hospitals, and pharmaceutical companies (pharmaceuticals), are pivotal in controlling costs; as nearly all healthcare expenses are directly or indirectly reflected in their profits or gross incomes. In turn, their service to patients has an overwhelming authority to determine healthcare quality. It is therefore reasonable to increase efforts to improve the design and operation of payment systems for these crucial players, worldwide. Recent studies on financial incentives in healthcare confirm that implementing them could lead to the intended behavioral or cost changes. However, due to the limited number of randomized trials in the available empirical research relative to the complexity among healthcare systems, it would be difficult to explore those possible cost changes further and draw generalized conclusions (Chaix-Couturier, Durand-Zaleski et al. 2000). To Design a proper remuneration scheme can be a laborious and expensive process, which is subsequently heavily scrutinized. Thus, though substantial empirical evidence is needed to affirm scheme design and choices, comprehensive analytic research is also necessary to study financial incentive designs, their desired outcomes, and unintended consequences.

Aware of the significance of payment schemes for healthcare providers and necessity of decision tools to assist policy makers and hospital administrators while

designing financial incentives, Operations Research and Management Science (OR & MS) researchers have made noteworthy contributions to the improvement of financial incentives and payment schemes for hospitals and physicians. In this review, we summarize OR & MS studies on financial incentives for particular healthcare systems. The main problems within each geographic setting are illustrated, and also the OR & MS research methods. According to the types of healthcare providers, the rest of this paper is organized as follows. Section 2 describes major payment schemes for physicians with detailed analysis of their strengths and weaknesses; Section 3 focuses on hospital funding systems, including retrospective and prospective financial schemes covering for external sources as well as internal allocation of budget within hospitals (Section 3); and Section 4 covers pharmaceuticals, and mainly focuses on risk-sharing financing for drug manufactures, sales or purchasing to improve drug access. Limitations of existing literature, potential challenges and directions for future research are discussed in the final section.

2.2 Payment Schemes for Physicians

The design of financial incentives for physicians is critical for controlling costs and improving efficiencies in healthcare, because physicians generally have the greatest control in deciding the type, quantity, and quality of treatment services (Leger 2008, Institute of Health Economics 2009), and hence directly influence expenses. Recent empirical studies show that physician payment mechanisms not only influence how physicians determine the volume of health services, but can also provide incentives for efficient and effective preventive care, and chronic disease management. Hence, physician reimbursement schemes are of great interest to health policy makers (Institute of Health Economics 2009).

Payment mechanisms vary geographically. A significant majority of physicians in Canada and the United States bill directly to public or private healthcare insurers, under different payment schemes. In contrast, their colleagues in Europe are mostly salaried employees, contracted to clinics, hospitals, or health institutes. Variations in

population demographics around the world are reflected in local healthcare systems, and the specific nature of their incentives and payment mechanisms.

Broadly, five main payment schemes occur worldwide. They include FFS, capitation, salary/contract, P4P, and blended payment schemes. Studies in OR & MS analyze the strengths and weaknesses of different payment mechanisms, not only identifying optimal reimbursement mechanisms for various geographic regions, settings, and disease types, but also describing the impact of these mechanisms on healthcare service efficiency, quality, and resource allocation.

2.2.1 Fee-for-service (FFS)

Under FFS, physicians are reimbursed at a pre-determined rate for each service they provide. It has been used almost exclusively in Canada and the United States since the 1980s (Cutler 2002, Institute of Health Economics 2009). This scheme is intended to motivate physicians to provide the necessary healthcare services and proper treatments relative to the health status of individual patients. In practice, however, FFS provides financial incentives for physicians to prescribe a greater volume of services, i.e. increase the number of prescriptions and treatments, some of them being unnecessary. In other words, this scheme encourages physicians to over-produce care because it raises their incomes. For instance, The recent work of Adida and his colleagues (Adida et al 2016) adopted a model-based approach and their analytic results confirmed the presence of overtreatment under FFS, whereas it does not result in any patient selection nor expose any financial risks on physicians.The result is a waste of scarce healthcare resources. To avoid these unintended outcomes of FFS, alternative mechanisms have been developed (Leger, 2011, Adida et al 2016).

2.2.2 Capitation

Here, physicians are reimbursed at a fixed rate per patient. This system provides financial incentives to control costs by minimizing unnecessary services, thereby maximizing physician incomes (Tor and Hilde 2000). However, the fixed rate applies

regardless of the character of patients, or of differences in the enrolled population. This may be detrimental to proper treatment for patients with severe conditions. Another negative consequence of capitation is that it stimulates physicians to recruit a bigger patient panel than they can handle, or to select "the healthiest patients and avoid admitting more complicated cases in order to save effort under the universal visit fee" (Ellis 1998). The latter phenomenon is typical of "cream skimming" in the economic domain, and has become the most significant negative side effect of this reimbursement mechanism.

Therefore it is argued in the empirical work of (Hutchison, Hurley et al. 2000) that greater adjustments for patient factors should be included when setting up the rate of capitation. Using Canadian data, these researchers developed alternative capitation formulas to replace FFS for primary care physicians based on the population's relative needs, and demonstrated that the formulas would be both valid and administratively feasible under the current healthcare scheme. This study was motivated by the method implemented in the United Kingdom, which adjusts capitation for general practitioners based on age and sex of patients.

2.2.3 Bundle

This relatively new reimbursement refers to a fix payment for healthcare providers to cover relevant services to treat a specific medical condition per episode. Though this mechanism tends to reduce overtreatment and lower healthcare expenses, it can lead to negative patient selection. The analytic results of Adida et al. (2016) found that this negative patient selection under bundle payment could incur especially when the payment rate is lower or physicians are more risk averse. The higher financial risks born by physicians under bundle payment would potentially lead to the bankruptcy of physicians and consequentially reduce the quantity of healthcare providers, thus could generate detrimental problem for healthcare system in the long term (Adida et al. 2016).

In order to deal with physicians' financial risks exposed under bundle payment, Adida et al. (2016) further proposed a stop-loss mechanism, a modified improvement

of bundle payment aiming to enhance physicians' performance by spreading risks among both payers and providers.

Moreover, due to the newly evolved payment mechanisms, there might exist certain unkown but possible unconscious consequences. Therefore this mechanism should be cautiously implemented (Adida et al. 2016). In order to explore this payment mechanism, Center for Medicare and Medicaid initiated "bundle payments for care improvement" (BPCI) mechanism that selects and funds proposed bundle. Each propose defines the amount of bundle payment, services and treatment, as well as target care quality score. Though proposes with higher expected quality scores and lower costs should be selected, proposers tend to provide minimal discounts to gain more incomes. The work of Gupta and Mehrotra (2015) analyzed and confirmed that an uncertain mechanism of proposal selection is optimal, rather than a fixed selection mechanism, in dealing with the uncertain number of submitted proposes. They employed a normative model, and further incorporate different types of proposers' private information and multiple proposers with competition. Moreover they figured out the current selection mechanism may not be optimal, leading to a lower quality score, and potentially impeding the original motivation of better service coordination.

2.2.4 Contract mechanism

Physician pay is based on a pre-negotiated amount over a certain period regardless of the number of services provided and the complexity of patients. In this contract system, physicians are employed by hospitals or clinics and paid a salary for all services. Unlike FFS, there are no financial incentives in this scheme to provide additional unnecessary services. The similarly fixed amount of income paid under capitation may actually reduce physician productivity and cultivate bureaucracy. This may lead to inadequate access to healthcare services, and potentially reduce healthcare quality (Robinson 2001, CIHI 2012).

2.2.5 Pay-for-Performance (P4P) and Outcome-Adjusted Payment (OAP)

Using various criteria such as health outcomes, access to care and patient satisfactions, a framework is developed to incentivize appropriate levels of high quality care. This approach has frequently been paired with an existing payment mechanism; physicians are also rewarded bonuses (OAP) for achieving certain quality benchmarks, such as meeting quotas or target levels for specific procedures or programs. This encourages physicians to commit their time and effort to particular activities. Essentially, P4P is the same as OAP, but the latter focuses on issues of quality that have plagued healthcare systems (Institute of Medicine 2001).

The payment schemes above (FFS, capitation, and contract) differ most from P4P and OAP regarding uncertainty about physician total income, because P4P and OAP link reward to measures of treatment outcomes, and so can be categorized as "prospective". Difficulties in designing and implementing this scheme have drawn much attention in OR & MS domains.

The most important advantage of a properly designed P4P is that it incentivizes a high quality of care in many health specializations as well as geographic areas (Institute of Medicine 2007, Leger 2011). Fuloria and Zenios (2001) proposed an OAP system using a dynamic principal-agent model, which is originating from economic studies. Principal-agent models focus on situations where a principal ("she" hereafter) delegates her task to an agent ("he" hereafter) she pays, rather than do it herself. The principal wants both a task in good quality and to minimize the fee paid to the agent; the agent, however, wants to maximize his own earnings. Therefore, the interests of the players conflict, causing many researchers to focus on the problem of aligning their goals. Both economic and non-economic strategies have been designed, which motivate an agent to prioritize the principal's goals rather than merely his own. Moreover, it is common for the principal to possess only partial information, such as the final output but not the agent's effort, and thus the principal may fail to gain complete information for reimbursing the agent. This asymmetry of information is

the source of the problem.

In the model of Fuloria and Zenios (2001) a prospective payment per patient is combined with a retrospective payment adjustment that is based on adverse short-term patient outcomes. The model's aim is to determine an optimal payment system that reimburses a physician according to observed patient outcomes while also inducing physician choices that maximize total social welfare. Using the context of end-stage renal disease, this research compares the OAP system with the most common scheme of payment-per-treatment and capitation systems. The OAP outperforms the other two models, and indicates that this system would improve patient life expectancy without incurring higher costs.

However, choosing the best criteria to measure performance is among the biggest obstacles to effective implementation of this payment scheme. The case-by-case criteria used to measure the performance of certain treatments are difficult to identify, and improper proxies may directly reduce the effectiveness of this mechanism. The USA's first P4P system, Medicare's *End-Stage Renal Disease Quality Incentive Program*, was developed in 2010, and pays providers for compliance with measures of specific care processes (intermediate outcomes). Two researchers, Lee and Zenios (2012), found that Medicare's limited set of intermediate measures was insufficient to support payment schemes dependent on them. Their work also incorporated interaction between Medicare (the principal) and diagnosis providers (the agents). Because the sole objective of providers is profit maximization, they control the effort they invest in treatment. By contrast, Medicare aims at better outcomes. However, both the final and intermediate outcomes are uncertain, and depend partly on patient character and provider efforts. Medicare cannot observe patient conditions or physician effort. Insights from this study enabled Medicare to design reimbursement contracts based on a desired set of outcomes, and thereby successfully induced physicians to spend more effort on treatment. Specifically, they investigated the merits of Medicare switching from a per-treatment system to a pay-for-compliance system based on the intermediate measures of dialysis adequacy and anemia control. Moreover, despite these improvements of OAP by Fuloria and Zenios (2001), they recommended a capi-

tation system due to its robustness. They advised against implementation of an OAP system due to its heavy reliance on information that may not be practically available. The work of Shwartz et al (2016) compared different ways to measure healthcare performance and studied their impact of those methods on the P4P scheme. The authors incorporated Data Envelopment Analysis (DEA) in composite measures, and then compare the results with other composite measures, namely opportunity-based weights and a Bayesian latent variable model. They found that DEA led P4P tend to identify the fewest top performers but with higher rewards, among P4P contracts results from these three methods.

Jiang, Pang et al. (2012) endeavoring to align the goals of healthcare purchasers and providers, proposed an optimal "threshold penalty performance-based contract". A national healthcare payer acting as the principal, aims both to shorten the waiting time in the system and to minimize total service funding costs. Providers allocate capacity based on appointment requests from a national online booking system (CaB) that allows patients to make same-day and advance service appointments. Patients are modeled in two categories:"dedicated", who insist on having their service provided by a particular hospital, regardless of whether CaB shows any appointments available in that hospital; and "flexible", who will select any available service provider. Based on an M/D/1 queue model, the authors endeavor to determine the payment contract terms that would incentivize providers to act optimally to achieve a first-best solution in different settings: with complete information, with asymmetric information, or with private agents. They compare capitation, FFS, and payment-by-results (PbR). The PbR contract incorporates service quality measures (maximum wait time for outpatients to see a specialist, in this case) and is able to achieve first-best results. Moreover, in order to attain second-best results for dedicated patients as well, PbR was modified into a threshold-penalty contract, where providers receive a fixed payment (like a contract salary) and are penalized by a fixed amount if the target waiting time is not achieved.

Finally, P4P and OAP may cause providers to concentrate on activities that achieve merit in performance measurement and to skim other services that do not

(Feasby and Gerdes 2006, Leger 2011).

2.2.6 Blended Payment Schemes

Blended payment systems combine multiple mechanisms in practice, using the robustness of one to offset the weakness of another, to provide the intended incentives for physicians. For example, the FFS scheme discussed earlier may incentivize the overconsumption of care, while its alternatives, like capitation, may encourage underconsumption. One potential solution to these distortions is a blend of FFS and capitation, which aims to incentivize physicians to consider proper amounts of treatment in relation to their own incomes. A typical example of blended payment proposed in Adida et al. (2016) is called "hybrid" scheme, which is essentially a combination of FFS and bundle payment, therefore inherent the benefit of both while balance off their drawbacks as well.

The advantages of blended payment schemes were demonstrated by Chu and his colleagues (Chu, Liu et al. 2003). They examined the immediate impacts of Taiwan's *Physician Compensation Program*, where physicians were paid a base salary plus incentives, rather than a salary based on seniority and rank. They concluded that this blended mechanism could not only induce physicians to enhance efficiency and team cooperation, but also increase overall hospital revenue.

However, the obvious concern for blended schemes lies in properly mixing multiple schemes within a specific environment of healthcare services. Using a method similar to that of Hutchinson and Hurley (Hutchison, Hurley et al. 2000), an empirical study of several Norwegian municipalities (Sorensen and Grytten 2000) called for implementing several blended schemes. Using a mixed scheme of FFS and per capita subsidies, the authors recommended a relatively low basic grant with a higher per capita subsidy, FFS payments for municipalities with low physician coverage. For municipalities with high physician coverage, the authors recommended a higher basic grant and lower per capita subsidy plus FFS payments. In other countries too, this formula may adequately distinguish physician coverage levels between rural and urban areas.

2.3 Financial Incentives for Hospitals

Hospitals can be defined as healthcare organizations that provide nursing, diagnosis, and therapy for patients as required by physicians, and certain hotel and social services (Fetter 1991). Hospitals should provide care of high quality that is widely accessible and cost-effective. Achieving this goal has increasingly challenged hospitals because their rate of cost increase has outpaced GDP growth and inflation. Even though their share of total healthcare spending has fallen noticeably over the last several decades, hospitals still account for the largest single percentage of health expenditures in most OECD countries (Sutherland 2011). Although most public hospitals are classified as nonprofit organizations (or not revenue-driven), they are cost centers that remain exposed to financial pressure and must at least break even. That is, income from all sources must cover their expenses, in order to maintain normal business (Verheyen 1998). Under certain reimbursement policies, hospitals usually receive funding to cover operating costs from the public sector, from for-profit or nonprofit organizations, health insurance companies, or charities including direct charitable donations. Ownership may impact hospital funding and further influence performance. Private hospitals may behave differently from their public peers, even in the same geographic and payment-policy setting. Private hospitals from Washington, USA, used so-called "cost shifting" across inpatient and outpatient services, which means that they raised prices for one type of service if the government lowered fees for other types. But Friesner and Rosenman (2004) could not find any evidence that this occurred in government-owned hospitals. An empirical study by Czypionka and his colleagues (Czypionka et al. 2014) investigated the impact of ownership, financing system and financial incentives on the efficiency of acute care sector and inpatient section of Austrian hospitals with an extensive dataset covering 128 public and private hospitals. Using DEA framework, they confirmed that private hospitals tend to be more efficient than their public peers in Australia. They also found the impact of financial incentives on hospital efficiency by comparing their study with a similar study on German hospitals. Because funding resources depend largely on

ownership, financial incentives ultimately become the main reason that hospitals or clinics perform differently.

Healthcare systems with both public and private hospitals co-existing tend to be very complicated. Hua and his colleagues (Hua et al. 2016) investigated such two-tier service system with two types of service providers offering similar service and targeting the same group of clients - public providers who offers free of charge service, and private providers charge clients for a possible higher service quality. In the context of healthcare providers, the public hospitals are generally funded by government but may incur longer waiting time; while patients can pay out-of-pocket and seek treatment from those private hospitals, where overcrowding is less severe. Their work first figured out the conditions under which both providers are able to exist in the same system, and they found a unique Nash equilibrium in the competition process for the common client in such a system. Moreover, they found that neither type of providers were able to achieve the social welfare goal. Public hospitals aim to maximize total customer utility under capacity constraints, and private peers attempt to maximize their profits. They proposed government intervention via tax, budget subsidy to align both types of providers to coordinate and hence increase social welfare.

Both public and private payers know that different funding methods may significantly impact hospital performance. For instance, Rosenman and Li (2002) find that grants and contracts have different effects than donations. More specifically, they observe that grants and contracts received by health clinics in California affect performance differently from donations received, with respect to quality enhancement. Donations may not trigger average expenses, in contrast to grants and contracts. After further empirical investigation, they conclude that grants and contracts were used as seed money to create quality. Therefore, they recommend rewarding those clinics that have already achieved high quality of care, rather than investing in new quality initiatives. This example indicates that wiser financial reimbursement strategies could improve the effect of limited healthcare funds.

In general, hospital reimbursements can be classified as retrospective or prospective payment systems. The main question is how these two main types of payments

impact hospital efforts to achieve crucial targets of healthcare. Subsidiary questions concern the popularity, complexity, and controversy of prospective payment mechanisms. The next section examines these issues.

2.3.1 Retrospective payment system

Hospitals and clinics are reimbursed for each service they provide, i.e. the allowable cost based on an agreed schedule of fees. Thus, almost all operational costs can be reimbursed without any uncertainty. Typically, FFS is the payment scheme for this system.

In most countries, retrospective payment has been gradually phased out in hospitals and replaced by prospective payment, due to its major disadvantage: reducing efficiency. This reimbursement system discourages optimal utilization of healthcare resources, leading to impairment of access to healthcare. This happens because hospitals have no financial motivation to increase the volume of admitted patients (Sutherland, 2011). Using outpatient data from hospitals in North Carolina, Morey and Dittman (1996) demonstrated empirically that lower efficiency was seen in hospitals where a higher percentage of costs had guaranteed reimbursement, compared with their peers that had a larger percentage of costs with unsecured reimbursement. Their work adopted the DEA that can simultaneously consider multiple inputs and outputs.

On the other hand, the advantage of this system is that hospitals or clinics are financially riskless, since the reimbursed amount covers almost all the service and treatment costs, and thus guarantees the essential operations of certain hospitals. This is the main reason FFS still exists in some specific situations. By studying the impacts of environmental factors, including Medicare and Medicaid reimbursement, hospital ownership, and market competition on the efficiency of critical access hospitals in rural areas of the USA, and further proposing a two-stage procedure using semi-parametric approach and bootstrapping, Nedelea and Fannin (2013) examined the case of the Critical Access Hospital (CAH) Program, where cost-based reimbursement, i.e. FFS, was adopted by Medicare to fund rural hospitals with small patient

volume. The CAH program was introduced to address the difficulties faced by these hospitals in covering their costs under a prospective payment system. The authors provided no conclusive evidence that the CAH program negatively impacted technical efficiency in these hospitals.

2.3.2 Prospective payment systems

Funding and capital for a hospital are not completely linked to the amount of services provided or actual direct cost of treatments. The amount of funds is usually negotiated and agreed upon by hospitals and payers before services and treatment take place. Various payment schemes use this system, including fixed price per DRG, activity-based financing, capitation, and fixed global budget.

In contrast to the retrospective payment system (in 3.1), the prospective payment system has shown a powerful stimulus on efficiency, by shifting the financial risk from healthcare payers to hospitals. Ankjær-Jensen, Rosling et al. (2006) concluded from their review of cost accounting used in Danish hospitals that a prospective case-mix payment system is able to stimulate higher productivity. Empirical work by Clement, Grosskopt et al. (1996) showed that hospitals engaged in selective contracting for patients under California's Medicaid program (Medi-Cal) are relatively more efficient than non-contracting hospitals. Such contracting hospitals were financed under a prospective payment mechanism, since they were reimbursed by a fixed unit reimbursement rate per Medicaid patient. The authors found closer agreement between relative shadow prices and relative reimbursement rates for the contacting hospitals, after calculating the shadow prices of contracting and non-contracting hospitals, and then comparing to actual relative reimbursement rates. Puenpatom and Rosenman (2008) studied the effect of a capitation-based payment mechanism in large public hospitals in Thailand and found that transition to a capitation scheme allowed immediate improvements in efficiency. More interestingly, their results showed that the hospitals in wealthy regions become more efficient than those in poorer areas, after both groups made this transition. They combine the method with a bootstrapping procedure to correct DEA efficiency scores.

When hospitals and clinics encounter financial risk, they may have to take immoral action for the sake of survival. The prospective payment mechanism thus also involves access problems including cream skimming and dumping. In hospice settings where healthcare providers serve patients near the end of their lives and offer palliative rather than curative care, Ata, Killaly et al. (2013) pointed out some unintended consequences of Medicare's prospective funding. These included the tendency for hospices to admit patients with relatively shorter remaining lifespans, and not admitting new patients near the end of payment cycles. They further studied Medicare's current funding policy of annual caps on total reimbursement based on the number of patients, as well as a daily cap for each patient treated. To overcome the negative consequences of this existing policy, the authors proposed a legacy policy with a fluid model of patient arrivals, and adjusted the accounting time benchmark of the accumulated cap. Their results were based on hospice research and could valuably be applied to outpatient settings, because a majority of hospices provided routine home care to patients.

The most popular prospective payment mechanism across the world is DRG, and the next subsection examines its advantages and limitations. Studies on other popular hospital-funding approaches, including global budgeting and activity-based funding (McKillop, Pink et al. 2001) are reviewed thereafter.

Diagnosis-related-groups (DRG)

A system to classify and quantify hospital outputs was developed in the USA in the early 1980s (Goldfield 2010). DRG quickly became one of the most popular case mix methods. It assigns individual patients to case mix groups by similarity of clinical features, and a given group has a cost-weight index determined by the mean relative cost. Using DRG, all hospitals get the same funding for treating patients in a specific DRG.

Over the past two decades, more than 20 countries have implemented variants of hospital payment strategies based on DRG and their national settings. Extending the DRG framework, several countries developed comorbidity (multiple illness) ad-

justments that assign patients to subgroups based on secondary diagnoses. Examples of subgroups defined by national clinical practice are the Medicare Severity DRG (MS-DRG) in the U.S., Germany's Diagnosis Related Groups (G-DRG), Case Mix Groups (CMG+) in Canada, Healthcare Resource Groups (HRG) in England, and Australia's Refined DRG (AR-DRG). Sutherland, Hamm et al. (2009) proposed an empirical Bayesian framework to adjust DRG reimbursement amounts for incomplete and inaccurate comorbidity information in the USA.

The popularity of DRG may be largely due to its positive impact on cost control, and enhancement of efficiency in hospital services. In his study on examining the process of developing DRGs, Fetter (1991) points out that, the tricky part of hospital management lies in isolation of providing service and treatments efficiently from effectively taking advantage of those service and treatments.

> *The effective utilization of a hospital's resources is primarily a function of its ability to treat specific kinds of illnesses.*

Indeed, Dismuke and Sena (1999) confirmed the positive impact of DRG on Portuguese hospital service productivity, particularly the efficient use of some diagnostic technologies. They proposed a two-stage procedure using both parametric and nonparametric frontier models. After German hospitals introduced a DRG payment system, Herwartz and Strumann (2012) confirmed the expected rise in competition for low-cost patients, a trend indicated by a significant increase of negative spatial spillovers, or equivalently, hospital efficiency, by incorporating comparative applications of DEA and SFA. DEA users often assume a deterministic production frontier. That is, all deviations from the frontier are regarded as technical inefficiencies. This is unrealistic. Those deviations may be caused by measurement errors or other stochastic impacts. On the other hand, stochastic frontier analysis (SFA) can distinguish between inefficiency and noise components, but at the cost of a more restrictive parametric approach. while studying prospective hospital reimbursement methods based on DRG.

Moreover, DRGs can monitor the quality of hospital services and operations, be-

cause it was originally developed to provide structured definitions for hospital outputs (Fetter 1991). Sharma (2008) concluded that the hospital sector modifies its case-mix in response to changes in relative cost weights, and further confirmed an improvement in the quality of care under the DRG system. The study adopted a stochastic kernel approach to analyze the distribution of declines in length of stay after elective surgeries in an Australian hospital, where DRG-based funding is adjusted for patients with unusual lengths of stay (whether over or under the average).

In addition, DRGs can serve as fundamentals for hospital budgeting (Fetter 1991). Woodbury and his colleagues (1993) propose a quadratic programming model to allocate a national budget to different hospitals, by calculating specific DRG cost weights. The resulting weights prohibit hospitals from operating at either a loss or a profit, and thus minimize the deviation of each predicted budget item from observed expenditures. To estimate DRG marginal costs, they use the model to predict the hospital's budget based on its patient volume, case-mix structure, and the function of the hospital. The DRG methods used to set and update prices for inpatient services in Hungary are discussed by Gaal, Stefka et al (2006), while Epstein and Mason (2006) examine another extension of DRG, in the structure of the UK National Health Service's HRG tariff. They describe how costs are determined, analyze the extent to which prices reflect costs, and review the results of an early evaluation of the system. In Italy, Fattore and Torbica (2006) compared the DRG tariff systems applied to inpatient services at the regional and national level. Bellanger and Tardif (2006) reviewed the changes made to the French reimbursement system for acute care, which was transitioned to DRG for public and private hospitals, as well as the price setting mechanisms and methods assisting this transition. DRGs have played an important role in deriving hospital-funding data from clinic-featured costs. A close link with clinical factors enables hospitals and clinics to hedge unnecessary financial risks. Therefore, DRGs and their derivative mechanisms are attractive to hospital and clinic managers. However, determining proper rates for each group is not an easy task. For instance, Sánchez-Martínez, Abellàn-Perpiñàn et al. (2006) analyzed the DRG related reimbursement system for hospitals in the Spanish National Health System, and found

that price setting does not reflect actual costs of providers that are reimbursed by public funders based on historical tariffs. Thus, this pricing mechanism has no incentive to implement cost control accounting systems. For the German DRG system, Schreyögg, Tiemann et al. (2006) found that data samples used in determining rates did not have qualified repetitiveness, and pointed out major challenges to improving the DRG system, particularly in data accuracy.

To overcome these obstacles, a lot of research has focused on designing and improving DRG mechanisms. In New Zealand, the difficulties surrounding the methodical development and implementation of a national pricing framework for hospitals using Data Envelopment Analysis were chronicled by Rouse and Swales (2006). After Medicare instated a prospective payment system in the USA, Shwartz and Lenard (1994) attempted to ascertain whether an alternative method of price setting would provide better financial incentives than the average cost calculation under this system. They propose two linear programming models that use the number of patients in a patient-type ("groups of patients resulting from the aggregation of DRGs about which management decisions might reasonably be made" (p.782) being treated at each hospital as the decision variable, resulting in a price for each patient. The first model, assuming hospitals operate in an environment of pure competition, gives the competitive equilibrium allocation, i.e. the reallocation of patients in order to minimize costs. The second model is run under the constraint of market boundaries, which assumes that competition from other hospitals is limited to an area of reasonable travel distance for patients, and therefore determines what mix of patients would maximize profit, given set prices. Using data from hospitals in eastern Massachusetts, the equilibrium prices derived from the two models are then empirically compared to an estimation of the average cost pricing. This is used to treat patient types in an "all-payer" system that includes all patients regardless of their third-party payer. To assess the performance of the pricing systems, the authors define a disincentive index, which is aimed to be reduced to zero for pursuing efficient behavior. The results indicate that equilibrium prices occurring in a single market model are also the optimal prices under the constraint of market boundaries, and that this pricing is superior to

the average cost pricing in use.

Global budget funding

This model has been predominant in Canada and public hospitals in the United States (Sutherland 2011). Under this system, a fixed amount of funding is allocated among hospitals based on various criteria, including previous budgets, inflation rate, and major investments in the upcoming years. Allocation is independent of the volume and intensity (the amount of care required) of patients in a hospital. This mechanism functions primarily to control costs, and does not provide any financial incentives to shorten wait times or length of stays, nor to increase quality of care or volume of patients. Peacock and Segal (2000) discuss with the help of economic analysis the feasibility of implementing a weighted capitation (global budget) formula in the Australian health system at the hospital level as a way to enhance efficiency, equity and accountability.

Activity based funding (ABF)

This recently-developed hospital funding model is based on both the type and volume of the services (hospital outputs), and also on the intensity of the patients (Moreno-Serra and Wagstaff, 2009). For the Norwegian hospital sector, Biorn et al. (2003) showed that the introduction of ABF has improved technical efficiency, which is defined as an increase in output requiring a corresponding decrease in another output or an increase in input. Inpatient and outpatient care are defined as the outputs of physician and labor full-time equivalents, plus the hospital inputs of medical expenses and total running expenses. Later, Bi, Hagen et al. (2009) confirmed ABF's positive impacts on hospital efficiency even when taking into account hospital heterogeneity regarding the disutility of effort, with the help of a new DEA frontier from pseudo observation on top of bootstrapping and kernel density estimates. In contrast, Sommersguter-Reichmann (2000) showed that even though significant changes in healthcare performance are observed, such as improvements in technology, a new activity-based scheme in Austria had little immediate impact on technical efficiency.

They used Malmquist indexes, defined as ratios of distance function, to measure technical efficiency over time. Using these indexes, DEA can obtain a simple efficiency score representing the ability of units to maximize outputs while keeping the input fixed, or to minimize inputs given constrained outputs. However, the general applications of DEA include a two-stage approach. The first uses DEA to estimate efficiency, and the second features a regression equation using the estimated efficiency as a dependent variable.

Internal cost allocation

After receiving external funds, hospitals must decide allocation between salaried physicians and internal departments. In this context, Verheyen and Nederstigt (1992) developed an integrated cost-information system for both inpatient and outpatient hospital internal budgeting, aiming to synthesize the Dutch external model of lump-sum capitation with internal DRG based budgeting. Later, Verheyen (1998) examined a system for internal fund allocation in nonprofit Dutch hospitals that eases the potential internal financial tension between physicians and hospitals. Using Verheyen's proposal, hospitals maintain a high level of autonomy regarding budget allocation, and get external funding as a lump-sum payment based on the size of population being serviced, hospital capacity, and production indicators (such as the number of admissions). Internally, the hospital then allocates budgets to departments providing direct care to patients. DRG prices are used to assess the direct care tasks. Based on the DRGs, the direct care departments pay those departments that provide indirect care. This "budget/price" method ensured that both administrators and physicians work towards the same goals in providing patient care.

It is both obvious and logical to distribute funds internally according to the actual costs of different departments. Taking advantage of dual mathematical programming and shadow prices, several studies estimated the marginal costs by computing dual multipliers as shadow prices. For instance, by using the perspective of a hospital planner, Morey and Dittman (1984) constructed mathematical models to analyze the impacts of Medicare reimbursement, imposing total revenue ceilings and allocating

costs between departments under the assumption that all patients, under Medicare or not, would be treated at the facility. The shadow prices derived from nonlinear and linear programming models can usefully distinguish the costs of different departments. The objective function of this model, however, is only to maximize profit, and this may no longer be the sole aim of hospitals.

2.4 Funding Pharmaceuticals

Although non-medical costs funded by the prospective payment system have recently decreased, due to the continuing transition in healthcare services from inpatient care to outpatient care, drug prices over almost the same period have risen, and accordingly have contributed to the overall growth of healthcare expenses (Health Care Financial Review 1996, Kolassa 1997). The rising drug prices also leads to the accessibility problems in certain poor countries, where patients cannot afford basic drugs. Studies in the OR & MS literature attempt to reduce this social problem by proposing economic incentives for drug usage and supply streamlining.

The case of influenza vaccinations illustrates the need to understand how financial incentives affect pharmaceutical fund allocation. To fight influenza, vaccinations are considered the primary weapon, and therefore are widely produced around the world. However, this is constrained by transportation problems, limited raw materials, and the costs of production, research, and storage. Originally, the manufacturers bore all production risks, which they were forced to mitigate by producing smaller amounts. This caused insufficient vaccine supply. In this context, Chick, Mamani et al. (2008) endeavored to align the coordination of vaccine manufacturers and buyers operating in the setting of government health services. This study proposed to optimize the vaccine supply chain using cost-sharing contracts that made buyers share some risks. This would create greater available quantities. Sun, Yang et al. (2009) adopted a game theory framework to analyze country heterogeneity. Simply put, countries are either have or have-not in terms of vaccination stocks. Some have higher vaccination production and stocks, others have little or none. Since influenza

is epidemic, countries with vaccination production may donate some stocks to those with insufficient supply. This indirectly protects their own populations while helping to reduce global losses. The paper showed that the decisions made by individual countries would be different from the optimal allocation by a centralized resource decision-maker. Centralization could reduce infections. Similarly, Mamani, Chick et al. (2013) proposed a contract to allocate limited amounts of influenza vaccinations among countries in order to maximize influenza prevention with optimal cost savings. Based on a model of the transmission of disease between countries, the contract results in better prevention with fewer expenses. A game model is applied to reach the equilibrium where governments minimize their perceived total cost of an outbreak. From the perspective of coordinated decision makers, however, a system model would minimize the overall financial and health costs of all nations. Finally, the study proposed a coordinating contract to resolve the "misaligned incentives" by incorporating the differences between the game and system models. The goal of the research above is finding the global equilibrium that is a proxy for optimal allocation of healthcare resources, increased service accessibility, and maximum social welfare. However, designing a financial contract that is both rational and practical is the key to motivating coordination of different players.

Another study, by Malvankar-Mehta and Xie (2012) considered prevention resource allocation for HIV/AIDS by multi-level decision makers. They investigate the optimal way of allocating budgets to regional governments to maximize the number of infections avoided. There are three levels of players in this specific model, incorporating two fund allocations; first, the upper-level decision-maker (UD) allocates to its lower-level decision-makers (LD); then the LDs distribute further into end users. The UD seeks to maximize its utility function by choosing its level of incentive (i.e. number of infections avoided), and the LDs then maximize their own utility functions based on that decision. The UD has to incorporate equity in order to encourage effective utilization of limited resources.

Besides literature on preventive drugs, there has also been MS & OR studies dedicating to solve the accessibility problem of responsive drugs. In the perspective of fund

donors, Taylor and Xiao (2014) attempted to seek for the optimal way to improve the accessibility of malaria drugs in the regions where patients cannot afford those drugs. Fund donors may face options to fund the drug sales, or to subsidize purchases. With the framework of game theory, they concluded that it is always optimal to subsidize drug purchases only in order to increase the numbers of patients who actually take the drugs, especially the long shelf life drugs. Besides, they argued that funding both drug sales and purchases can be optimal under certain conditions. Another study by Levi and his colleagues (Levi et al. 2016) aimed to maximize the consumption of malaria drugs from the perspective of a central planner, who currently adopts a simple and perceived fair uniform subsidy to drug producers. Using mathematic programming with equilibrium constrains, the authors confirmed the effectiveness of this subsidy policy, that is, subsidize the same unit rate to every producer regardless of different producers' cost efficiency. They found that this uniform subsidy is effective with the presence of producers with varied efficiencies, and even when the planner has no idea about market conditions. Moreover, they figure out that this uniform subsidy can achieve maximal social welfare under certain circumstances. However, this uniform subsidy may not work if producers face a fixed market entry cost.

Due to uncertainty of drug sales, cost-effectiveness and risks in their manufacturing and storage, financial incentives in pharmaceuticals have attempted to spread all sorts of risks among healthcare players, including drug manufacturers, payers and clinics, and thus ensure certain supply and accessibility of drugs. Zhang et al. (2011) studied price-volume contracts between drug manufacturers and third-party payers. These types of contracts come up for negotiation when a payer decides to add a new drug to their list of those eligible for reimbursement. While drug manufacturers must often submit a budget impact analysis, which estimates the total cost for the payer if they approve the drug, the manufacturer has access to more information, and the cost is difficult to verify until the contract has been implemented. Therefore different mechanisms for risk mitigation have been developed. One is the price-volume agreement, where the payer receives a rebate after a certain number of sales. A principal-agent model is applied to determine the optimal conditions, including rebate

rate, price, and profit of these contracts for payers and manufacturers respectively. In certain cases a rebate was sub-optimal, while in other scenarios the payer incurred heavy costs when the rebate was either 0 or 100% of excessive sales beyond threshold. The benefits of this type of contract include risk sharing between manufacturer and payer, as well as providing patients with access to new drugs that may otherwise be prohibitively expensive.

Mahjoub et al. (2014) analyzed a P4P contract between drug manufacturers and healthcare payer, and this specific contract with a pre-determined rebate rate and unit price has a risk-sharing feature. This contract is expected to mitigate drug manufacturers' risks of the effectiveness of the drug. Due to the fact that performance of the drug is mainly measured by the patients' response, manufacturers' profits depend on drugs' realized effectiveness, patients' response rate and the rebate rate. The uncertainty of oncology progression is studied using a Markov model. They found out the relationship of those parameters so that drug manufactures would make profits under this contract. So and Tang (2000) modeled a reimbursement scheme for the prescription drug Epogen under a proposed policy by the USA's federal *Health Care Finance Administration*. Clinics purchase the drug from a pharmaceutical company, prescribe it to patients, and then file for reimbursement from the healthcare insurer. However, the insurer pays only if the health of the drug recipient is below a certain threshold. The clinic therefore takes all the financial risk up-front. The simple dynamic model developed by the authors to determine a patient health ("well-being") score, before and after drug treatment, examines how an outcome-oriented reimbursement affects a clinic's prescription policy, profitability, patient health outcomes, and the pharmaceutical company's revenue. Interested readers could refer to book chapter of Zaric et al. (2013) for review of risk sharing contracts in healthcare literature.

2.5 Conclusions and Future Research

We have discussed pros and cons of each existing financial or payment mechanism by reviewing relevant literature in OR & MS domain. Existing studies are categorized

according to different types of healthcare provides who receive those funds. First we analyzed different payment mechanism for physicians, including both positive and negative consequences of each mechanism. Second we considered two categories of external funding methods for hospitals or clinicals - retrospective and prospective financial incentives - and demonstrated their envolvement, conscious and unconscious impacts with existing studies. Internal allocation of budget within hospitals or clinics and relevant literature have been discussed at the end of Section 3. Third we demonstrated contributions of OR & MS literature to increase availability and access of both preventive and responsive drugs. Popular methods (e.g. game theory, principal and agent framework, DEA etc.) have been highlighted throughout the review, and we also analyzed critically the limitations of literature.

OR & MS literature has profoundly impacted the design of contracts throughout the supply chain in areas including wholesale pricing, cost-sharing or revenue-sharing contracts and variants of newsvendor models. The optimal contracts proposed in the domain of healthcare are similar with respect to risk-sharing and motivating coordination. For example, Chick, Mamani et al. (2008) developed a variant of a cost-sharing contract for government (payers) to share yield risk with manufacturers in order to align the goals of both parties to achieve global cost effectiveness. Zhang, et al. (2011) designed an incentive compatible contract called price-volume agreement to share risks among drug purchasers and manufacturers. Mamani et al. (2013) proposed a coordinating contract for multiple countries to fight pandemic influenza more efficiently. The P4P contract proposed in Mahjoub et al. (2014) features risk-sharing between drug manufacturers and healthcare payers. However, due to unique circumstances in healthcare, the proposed contracts are different from those applied in traditional supply chains. First, in contrast to the linear cost and profit (or revenue) functions in general contracts of supply chain, the costs and benefits in healthcare settings are not limited to the monetary investments and profits of services and drugs. Social costs of producing drugs and infection of pandemic diseases have also been considered as costs in the literature (Chick et al. 2008, Mamani et al. 2013). Measures of benefits, a corresponding concept of profit or revenue in traditional supply chain,

would be even more comprehensive. The improvement of efficiency, effectiveness and quality all fall into measures of benefits. Hence the resulting cost and benefit function in healthcare ends up showing a more complicated and non-linear formula, leading to more complicated contracts for healthcare providers and drug manufacturers. In fact, Chick et al. (2008) detailed the differences of modeling with comparisons of linear and non-linear values of sale. Although non-linear penalty contracts are expected to correct the asymmetry of information in Principal-Agent contract designs, simplified linear penalties have been adopted, as in Jiang et al. (2012). Investigating more complex contract structures would be the next step. Moreover, the measures of quality for the effectiveness of a drug or a treatment tend to be very complicated and multi-dimensional, which are large obstacles for OR & MS studies.

Second, drug producers and healthcare providers are exposed to more rigorous regulations than other industries. As So and Tang (2012) pointed out the conventional risk-sharing scheme with its price rebate property was not perceived to be legal by certain health insurance agents, therefore they proposed an alternative outcome oriented reimbursement policy to replace the newsvendor model and general risk-sharing contract.

Finally, the impact of payment schemes in healthcare can be profound and extensive, ranging beyond the borders of countries and bounds of industries. As Zhang et al. (2011) mentioned, lowering official prices in a country could influence the future profitability of similar drug manufacturers worldwide, due to the international reference pricing adopted in many jurisdictions.

There are numerous opportunities for future OR & MS research on the design of financial incentives in healthcare, particularly on filling the gaps between expected or theoretical outcomes and observable results. First, all the challenges faced due to the special settings of healthcare in the design of contracts are important issues that OR & MS researchers must address as future research. Selecting proper contracts in different scenarios has been a challenging task and worth exploring (Gupta and Mehrotra 2015). Additionally, more work is needed to identify reliable outcomes for payment schemes. Fuloria and Zenios (2001) pointed out that designs for a more

effective fund allocation should be based on observed patient outcomes. Difficulties arise from contracts based on downstream outcome, and more effective incentives may instead result from measuring payment based on intermediate results, particularly when they can easily be obtained.

Another important research area that must be addressed is the design of contracts for integrated healthcare systems. Recently, integrated healthcare delivery systems have been seen as a promising solution for significantly improving quality and efficiency. However the design of incentive contracts and payment schemes remain one of the most critical problems; particularly cost, revenue and risk sharing among healthcare providers and payers as well as internal budget allocation, i.e. allocating resources within a team to incentivize better cooperation.

So far, we have seen most works in static settings, i.e. they are limited to a one-period time horizon (e.g. Lee and Zenios 2012, Gupta and Mehrotra 2015). Although Sun et al. (2009) considered two periods of the initial onset of pandemic influenza, the design of financial contracts would be more difficult but more promising if considering further spreads of a longer time horizon. The outcome-oriented reimbursement policy developed in Fuloria and Zenios (2001) showed a good example of an optimal contract that penalized short-term adverse results while encouraging long-term benefits. A dynamic model would also be of interest when incorporating the learning curve of healthcare providers. For instance, agent learning may be worth incorporating into the Principal-Agent model, because the immediate response to a payment scheme can trigger dynamic incentive decisions, leading to a different optimal decision policy. A dynamic model extension would be particularly interesting in this case, because a data-driven reimbursement system depends on previous provider responses.

For simplicity, most articles treat the risk attitudes of hospitals and physicians homogenously, which is not the case in reality. Progress by Shumsky and Pinker (2003) in studying two types of agents could be extended to heterogeneous physicians. Single-dimension models are widely used to measure patient health (So and Tang 2000), as are two-dimensional models in the case of Jiang, Pang et al. (2012). In reality, however, a patient's health measure would be affected by multiple factors

such as diet, drug usage, and stress, with a good deal of fluctuation. To reflect these factors better, a multiple score with a multidimensional model is needed.

Existing studies consider healthcare providers as profit maximizers (e.g. Fuloria and Zenios 2001, Lee and Zenios 2012), i.e., the ultimate goal of a hospital or a physician is to maximize his/her monetary income. This is not completely realistic, since hospitals do consider multiple objectives, such as quality, access, efficiency, effectiveness and several specific goals in operations, like shortening length of stays or waiting time. Similarly, physicians consider patients' benefits from certain treatment, liability issues, and efforts invested in treatments and their reputations, in combination with monetary incentives. Utility functions of healthcare providers would definitely be more promising but complicated when accommodating multiple objectives.

The patient-mix is considered homogenous in most studies. Although several works incorporated varied types of patients, the number of patient classes was limited to two (e.g. Morey and Dittman (1984)) or three (e.g. Jiang et al. (2012)). Single-dimension models are widely used to measure patient health (So and Tang 2000), or two-dimensional models in the case of Jiang, Pang et al. (2012). In reality however, a patient's health measure would be affected by multiple factors such as diet, drug usage, and stress, with a good deal of fluctuation. Future studies are expected to incorporate more patient groups, because the characteristics of patients are apparently multi-dimensional. Moreover, varied levels of healthcare services or treatments may be necessary for different clusters of patients. Therefore, the design of financial incentives should take that into account for such cases.

Finally, existing research has considered passive patients while modeling the behaviors of health care providers in the design of payment incentive schemes, Fuloria and Zenios (2001); however, in reality patients are actually active. They would like to select physicians and hospitals they prefer or leave a physician if they are not satisfied with the care received. Moreover, they may not perfectly conform to the decision made by their physicians. All those behaviors can create an indirect impact on implementing payment schemes. We believe that future works incorporating active patients would be more promising among OR & MS researchers, and welcomed by

healthcare managers and professionals.

After this comprehensive literature review of financial incentives in healthcare, the following chapter focuses on designing an incentive based reimbursement policy for physicians in the setting of maternity care.

Chapter 3

On Reducing Medically Unnecessary Cesarian Deliveries: The Design of Payment Models for Maternity Care

3.1 Introduction

Cesarean section (CS) is one of the most frequently performed types of major surgery in both developed and developing countries (Spong et al., 2012; World Health Organization, 2015). Although it is a proven surgical procedure, with significantly improved maternal and neonatal outcomes for high-risk pregnancies, there is no evidence that either mothers or newborns benefit from this practice in low-risk cases. Moreover, CS is associated with short- and long-term risks, including a higher likelihood for the mother of requiring further surgery, a hysterectomy, of experiencing infection or deep vein thrombosis haemorrhage, and for the newborn, of having respiratory distress syndrome, pulmonary hypertension, or refusing to breastfeeding (Knight et al., 2008; Goer et al., 2012). In addition to the potential negative clinical effects, CS places a heavy economic burden on the health care system. According to a 2013 report by Truven Health Analytics (2013), the gross hospitalization costs for CS were almost 50% higher than for natural births (NBs), for both public and private payers. This disparity would be even more significant if the costs of hospital readmissions and post-discharge follow-up care were taken into account.

Nevertheless, CS rates have been increasing constantly for both high- and low-risk pregnancies around the world. Approximately one-third of births in the US are delivered by CS, accounting for more than 1.3 million surgeries each year (Center of Disease Control, 2014). Moreover, despite the fact that the low-risk cases do not benefit from CS, the rates for this group have risen progressively, reaching a high of 28% in 2013. In Canada, the overall rates have also grown steadily, from 5.7% in 1970 to 28% in 2014, while the rate for low-risk births is now almost 15% (Canadian Institute for Health Information, 2016). Due to the dramatic increase in CS rates, the "Healthy People 2020" initiative launched by the Centers for Disease Control and Prevention set the explicit goal of reducing the cesarean birth rate and identified 23.5% as the United States' target for cesarean deliveries (U.S. Department of Health and Human Services, 2015). The existing literature strongly suggests that the physicians have other motives besides the patient's clinical characteristics while

medical decision-making during childbirth and economic incentives is seen as one of the most important factors (Taljaard et al., 2009; Johnson et al., 2016). Likewise, in the comprehensive report on evidence-based maternity care by Sakala and Corry (2008), the misaligned or perverse incentives of payment system have been described as one of the pervasive barriers to reducing the cesarean rate. Given the role of economic incentives in the decision process of physicians, this paper focuses on the design of financial incentives in order to reduce unnecessary C-sections, resulting in enhanced birth quality with alleviated economic burden for overall health care system.

Maternity care typically comprises three stages: prenatal, delivery, and postpartum, all of which are under the financial responsibility of the health care payer. The payment model most commonly used by public and private payers is fee-for-service (FFS), where physicians receive a fixed rate for each service they provide. A significant portion of the obstetrics fee under the FFS model is associated with delivery; hence, physicians providing prenatal care are incentivized to deliver their own patients. This is unlikely to occur with natural deliveries, since care in hospitals is often provided by a team of physicians working on a rotation basis. Furthermore, the fees for cesarean delivery are almost 50% higher than those for natural delivery (Thomson Healthcare, 2007; BC Health Authority, 2016), which further encourages physicians to perform planned CS. In addition to offering physicians a higher payment, CS also has a lower opportunity cost. Since NBs often involve a long labor (i.e., requiring an average of twenty hours of medical attention, as compared to two hours at most for a CS) and a great deal of uncertainty, they may impede the physicians' ability to perform their other duties (Sakala and Corry, 2008). This can lead to physicians not investing their full effort and attention in monitoring labor until a NB occurs, which in turn may result in unnecessary CSs being performed. Indeed, recent studies have shown that limited resources, a high workload and inadequate financial incentives increase the pressure to move patients through the system faster, which may lead to increased CS rates (?Ariadne Lab, 2017). For example, Spong et al. (2012) find that "If labor occurs during night or weekends, physicians are more likely to decide on emergency CS rather than waiting for completion of the labor due to an appetite for

convenience".

Drawing on the realization that better-aligned financial incentives could help drive down the increase in cesarean deliveries, an array of local, state, and federal initiatives are underway to improve maternity outcomes through payment reform (CPR, 2012). The core of these reforms is offering alternative payment models to FFS, including blended and bundled systems.

Under the blended systems, rather than having different delivery fees for CS and NB, physicians receive a single rate per delivery, regardless of delivery mode (Main et al., 2011). In theory, a blended payment system removes the financial incentive for CS by providing one rate for all types of delivery. However, in practice, an equal payment for all deliveries might not fully compensate for the increased opportunity cost of natural deliveries.

The Health Care Incentives Improvement Institute has proposed a different model based a bundled payment for maternity care. This restructured payment method bundles the payment for the full extent of care for women and newborns (Child Birth Connection, 2011). Under this system, a set payment is received for each registered pregnancy, including all prenatal consultations, lab tests, and ultrasounds, the actual delivery, as well as the post-delivery hospital stay for both the mother and newborn, regardless of delivery mode, and regardless of the resources expended (CPR, 2012). A bundled payment structure shifts the financial responsibility for care management to the providers and creates financial incentives to reduce resource costs. In addition to its potential benefits, the bundled payment structure also introduces several new challenges. For instance, providers may struggle to predict the complexity of a pregnancy, and that complexity may change throughout an episode of care. Therefore, the actual cost at the end of an episode of care could be much higher than the price of bundled care, which places a high financial risk on the shoulders of physicians. Another issue pertains to the choice of necessary care. Given that natural delivery is less costly than CS (leaving more of the bundled payment available as profit), under the bundled payment structure, physicians are incentivized to delay making a CS decision. (Warrington and Brunkow, 2011)

These payment mechanisms can be paired with other incentives, i.e.complementary payments. Specifically, pay for performance (P4P) bonuses further motivate certain physician behaviors in order to offset some the disadvantages of the alternative schemes mentioned above. Therefore, the P4P design is also seen important element of payment reform initiatives for maternity care. P4P refers to "compensating physicians according to an evaluation of their performance on defined metrics, typically a potential bonus on top of other payment schemes" (American Medical Association, 2015). These performance metrics can be based on process quality and efficiency, outcome, or cost. Although P4P programs have been in force in developed countries since late 1990s, they were rarely applied to maternity care until recently. One of the current initiatives involves a bonus offered to a group of physicians in the event that their overall CS rate is below a certain level (Das et al., 2016). A comprehensive review of a number of P4P implementations in practice concludes that P4P effectiveness depends greatly on the program's design (Eijkenaar, 2013). In this context, policy designers face two critical questions: (a) what to incentivize: which performance metrics should be chosen? (b) whom to incentivize: individuals or groups?

Although payment mechanisms and incentives have been studied in the literature mainly in the context of primary care, chronic care and several surgical procedures, the relevant literature in the area of maternity care is still in an early stage of development. In practice, there are only a small number of payment reform programs across the US and Canada. Policy makers all agree that developing the best model for maternity care is a complex process that requires detailed analyses of each model's outcomes before it can be implemented widely. In this paper, we fill this gap by proposing an analytical framework to study the impact of alternative payment systems in maternity care and compare their performance under different criteria. This enables us to determine the payment scheme that induces physicians to deliver the most appropriate maternity care.

We compare payment systems in relation to a variety of performance metrics, within a two-level hierarchy: (i) payment models that define a base fee payment and (ii) bonuses complementary to base payment. Using this hierarchy, we answer the

following research questions:

1. Does the payment scheme under consideration provide incentives for improved quality of care and lead to cost reductions, as compared to FFS?

2. Does the addition of the P4P model (i.e., metric and payee) under consideration have the desired impact on cost and quality of care?

Our modeling framework focuses on a group of physicians and a single healthcare payer. We base the payer-physician system on the principal-agent model. The payer, as the principle ("she" hereafter) delegates the task of maternity care to the physician. She aims for both a good quality of care and the minimization of overall maternity care expenses, including physician reimbursements. The physician ("he" hereafter) however, maximizes his own utility, including his own earnings. Moreover, the payer possesses only partial information (the delivery mode), not complete information about the pregnancy's complexity, the labor complications, or the physician's effort. In accordance with the medical literature, we represent the physician's utility via three components: benefits accrued by the patient, expected effort, and income as a function of the delivery mode (Eggleston, 2005). Through our analytical framework, we analyze the impact of hidden efforts and estimate their consequences on unnecessary planned and emergency CS rates. The payer's problem is to design a payment system that links the physician's observable actions with her objective of maximizing the value of care for patients, that is, to achieve the best health outcomes at the lowest cost (Porter and Lee, 2013).

Our data-driven research is based on approximately 16.2 million individual birth records from 2011 to 2015. The source of our data is the "Natality" database published by the *National Bureau of Economic Research (NBER)* `http://www.nber.org/data/vital-statistics-natality-data.html`. This database from the National Vital Statistics System of the National Center for Health Statistics provides demographic and health data for births occurring during the calendar year, i.e., approximately 4 million births each year. It is based on information abstracted from birth certificates filed in the vital statistics offices of every state across the US. In

analyzing the alternative payment mechanisms one of the key challenges is defining the pregnancy complexities via a quantitative metric and identifying the group of patients for whom CS is medically appropriate based on this metric. Although there are several clinical guidelines on defining the set of patients that CS is medically indicated, there is no such quantitative framework in the literature. This very large data set enables us first, to measure patient's complexity, then to rank patients according to their pregnancy complexities, and finally to characterize a threshold between a spontaneous birth (i.e., NB and emergency CS) and a medically appropriate, planned CS. Furthermore, we estimate the probabilities for a set of post-delivery complications under planned and emergency CSs as well as natural deliveries for patients with a given complexity level; therefore, we are able to accurately estimate the cost of delivery and postpartum care for alternative delivery modes for a given patient complexity.

The proposed analytical framework enables us to show that none of the base payment mechanisms is not sufficient to perform at the desirable levels in both quality and cost simultaneously. We also propose a set of complementary incentives based on reliable and tractable metrics of quality of maternity care. Acting as an add-on to the payment mechanism, these incentives are capable of discouraging hidden efforts in both prenatal and delivery stages of maternity care. Furthermore, we propose an easily implementable and robust two-level payment model, i.e. blended payment and a process-oriented bonus, that results in risk sharing between payer and physicians, and coordination among the group of physicians. We empirically verify our analytic results in the numerical study based on our data set, and demonstrate that the ensuing quality and expense of our proposed two-level payment mechanism outperform traditional mechanisms. Specifically, our analysis shows that this recommended policy proposes 3% reduction in average birth related costs and 27% decrease in overall CS rate compared to those under FFS system.

The rest of the essay is organized following: the related literature is discussed in Section 3.2. We outline our data-driven approach for modeling pregnancy complexity in Section 3. The models regarding physician's decision making in childbirth

Table 3.1: Descriptive Statistics

Calendar year	2011	2012	2013	2014	2015
Total live births excluding in territories	3,961,220	3,960,796	3,940,764	3,998,175	3,988,733
% Births in hospital	98.70	98.61	98.54	98.47	98.42
% of overall CS	34.75	34.92	34.85	34.23	32.87
$\geq$ 37 gestational weeks	3,029,917	3,098,879	3,162,890	3,421,630	3,479,307
Full records of labor and delivery	3,008,781	3,079,833	3,146,062	3,406,120	3,479,307

and healthcare payer's problem are presented in Sections 4 and 5, respectively. Section 6 and 7 discuss our analytical results on base payment schemes and proposed complementary incentives, respectively. Our proposed two-level payment model is demonstrated in Section 8. Numerical results are presented in Section 9, followed by conclusions and limitations in Section 10. Supplementary statistical results, parameter estimation for numerical analyses and sensitivity analyses are provided in Appendix.

3.2 Literature Review

This work falls in two main research streams: financial incentives in the area of health care, and game and contract theory in operations management. Given that the payment reforms in maternity care by introducing alternative base and complimentary payment models are currently tested, the literature addressing the financial incentives directly in this domain is really limited. However, the subject of financial incentives applied to various settings in the health care system has been studied extensively in health economics, health policy and operations management literatures. Interested readers can refer to the comprehensive review paper by Kucukyazici and Zhu (2017) for the relevant works in operations management. We mention the most relevant research in operations management hereafter.

Our work relates to research on base payment policies for health care providers

as well as the performance-based reimbursement mechanisms. In the context of base models, the analytical models developed by Adida et al. (2017) compare a traditional payment scheme FFS and the more recently designed bundled payment regarding varied performance measures, including patient selection, treatment levels selected by the physicians, financial risk born by the healthcare providers, and overall payoff for the healthcare system. Their analytical results also reveal the impact of providers' risk aversion. The authors provide two possible ways of improvement - a stop-loss mechanism to offset the drawbacks of the bundled payment, and a hybrid scheme combining both payment systems in order to coordinate health care payers and providers to a system optimum. **?** also study the models of FFS and bundled payment by using a three-stage Stackelberg game. Through this modeling framework, they investigate the possible impacts of these reimbursement schemes on patients' welfare, readmission rate and waiting time in a public healthcare system. Compared to these papers focused on general healthcare settings, this work aims to provide analytical analyses as well as managerial insights on three base payment mechanisms - FFS, blended and bundled payments in the specific circumstances of the maternity care.

Besides the studies on base payment schemes, there have been evolving initiatives of performance-based reimbursement mechanisms aimed at improving the quality of health care: performance- or outcome-based incentives. P4P and outcome-adjusted payment (OAP) have been popular in practical and theoretical works, after the unintended but disappointing impacts of commonly used base reimbursement mechanisms, including FFS, capitation, and fixed salary policies. These models are outcome or process oriented and only reward the successful achievement of certain quality benchmarks; hence, their intent is to encourage physicians for their commitment to care quality. As one of the early works, So and Tang (2000) model an outcome-oriented reimbursement program for the drug industry. Fuloria and Zenios (2001) introduce an OAP as an add-on of existing retrospective payment adjustments, based on the adverse short-term outcomes of patients with end-stage renal disease. Moreover, in the setting of Medicare's End-Stage Renal Disease Quality Incentive Program, Lee and Zenios (2012) design a pay-for-compliance system based on the intermediate mea-

sures of dialysis adequacy and anemia control. Jiang et al. (2012) propose an optimal "threshold penalty performance-based contract", derived from a payment-by-result contract, to motivate health care providers to shorten waiting time. More recently, **?** analyze the performance-based reimbursement in the context of cancer treatment. Jiang et al. (2017) study the joint impact of performance-based incentives and competition on the healthcare service providers. In our study, we particularly propose four types of process- or outcome-oriented incentives to act as add-ons to base payment schemes. These are selected based on opinions of those physicians we approached as well as a literature review, including the clinical guidelines, on targeted performance metrics for maternity care. Moreover, we examine the effects of different possible recipients for these complementary incentives in the context of maternity care.

Our work is also closely related to principal-agent framework in contract theory. Typically in healthcare systems, the payer or the principal, i.e. public or private insurer, fails to fully observe patients' true health conditions, which are detected by the healthcare providers (commonly the physicians), or the agents. As the principal cannot provide professional health care herself, she delegates these services to physicians, which might lead to a misalignment of priorities between the principal and the agents. Therefore, the payer has to rely on reimbursement contracts to align her goals with the physicians' aims, as presented by So and Tang (2000), Fuloria and Zenios (2001), Lee and Zenios (2012), and Jiang et al. (2012). The most recent application of this framework in health care domain is presented by Zorc and his colleagues (2017) in chronic care setting. Their work focuses on comparing different contracts with individual or group physicians, and proposing a payment policy that minimizes the chance of adverse effects. In contrast to those papers where the moral hazard (i.e. hidden effort) problem occurs in a single epoch, our analytical framework considers moral hazard problems in two stages of maternity care. First in the prenatal care stage, the physicians tend to prescribe planned CS for the sake of earning a guaranteed income at regular business hours. Second in the delivery stage, there is a hidden effort problem, since the physicians may not spend their full efforts for monitoring the ongoing labor, especially during non-business hours, resulting in the unnecessary

emergency CS cases, which can be avoided with full monitoring efforts.

Our framework also falls under gatekeeping problems, wherein gatekeepers have the option to either keep serving clients or refer them to specialist colleagues. Gatekeepers are common in call centers, while in a health care system, family physicians or primary care providers are considered as the "gatekeepers". Gatekeepers can serve those relatively easy or low-risk patients themselves, whereas need to refer those more complicated or high-risk cases to specialists. The efficiency of generalist physicians' compulsory referral mechanism has been an area of concern in health economics, especially in a system where physicians can have a dual role, in other words, physicians can be both gatekeepers and specialists, and therefore they can refer their own patients to other services provided by themselves (e.g., Gonzalez, 2004; Biglaiser and Ma, 2007). Such case is also called physicians' self-referral, and is considered as one of six conflicts of interest in medicine, according to Rodwin (1993). In operations management, **?** conducts an empirical study to show the impact of workload on the gatekeepers' decision making in a maternity unit. As a more relevant study to ours, Shumsky and Pinker (2003) study a gatekeeper model without self-referral; the gatekeeper's chance of successfully solving a problem decreases as the problem becomes more complicated. Thought gatekeepers try to solve as many problems as possible, this may detriment clients' satisfaction in a service system. This work evaluates the performance of contracts, and provides managerial insights in contract design for heterogeneous gatekeepers. In contrast to this paper, our model of gatekeepers comprises the physicians' self-referrals or dual roles as well, that is, the physicians might serve their own patients in the delivery stage.

3.3 A Data-Driven Approach for Representing the Level of Pregnancy Complexity

Prenatal care is provided by a consulting physician who, through a number of visits, monitors the pregnancy and assesses its level of complexity and the associated risks.

By the end of prenatal care, the consulting physician decides on the delivery mode: spontaneous birth (SB) or planned CS. For a planned CS, the consulting physician schedules the operation and performs it, thereby becoming the delivery physician as well. However, in the case of an SB, the patient goes to the hospital's birthing center once labor begins. Consulting physicians and their colleagues serve at the birthing center on a rotational basis. Therefore, the on-call physician, who can be anyone within the group, is responsible for the delivery. During labor, the on-call physician can also order a CS (emergency CS) for various reasons, including problems with the umbilical cord, sudden changes in the baby's heart rate, and prolonged labor.

The patient's pregnancy complexity level, which we denote by x, plays a key role in the physician's ultimate choice between an SB and a planned CS. In the context of this research, we use the likelihood of a CS as a proxy for pregnancy complexity. In this section, we describe how we estimate x and its distribution using our dataset, and define a pregnancy complexity threshold (i.e. x^*) for a medically necessary, planned CS, such that, if the complexity level is higher than this threshold, having a planned CS is clinically appropriate, through our large data set. Our data set includes the birth records of all live births with full-term deliveries, meaning births with 37 or more gestational weeks, and served by physicians in hospitals of U.S. during each calendar year from 2011 to 2015. It consists of detailed individual-level records of approximately *16.2 million* births, which contain i) the mother's demographic information, ii) pregnancy history, iii) clinical risk factors of the current pregnancy, iv) delivery information, including birth date and time and method of delivery, and v) postpartum information including post-delivery complications.

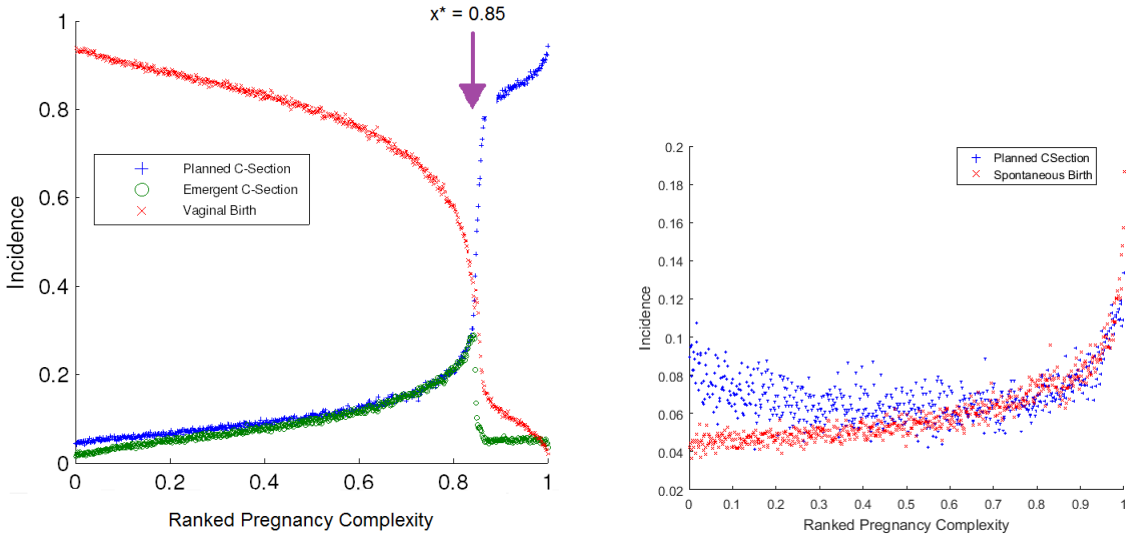
Using this data set, we first estimate the probability of having a planned CS through a logit model, given medical risks and patient characteristics observed in prenatal care. In our statistical model, we use births occurring over a whole calendar year for in-sample modeling, and the births in the following year as out-of-sample. That is, we run our logistic regression model for births in the years of 2011, 2012, 2013, and 2014; and demonstrate the forecasting power for deliveries in 2012, 2013, 2014, and 2015, respectively. Our results show that the model predicts well, and that

the area under the ROC curve for out-of-samples are quite similar across four different years, ranging between 83.29% and 83.76%. Therefore, our explanatory variables have high discriminatory power. The results of logit models are presented in Table EC.1 in Electronic Company. There are other non-clinical factors that may be associated with the decision to have a planned CS, such as differences in physicians' diagnostic skills and practice styles. However, as was argued by Currie and MacLeod (2013), the pooled birth records, with decisions by thousands of physicians, offset the impact of these non-clinical factors.

Next, we rank women according their probability of CS estimated in the logit models, and then normalize them into the standard values of zero and one $[0, 1]$. By this approach, pregnancy complexity x represents the percentile of CS likelihood. For example, for a pregnant woman with a complexity level of 0.4, there are 40% of women in this population with a lower risk than her and 60% with a higher chance of having a CS (see Figure 3-1a). In Figure 3-1a, we observe that, as expected, the probability of having an NB decreases as x increases, and that this decrease is quite sharp at the complexity level of 0.85, suggesting that the ordered ranking of pregnant complexity shows two different regions as regards the medical appropriateness of planned C-sections. Although the rate of SB decisions during prenatal care is almost 90% for patients with a complexity level lower than 0.85 (i.e., low-risk group), it is only 17% for those with $x > 0.85$ (i.e., high-risk group). Likewise, the chance of having an emergency CS under the decision of SB is three times higher for the high-risk group. We also run sensitivity analyses around the complexity level of 0.85, (i.e., testing 0.75, 0.8, and 0.9 as cut-off points) and conclude that, in terms of differences between two groups, the most distinct categorization is provided by x of 0.85.

Moreover, birth outcomes present different patterns for these two groups. We use post-delivery (i.e., postpartum) complications as a proxy for birth quality. The birth records contain over 20 variables relating to the existence of major severe post-delivery complications, including maternal complications (such as excessive bleeding and hysterectomy) and abnormal conditions in the baby (such as brachoplexis, fractures, meconium, birth injuries). We define an undesirable health outcome as involving at

least one maternal or neonatal complication, and the incidence of post-delivery complications as the percentage of deliveries with undesirable outcomes. Our empirical analysis shows that the incidence of post-delivery complications following a planned CS is independent from x, the complexity level of the patient diagnosed in prenatal care; however it has a significant positive correlation with x under SB, with a pseudo R-squared statistics of 90% or higher for in-sample study of all four years'. Further analysis on complication incidences shows that, for $x < 0.85$ the average risk of a postpartum complication under planned CS is higher than that of SB. After this point, however, the risk of a complication following a SB is higher. Therefore, it is essential to perform a planned CS for cases with $x > 0.85$ and decide on a SB for the ones with $x <= 0.85$ in order to minimize the incidence of post-delivery complications (See Figure 3-1b). Based on the observed differences regarding the rate of SB decisions, emergency CS rates, and complication incidences, we propose that x^* represents the cut-off point, above which a planned CS is medically more appropriate. Alternatively, an unnecessary C-section refers to the prescription of a planned CS for a woman with a complexity level of less than x^*.



(a) (b)

Figure 3-1:

Table 3.2: Logistic Regression Results

Variables	Coefficients	
(Intercept)	-5.4857	***
age	0.04263	***
prior other termination	0.29614	***
live birth order	-0.6599	***
previous CS	2.21392	***
eclampsia	0.52242	***
month of prenatal care began	0.03094	***
infertility treatment	0.25569	***
total birth order	-0.2557	***
weight gain	0.01789	***
cigarette record	0.21428	***
diabetes	0.73514	***
gestational diabetes	0.32568	***
previous preterm birth	-0.0271	
chlamydia	0.16002	***
hepatitis B	0.31533	***
hepatitis C	0.08287	
BMI	0.06996	***
plurality	1.08439	***

Notes
1. *** denotes for the significance level less than 0.001;
2. This table summarizes the logistic regressions results of in-sample for the year of 2013.

3.4 The Physician's Decision: The Mode of Delivery

In this section, we start developing the modeling framework by focusing on the physician's decisions. We will represent them as a set of constrains in the payer's decision model in Section 5. We start by discussing the modeling framework, and then present our formulation concerning the physician's best response strategy. The notation is summarized in Table 3.3.

We consider a population with a finite number of pregnant women, a single group of J physicians and a single health care payer. We assume that each physician has the same diagnostic skills, the same preferences over delivery procedures, evenly shares on-call time in hospital, and an equal number of pregnant women (at similar levels of complexity) registered to his panel. Note that we relax this homogeneity assumption in Model Extension (Section 3.9) without loss of generality. Empirical studies show that the physicians tend to be influential on their patients with regards to the decisions concerning delivery mode; consequently, it is assumed that the patients are in compliance with the consulting physician's decision (Fabbri and Monfardini, 2008; Grytten et al., 2013). In our framework, the physician's utility consists of three components: a patient's benefits, effort spent and income gained by the physician as a function of the realized delivery mode. Patients' benefits and physicians' effort are exogenous factors in reimbursement policies.

We consider a population with a finite number of pregnant women, a single group of J physicians, and a single health care payer. We assume that each physician has the similar diagnostic and procedural skills, similar preferences regarding delivery procedures, the same share of on-call time in the hospital, and an equal number of pregnant women at similar levels of complexity registered to his panel. Note that we relax this homogeneity assumption in Section 8.2 without loss of generality. Empirical studies show that physicians tend to hold influence with their patients as regards decisions about delivery mode; consequently, it is assumed that the patients comply with the consulting physician's decision (Fabbri and Monfardini, 2008; Grytten et al., 2013). We also assume that the physicians are rational decision-makers, i.e., they

Table 3.3: Summary of Notations

	Decision Variables
s	Threshold of pregnancy complexity between spontaneous birth and planned CS;
λ	Physicians' effort level of serving a delivery on their shift;
P^N	FFS rate per each NB;
P^{EC}	FFS rate per each emergency CS;
P^{PC}	FFS rate per each planned CS;
P^{BP}	Payment rate under blended payment;
P^{BL}	Payment rate under bundled payment;
m_D	Physician's income under decision D;
B^{PO}	Postpartum outcome-oriented bonus;
B^{CO}	Complexity add-on;
B^{NB}	NB rate bonus ;
B^{TH}	CS threshold bonus;
	Other Parameters
r	overall CS rate;
J	the number of physicians in a group, i.e. the group size;
x	$\in [0,1]$ ranked complexity of pregnancy;
$f(\lambda, x)$	expected successful rate of a NB for complexity of x and physicians' delivery effort level λ;
x^*	clinically optimal threshold of complexity for planned CS;
α	a physician's benevolence level;
$\Pi^E(\cdot)$	payers' total birth related economic costs;
$\Pi^Q(\cdot)$	payers' quality objective function;
	Prenatal Stage
D	physicians' decision in the consulting process, $D \in \{SB, CS\}$;
SB	physicians' decision of spontaneous birth;
CS	physicians' decision of planned CS;
$b_D(x)$	benefit of decision D for a patient with complexity of x;
$u_D^I(\lambda, x)$	physician's expected utility under decision D by the end of prenatal care;
$U_D(\lambda, x)$	overall utility of physicians if decision D is made for a patient with complexity of x and their agreed effort λ;
$I_D(\lambda, x)$	incidence of postpartum maternal and neonatal complications for pregnancy complexity level x and effort level λ;
	Delivery Stage
e^C	effort of serving a C-section;
e^N	effort of serving a NB;
e^{MN}	effort or inconvenience factor for monitor labor in spontaneous birth;
	Payers
c_H^C	facility fee for a CS;
c_H^N	facility fee for a NB;
C	average unit cost of treating postpartum complications;
$CH(\lambda, s)$	total facility fee depending on threshold s and effort level λ;
$CI(\lambda, s)$	total postpartum treatment costs dependent on threshold s and effort level λ;
$g(\cdot)$	Intensity of actual population in terms of pregnant complexity.

choose the best action possible given the clinical information concerning the patient as well as their own utilities.

3.4.1 Physician's Utility

In our framework, the physician's utility is made up of three components: (i) the patient's benefits, (ii) the effort expended by the physician, and (iii) the physician's income as a function of the realized delivery mode. Note that the patients' benefits and the physicians' effort are exogenous factors in reimbursement policies.

Patient's Benefits

From a quality-of-care standpoint, x^* can be seen as an important marker. Avoiding a planned CS for patients with pregnancy complexity below x^* enhances the quality of care by reducing the risks of post-delivery complications. The decision of an SB for women with a pregnancy complexity higher than x^*, however, leads to potential under-treatment and jeopardizes the well-being of both mother and child. As discussed in Section 3, the existence of a threshold x^* in our empirical study allows us to evaluate decisions made about the method of delivery during prenatal care. To this end, we use the choice distance model, i.e., we capture the benefits (costs) of a pregnant woman with pregnancy complexity x with the distance between x and x^*. For instance, in light of the $x^* = 0.85$ identified in our empirical study, where a planned CS is performed for a woman with $x = 0.7$, the benefit of this decision is $b_{CS}(x)$, is -0.15, whereas the opportunity benefit of an SB, $b_{SB}(x)$, for the same case is 0.15. That is,

$$b_{SB}(x) = x^* - x, \quad b_{CS}(x) = x - x^*. \tag{3.1}$$

In the decision on a delivery mode, we assume that each physician is fully able to diagnose the patient's complexity level x, and is informed about the clinical cut-off point x^*.

Physician's Efforts

We model the physician's efforts in terms of the monetary value of the time he would spend performing the tasks in two stages: monitoring the labor and delivering the

baby. Since labor monitoring is only required for SBs, it is not included in the utility of planned CSs. We denote the efforts expended by physicians while performing an NB and a CS by e^N, e^C, respectively. Furthermore, we define $\bar{e}^{MN}$ to represent the effort of fully monitoring labor from onset to an NB or a medically necessary emergency CS.

Let λ, ($\lambda \in [\underline{\lambda}, 1]$) be the proportion of the full effort that a physician has spent from the onset of labor to the delivery. Therefore, the actual effort of monitoring labor, $e^{MN} \triangleq \lambda \bar{e}^{MN}$, is non-decreasing with respect to the effort level λ. We assume that physicians are aware of their effort level. Note that $\lambda = 1$ indicates the full labor monitoring. Let $\underline{\lambda} > 0$ denote the lower bound of the efforts.

Given that the average NB requires about twenty hours of medical monitoring, which can take place anytime, including nights and weekends that elevates the level of inconvenience and, as a result, the amount of effort required by physicians. It is assumed that the cumulative effort to serve an SB (either NB or emergency CS) are higher than for a planned CS; whereas, we assume $e^C > e^N$, since CSs involve a surgical procedure. Specifically,

$$e^N < e^C \leq e^N + \underline{\lambda} e^{MN}. \tag{3.2}$$

We define $f(\lambda, x)$ as the chance of having an NB for a given prenatal complexity level of $x \in [0, 1]$ and an effort level λ, following the SB decision. By definition, $f(\lambda, x)$ is monotonously increasing with respect to λ. As verified by the empirical analysis on our data set, this function is monotonously decreasing with a pregnancy's complexity x: those with a lower complexity level are more likely to have an NB.

The physician's expected effort following an SB decision for a patient with a complexity level of x can be written as $f(\lambda, x)e^N + (1 - f(\lambda, x))e^C + e^{MN}$.

Finally, we assume that the amount of effort saved in labor monitoring dominates the difference between the two procedures. Specifically,

$$(e^C - e^N)\frac{\partial f(\lambda, x)}{\partial \lambda} \leq e^{MN}. \tag{3.3}$$

Physician's Income

As discussed before, the consulting physician makes a decision D on the delivery mode by the end of prenatal care. When the recommendation is a planned CS, he performs the delivery himself and receives a reimbursement of m_{CS}. SB cases however are taken on by the on-call physician (among the group of J physicians), who claims the reimbursement of m_{SB}. As discussed, the on-call physician can also order an emergency CS during labor, and therefore, the attempted NB may be followed by an emergency CS. Without abuse of notation, we denote the reimbursement $m_D \triangleq m_D(\lambda, x)$ where $\forall D \in \{SB, CS\}$, that is, the physician's income is a function of his actual effort and the patient's level of pregnancy complexity. Its exact formulation depends on the reimbursement policy: we will specify the exact policy-dependent form of m_D in Section 6.

Physician's Utility Function

Let $u_D^I(\lambda, x)$ be the expected utility of any physician by the end of prenatal care, solely from the point of view of his own benefit (i.e., monetary income and the dis-utility of efforts), under his decision $D \in \{CS, SB\}$ and for a patient with a complexity x. Accordingly, the physician's utility is estimated as follows:

$$\begin{aligned} u_{CS}^I(\lambda, x) &= m_{CS}(\lambda, x) - e^C \\ u_{SB}^I(\lambda, x) &= m_{SB}(\lambda, x) - [f(\lambda, x)e^V + (1 - f(\lambda, x))e^C + e^{MN}]. \end{aligned}$$

Once we incorporate the benefits of a patient with a pregnancy complexity x into the physician's decision-making process, his expected utility under the decision of a planned CS can be written as

$$U_{CS}(x) = \alpha b_{CS}(x) + u_{CS}^I(\lambda, x),$$

where α refers to a physician's benevolence level, that is, the weight given to the patients' benefit in a physician's mind. The more benevolent the physician, the more he values the quality of care and the benefit of his patients, and the higher will α be. Note that $\alpha \to \infty$ indicates that a physician values only quality of care and completely ignores his own utility: he makes his decisions from a purely clinical perspective, at the cost of his own benefit, which would eliminate unnecessary planned as well as unnecessary emergency CSs. In our framework, we assume that, while physicians are not perfectly altruistic, they all have the same considerably high level of benevolence.

On the other hand, a decision for an SB factors in that each physician has a hospital rotation dedicated to deliveries and is on call for an equal amount of time, and therefore, all physicians in the group have an equal chance of performing this delivery. Hence, they have a $1/J$ chance of gaining the expected utility for an SB, such that

$$U_{SB}(\lambda, x) = \alpha b_{SB}(x) + \frac{u^I_{SB}(\lambda, x)}{J}.$$

Therefore, for an individual patient with a pregnancy complexity x, a rational physician would decide on an SB if $U_{SB}(\lambda, x) \geq U_{CS}(x)$, and on CS otherwise.

3.4.2 Physician's Best Response Strategy

We start the analysis by defining a pregnancy complexity threshold s that maximizes the consulting physician's overall utility over a population of patients. Assuming that the physician behaves according to the utility described in the previous subsection for the population of all patients registered to his panel, this decision is equivalent to setting up an optimal threshold s in order to maximize his total utility. Note that the population density is uniform across the spectrum of complexity levels due to the way x is estimated from our dataset, as explained in Section 3. Therefore, we also normalize the total population to 1. More specifically:

Lemma 3.1 *If a physician aims to maximize his own overall utility, i.e., U, he should decide on a planned C-section by setting an optimal level of s in the prenatal stage,*

i.e.,

$$s = \underset{[0,1]}{arg\ max} \int_0^s U_{SB}(\lambda, x)dx + \int_s^1 U_{CS}(x)dx,$$

which is equivalent to setting s *as*

$$U_{SB}(\lambda, x) \geq U_{CS}(x), \quad \forall x \leq s,$$
$$U_{CS}(x) \geq U_{SB}(\lambda, x), \quad \forall x \geq s.$$

This lemma shows that in order to maximize his overall utility, the physician will set a threshold s in such a way that he will recommend an SB for patients with a clinical complexity x lower than s and will prefer a planned CS for the rest. However, the threshold s set by the physician is influenced by the reimbursement policies and may not necessarily be equal or close to x^*.

Next, we focus on a group of physicians: There are J physicians in the group, and each physician j selects a decision from a set of strategies $D_j = \{CS, SB\}$ and a payoff function $U_j(D_1, D_2, \cdots, D_J)\ \forall j \in \{1, 2, \cdots, J\}$, in the prenatal care stage. They all agree on the same λ under a given reimbursement mechanism m_D. This is a finite symmetric game $< J, D, U >$ given $D = D_1 = D_2 = \cdots = D_J$, and $\forall i, j \in \{1, 2, \cdots, J\}$

$$U_j(CS, d_{-j}) = U_i(CS, d_{-i}), \text{ for } d_{-i} = d_{-j},$$
$$U_j(SB, d_{-j}) = U_i(SB, d_{-i}), \text{ for } d_{-i} = d_{-j},$$

where d_{-i} refers to the decisions of all physicians other than physician i. We present existence and uniqueness of the equilibrium for this game below.

Lemma 3.2 *Each physician should make the same decision for a patient with a complexity level of* $x, \forall x \in [0, 1]$ *at the Nash equilibrium.*

Finally, we present a closed form to calculate the overall CS rate (i.e., planned and emergency CSs). For simplicity, we assume that the physicians are aware of the

function $f(\lambda, x)$ while making a decision D on the delivery mode during prenatal care ($D \in \{CS, SB\}$). The physician's expected effort for all his patients becomes a function of his threshold s and effort level λ and can be written as

$$E(\lambda, s) = \int_0^s \left[f(\lambda, x)e^N + (1 - f(\lambda, x))e^C + e^{MN} \right] \mathrm{d}x + \int_s^1 e^C \mathrm{d}x.$$

Let r be the resulting overall CS rate of a given population. We then set up a one-to-one mapping relationship between s and r by the following lemma.

Lemma 3.3 *Given that $f(\lambda, x)$ is a decreasing function of the pregnancy complexity x, the overall CS rate r can be expressed by the planned CS threshold s*

$$r = 1 - \int_0^s f(\lambda, x)\, dx.$$

Clearly, the overall CS rate r monotonously decreases as the planned CS threshold s increases.

3.5 Health care Payers' Problem

In the context of our work, the term "payer" refers to a private or a public insurer who reimburses the maternity care expenses. In general, the payer aims to maximize the value of care for the patients by achieving the best health outcomes at the lowest cost. This amounts to a two-dimensional objective: maximization of quality and minimization of costs. We start the section by focusing on the payer's goals, and then we study the problem under a perfect information setting, assuming that the payer can fully observe the patient's pregnancy complexity level by the end of prenatal care, as well as the physician's efforts during the labor. This sets a benchmark for our analysis. We end this section by examining the payers' objectives in the more realistic asymmetric information setting within a principal-agent framework.

3.5.1 Payer's Objectives: Maximization of Value for the Patient

Here we introduce a two-dimensional objective function, a weighted sum of economic and quality goals, which is aligned with the models presented by Hua et al. (2016) and Levi et al. (2016).

$$\Pi^{VM} = \beta\Pi^{Q}(s) + \Pi^{E}(s, \lambda, m_D), \tag{3.4}$$

where β is the weight of the quality objective, with a monetary unit. A higher amount of β indicates that a greater importance is given to quality in the payers' policy design. Specifically, in the event that $\beta = 0$, Π^{VM} becomes the sole economic objective; and if $\beta \to \infty$, then Π^{VM} is equal to the quality objective. The physician's optimal threshold is denoted by s^E and s^Q in these two special cases, respectively.

Quality Perspective: Maximization of Care Quality

Birth quality is the most important objective from the perspective of social welfare and is accordingly an essential concern for the payers. Recall that the cut-off point x^* represents a clinically appropriate threshold for a planned C-section. We consider the distance $s - x^*$ a measure of quality of care. More specifically, in the event of $s > x^*$ under-treatment occurs for complex pregnancies that should have been planned CSs. By contrast, $s < x^*$ indicates overtreatment, or the inappropriate selection of a planned CS for low-risk pregnancies that should have been SBs. Therefore, the payer's goal in the context of care quality can be expressed to minimize

$$\Pi^{Q}(s) = |s - x^*| \tag{3.5}$$

which is independent of any reimbursement mechanism for physicians. Evidently, x^* is the resulting optimal quality threshold satisfying unconstrained Eq.3.5.

Economic Perspective: Minimization of Costs

The cost of maternity care includes all expenditures involved in the prenatal, delivery and postpartum stages of care. Given that prenatal care costs (consultation provider

fees plus the cost of imaging and laboratory tests) are independent of the decision about a delivery mode, they are outside the scope of this work. Therefore, in the context of our study, maternal expenditures consist of all payments for delivery and postpartum care.

Cost of delivery care captures all the expenses of the delivery and post-delivery hospital stay, for both mother and newborn, and we categorize them into two groups: facility fees and physician charges. The former includes payments for the physical facility (e.g. delivery room, operating theater, post-surgery recovery room, etc.) as well as for nursing, anesthesiology, radiology/imaging, laboratory, and pharmacy services. Let c_H^C and c_H^N represent the facility fees for a CS and an NB, respectively. Because of the surgical nature of the CS, and the longer in-hospital stay that follows this operation, its associated facility, hospitalization and nursing costs are significantly higher than those for an NB, i.e., $c_H^C > c_H^N$. c_H^C is considered to be the same for both planned and emergency CSs, because of the similar resource requirements.

For a given population of patients, the expected facility fees $CH(\lambda, s)$ depend on the threshold s determined by the physicians, and can be written as

$$CH(\lambda, s) = \int_0^s \left[f(\lambda, x) c_H^N + (1 - f(\lambda, x)) c_H^C \right] \mathrm{d}x + \int_s^1 c_H^C \mathrm{d}x.$$

The expected total physicians' fee is also dependent on the threshold s, and can be expressed as follows:

$$M(\lambda, s, m_D) = \int_0^s m_{SB}(\lambda, x) \mathrm{d}x + \int_s^1 m_{CS}(\lambda, x) \mathrm{d}x.$$

The cost of postpartum care takes into account expenses resulting from the treatment of post-delivery complications, including re-admissions, as well as follow-up care provided in the three months after childbirth. Both the medical literature (e.g. Knight et al., 2008; Goer et al., 2012; Villar et al., 2007) and our empirical study confirm that the risk for mother and baby of having post-delivery complications varies significantly with the pregnancy's complexity and the mode of delivery. We use $I_D(\lambda, x)$ to denote the incidence of post-delivery complications under decision D for a pregnancy com-

plexity level of x with an effort level of λ in the delivery stage. Our empirical analysis reveals that, for a given complexity level of x, $I(CS, x) \triangleq I_{CS}$, as it is independent of x, and $I(SB, x)$ is an increasing linear function of x. In light of this information, we further assume that incidence decreases as effort increases in the delivery stage, since a higher effort will lead to lower rates of unnecessary emergency CSs and of complication risks. Specifically,

$$\frac{\partial I_{SB}(\lambda, x)}{\partial x} > 0, \quad \frac{\partial I_{SB}(\lambda, x)}{\partial \lambda} > 0. \tag{3.6}$$

Let C be the average treatment and re-admission costs for post-delivery complications per case, for the mother and baby. Then, the postpartum expenses for the overall population can be expressed as

$$CI(\lambda, s) = C \int_0^s I_{SB}(\lambda, x)\mathrm{d}x + C \int_s^1 I_{CS}\mathrm{d}x.$$

This portion of the expenses can be incurred as extra charges to payers during the original hospitalization or over the short term after discharge. It is an essential component of birth-related expenses for payers, but has been consistently underestimated by both payers and policy makers (Truven Health Analytics, 2013).

From an economic perspective, the payer aims to minimize her total maternity care costs by minimizing the following objective function

$$\Pi^E(\lambda, s, m_D) = M(\lambda, s, m_D) + CH(\lambda, s) + CI(\lambda, s). \tag{3.7}$$

Lemma 3.4 *If $M(\lambda, s, m_D)$ is a convex function of s, Π^E is convex with respect to s.*

Through this lemma, we show that this objective function is convex in terms of physicians' threshold, determined in the consulting stage, under any reimbursement mechanism.

3.5.2 Benchmark: Payer's Objectives under Perfect Information

Before beginning our analysis of how the physicians' actions to maximize their own utilities affect the cost and quality of care, we first present the benchmark, in which the payer can fully observe the patients' health conditions or pregnancy complexity, as well as the physicians' efforts. This allows the payer to set a threshold for physicians for ordering a CS during the stage and enables her to require that physicians expend a full effort during the delivery stage. Specifically, under a setting of full information transparency, the payer's problem can be written as follows:

$$
\begin{aligned}
Z_{BM} = \min_{s, m_D} \quad & \Pi^{VM} \\
\text{subject to} \quad & u^I_{SB}(\lambda, x) \geq 0, \quad \forall x \leq s && \text{(PCN)} \\
& u^I_{CS}(\lambda, x) \geq 0, \quad \forall x \geq s && \text{(PCC)} \\
& \lambda = 1.
\end{aligned}
$$

The first two constraints, PCN and PCC, are motivated by Lemma 3.1 and ensure that the compensation for the NB and CS is sufficient for the physicians' efforts, so that they engage in both forms of delivery. Only under these constraints the payer is free to set any threshold s and m_D to achieve her objective. The third constraint requires that the physician monitor the full labor unless an emergency CS is medically necessary.

Similarly to the unconstrained representation of $\Pi^Q(s)$ in Section 5.1.1, the optimal s^Q is equal to x^* for the constrained problem under a full information setting. However, as Proposition 3.1 indicates below, the optimal s^E needed to minimize costs is not necessarily equal to x^*.

Proposition 3.1 *If the payer observes a certain effort level λ achieved in the delivery stage under a reimbursement mechanism m_D, and both $U_{CS}(\lambda, x) \geq 0$, $\forall x > s^E$ and*

$U_{SB}(\lambda, x) > 0$, $\forall x < s^E$, s^E can be calculated as solution of equation

$$f(\underline{\lambda}, x)(c_H^N + e^N - c_H^C - e^C) + C(I_{SB}(\underline{\lambda}, x) - I_{CS}) + \bar{e}^{MN} = 0, \tag{3.8}$$

if and only if s^E is in $[0, 1]$; Otherwise, $s^E = 1$.

Note that the optimal values of s^Q and s^E represent the boundaries of the region of the optimal physician threshold s. Depending on the reimbursement mechanism in effect, s^E could be on either side of s^Q. Increasing the weight of care quality (i.e., β in Eq. 3.4) would move s towards s^Q and away from s^E. That is, a quality improvement would come at an increased maternity care cost. Considering the many instances in which increased health care expenses do not necessarily improve quality of care, this is presumably more palatable for the payer. Proposition 3.2 presents a formal statement of this window of opportunity, where there is a clear trade-off between the quality and cost of maternity care.

Proposition 3.2 *The optimal value of s for the payer is between s^E and x^* under perfect information.*

For any s outside this region, the additional expenditures for maternity care may not lead to an increase in quality of care, and hence, this region constitutes a benchmark for us.

3.5.3 Payer's Objectives under Asymmetric Information

Let us now enhance the model to represent the asymmetric information setting, where only physicians are able to observe the patient's level of complexity at the prenatal stage, and the progress of labor during the delivery stage. Therefore λ and s are in fact decided by the physician, whereas the payer's only lever to incentivize the physicians to achieve the desired threshold s is m_D. Under this framework, we formulate the payer's optimal problem as follows:

$$
\begin{aligned}
Z_P = \min_{m_D} \quad & \Pi^{VM} \\
\text{subject to} \quad & u^I_{SB}(\lambda, x) \geq 0, \quad \forall x \leq s && \text{(PCN)} \\
& u^I_{CS}(\lambda, x) \geq 0, \quad \forall x \geq s && \text{(PCC)} \\
& \lambda = \underset{\underline{\lambda},1}{\operatorname{argmax}} u^I_{SB}(\lambda, x), \quad \forall x \leq s && \text{(ICE)} \\
& U_{SB}(\lambda, x) \geq U_{CS}(\lambda, x), \quad \forall x \leq s && \text{(ICN)} \\
& U_{SB}(\lambda, x) \leq U_{CS}(\lambda, x), \quad \forall x \geq s && \text{(ICC)},
\end{aligned}
$$

where the first two constraints, PCN and PCC, are exactly the same as those in Problem Z_{BM}. In the third constraint, however, the materialized effort λ is determined by the physician, according to his own utility $u^I_{SB}(\lambda, x)$ rather than being set by the payer to 1. The constraints of ICC and ICN refer to the physician's decision of a delivery mode by the end of prenatal care and ensure the maximization of the physicians' utility, as presented in Lemma 3.1, in Section 4.2.

By incorporating the ICN and ICC constraints into this asymmetric setting, we first examine the impact of the physicians' benevolence α and the group size J on the physician's threshold s and on the quality of care. As expected, the deviation from the threshold for a planned C-section s under a given reimbursement mechanism m_D and x^* is non-increasing as α increases. We present our findings analytically in Section 9. Moreover, s is sensitive to the group size J in the asymmetric setting.

Lemma 3.5 *If $U_{SB}(x) \geq 0$ $\forall x \in [0,1]$ under given reimbursement mechanism m_D, s is non-increasing as J increases.*

For a reimbursement mechanism that leads to a threshold of s less than x^*, a smaller group may be preferable, as it increases the threshold closer to x^*, given that the physicians are more likely to serve their own patients in cases where a decision for an SB is made within a smaller group. Under a reimbursement mechanism that motivates a preference for SBs (even where a planned C-section could be more medically

appropriate, i.e., $s > x^*$), a larger group has the advantage of lowering the threshold, that is, of avoiding the SB for high-risk patients. This finding could provide important managerial insights for a payer dealing with physician groups of different sizes and different contract types.

Next, we study the characteristics of a feasible solution to Problem Z_P.

Corollary 3.1 *Suppose the range for the set of feasible solutions of threshold s to Problem Z_P is $[\underline{s}, \bar{s}]$. If this range is completely exclusive of the interval between s^E and x^*, then, under the asymmetric information setting, the optimal cost-minimization threshold is equal to the optimal quality threshold, which is equal to*

- *$\bar{s}$, if $\bar{s} < min\{s^E, x^*\}$;*
- *$\underline{s}$, if $\underline{s} > max\{s^E, x^*\}$.*

This result implies that there is an equivalent optimal solution for the payer's economic and quality objectives under the asymmetric information setting. However, the reimbursement policy that results from this situation is suboptimal from the point of view of value maximization and should definitely be avoided, since such policies erode the quality of care while also increasing the related expenses. Consequently, the payer is disadvantaged by a double-layered "information rent" in the form of reduced quality and expanded costs.

3.5.4 Payer's Objectives Under Asymmetric Information

In reality, though the payer reimburses physicians for the sequential procedures in the delivery stage, he has no direct access to patients' health condition or pregnancy complexity; and only physicians can actually observe the progress of patients' pregnancy. This is a typical setting of asymmetric information with moral hazard (or hidden actions), where a physician can take advantage of all the information he observes, and decide his threshold for planned CS in consulting stage, and effort level during delivery stage. Therefore, the payer (the principal) relies on the payment to incentive the physicians (the agent) to achieve the desired threshold s in his economic

or quality objective ; since the payer cannot control the threshold directly and explicitly. Under this typical principal-agent framework, we formulate the payers' optimal problem under this setting by incorporating three incentive constraints to physicians, in additional to participating constraints.

$$\begin{aligned} &\min_{m_D} \Pi_P(s, \lambda, m_D) \\ \text{subject to} \quad & u^I_{SB}(\lambda, x) \geq 0, \quad \forall x \leq s \quad \text{(PCN)} \\ & u^I_{CS}(x) \geq 0, \quad \forall x \geq s \quad \text{(PCC)} \\ & U_{SB}(\lambda, x) \geq U_{CS}(x), \quad \forall x \leq s \quad \text{(ICN)} \\ & U_{SB}(\lambda, x) \leq U_{CS}(x), \quad \forall x \geq s \quad \text{(ICC)} \\ & \lambda \in \underset{\underline{\lambda},1}{\text{argmax}} u^I_{SB}(\lambda, x), \quad \forall x \leq s \quad \text{(ICE)}, \end{aligned} \tag{3.9}$$

where Π_P can be any objective Π^Q_P, Π^E_P, Π^{VM}_P as we mentioned in section 3.5. The first two constraints (PCN and PCC) are exactly the same as those in Problem **??**. The last three constraints (ICN, ICC and ICE) are the incentive constraints typically in the setting of asymmetric information. Lemma 3.1 implies that the combination of ICC and ICN ensures the maximization of physicians' utility. Furthermore, the incentive constraints ICC and ICN are typical in consulting stage with a twofold impact: (i) physicians eliminate unnecessary CS for patients with lower risks; (ii) the planned CS should be retained for those with higher complexities. ICE is typically set up for the delivery stage, where physicians select the most desirable effort level of providing a care during delivery through their shift in hospital - the one that maximizes their expected utility $u^I_{SB}(\lambda, x)$.

With ICN and ICC in this asymmetric setting, we first examine the impact of physicians' benevolence α on defining their threshold s, and consequential impact on quality of care.

Lemma 3.6 *If a reimbursement mechanism $m_D(x)$ $\forall D \in \{SB, CS\}$ leads to a consequent threshold of planned CS s, quality of care increases with respect to his benevolence. That is, the deviation from the clinical cutoff of planned CS $|s - x^*|$ is non-*

increasing as α increases.

Consider the worst scenario where physicians fully ignore patients' benefits, or the most selfish physicians, i.e. $\alpha = 0$, a certain reimbursement policy leads physicians to determine the threshold of s. The deviation from the clinically optimal rate of planned CS is no larger than $|s - x^*|$ for physicians with $\alpha > 0$. However, the clinically optimal threshold can be gained under any reimbursement,i.e.$s = x^*$ if and only if $\alpha \to \infty$.

Moreover, the following result states that the quality objective is sensitive to the group size J in the asymmetric setting, given the fact of sharing the tasks to serve SB in hospital.

Lemma 3.7 *If a reimbursement mechanism $m_D(x)\ \forall D \in \{SB, CS\}$, satisfying $u^I_{SB}(x) \geq 0\ \forall x \in [0,1]$, it leads to a consequent threshold of planned CS s, which is non-increasing as J increases.*

Lemma 3.7 demonstrates a two-fold impact of group size. For a reimbursement mechanism that leads to a threshold lower than x^*, a smaller group may be preferable as it increases threshold closer to s. Intuitively, physicians are more likely to serve their own patients in the case of deciding a SB, in a smaller group. Under a reimbursement mechanism that motivates insufficient planned CS, i.e. $s > x^*$, a larger group has the advantage of lowering the threshold, reducing the mis-application of SB for highly risky patients. The extreme example is when $J \to \infty$, the group is very large, and hence physicians have little chance to serve any NB during their shifts in hospital. Consequently, they have no chance to increase their utility from delivery stage, leading them to an induction of gaining the certain overall utility by deciding on planned CS.

Recall s^E is the resulting economically optimal threshold that satisfies the cost minimization objective function in Eq.3.7 in the setting of perfect information (with constraints in Problem **??**), and then we have the following corollary, interpreting characteristics of feasible solution to Problem 3.9.

Corollary 3.2 *Suppose the feasible solution of threshold to Problem 3.9 is* $[\underline{s}, \bar{s}]$. *If it falls outside of the interval between* s^E *and* x^*, *the optimal cost-minimization threshold is equal to the quality optimal threshold. That is, both optimal thresholds become*

- $\bar{s}$, *if* $\bar{s} < min\{s^E, x^*\}$;
- $\underline{s}$, *if* $\underline{s} > max\{s^E, x^*\}$.

Which leads to the fact that payers' economic objective to minimize Π_P^E *is equivalent to maximize their quality objective* Π_P^Q *in this case.*

It implies an equivalent optimal solution for both economic and quality objectives of payers in the asymmetric information setting. However, the reimbursement policy that results in this situation is sub-optimal and should be definitely avoided, from the perspective of value maximization, according to Proposition **??**, and hence detriments the quality and increases related expenses. Consequently, payers suffer from double layered "information rent" in the form of reduced quality and additional costs.

3.6 Payment Models - Level 1: Mainstream Payment Schemes

In this section, we study payment schemes, namely FFS, blended and bundled payments, in the context of maternity care through our modeling framework and we discuss our findings.

3.6.1 Payment Scheme Descriptions

Given that each payment model may have different interpretations for different specialties, we first would like to provide the description of the payment schemes in a maternity care setting. The specific formula under each model is summarized in Table 3.4.

Fee-for-Service (FFS): A physician gets a payment of a fixed rate of P^{PC} for performing a planned CS delivery, and P^{N} for an NB or P^{EC} for performing an emergency CS. In practice, the rates for emergency and planned CSs vary, yet both are higher than the rate for an NB, due to the surgical nature of a CS (Faloon, 2012; Optum, 2013). That is, $P^{EC} > P^{N}$, $P^{PC} > P^{N}$. Moreover, the rate for an emergency CS may be a little higher than for a planned CS; however, the difference is not significant (MSC Payment Schedule, 2016; AHCIP, 2016; Ontario Health Insurance Plan, 2016). Specifically, we assume that $P^{PC} \leq P^{EC} < e^{MN} + P^{PC}$.

Blended Payment: A single rate P^{BP} is paid for a delivery, regardless of mode.

Bundled Payment: A fixed amount P^{BL} is paid for each registered pregnancy, including prenatal care (i.e., consultations and ultrasounds), delivery, and the post-delivery hospital stay, regardless of the delivery mode (CPR, 2012). The portion to be paid for prenatal care is not included in our current analysis in order to keep expenses comparable with those of other payment methods. There are alternative approaches for sharing the risks and gains between hospital and the physician group under this program. In the context of this research, we propose a full gain/risk sharing for the physicians. This maximizes the accountability of the physicians regarding to care they provide and the coordination among the physicians. Under this model, since the delivery cost for a CS is higher than for an NB, the physicians' marginal income following a CS is lower than with an NB.

For the FFS and blended methods, the amount of the fee and the payee are determined retrospectively, after the delivery of the baby. This is in contrast to the bundled payment scheme, which pays a pre-established amount, i.e. prospective reimbursement.

Table 3.4: Specific Notations of Different Payment Policies

Policy	$m_{SB}(\lambda, x)$	$m_{CS}(\lambda, x)$
FFS	$P^N f(\lambda, x) + P^{EC}(1 - f(\lambda, x))$	P^{PC}
Blended	P^{BP}	P^{BP}
bundled	$P^{BL} - \frac{c_H^N f(\lambda,x) + c_H^C (1-f(\lambda,x))}{J}$	$P^{BL} - \frac{c_H^C}{J}$

Table 3.5: Impacts of Payment Methods on Cost, Quality of Care, Financial Risks and Accessibility

	FFS	Blended	Bundled
Incentive for Quality of Care	None (Proposition 3.3, Lemma 3.8)		None (Proposition 3.4)
Incentive for Cost Control	None (Corollary 3.3)		High (Corollary 3.4)
Physician's Financial Risks	None	None	High (Corollary 3.4)
Potential Accessibility Problem	None	None	High (Proposition 3.5)

3.6.2 Analytical Analysis on Payment Schemes

Next, we investigate the impact of payment schemes on quality of care and on the overall maternity cost. We also study potential problems, such as the financial risks taken on by physicians and the accessibility to physicians under these models. A brief summary of our analytical findings is given in Table 3.5.

FFS and Blended Models

Since they use a retrospective payment approach, the FSS and blended payment schemes share similar characteristics. First, we investigate the threshold under these payment schemes between an SB and a CS, as determined by the physicians by the end of prenatal care, and next we study their impact on the physicians' efforts while monitoring labor under SB decisions.

Proposition 3.3 *Under the FFS and blended payment models, the optimal threshold s determined by the physicians in Problem Z_P satisfies $s < x^*$. Moreover, under the blended payment model, s is monotonically decreasing with respect to P^{BP}.*

Lemma 3.8 *Under the FFS and blended payment models, $\underline{\lambda}$ is optimal for physicians at the delivery stage, following an SB decision.*

Proposition 3.3 implies that both payment schemes lead physicians to choose CS over SB, although it may not be medically necessary for some patient groups. Moreover, it shows that offering too low of a blended rate would continue to encourage cesarean deliveries. Likewise, as presented in Lemma 3.8, even under an SB decision, these payment mechanisms motivates the physicians not to give their full effort in the delivery stage. Hence, the desired rate of NBs may not be realized under these payment mechanisms. Our finding that FFS incentivizes physicians in favor of overtreatment through CSs is consistent with existing empirical studies in the literature (Gruber et al., 1998). Additionally, we conclude that equalizing the fees for NBs and CSs through a blended model has a limited impact on controlling CS rates. The blended model eliminates the direct financial incentives for preferring CS as a procedure but does not provide any incentives regarding the physician's desire to get the delivery fee for his own patients or avoid the inconveniences of NBs.

From an economic perspective, we first show that the function representing the total amount of reimbursement transferred from the payer to the physicians under these two mechanisms is non-concave with the following lemma.

Lemma 3.9 *Under the FFS and blended payment systems,* $M(\lambda, s, m_D)$ *is non-concave with respect to* s*.*

By using this property, we next show the following:

Corollary 3.3 *Under the FFS and blended payment systems,* s^E *presented in Corollary 3.2 is an infeasible solution for Problem* Z_P*.*

This result implies that the feasible solutions to Problem Z_P under these two mechanisms are outside the interval between s^E and x^* (or x^* and s^E), as presented in Corollary 3.2. In other words, under these payments schemes, the physicians are overpaid for the level of effort they invest and for the quality of care resulting from their effort. Intuitively, physicians can always save a certain amount of effort in the delivery stage by performing an emergency CS while receiving at least the same payment. Therefore, physicians get a higher margin—the difference in income and effort—from

planned or emergency CSs under these models, which may result in a higher number of unnecessary CS cases.

Bundled Payment Model

As we discussed before, the bundled payment approach is a prospective reimbursement model, in which a fixed up-front payment is received for each patient, regardless of the actual delivery mode used for that patient. Therefore, under this payment model, the net transfer of funds from the payer to the physicians has a concavity feature with respect to the threshold decided upon by the physicians by the end of the prenatal stage.

Lemma 3.10 *$M(\lambda, s, m_D)$ is concave with respect to s under the bundled payment model.*

We first consider the impact of bundled payments on the physicians' decision to plan a CS. This payment scheme's structure aims to provide incentives for better outcomes, specifically by avoiding over-treatment by shifting the financial responsibility to the providers. On the other hand, since this payment model reimburses regardless of the resources used, it may jeopardize the quality of care by increasing the desire to keep costs low, which may lead to physicians not prescribing a planned CS where medically required (i.e., undertreatment) (Feder, 2013; Adida et al., 2017). Proposition 3.4 confirms undertreatment under the bundled payment scheme, as compared to the retrospective payment mechanisms discussed above.

Proposition 3.4 *Under the bundled payment model where a physician's facility costs dominate the monetary value of the physicians' effort invested in servicing a delivery, specifically,*

$$f(\lambda, x)\left(\frac{c_H^N}{J} + e^N\right) + (1 - f(\lambda, x))\left(\frac{c_H^C}{J} + e^C\right) + e^{MN} \leq \frac{c_H^C}{J} + e^C, \forall \lambda \in (\underline{\lambda}, 1), \forall x \in (0, 1),$$

the physician's threshold $s > x^*$ in the prenatal care stage.

Our findings are consistent with the existing literature, which has demonstrated that this payment model discourages physicians from overusing surgical procedures

(Ransom et al., 1996; Lally, 2013).

Corollary 3.4 *Under a bundled payment model where a physician's facility costs dominate the monetary value of the physicians' effort invested in servicing a delivery, increasing bundled rate can motive consulting physicians to set up the quality-maximized threshold x^*.*

This result highlights the fact that the bundled payment approach result in higher expenses due to the required risk premium for physicians, to motivate the adoption of this payment scheme and to guarantee a certain care quality level. Otherwise, the high level of financial risk may lead to a great deal of resistance to adopting this model (Adida et al., 2017). Or, in order to reduce their costs significantly, and thus alleviate the financial risks they face, physicians may be inclined to under-treat a significant number of cases by preferring NB even where a CS is more medically appropriate. Moreover, it may also result in patient selection, also known as "cherry picking". Physicians may refuse to serve high-risk women, since this group is potentially more costly due to the higher chance of a planned or emergency CS. We highlight the link between the potential for patient selection and the bundled payment rate in the following proposition.

Proposition 3.5 *Under the bundled payment model, the lower bound of the bundled rate P^{BL} is $e^C + c_H^C/J$. Moreover, if $P^{BL} < e^C + c_H^C/J$, physicians may refuse patients with complexity x as long as*

$$x > f^{-1}\left(\frac{J(e^C + e^{MN} - P^{BL}) + c_H^C}{J(e^C - e^N) + c_H^C - c_H^N}\right)$$

where $f^{-1}(x)$ is the inverse function of $f(1, x)$.

This implies that there is a higher chance of cherry picking, that is, of physicians choosing low-risk patients over high-risk ones, under a lower bundled rate. Moreover, we have several interesting observations about the formula on the likelihood of refusal. Given a fixed e^C, the effort for performing a CS, a lower effort in attending to a NB, i.e. e^N, leads to higher number of patients being refused; similarly, increasing difference

between facility fees of the two procedures, i.e. c_H^C and c_H^N, results in more patient selection.

In practice, there might be a higher bundled rate for high-risk patients. However, this does not alleviate the moral-hazard problem for the following two reasons: First, physicians will still recommend an SB for some high-risk patients, since this is the "optimal" way for them to minimize their expenses, and hence, maximize their utility. This again reflects the typical dilemma of a moral-hazard problem. Second, a higher bundled rate for high-risk patients does not eliminate the "cherry picking" of patients. Indeed, certain intermediate-risk patients may be discriminated against, since they are more likely to have a CS than low-risk patients but physicians will not receive a higher payment for treating them. Therefore, this group of patients offers the least utility to physicians, as compared to low- or high-risk patients. In our numerical analyses, we allow for the definition of two separate rates, for high-risk and low-risk pregnancies, presented in Section 3.9.

3.7 Payment Models - Level 2: Complementary Bonuses

In the payment reform of maternity care, complementary payments play an important role since they may offset some of the disadvantages of the payment schemes discussed above. The effectiveness of these add-on bonuses depends greatly on their design: (i) the performance measure that will be incentivized, and (ii) the person(s) to be incentivized. Therefore, in this section, we first present our proposed add-on bonuses, i.e., performance metrics and distribution mechanisms for maternity care. Then, we discuss the analytical properties for these bonuses.

3.7.1 Proposed Add-on Bonuses for maternity care

In terms of the performance measures to be incentivized, we propose four types of process- or outcome-oriented bonuses to act as add-ons to payment schemes at level 1. These are chosen based on our conversations with physicians and hospital administrators, and on a detailed literature review on targeted performance metrics for

maternity care. They are all based on simple metrics that are easy to observe, and thus, are easy to implement in practice. We specify the formula of this set of bonuses as one of multiple components in $m_D, D \in \{CS, SB\}$ and $M(\lambda, s, m_D)$ in Table 3.6.

Complexity bonus. Bonus B^{CO} is paid if the physician prescribes a planned C-section for those with $x > x^*$, and an SB to patients with $x < x^*$ by the end of prenatal care. This bonus policy would be implemented under the assumption that the clinically optimal cut-off point x^* is determined and set by the payer for use as the patient pregnancy complexity threshold.

Postpartum outcome bonus. Bonus B^{PO} is paid in the case that neither the patient nor the baby has any post-delivery complications.

NB bonus. Bonus B^{NB} is paid so long as the patient has a NB.

CS threshold bonus. Bonus B^{TH} is paid to every physician in the group when the overall CS rate for their patients is below a threshold.

In terms of the person(s) to be incentivized, to parallel the two stages in which physicians are involved in deciding on delivery modes, we provide possible compensatory methods that cover both the prenatal and delivery stages. Specifically, we recommend four alternative recipients for the proposed bonuses.

Consulting only: The physician responsible for prenatal care;

Delivery only: The physician responsible at the delivery stage;

Relevant parties: The responsible physicians at both the prenatal and delivery stages;

Group: All physicians in the group when a single birth meets a certain criteria of outcome metric.

Although the Complexity Bonus and the CS Threshold Bonus are only applicable to consulting physicians and to the group of physicians, respectively, the rest can be

Table 3.6: Specific Components of Different Bonus Policies

Policy	$m_{SB}(x)$	$m_{CS}(x)$	$M(\lambda, s, m_D)$
B^{CO}	$B\mathbb{I}_{x<x^*}$	$B\mathbb{I}_{x\geq x^*}$	$B(1-\lvert s-x^*\rvert)$
B^{PO}	$B(1-I(SB,x))$	$B(1-I(CS,x))$	$B(1-\int_0^s I(SB,x)\mathrm{d}x-\int_s^1 I(CS,x)\mathrm{d}x)$
B^{NB}	$Bf(x)$	$Bf(x)$	$B\int_0^s f(\lambda,x)\mathrm{d}x$

Table 3.7: Distribution Mechanisms for Proposed Complimentary Payments and the Relevant Analytical Findings

Policy	Consulting only	Delivery only	Relevant Parties	Group
Complexity B^{CO}	Proposition 3.7	Not Applicable		
Postpartum B^{PO}	Proposition 3.6	Proposition 3.8	Proposition 3.6 and 3.8	Proposition 3.11
NB B^{NB}				
CS threshold B^{TH}	Not Applicable			Proposition 3.12

provided through all four distribution mechanisms. The associated expressions for the proposed bonuses under different distribution mechanisms are specified in Table 3.9.

Table 3.8: Applicability of Distribution Mechanisms for Outcome oriented Bonuses

Policy	Consulting only	Delivery only	Relevant Parties	Group
Complication B^{CO}	Proposition 3.7	Not Applicable		
Postpartum B^{PO}	Proposition 3.6	Proposition 3.8	Yes	Proposition 3.11
NB B^{NB}				
CS threshold B^{TH}	Not Applicable			Proposition 3.12

Table 3.9: Model of Different Bonus Distribution Mechanisms

Policy	$m^I_{SB}(\lambda, x)$	$m^I_{CS}(x)$	$m_{SB}(\lambda, x)$	m_{CS}	$M(\lambda, s)$
Complication Bonus					
Consulting Only	P	P	$\frac{P}{J} + B\mathbb{I}_{x<x^*}$	$P + B\mathbb{I}_{x\geq x^*}$	$P + B(1 - \lvert s - x^*\rvert)$
NB Bonus					
Consulting Only	P	P	$\frac{P}{J} + Bf(x)$	P	$P + B\int_0^s f(\lambda, x)\mathrm{d}x$
Delivery Only	$P + Bf(x)$	P	$\frac{P+Bf(x)}{J}$	P	$P + B\int_0^s f(\lambda, x)\mathrm{d}x$
Relevant Party	$P + Bf(x)$	P	$\frac{P+(J+1)Bf(x)}{J}$	P	$P + 2B\int_0^s f(\lambda, x)\mathrm{d}x$
Group	$P + Bf(x)$	P	$\frac{P}{J} + Bf(x)$	P	$P + JB\int_0^s f(\lambda, x)\mathrm{d}x$
Postpartum Outcome Bonus					
Consulting Only	P	P	$\frac{P}{J} + B(1 - I(SB, x))$	$P + B(1-$ $I(CS, x))$	$P + B(1-$ $\int_0^s I(SB, x)\mathrm{d}x - \int_s^1 I(CS, x)\mathrm{d}x)$
Delivery Only	$P + B(1-$ $I(SB, x))$	$P + B(1-$ $I(CS, x))$	$\frac{P+B(1-I(SB,x))}{J}$	$P + B(1-$ $I(CS, x))$	$P + B(1-$ $\int_0^s I(SB, x)\mathrm{d}x - \int_s^1 I(CS, x)\mathrm{d}x)$
Relevant Party	$P + B(1-$ $I(SB, x))$	$P + B(1-$ $I(CS, x))$	$\frac{P+(i+J)B(1-I(SB,x))}{J}$	$P + B(1-$ $I(CS, x))$	$P + 2B(1-$ $\int_0^s I(SB, x)\mathrm{d}x - \int_s^1 I(CS, x)\mathrm{d}x)$
Group	$P + B(1-$ $I(SB, x))$	$P + B(1-$ $I(CS, x))$	$\frac{P}{J} + B(1 - I(SB, x))$	$P + B(1-$ $I(CS, x))$	$P + JB(1-$ $\int_0^s I(SB, x)\mathrm{d}x - \int_s^1 I(CS, x)\mathrm{d}x)$
CS Threshold Bonus					
Group*	$P+$ $B\mathbb{I}_{\int_0^s f(\lambda,u)\mathrm{d}u<r^*}$	P	$\frac{P+B\mathbb{I}_{\int_0^s f(\lambda,u)\mathrm{d}u\geq 1-r^*}}{J}$	P	$P+$ $B\mathbb{I}_{\int_0^s f(\lambda,u)\mathrm{d}u\geq 1-r^*}$

3.7.2 Analytical Properties of Proposed Bonuses

We demonstrate the analytical properties of these bonuses, for which our findings are summarized in Table 3.8. These analytic properties provide crucial managerial insights, given that most of these bonuses have not yet been implemented in the health care system. We verify our analytical findings with numerical experiments, which are presented in Section 3.9.

First, we show the impact of alternative bonus types on physicians' decisions to perform a planned CS in the prenatal stage. By design, these would be the NB, Postpartum Outcome and Complexity Bonuses that are provided to the consulting physicians.

Proposition 3.6 *If $\bar{s} < x^*$ under the original payment mechanism, an NB Bonus B^{NB} increases $\bar{s}$ in the prenatal stage. If $\bar{s}$ is smaller than the intersection of I_{CS}) and $I_{SB}(\lambda, x)$, the incidence of complications under CS and SB respectively, a Postpartum Bonus B^{PO} also increases $\bar{s}$ in the consulting stage.*

This result implies that the *NB* and *Postpartum Outcome* Bonuses simply reduce the chance of a planned CS decided on in the prenatal stage. For the former one, this may lead to under-treatment, i.e., not prescribing a CS although it is medically appropriate, in some cases depending on the monetary value of the bonus and base payment model. On the other hand, a *Complexity* add-on may be more effective in discouraging either over- or undertreatment. Intuitively, it provides motivations to prescribe both SBs for patients with low risk and planned CSs for medically appropriate cases. This is formulized in the result below.

Proposition 3.7 *A Complexity add-on reduces deviation from the clinical cut-off point, as compared to the same reimbursement mechanism without the add-on.*

This proposition shows that complexity-related add-on motivates physicians to align with the payer's quality objective in Eq.3.5.

Second, we show the advantage of proper add-on to motivate the physician's full effort when performing a delivery under an SB decision. These would be the NB and Postpartum Outcome Bonuses to be paid to the delivering physicians.

Proposition 3.8 *Providing an NB rate or a Postpartum Outcome Bonus to on-call physicians can increase their effort level λ. Specifically, the lower bound of the NB rate add-on or Postpartum Outcome Bonus is*

$$\underline{B}_{NB} = \frac{1}{\nu}\bar{e}^{MN} - (e^C - e^N), \text{ where } \nu = \min_{\lambda,x} \frac{\partial f(\lambda, x)}{\partial \lambda};$$
$$\underline{B}_{PO} = \frac{1}{\tau}\bar{e}^{MN} - \frac{\upsilon}{\tau}(e^C - e^N), \text{ where } \tau = \min_{\lambda,x} \frac{-\partial I_{SB}(\lambda, x)}{\partial \lambda}, \upsilon = \max_{\lambda,x} \frac{\partial f(\lambda, x)}{\partial \lambda};$$

Proposition 3.8 implies that the NB and Postpartum Outcome Bonuses may lead to a reduction in the number of unnecessary emergency CSs. These two types of bonuses would motivate physicians to choose the most appropriate and efficient procedure at the delivery stage by giving their best effort. The Postpartum Outcome Bonus can also be interpreted as the combination of an upfront payment with a penalty for post-delivery complications, where the penalty would discourage improper procedures and underutilization of efforts at the delivery stage, resulting in an improved quality of care.

Next, we study the alternative bonuses from the cost-effectiveness perspective. The following proposition reveals the difference of Complexity bonus and NB bonus regarding to their costs.

Proposition 3.9 *Given the same effort level in the delivery stage, regarding consulting only bonuses to achieve the same feasible level of quality, Complexity Bonus costs less than NB Bonus.*

Although Complexity Bonus performs better in reducing costs with the similar level of care quality, it has its own drawbacks. Note that, it is very challenging to ascertain the true complexity level of a patient, because this is an assessment done by the physician. Some physicians may overestimate x that would result in an increase in the bonus payments. Therefore, the implementation of a complexity bonus involves a potentially expensive monitoring and auditing mechanism.

Similarly, the following proposition shows the advantage of postpartum bonus in terms of cost savings.

Proposition 3.10 *Given the same effort level in the delivery stage, regarding consulting only bonuses to achieve the same feasible level of quality, Postpartum Outcome Bonus costs less than Complexity Bonus.*

There are also several obstacles in implementing Postpartum Outcome Bonus. First, it is effort intensive in the selection of an appropriate set of postpartum metrics, and then in follow-up and reporting. Our analysis shows that the quality of care resulting from this bonus type is quite sensitive to the list of postpartum complications included in the scope of the bonus. The monetary value of this bonus is a function of the frequency of different complication types. Therefore it can vary significantly with the complications included, as presented in Proposition 3.8. Moreover, there could be a significant time lag between childbirth and potentially experiencing at least one of these postpartum complications. Theoretically, the longer the period after childbirth, the more metrics can be included, which increases the policy's efficacy. However, an extended post-childbirth period adds more difficulties to monitor.

Providing bonuses to both relevant parties reinforces the impact of bonuses on both the prenatal and delivery stages, resulting in reductions in emergency and planned CSs. The *Group* mechanism not only has the same advantage as the *Relevant Party* one, but also involves "peer pressure", which offers a further motivation in addition to financial incentives. Specifically, physicians who do not practice properly are very likely to be pressured by their colleagues because their decision or effort level negatively impacts on their colleagues' incomes, in addition to their own. However, a bonus to the whole group may not be an economical option from the payer's perspective, i.e., the marginal benefit may not be as high as the increased expenses.

Proposition 3.11 *A Group bonus leads to higher expenses for the payer, as compared with a Relevant Party bonus with the same impact.*

This result implies that a *Relevant Party* mechanism is preferable to a *Group* mechanism from a cost-saving perspective. Moreover, a *CS Threshold* add-on based on the delivery mode of the pooled patients for the group of physicians may be problematic from the perspective of quality of care as Proposition 3.12 specifies.

Proposition 3.12 *Suppose that any group of physicians is eligible for the bonus B^{TH} if the overall CS rate for their patients does not exceed r^*, which is associated with a desired cut-off of s^* derived from Lemma 3.3. Let $g(\cdot)$ be the intensity distribution of a certain population in terms of pregnancy complexity. The impact of B^{TH} on the physicians' actual decision s at the prenatal stage can be*

- $s > s^*$ *if high risk population* $\int_0^x g(u)du \leq x$;
- $s < s^*$ *if low risk population* $\int_0^x g(u)du \geq x$.

Proposition 3.12 demonstrates the existence of under- ($s > s^*$)or overtreatment ($s < s^*$) with the *CS Threshold* bonus. After all, it is impossible for the universal threshold rate to work appropriately for all physicians with different patient case-mixes. Indeed, some physicians may have more high-risk patients than others. Physicians with relatively more high-risk patients would have to avoid using clinically necessary CSs in order to achieve the desired CS rate, and hence, avoid financial losses. By contrast, physicians with fewer high-risk patients would enjoy those bonuses but still implement unnecessary CSs. In addition, the demographic characteristics of a population may vary over time, but this static threshold cannot adapt to dynamic demographic shifts. Therefore, this bonus can place quality of care at risk. However, this sort of bonus has been considered the most popular mechanism in recent P4P initiatives in maternity care, due to the fact that it is easier to monitor and record the aggregated results of a group of physicians than the separate records of individuals.

3.8 Proposed Reimbursement Policies

In this section, we propose a two-level payment model for maternity care. First, we present our proposed policy and then discuss its performance when we incorporate the physicians' heterogeneity in the medical decision making process of childbirth.

3.8.1 Proposed Model

We propose a simple reimbursement policy for maternity care: a blended model as base payment and an NB add-on as a complementary incentive, which is effective at improving the quality of maternity care and reducing overall expenses. While proposing a reimbursement policy among several alternatives, we take into account the factors of (i) being easily implementable in practice, (ii) being robust to the different parameters of maternity care, and (ii) still perform good once we incorporate the physicians' heterogeneity in the medical decision making process of childbirth

Definition 3.1 *The proposed reimbursement policy involves a blended rate P^{BP} as a base payment plus an NB bonus rate B^{NB} for physicians who serve the delivery.*

In terms of the recipient of the bonus part, although it could be paid to either consulting or delivery physicians or both of them we propose to be paid to the delivery physician. Although we do not capture it explicitly in our model, the literature suggests that proving a bonus of NB during SB will create "peer pressure", which offers a further motivation in addition to financial incentives. Specifically, physicians who have a tendency to perform planned CSs for the cases that NB will be medically more appropriate are very likely to be pressured by their colleagues because their decision negatively impacts on their colleagues' incomes as well. Therefore, an NB bonus paid to delivery physicians serves as a dual incentive for physicians: it works indirectly toward having them prescribe an SB during the prenatal care, and directly to promote a full effort during the delivery stage. With the effect of peer pressure, this alternative could result in similar level of care quality with less maternity costs (**??**).

As discussed before, bundled model performs best regarding to minimizing the deviation from a clinical cut-off point. By the following lemma, we first show that our proposed payment policy is a special case of bundled payment.

Lemma 3.11 *A bundled payment P^{BL} is equivalent to the combination of a blended payment with a blended rate $(P^{BL} - \frac{c_H^C}{J})$ and a NB bonus of $\frac{c_H^C - c_H^N}{J}$.*

Recall that the bundled payment model also motivates physicians to give their best effort during deliveries occurring in their shifts. The NB bonus works in a similar way to a bundled payment. The blended rate and the NB add-on are equivalent to a linear combination of blended and bundled payments. We show the equivalence when both are combined in a blended payment in Proposition 3.13 below.

Proposition 3.13 *A blended payment with an NB bonus* (P^{BP}, B^{NB}) *is equivalent to the linear combination of a blended payment with a blended rate* $(1-\theta)P^{BP}$ *and a bundled payment with rate* θP^{BL}, $\forall \theta \in [0,1]$, *where* $B^{NB} = \frac{\theta}{J}(c_H^C - c_H^N)$, *and* $P^{BL} = \frac{c_H^C}{J} + P^{BP}$.

The linear combination of blended and bundled payments shows the cost-sharing feature of the optimal reimbursement scheme, where physicians share part of the delivery cost with the payer. When $\theta = 0$, both are pure blended payment schemes. While $\theta = 1$, they both become bundled system. The lower bound of B^{NB} in Proposition 3.8 indicates a lower bound of θ, the minimum effective portion of the delivery cost that a physician should bear in order to motivate a full effort during the delivery stage. Physicians do not bear all the financial risks in this proposed scheme, unlike they do in a bundled payment model; therefore, the payer provides a lower risk premium than in a bundled payment, leading to a lower maternity care cost.

The blended rate acts as a base amount that physicians receive regardless of the delivery mode. This base rate is supposed to be high enough to cover the efforts of the least effort-intensive mode, i.e., a planned CS. Thus, it guarantees that certain planned CSs will be used for high-risk patients. The second bonus encourages NBs at both the prenatal and delivery stages. We further examine the features of this model below, by breaking down the physicians' average rates into different delivery procedures.

$$P^N(\lambda, x) = P^{BP} + B^{NB} f(\lambda, x);$$
$$P^{EC}(x) = P^{PC}(x) = P^{BP}.$$

This implies a modified outcome-dependent FFS mechanism. The CS and NB rates vary with the actual delivery procedures. Clearly, a successful NB for a low-risk patient leads to the highest marginal income, which is different from the traditional static FFS. Moreover, payers can flexibly adjust the overall CS rates by setting up proper rates in Def. 3.1.

Lemma 3.12 *Under the reimbursement scheme expressed in Def. 3.1, the overall CS rate increases as B^{NB} decreases, or as P^{BP} increases. Moreover, $M(\lambda, s, m_D)$ is concave with respect to s.*

The following proposition states that a global optimal solution exists to Problem Z_P with the specific rates outlined in Definition 3.1.

Proposition 3.14 *There exists at least one global optimal solution to the optimization problem Z_P under the proposed policy in Definition 3.1.*

The desired maternity care outcome therefore exists under this two-level payment mechanism. Moreover, the value maximization solution in Proposition 3.2 is achievable, though the total expenses would be higher than those in the benchmark.

Corollary 3.5 *The value maximization solution to Problem Z_{BM} is a subset of feasible solutions to the optimization problem Z_P under the payment scheme in Def. 3.1.*

We show our numerical analysis on the optimal threshold and the associated expenses in Section 3.9. The robustness of this scheme is examined and verified through a sensitivity analysis over various parameters including group size, altruism level of physicians, clinical threshold of x^* and effort levels for alternative delivery modes, presented in EC5 in Electronic Company.

3.8.2 Incorporating Physician Heterogeneity

This section relaxes the typical assumption of physician homogeneity in the basic model and study the proposed reimbursement policy in physician heterogeneity con-

text. In reality, physicians tend to have different patient mixes (i.e., different distributions of patient complexity), and they themselves vary according to their preferences, experience and skills. For example, O'Neill and Kuder (2005) finds that physicians' personal characteristics, practice settings and patient populations contribute to variations in the likelihood of prescribing a service in three specified clinical scenarios. Feinstein et al. (2013) studies the impacts of patient and physician factors, apart from regional variations, on the utility of radiation therapy with a retrospective cohort design. Nevertheless, the features of optimal reimbursement mechanisms may not be mitigated by physician heterogeneity. Moreover, when properly designed, reimbursement mechanisms are able to motivate physicians to enhance their professional skills, in addition to achieving the main goal of reducing unnecessary CSs.

Heterogeneous Patient Mix

First we study physicians with a heterogeneous patient mix. Typically, we consider a group of two physicians: one with riskier patients and the other with fewer high-risk patients.

Proposition 3.15 *The proposed reimbursement mechanism is independent of different complexity distributions. Specifically, each physician's total income is independent of his patient mix.*

This proposition implies another advantage of the proposed policy. It creates a mechanism for physicians to share patients - to "exchange" patients between them - to support NB at the actual delivery stage. Physicians may therefore be indifferent to the possibility of having a different patient mix than their colleagues.

Heterogeneous Diagnosis Skills

The study by Ghaffarzadegan et al. (2013) finds that physicians who have been practicing longer are more likely to decide on an arranged CS, based on their system dynamics simulation model of physicians focusing on experiential learning. Suppose a given original existing reimbursement mechanism induces physicians to set up an optimal threshold as s_0. Denote the actual pregnancy complexity as x. Then, physicians with higher-qualified diagnosis skills may have a x_d very close to x. However,

diagnosis skills make a difference only if x_d and x fall onto different sides of s_0. More specifically, if $x < s_0$ and the woman should be prescribed an SB, but a physician diagnoses the pregnant woman and assesses a $x_d \geq s_0$, and accordingly prescribes an arranged CS, then the patient will suffer more damage than benefit, impairing the quality of care; and eventually, the physician would suffer a loss of his total utility. However, so long as the physician can diagnose a $x_d < s_0$ and prescribe an SB, the diagnosis can be considered proper. Therefore, we model the physicians' diagnosis skills as the probability of x_d and x falling onto the same side of s_0. In other words, the chance of correctly prescribing an arranged CS CS for higher-risk pregnant women H with $x \geq s_0$, due to the diagnosed $x_d \geq s_0$ is assumed the same as the probability of prescribing an SB SB for low-risk patients L with $x < s_0$ and $x_d < s_0$. Suppose that the probability of a correct diagnosis is a, following the framework of Allard et al. (2011).

$$\Pr(CS|H) = a, \quad \Pr(CS|L) = 1 - a$$
$$\Pr(SB|L) = a, \quad \Pr(SB|H) = 1 - a$$

Assume that physicians differ only in terms of diagnosis skills, and that the other facets remain the same for all physicians. The following proposition shows the relationship between diagnosis skills and the likelihood of an improper decision.

Proposition 3.16 *Physicians may lose* $\Pr(L|CS)$ *of their incomes from decision of a planned CS, and lose* $\Pr(H|SB)$ *of their income from a decision of an SB. The losses are non-increasing with respect to* a*.*

Proposition 3.16 indicates that there is less chance of a loss of utility if diagnosis skills have been enhanced or a becomes larger. Therefore, this proposed reimbursement mechanism contributes to motivating physicians to improve their diagnosis skills since this can lead to a larger total utility.

Heterogeneous Procedural Preferences

Physicians may have different preferences or treatment styles (Epstein and Nicholson, 2009); for instance, some may be more confident with a natural birth, while others

may be better at CS surgery. Goyert et al. (1989) finds that physicians' practice styles are more likely to contribute to large variations in the CS rate than other physician factors like medical and legal experience.

We would like to depict these types of preferences or procedural skills by adding a preference factor $pf > 0$, such that the physicians who prefer CSs consider their CS effort as $e^{EC} - pf$. On the other hand, those with a preference for natural births will view the effort of natural birth as $e^{N} - pf$. We study the impact of heterogenous procedural preferences on proposed scheme through the change of physicians' effort in the sensitivity analysis, and show that our proposed scheme is relatively the most robust regarding this issue; although financial incentives tend to be very weak to impact physicians' preferences.

3.9 Numerical Analysis

We verify our major analytical results, undertake a comparative study of the base and complementary payment schemes and assess the performance of the proposed policy on the same data set described in Section 3. Our methodology to develop a quantitative metric for measuring the pregnancy complexity and identifying a threshold x^* is described in Section 3. For our numerical analyses, we estimate the probability of having a NB under SB decisions for given x as well as the probability of having a postpartum complication for given x and the delivery mode by analyzing the same data set. The related cost figures are calculated by using detailed published reports on the cost of childbirth. Further details on the parameter estimation methods are provided in Electronic Companion. The main results of our numerical analyses are given in in Table 3.10 and 3.11.

In this section, we highlight the major findings of the numerical study. Table 3.10 and Table 3.11 illustrate the resulting CS rates and the average cost per delivery under different reimbursement mechanisms.We assess the performance of an incentive mechanism in terms of three factors: (i) deviation between clinical cut-off point and s, (ii) overall CS rate r and (iii) the expected cost.

Table 3.10: Compare Optimal Rates

Policy	Cost			Quality		
	Threshold s	$\Delta\Pi$ (%)*	r (%)	Threshold s	$\Delta\Pi$ (%)*	r (%)
Benchmark	0.98	-	21.87	0.85	-	24.91
FFS	0.72	4.87	34.22	0.72	3.99	34.22
Blended	0.76	3.78	31.01	0.76	2.90	31.01
Bundled	0.99	1.97	19.89	0.85	3.57	23.51
Bundled*	0.83	2.96	25.90	0.85	3.75	24.91
Proposed	0.83	2.07	25.90	0.85	2.18	24.91

Notes
1. $\Delta\Pi$ (%) are the percentage change from benchmark;
2. Bundled* refers to the bundled payment with different rates for low and high risk patients.

Observation 1: Under FFS, the ideal clinical cut-off point is not achievable and the average birth-related costs are approximately 5% higher than the benchmark. We estimate the average overall CS rate under this payment mechanism as 34%, which is really close the current CS rates in US. The blended model provides certain improvements in average cost and CS rate over FFS i.e., 1.09% and 9.40% respectively. However, we also numerically confirm that Corollary 3.2 holds both for FFS and blended models, such that the set of feasible solutions for s under these models is outside theare region defined by x^* and s_E. Therefore, although blended model might offer some improvements in the system, it would be still suboptimal with a chance that increasing cost would not necessarily improve the quality of care.

Observation 2: On the other hand, under the bundled system, the average costs can be reduced by 3%, a significant improvement over FFS under cost minimization objective. However, the deviation between the threshold for medically appropriate planned CS and the physicians' threshold s is pronounced, indicating the tendency of physicians to under-treat patients under bundled payment. We numerically verify Corollary 3.4 that increasing bundled rate can motivate physicians to set up the quality-maximizing threshold x^*; however in this case, the estimated improvement in the average cost is only 0.4% over FFS. Our experiments show that although it

Table 3.11: Compare Different Bonus Mechanisms

Policy	Cost			Quality		
	Threshold s	$\Delta\Pi$ (%)*	r (%)	Threshold s	$\Delta\Pi$ (%)*	r (%)
Complexity Bonus						
Consulting Only	0.84	2.74	25.27	0.85	1.93	24.91
NB Bonus						
Consulting Only	0.82	2.85	26.57	0.85	3.30	24.91
Delivery Only	0.83	2.07	25.90	0.85	2.18	24.91
Relevant Party	0.82	3.63	26.57	0.85	7.54	24.91
Group	0.76	3.78	31.01	0.85	20.91	24.91
Postpartum Bonus						
Consulting Only	0.76	3.78	31.01	0.77	4.22	30.23
Delivery Only	0.91	0.30	23.13	0.85	0.49	24.91
Relevant Party	0.83	3.50	25.90	0.85	2.67	24.91
Group	0.76	3.78	31.01	0.85	13.95	24.91
CS Threshold Bonus						
Group	0.85	13.82	24.91	0.85	12.86	24.91
High risk population	0.87	13.77	24.22	0.85	13.19	24.91
Low risk population	0.83	14.20	25.90	0.83	13.24	25.90
Complexity Bonus for Consulting + NB Bonus for delivery only						
Combined	0.83	2.07	25.90	0.85	1.93	24.91

$\Delta\Pi$ (%) are the percentage change from the corresponding cost-minimum or quality maximum benchmark. All bonuses are complimentary to blended payment. For CS Threshold Bonus, the threshold is 25 % overall CS rate.

is possible to further improve quality of care by offering different bundled rates for low and high risk patients, this offsets 1.9% of the cost advantages of using a single bundled rate. We also observe that smaller physician groups are more likely to under treat patients, whereas larger physician groups tend to be less sensitive about the resource utilization.

Observation 3: Complementary payments are quite effective to offset some of the disadvantages of the base payment models, if properly designed. For almost all alternative combinations (i.e. bonus type plus recipient) of complimentary payments, the feasible solution for s includes at least one of the x^* and s_E values. That is a preferred solution as discussed in Corollary 3.2. The only exceptions are postpartum bonus for consulting physicians and group threshold bonus for low risk population.

Observation 4: Regarding the complimentary payments offered for the consulting physicians, a Complexity add-on is a more effective way to motivate physicians to make the proper decisions compared to other alternatives, with estimated improvements of 2.1% and 26% for average cost and CS rates respectively, over FFS. It is followed by the NB bonus, with 2% decrease in average cost and 22% decrease in CS rates when compared to those for FFS. Under care quality maximization objective, Complexity and NB bonuses reported the same planned and overall CS rates, where the former imposes the same impact with less average costs, approximately 1.4% less, which confirms our analytical finding presented in Proposition 3.9.

Observation 5: Our results suggest that Postpartum Outcome Bonus offered to consulting physicians is not really efficient in providing necessary incentives. This complementary payment model, however, is quite effective if it is offered to delivery physicians. It deviates only 0.3% and 0.49% from the cost minimization and the quality of care maximization objectives of benchmark problem, respectively. On the other hand, NB bonus is more effective if it is offered to delivery physician compared to being offered to consulting physicians (at least the same s and r with lower cost).

Observation 6: For the add-ons that can be paid to either of the relevant parties or offered as group bonus, the former outperforms the latter. The numerical experiments confirm our analytical findings that complementary payments offered to all physicians

in the group is less cost-effective compared to other alternative recipients.

Observation 7: In parallel to our findings discussed above, CS Threshold policy is quite expensive compared to other alternatives. For instance, the average maternity cost under Π^Q with CS Threshold bonus is 8% higher than that with NB bonus offered to delivery physician, whereas they result in the same s and r.

Observation 8: Our recommended policy proposes 3% reduction in average birth related costs and 27% decrease in overall CS rate compared to those under the FFS system. Please note that regarding the performance of alternative complimentary payment model, under both objective functions, Postpartum Outcome Bonus given to delivery physicians performs best based on the deviation from the benchmark for the related objective. However, among all these incentive models with similar impacts on overall CS rates, our recommended policy is the most robust with regard to the important parameters of the maternity care including the group size, the physicians' altruism levels, a possibly varied clinical cut-off point, and the physicians' heterogeneous procedural preferences. Moreover, the impact of Postpartum Outcome as well as Complexity bonuses depend on the intensity distribution of a certain population in terms of pregnancy complexities. However, it is not the case for our proposed policy. Detailed graphs and results of the sensitivity analysis are presented in the Appendix.

3.10 Limitations and Conclusion

This work focuses on the design of financial incentives in order to reduce unnecessary C-sections. Through our modeling framework, we first analyze different base payment mechanisms, and then alternative complementary incentives comprehensively. In the context of basic payment mechanisms, both FFS and blended payment schemes lead to increased CS rates. Although blended model does not give direct economic benefits to perform CS, it fails to provide incentives to physicians in order to give full effort while monitoring the prolonged labor or eliminate the tendency to deliver their own physicians. While bundled system provides the best solution for minimizing the maternity care cost from payer's perspective, in this model the physicians face a

high financial risk, which leads to under-treatment and patient selection. Likewise, assuring certain level of care quality might require a high-risk premium for physicians to take on all associated financial risks. Among alternative add-ons, we show that the Group Bonus mechanism, and therefore the CS Threshold Bonus, is quite costly compared to the other options. Although a Complexity Bonus seems to be the best alternative for avoiding unnecessary planned CSs, it fails to motivate best practices during the delivery stage. The bonuses for Postpartum Outcome and NB have similar impacts on CS rates; however, the former has major drawbacks from an implementation perspective, including, first, selection of the proper set of post-delivery complications and then monitoring.

As a conclusion of our analyses, we propose a two-level payment scheme for maternity care. This policy involves a blended base payment and a bonus for NB. This typical contract inherits the feature of risk-sharing, or cost-sharing, from the traditional pricing contracts of supply chains. This proposed mechanism succeeds in aligning the physicians' priority of maximizing utility with the payer's value maximization objectives. Moreover, the proposed bonus linked with the incidence of successful NBs, contributes to the coordination among physicians in the same group. With the potential to motivate peer oversight, this policy tends to incentivize proper birth plan decisions and best practices at the delivery stage. Furthermore, it does not require any advanced information collection and monitoring, therefore it is really practical to implement and the administrative cost of our proposed measurement is expected to be low (Cachon and Lariviere., 2001).

Our study has several limitations. First, a number of existing empirical works have found that hospitals' guidelines and capacity issues might have an impact on the abuse of CSs (Smith et al., 1992; Font, 2009; Brick and Layte, 2011). Our study focuses on the decision making in childbirth from the physician's perspective; however, extending our model by incorporating hospital-physician interactions could provide important managerial insights as well. Second, our model is in a static setting. The decision-maker decides on a financial mechanism and then the physicians determine the delivery procedures in the same period. Though we do consider the impacts of a

well-designed reimbursement policy on constantly motivating enhancement of physicians' professional skills, as a reaction from contract takers, it would be valuable to find a way to study the physicians' dynamic reaction. Finally, for reasons of simplicity, our work makes the assumption of passive patients, i.e., who comply perfectly with the physicians' decisions. However, this might not be the case in reality. Patients tend to have various levels of reaction to their own treatment. They can shift to alternative care providers or follow their own preferences in choosing hospitals or physicians (e.g. Fabbri and Monfardini (2008)). One of the possible extensions of this study could be studying the interactions between patients and physicians.

Chapter 4

Design of Specialist Responsible Policies to Reduce Waiting Times in Emergency Departments

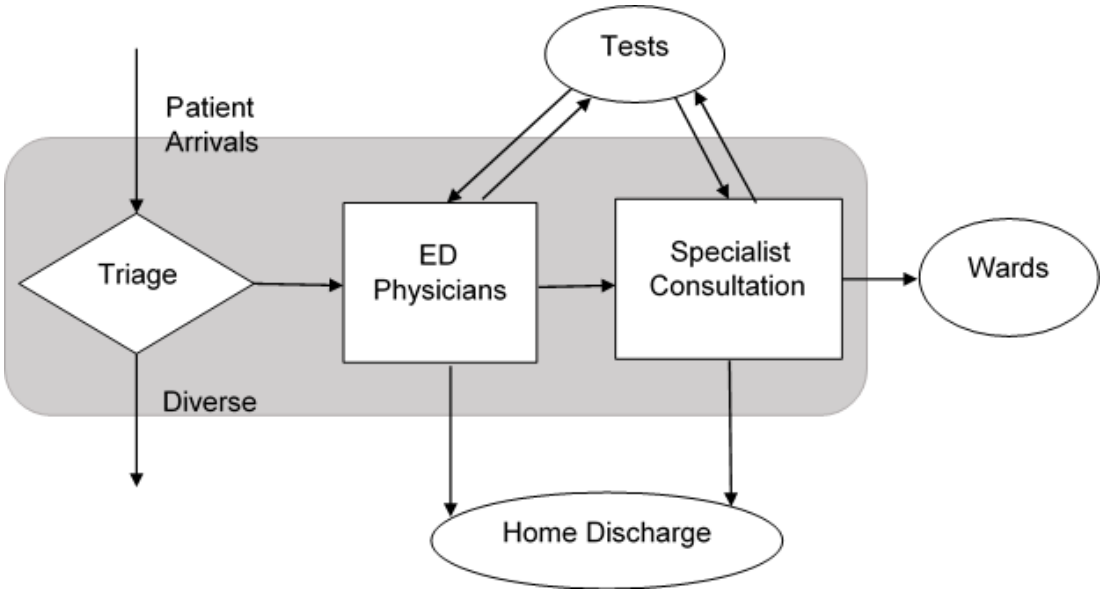
4.1 Introduction

Emergency department (ED) overcrowding is a widely used term referring to a situation where the demand for ED services exceeds the ability to provide care in a reasonable amount of time (Ospina et al., 2006). ED overcrowding has been a key issue in Quebec for more than 40 years. Despite increased political, administrative, and public awareness, ED overcrowding continues to rise in frequency and severity (Bond et al., 2007; Roberge et al., 2010). International comparative studies have found that the Quebec population had not only the highest rate of ED visits but also the longest waiting times to receive care in ED (Roberge et al., 2010). In Quebec, despite the established targeted ED average length of stay (LOS) being 12 hours, the average stay for stretcher patients reached 17.6 hours in 2011. Moreover, around 25% of those patients have had to stay more than 24 hours, exceeding the 10% target set by the ministry (MSSS, 2011). In addition, the LOS has been more than 48 hours for up to 10% of the stretcher patients (MSSS, 2011, 2010). Furthermore, a substantial body of the literature has linked the increased ED overcrowding, and ED LOS accordingly, with adverse patient outcomes (Sun et al., 2013; Carter et al., 2014). For instance, increased stretcher occupancy is associated with increased incidence of 30-day adverse patient outcomes (i.e. mortality and a return ED visit with hospitalization)(McCusker et al., 2014).

ED overcrowding is a complex, multi-dimensional health services problem, whose root causes extend beyond the walls of EDs. Using the well-established paradigm of Operations Management, this problem has been conceptualized using the input-throughput-output model (Schull et al., 2002; Asplin et al., 2003). Input factors reflect to any condition or characteristic that contributes to the demand for ED services, such as non-urgent visits and "frequent flyers", which, in general, refer to patients who have 4 or more annual visits to ED (Holt and Aronsky, 2008). In the context of throughput, there are two phases. The first phase focuses on ED care processes including triage, stretcher placement, and ED physician evaluation. The second phase includes use of hospital resources: diagnostic testing and specialist consultation. Output factors

reflect to efficient disposition of admitted and discharged patients out of ED. Figure 4-1 shows the typical patient flows in ED.

Figure 4-1: Patient Flow in ED



Contrary to popular perceptions and media attention, which have highlighted input factors such as inappropriate use of the EDs by high numbers of lower acuity patients, the vast majority of the delays occur in the second phase of the throughput (i.e. lab testing, diagnostic imaging, specialist consultation), as well as the output side of patient flow (i.e. admitting to hospital, discharge to home). Therefore these are the most significant factors causing ED overcrowding, and consequently longer ED LOS (Affleck et al., 2013; Canadian Institute for Health Information, 2014). Among all these "second phase of the throughput and output" related factors delays for specialist consultation (SC), i.e. the time between sending out an SC request and the arrival of the specialist, plays a key role.

Approximately 20% ED visits requires at least one SC. Moreover only specialists, not the ED physician, have the authority to admit patients into hospital wards. Moreover, for patients who need an SC, the discharge decision also has to be consulted to the specialist. In other words, the patients cannot be discharged home or admitted to hospital wards before seeing a specialist. Furthermore, a significant portion of

the blood and imaging tests are asked by the specialist. Thus, the specialists have a fundamental position in ED processes since they have direct influence on imaging, lab, discharge as well as admission delays.

In a descriptive study, Lee and her colleagues (Lee et al., 2013) show that consultation process time, including waiting for the consultation, is highly variable even in the same institution, and has an important impact on ED LOS. Our empirical analysis on all ED visits to one of the medium sized community hospitals in Montreal in one calendar year also shows an average of seven hours waiting for SC, from sending out a consultation request to arrival of the specialist; and this delay can be over 2 days. It is mainly because the specialists are generally busy with patients in hospital wards, walk-in clinics and operating rooms during weekdays and may not be on-call after business hours and over weekends. This contributes to longer LOS, and an overcrowding, accordingly, in EDs. Unfortunately, as far as we know there have been no systematic and practical rules for specialists to follow in ED. Our empirical study also shows that specialists can arrive at the ED at anytime, although they arrive more frequently in business hours. Motivated by this prolonged SC delays in ED, our study aims to reduce the average LOS in ED by designing optimal schemes for specialists' response to SC requests. We would like to address the following research questions:

1. What are the characters of potential rules to regulate specialists' response to ED requests?

2. Which rule is optimal for a certain specialist to follow?

3. How can those optimal rules for SC be best integrated into current triage in ED?

As an inevitable and critical part of ED flow streaming, specialist scheduling can benefit significantly from coordination with other processes in ED. To be more specific, patient prioritization based on the joint consideration of critical conditions and potential resources requirements (e.g. specialist, lab, etc.) can improve the overall

performance of EDs significantly. For example, under such a coordinated system, a patient with higher chance of SC will have a higher priority over a patient with same level of critical conditions, i.e. triage code, but with much lower chance of SC. Accordingly, the former type of patients will have access to ED physician assessment earlier, so the SC request for these patients will be sent sooner , resulting in reduced SC waiting time. Hence, through the policies we propose in this study, the delay for SC will be shorter. Thus patients will have a much shorter LOS, which will alleviate the overcrowding significantly.

Recently, a resource-based triage Emergency Severity Index (ESI) has been proposed by Gilboy et al. (2011), which recommends that non-crucial or life-threatening patients should be prioritized with their triage codes, as well as the predicted resource requirements of these patients in ED. As "resource" they refer to tests, SCs and hospital beds.

By analyzing the same data set mentioned above, we demonstrate that a patient's probability of requesting an SC can be predicted at the triage with high accuracy. ED triage can be more accurate and effective by considering both the patient's medical conditions and potential demand of SC. Thus, revised patient prioritization policy in triage, which incorporates the probability of SC request, can lead to improvements in ED overcrowding compared to current triage policies. In this study, by designing such a modified triage policy involving the prediction of a patient's SC request, we propose to streamline the ED patient flow from triage to SC. Our aim is to facilitate both specialists' schedules and ED administrators' management of patient flow with a systematically optimal strategy.

Although scheduling for ED operations, i.e. scheduling of ED physicians and nurses, has been studied extensively in literature, none of them has considered the impact of SC delay on LOS in ED. Most studies in healthcare literature assume that the patients get consultation without any delay. Chan et al. (2016) consider a similar problem in the setting of patient's discharge from hospital wards. Although patient discharge is normally delayed by the physician's inspection time and frequency, they focus on the optimal inspection frequency during a day. However, there are significant

differences between this paper and ours in the context of the problem setting. For instance, our empirical study shows that patients have requested over ten different types of SCs in ED, and the delay of specialist arrivals has a larger scale of impacts in ED context. It is due to the fact that the LOS in ED is normally measured by hours, whereas LOS in hospital wards by days. In this chapter, we consider the impact of different policies for specialists' arrival according to the patterns and volume of each SC demand, and examine the potential reduction in expected LOS when implementing resource-based triage in the end.

In order to capture the dynamics of ED patient arrivals, we study a time-varying queueing model, unlike most of literature that considers 3 types of regimes (namely overload or heavy traffic, critically loaded and under-loaded regime) separately, due to the fact that each regime features distinctive methods. Actually, according to the data set of ED visit records we analyze, patient flows in ED experience all three regimes during a typical day. The detailed analysis of our proposed model provides several guidelines of setting up optimal policies for specialist's response to SC request based on the volume and patterns of ED patients. Besides, we consider uncertain service time as well as the inaccuracies in prioritization at the triage in our modeling framework (Li and Glazebrook, 2010). Therefore, our analytical model incorporates inaccurate estimation of classification at triage where forecast of SC request with incomplete information (signals) is not perfectly precise (Argon and Ziya, 2009).

This chapter consists of three main parts. The first part focuses on analyzing alternative policies for specialist's arrival to ED for SCs, i.e. fixed time (FT) and timeline (TL) policies, and studying the proper application of each policy for varied SC demand, through queueing models with time-dependent arrivals. The second part focuses on the integration of the modified resource-based triage with the optimal specialists' arrival strategies. Then we use an empirical model for predicting the probability of each patient's consultation requests, according to patients' information collected at triage, through statistical learning methods. Finally, based on the forecast of consultation requests for each patient, we conduct a comprehensive simulation model to evaluate potential scenarios of optimal specialist arrival policies with and

without the modified triage policy.

4.2 Literature Review

In this study, as discussed before we consider a queueing model with time-varying arrival rates. Current literature tends to focus on different regimes of a time-varying queueing model, and develop typical methods to deal with each regime. Overload or heavy traffic regime, where customers congesting the queue wait to be served, has been prevalent in literature; since it is the most suitable approach to model systems with overcrowding. It is also the most difficult to solve. Fluid model is applied for this regime, and (generalized) $Gc\mu$ rule is proposed as an optimal, i.e. prioritize the class with the largest holding cost and service rate studied by (Huang et al., 2015), where this rule can incorporate the arrival rate and abandonment rate (Atar et al., 2010). Under due-date constraints, $Gc\mu$ is equivalent to prioritize generalized longest queue (GLQ) and generalized largest delay (GLD) rules (Van Mieghem, 2003). Critically load regime refers to a queueing system with moderate amounts of customers, and servers are occupied most of the time. Diffusion modeling approach or dynamic control is applied for this regime (Down et al., 2011). Under-loaded regime, where a queueing system has overstaffed servers, is very rare in reality. However it is important to balance the issue of server idleness and cost reduction. Dynamic control is applied (Down et al., 2011). In the setting of ED, eliminating the time-variation or focusing on certain regime may be unable to capture the time-varying performance, because all above regimes exist, link and impact each other, which leads to the failure of steady-state distribution in a time-homogeneous queue.

Our study falls into the field of patients' streaming and prioritization in ED. Traditional triage determines the prioritization of patients according to their medical conditions from clinical perspective. Those patients who are considered in critical or life-threatening conditions cannot wait long in ED, so they are treated before those who are in less critical conditions. Literature in operations research and management science tends to tackle this problem from the perspective of efficiency. Assuming the

homogeneous clinical conditions for all patients, index policies aim to minimize the average waiting time by prioritizing patients in the longest queue (Van Mieghem, 2003; Atar et al., 2010; Huang et al., 2015). Besides, classifying of patients and scheduling different groups are popular among literature. For example, Hu and Benjaafar (2009) showed that partitioning can be significantly beneficial to the queue system via approximation with fluid model and simulation. However, this work demonstrated that the benefit is realized at the expense of other customer classes, that is, it is impossible to have improvement for all customer classes in such a system. Joustra et al. (2009) examined whether or not pool urgent and regular patients waiting for consultation in the context of a radiotherapy outpatient department. They used queueing theory and discrete event simulation, and concluded that pooling does not always provide benefit to urgent patients. They also found that separation of those queues could reduce the capacity requirement while meeting the waiting time criteria for all patients.

More recently, it is proposed that A/D streaming is another way to reduce LOS in ED (Saghafian et al., 2012, 2014). A/D refers to a system where ED patients and resources are divided into two streams: one for those who are likely to be discharge home (D) and the other for those who are likely to be admitted to hospital (A). In their papers, Saghafian and his colleagues compared A/D streaming with pooling and incorporated sequence with feedback (i.e. prioritize new patients or old ones). They also proposed a virtual streaming, that is, switching the resources of one type to the other if they are idle. Omar and Okundan Kremer (2016) introduced a new dynamic patient grouping and prioritization algorithm based on patients' dissimilarity that are resulted from detailed triage raw information (age, gender, pain level, vital signs, temperature etc). In a general setting, Afeche (2013) differentiated among customer types, and implemented a strategic allocation based on revenue. (Baron et al., 2014) studied a set of threshold-based policies that strategically idle first station in a tandem queue. However, all those works are based on queueing system with time homogenous arrivals. Our work considers time varying arrivals and heterogeneous patients with different types of specialist requirements, thus we figure out the best specialist response policy for each type of specialist, and then test their scheduling

among all ED patients with their triage information.

Time-dependent queueing models have been studied in data fitting, staffing and capacity management policies. Interested readers can refer to Whitt (2016) for a detailed bibliography on existing work of queue systems with time-varying arrival rates. Chan et al. (2016) considers the frequency of inspection in hospital wards and its impact on the number of customers waiting in the system. They focused on the number of customer in system, the probability of waiting under time-varying arrival rates impacted by inspection time in hospital wards. Though expected waiting time is not as critical as the former two performance measures in their study, they found numerically that the careful choice of one inspection time per day depends on the magnitude of arrival rate variation. Similarly focusing on the discharge delay in hospital wards, another recent work of Dai and Shi (2017) studied a time-varying queue system with periodic Poisson arrival process. The processing time consists of two components: 1) length of stay; 2) departure time, referring to the discharge hour on the discharge day. They developed a novel midnight customer count process and further analyze its stationary distribution in order to approximate time-dependent customer count process and calculate multiple performance measures. They proposed to advance the discharge time to alleviate the overcrowding of peak arrivals. Our work is different from those studies. We focus in the setting of ED. In the queueing model with time-varying arrival rates with a daily cycle, we analytically prove the optimal fixed arrival time if specialists come to ED once a day, leading to the minimal average per-patient waiting time for SC delay.

This study is also relevant with the literature of batch scheduling, to be more specific with the integrated scheduling models of production and transportation. Interested readers can refer to Chen (2010) for models explicitly considering both production and distribution time or cost. Our work is specifically relevant with those deterministic scheduling problems, in which a series of delivery dates are fixed before those jobs are processed. Hall et al. (2001) provided an efficient algorithm for such models, and showed that the algorithm may not work for certain types of problems. Cheng and Kovalyov (2001) considered the batch scheduling of jobs with fixed due

dates or processing times. They presented dynamic programming algorithms to minimize lateness, the number of late jobs, the total delays and so on for both bounded and unbounded batches. They also developed more efficient algorithms for several special cases. In order to schedule a series of non-preemptive jobs with varied delivery dates on a single machine and a non-stepwise payoff function based on cumulative number of jobs processed before each job-independent delivery date,Seddik et al. (2013) found the complexity of this problem and provided a pseudo-polynomial time algorithm for the problem with two delivery dates based on dynamic programming. Seddik et al. (2015) further proposed a polynomial time approximation algorithm to meet both absolute and relative performance guarantees for this problem. Although the work of Janiak and Krysiak (2007) did not explicitly consider fixed due dates, the value of a job follows a stepwise non-increasing function in their model. Hence, the scheduling has a big impact on the total values of all jobs completed. They proved that such a problem could be equivalent to the NP-hard problem of minimizing weighted number of late jobs. They further designed a dynamic programming based pseudo-polynomial algorithm for jobs with common moments of value changes, and several heuristic algorithms to solve specially extended cases. Several papers have studied the typical scheduling problems with fixed-interval due dates. Chhajed (1995) considered jobs assigned to two due-dates with constant intervals. They found that the problem of minimizing a linear due-date penalty is NP-hard. Lee and Li (1996) developed a pseudo-polynomial dynamic programming algorithm for such a problem with a bounded amount of due-dates. Liu and Hsu (2015) analyzes three types of dispatching rules in a system with fixed interval delivery dates based on simulation. The finished jobs can be only delivered on the earliest delivery date, incurring both earliness and due-date costs for the producer. The study proposes a simple and feasible dispatching policy without parameter estimation to minimize the total earliness and due-date cost for this dynamic system. The Fixed Time (FT) policy, one of the proposed specialist policies, can possibly be modeled as a scheduling problem with fixed due-dates. However, our problem is more complicated, due to the fact that different specialist policies mix in ED. Hence we need to consider the

more comprehensive scheduling problem with both fixed due-dates and other types of constrains.

4.3 Optimal Specialist Response Policy

Current ED triage has become a mature system , due to the fact that it provides a rule for prioritization of patients and a limit for the acceptable delay between triage and ED physician assessment for each triage code. In contrast, there are no specific policies or rules for specialists to respond to SC . As a result, specialists themselves decide whether and when to provide consultation to ED patients upon receiving a request. Currently hospital administrations are considering to set up certain polices for specialists in order to reduce extended SC delays. Assuming specialists will comply with the rules, this section focuses on alternative policies for specialists from the perspective of efficiency and feasibility. Although financial incentives meant to motivate specialists to comply with these policies are an important component for the implementation of such policies, they are out of scope of this study. We would like to propose the optimal rules that are most convenient for specialists to follow, specifically the rules that require the least frequency of specialist visits and the maximal certainty, facilitating their scheduling on other tasks.

We consider the potential specialist response policies, and their impact on the LOS in ED. Each type of specialist response policies are explained as below:

Benchmark. Specialists arrive within 2 hours after request, 24/7. This is the ideal scenario, yet is impossible to launch unless the ED has a very strong bargaining power with its high volume of patients. Moreover, with the large scale of consultation demands, ED may have the capacity to hire its own specialists who are present all the time.

Fixed Time (FT). Specialists arrive at certain time every day. Specialists may have to make multiple short visits every day, depending on the corresponding demand flows. Multiple FT policies are preferred if the resulting performance

measure is similar with those under TL policies, because FT policies are certain, more feasible and convenient for specialists. -The specialists are aware of their visit times and of their expected length of visit in advance, so they are able to schedule other tasks accordingly.

Timeline (TL). Specialists arrive within 4 hours after request during certain periods,and arrive within 6 hours at rest periods. TL policies are preferred if their performance measures significantly outperform FT policies, for example frequency of visits are low, though uncertain.

For proposing the optimal specialist response policy, we compare the alternative policies by using the performance measure of average waiting time per patient.

4.3.1 Performance Measures

In this subsection, we first set up the model of SC demand arrivals, and then describe the performance measure within the modeling framework. We consider the flow of SC requests for a certain specialist. Let T_j be the arrival time of the jth patient who requires an SC. $N(t)$ denotes the number of arrivals in $(0, t], \forall t > 0$. $\lambda(t)$ is the arrival rate at time t. The properties of the time-dependent arrivals are highlighted below.

Property 4.1 (Time-dependent arrivals) *In an overtake-free system (FIFO) with no group/batch arrivals, time-dependent arrivals have the following properties:*

- $T_{i+1} - T_i$ *is independent of* $T_{j+1} - T_J, \forall j < i$ *given* T_i. *i.e. old arrivals have no impact on future arrivals;*
- $N(t)$ *is a renewal function dependent on* t*;*
- $N(t)$ *is right-continuous, differentiable;*
- $\lim_{\Delta t \to 0} \mathbb{E}[N(t) - N(t - \Delta t)]$ *exists, and is equal to* $\lambda(t)$*;*
- $\lambda(t)$ *is the derivative of* $N(t)$*;*
- $\lim_{\Delta t \to 0} \lambda(t)\Delta t$ *is the probability of one arrival in* $(t - \Delta t, t]$.

Let $L(t)$ be the number of patients waiting in the system, $S_j(t)$ jth patient's system time, and $W(t)$ total waiting time at time t of all patients arriving by t.

Note. In contrast to $M/G/c$ system, $L(t)$ and $W(t)$ depend on the initial conditions at time 0; and the mean value in a stationary system does not work in the time-varying case.

We consider the transient laws of $L(t)$ and $W(t)$ in the time-varying case, since the stationary system is not representative. The policy with the minimal average waiting time for a certain arrival pattern is the optimal for that specific cluster of patients.

$$\bar{W}(t) = \frac{W(t)}{L(t)}, \tag{4.1}$$

where $L(t)$ is the number of patients waiting in the system, and $W(t)$ total waiting time at time t of all patients arriving by t.

By the end of time t, $L(t)$ has the exact formula according to Bertsimas and Mourtzinou (1997).

Lemma 4.1 (Total Queue Length) *The expected total queue length (i.e. the number of patients waiting in the queue) until epoch t is*

$$\mathbb{E}L(t) = \int_0^t \lambda(\tau)d\tau. \tag{4.2}$$

4.3.2 FT Specialist Response Policy

In this subsection, we aim to find the optimal time for a specialist to arrive under FT policy, so that the average per person waiting time is minimized. Given the time dependent arrival of specialists' demand, the timing of a specialist's arrival can have a significant impact on patients' waiting time as shown in the left of Fig. 4-2. The red areas represent the sum of total waiting time for the same amount of patients who arrive in the two time windows of an equal length. Specifically, the vertical axis is the total number of patients' arrivals, and the horizontal line is the time. For the same length of two cyclical time patterns, the total amount of arrivals are the same

in both cycles. However, due to the different start time, the areas showing the total waiting time are different; the latter cycle (the right upper area) has a longer average waiting time per patient than the earlier one (the right below area). Therefore, the time of the specialists' arrival , that is exactly the start time of an SC session under the FT policy, determines the average waiting time per person in ED.

Consider the pure jump process $N(t)$, and $\{\mathcal{F}(t), t \geq 0\}$ is a filtration to which N is adapted. Let $M(t)$ be the compensated poisson process, specifically

$$M(t) \triangleq N(t) - \int_0^t \lambda(\tau)\mathrm{d}\tau. \tag{4.3}$$

Because $M(t)$ is a $\mathcal{F}$ martingale (Watanabe, 1964), the following is also a $\mathcal{F}$ martingale (Bremaud, 1981).

$$I(t) = \int_0^t H(\tau)\mathrm{d}M(\tau), \tag{4.4}$$

where $H(\tau)$ is any stochastic process depending on the past information, that is, it is a left continuous function. Hence, we have the following theorem regarding the expected total waiting time.

Theorem 4.1 (Total Waiting Time) *The expected total waiting time until epoch t is*

$$\mathbb{E}W(t) = \int_0^t \lambda(\tau)(t-\tau)\,d\tau \tag{4.5}$$

With the help of the martingale property, we first figure out the best timing when a specialist finishes his or her consultation session in ED.

Corollary 4.1 (Daily Optimal Time of Specialist Departure) *If hourly arrival rate $\lambda(t)$ is periodic with a cycle of 24 hours, that is, $\lambda(t) = \lambda(t+24)$, the optimal*

time of specialist departure (completeness of all service) T satisfies

$$24\lambda(T) = \int_0^{24} \lambda(t)\,dt, \tag{4.6}$$

$$\lambda'(T) \geq 0. \tag{4.7}$$

Then we can find the best arrival time for a specialist given any general distribution of SC duration with a finite mean.

Proposition 4.1 (Determine Daily Optimal Fixed Time) *Suppose hourly arrival rate $\lambda(t)$ is periodic with only a peak in each cycle of 24 hours, the optimal fixed time of specialist arrival T^* should be determined as*

$$T^* = T - \mathbb{E}(ST)\mathbb{E}[L(24)]; \tag{4.8}$$

where T is determined in Corollary 4.1. The corresponding minimal average waiting time per person is

$$\frac{W(T+24) - W(T)}{L(24)} - \frac{1}{2}\mathbb{E}(ST)\mathbb{E}[L(24]. \tag{4.9}$$

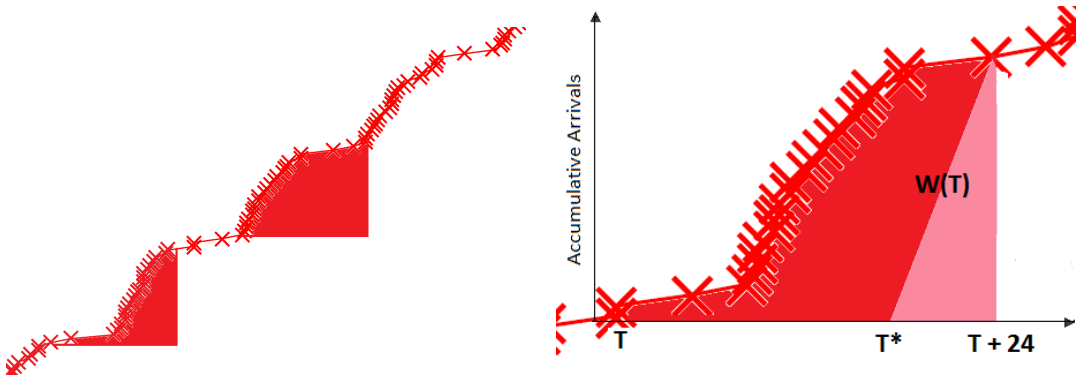
The proof of the above theorems is illustrated in the right of Figure 4-2. First we find the optimal T in the cycle of 24 hours for a specialist to finish his/her session. Then the waiting for the sequential specialist starts from T to $T + 24$, because we consider only one specialist arrival per day. Therefore, we first find the optimal T that minimize the total waiting time of all patients arriving this day in Corollary 4.1, and then figure out the optimal specialist's arrival time $T*$ in Proposition 4.1.

Proposition 4.1 demonstrates that three factors actually determine the arrival time of specialists under an FT policy:

- the volume of demands;
- the distribution of the demand arrival process;
- the mean of the duration of an SC.

Different types of specialists can have the same optimal timing for their SC, if the above factors are the same.

Figure 4-2: Decide Optimal Time



The left figure compares total waiting time under different timing; the right shows the stretch of proof.

The following corollary comes naturally after Proposition 4.1.

Corollary 4.2 (Sensitivity of Optimal Fixed Time) *With the same arrival pattern of SC demand, specialists should arrive earlier if*

- *the patient volume is higher;*
- *the specialist's consultation last longer.*

Moreover, average waiting time is shorter with a higher patient volume or longer consultation duration.

Because the departure time of specialist is determined by the arrival pattern, with the same arrival pattern, specialists stay in ED longer if more demands are present. This leads to an earlier arrival of specialists. Moreover, the average waiting time per person is shorter as specialists stay in ED for a longer period.

Although we do not take the crowding issue into account explicitly, the issue of ED crowding is still the concern of ED managers, and impacts the care quality in ED. The following corollary confirms that the peak of patients present in ED is irrelevant with the timing of SC sessions.

Corollary 4.3 (Maximal Amount of Patients) *Under daily periodic arrival rates, the maximal of total patient waiting for specialists increases with respect to $\lambda(t)$, decreases with ST. If $ST = 0$, the maximal of total patient waiting is $L(24)$ and is indifferent from T.*

That is, the ED overcrowding can possibly be eased by speeding up these SC sessions only. Normally shortening an SC can negatively impact the quality of care. So, the only way to ease ED overcrowding is to improve capacity of specialists if necessary, which will leads to the decrease of an SC duration.

Before the end of this session, we demonstrate the impacts of SC timing on waiting time per patient via two numerical examples. In order to verify our analytical results, we present numerical results with the time-varying arrival rates in the following examples in Table 4.1 and Figure 4-3.

Example Suppose

$$\lambda(t) = a\sin\left(\frac{\pi}{12}t + c\right) + b, \quad b \geq |a|. \tag{4.10}$$

and $\mathbb{E}(ST) = \mu$, where $24\mu b < 1$ for the sake of stability. Then the optimal specialist arrival time is

$$T^* = 12 - \frac{12}{\pi}c - 24b\mu. \tag{4.11}$$

And the minimal average waiting time per person is

$$\bar{W} = 12 - \frac{12}{\pi}\frac{a}{b} - 12b\mu. \tag{4.12}$$

Actually the daily specialist departure time is $T = 12 - \frac{12}{\pi}c$, which is also the optimal boarding time. Daily total patient amount is $\mathbb{E}[L(24)] = 24b$.

Example Suppose

$$\lambda(t) = b\min(\max(1, 3mod(x, 24) - 11), -\frac{1}{3}mod(x, 24) + 9), \tag{4.13}$$

that is, arrival rates follows a linear function

$$\lambda(t) = \begin{cases} 1, & t \in (0,4]; \\ 3t - 11, & t \in (4,6] \\ -\frac{1}{3}t + 9, & t \in (6,24]. \end{cases} \tag{4.14}$$

and $\mathbb{E}(ST) = \mu$, where $84\mu b < 1$ for the sake of stability. Then the optimal specialist arrival time is

$$T^* = 16.5 - 84b\mu. \tag{4.15}$$

And the mimimal average waiting time per person is

$$\bar{W} = \frac{823}{84} - 42b\mu. \tag{4.16}$$

Actually the daily specialist departure time is $T = 16.5$, which is also the optimal boarding time. Daily total patient amount is $\mathbb{E}[L(24)] = 84b$.

We compare the analytical optimal timing from our theories with numerical results in Table 4.1. There are two parts in this table, and one for each example above. Each row shows the different values of patient volume, indicated by the parameter b. The first two columns show the optimal arrival time and corresponding average waiting time per patient resulted from our analytical models. The middle two columns show the numerical results of optimal arrival time and corresponding average wait per person, when the SC session lasts for a deterministic duration. The last two columns are the paralleled numerical results when the SC session follows a stochastic distribution. We can see that our analytical results are fairly close to the numerical ones with both deterministic and stochastic service time.

We further show the impact of optimal timing on the waiting time per person in the left two plots of Figure 4-3. Specifically, for each of the above examples, different lines with color represent varied values of b, i.e. the parameter representing patient volumes. The horizontal axis is the specialist's arrival hour, and the vertical axis is

the average waiting time per patient. We can see that the minimal average waiting time per person tends to be 50% of the longest wait period regardless of the patient volumes b in both examples. Therefore, the decision of an optimal specialist arrival time is crucial to shorten waiting time for SCs under FT policies.

The right two graphs in Figure 4-3 exhibit the impact of different timing on the average waiting time per person, when there are two specialist arrivals per day. Two axes on the plane represent possible combinations of the specialists' two arrival times, and the third axis shows the average waiting time per person. There can be big differences on the average waiting time per person between the optimal specialists' arrival times and those sub-optimal timings.

Furthermore, we compare numerically the impacts of frequency under FT policies regarding the average waiting time per person summarized in Table 4.2. In each section associated with the specific example following the arrival function, the two columns under *Once per Day* show the optimal arrival time of specialists and corresponding average waiting time per patient; The rest columns show the optimal time of arrival and resulting average waiting time per person if there are two specialist arrivals per day. Although the extra arrival can largely reduce average waiting time per person, the marginal reduction will decrease as the frequency of specialists' arrivals increases.

Observation. The marginal benefits of arranging extra fixed time for SC decline as the frequency of consultation increases. Moreover, the marginal benefits decline as the volume of patients becomes higher. We observe the same numerical results in terms of the frequency of the SCs in each periodic cycle as Chan et al. (2016).

With the real data of ED visits, we first estimate the time-dependent arrival pattern $\lambda(t)$ for in-sample SC demand of each specialist type, and then decide the best optimal time for SC sessions. We will compare the analytical results and numerical ones later in Section 4.5.3.

Table 4.1: Compare Optimal FT Policy Once per Day

$\lambda(t) = b\left[\sin\left(\frac{\pi}{12}t\right)+1\right]$						
b	Analytical		Deterministic		Stochastic	
	Hour	Average waiting time (h)	Hour	Average waiting time (h)	Hour	Average waiting time (h)
1	10.56	7.46	11	7.4111	11	7.5082
2	9.12	6.74	9	6.7767	9	6.7965
3	7.68	6.02	8	6.0716	8	6.1824
4	6.24	5.30	6	5.3995	6	5.5289
5	4.80	4.58	5	4.7007	5	4.9680

$\lambda(t) = b\min(\max(1, 3mod(x,24)-11), -\frac{1}{3}mod(x,24)+9)$						
b	Analytical		Deterministic		Stochastic	
	Hour	Average waiting time (h)	Hour	Average waiting time (h)	Hour	Average waiting time (h)
0.2	15.49	9.2936	15.5	9.2435	15	9.2265
0.4	14.48	8.7896	14.5	8.7773	14	8.7279
0.6	13.48	8.2856	13.5	8.3031	13	8.2906
0.8	12.47	7.7816	12	7.7912	12.5	7.8501
1.0	11.46	7.2776	11.5	7.2764	11.5	7.3529

There are two sections in the table, and one for each examples above. Each row shows the different value of patient volume, indicated by the parameter b. The first two columns show the optimal arrival time and corresponding average waiting time per patient resulted from our analytical models. The middle two columns show the numerical results of optimal arrival time and corresponding average wait per person, when the SC session lasts for a deterministic duration. The last two columns are the paralleled numerical results when the SC session follows a stochastic distribution $\mathcal{N}(0.06, 0.0064)$.

Table 4.2: Compare Optimal FT Policy with varied Frequencies

$\lambda(t) = b\left[\sin\left(\frac{\pi}{12}t\right) + 1\right]$

b	Once per Day		Twice per Day		
	Hour	Average waiting time (h)	1st Arrival	2nd Arrival	Average waiting time (h)
1	11	7.5082	9	14	4.7817
2	9	6.7965	9	14	5.3937
3	8	6.1824	8	14	5.4060
4	6	5.5289	5	12	5.0364
5	5	4.9680	5	12	4.4502

$\lambda(t) = b\min(\max(1, 3mod(x, 24) - 11), -\frac{1}{3}mod(x, 24) + 9)$

b	Once per Day		Twice per Day		
	Hour	Average waiting time (h)	1st Arrival	2nd Arrival	Average waiting time (h)
0.2	15	9.2265	13	20	5.8105
0.4	14	8.7279	14	19	7.0911
0.6	13	8.2906	13	19	7.1909
0.8	12.5	7.8501	11	18	7.0816
1.0	11.5	7.3529	10	17	6.6573

Specialist treatment time follows $\mathcal{N}(0.06, 0.0064)$.

4.3.3 TL Specialist Response Policy

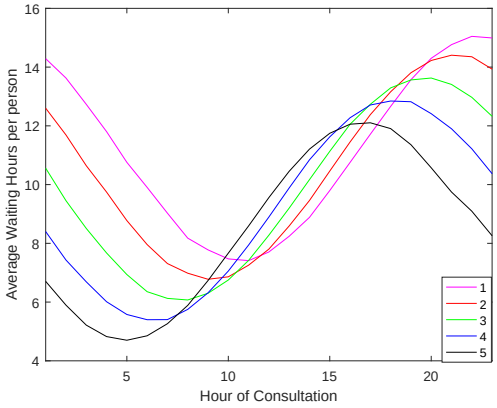
Under TL policies, specialists have to arrive within a certain time window after any consultation request. Therefore, their arrival time is uncertain. However, this stochastic feature of a TL policy results in the fact that specialists' arrival time may not impact on performance measure significantly. This is because the patient waits at most the length of the time window.

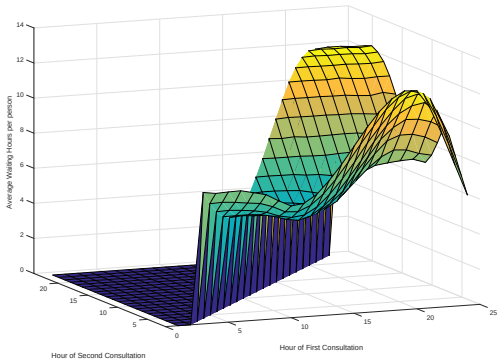
In fact, we conduct numerical experiments to show the features of this policy as in Figure 4-4. Each column shows the results for the example of arrival patterns specified on the top row. Different line colours represent different patient volumes. The first two plots in each column are for TL policies with the same "deadline". That is, the time window within which a specialist has to arrive is constant throughout a day. The first plot shows the average per patient waiting time versus different lengths of time window. And the following plots shows the frequency of specialists' arrivals

Figure 4-3: Optimal Timing under FT Policies

$\lambda(t) = b\left[\sin\left(\frac{\pi}{12}t\right) + 1\right]$

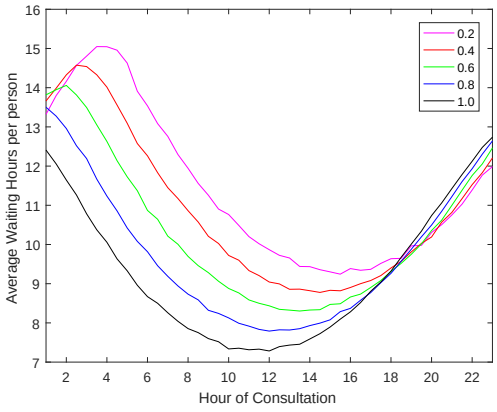
Once per Day | Twice per Day

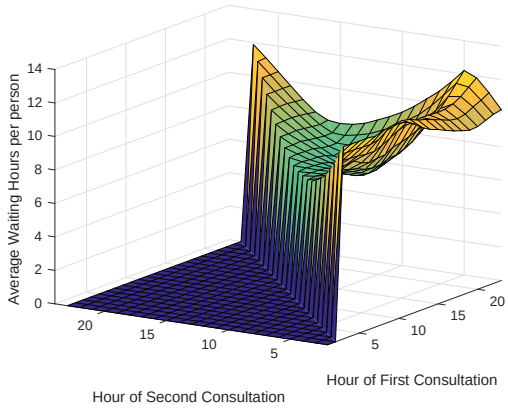




$\lambda(t) = b\min(\max(1, 3mod(x, 24) - 11), -\frac{1}{3}mod(x, 24) + 9)$

Once per Day | Twice per Day





Specialist treatment time follows $\mathcal{N}(0.06, 0.0064)$. In the left two plots, different lines with color represent varied values of b, i.e. the parameter representing patient volumes. The horizontal axis is the specialist's arrival hour, and the vertical axis is the average waiting time per patient. The right two graphs exhibit the impact of different timing on the average waiting time per person, when there are two arrivals per day.

during a day. The last plot in the column shows the average waiting time per patient versus two different time windows. Typically, specialists have a longer time window for their response to ED requests during non-business hours. From this figure, we observe longer delay of SC when specialists are allowed to arrive within a longer time window. Indeed, we have the following observation.

Observation. The average waiting time of specialists following a TL policy increases as

- the expected arrival windows become longer;
- the volume of patients decrease.

Intuitively, if the volume of demand is large, the on-call specialists can have the "economy of scale" and multiple patients are able to share the same specialist visiting the ED, leading to a significant reduction of waiting time. We illustrate this with the example of homogeneous case in the following.

Example: The Time-Independent Case $M/G/1$

Here we consider the corresponding canonical model with generally distributed specialist's treatment times and constant Poisson arrivals with rate $\lambda(t) = \lambda$. The first and second moments of the specialist's treatment time are denoted by $\mathbb{E}ST$ and $\mathbb{E}(ST^2)$, respectively. Because the specialists' arrival is uncertain, we denote $\mathbb{E}(B)$ and $\mathbb{E}(B^2)$ the first and second moments of the time until the arrival of the next specialist, where the time until the next specialist's arrival B is generally distributed.

In the time-independent case of $M/G/1$, mean value technique and Poisson Arrivals See Time Averages (PASTA) proper can be used to calculate the mean waiting time of all patients in the system Adan and Resing (2015).

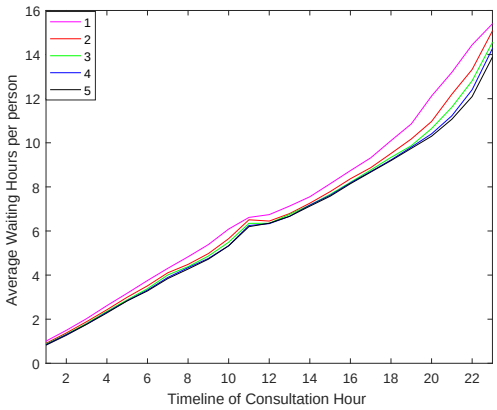
$$\mathbb{E}(W) = \frac{\rho}{1-\rho}\frac{\mathbb{E}(ST^2)}{2\mathbb{E}(ST)} + \frac{1/\lambda}{1/\lambda + \mathbb{E}(B)}\mathbb{E}(B) + \frac{\mathbb{E}(B)}{1/\lambda + \mathbb{E}(B)}\frac{\mathbb{E}(B^2)}{2\mathbb{E}(B)}, \tag{4.17}$$

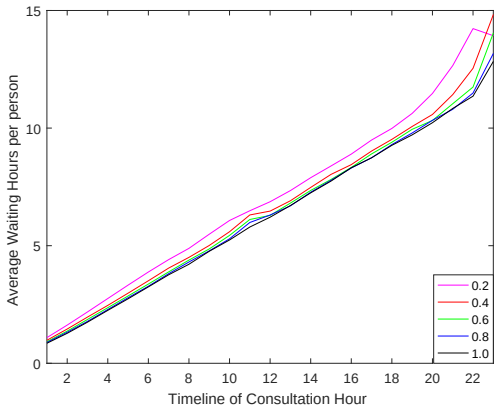
where $\rho = \lambda\mathbb{E}(ST) < 1$ due to the stability of the system. Therefore, we can tell the monotonicity property from the closed form formula as below.

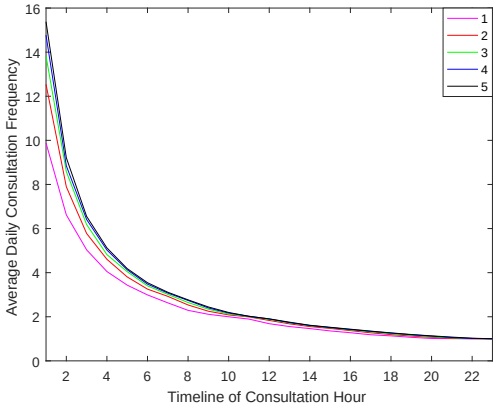
Figure 4-4: TL Policy

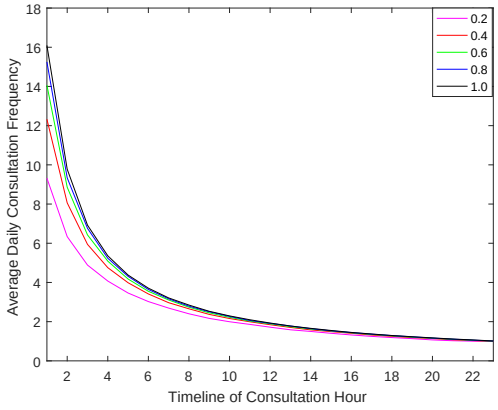
$\lambda(t) = b\left[\sin\left(\frac{\pi}{12}t\right) + 1\right]$	$\lambda(t) = b\min(\max(1, 3mod(x, 24) - 11), -\frac{1}{3}mod(x, 24) + 9)$

Constant TL Throughout A Day

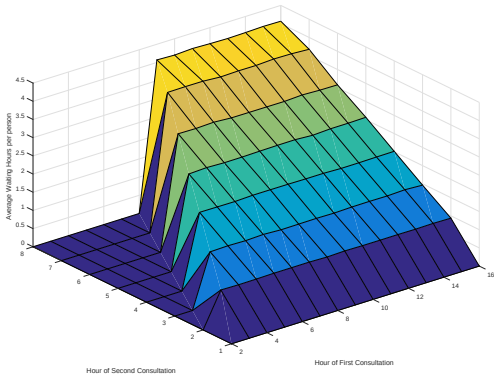


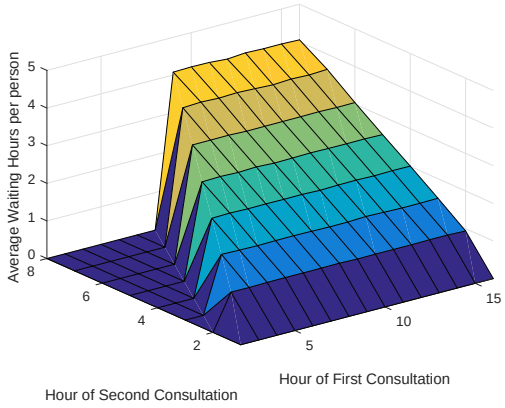






Two TLs Throughout A Day





Specialist treatment time follows $\mathcal{N}(0.06, 0.0064)$. Specialist arrivals follow the pearson system distribution with variance 1, skewness -1 and kurtosis 4.

Lemma 4.2 (Monotonicity of Average Time) *Given* λ, $\mathbb{E}(ST)$ *and* $\mathbb{E}(ST^2)$, *average waiting time* $\mathbb{E}(W)$ *increases monotonically as*

- *expected time until next specialist's arrival* $\mathbb{E}(B)$ *increases;*
- *second moment of time until next specialist's arrival* $\mathbb{E}(B^2)$ *increases.*

4.3.4 Determination of the Optimal Specialist Response Policy

Potential specialist response policies, namely FT and TL policies, have distinct characteristics. Table 4.3 summarizes the features of different specialist response policies for comparison purposes. Specifically, an FT policy specifies the certain time when specialists should show up in ED, therefore, it is more feasible for specialists to implement and easier to arrange their schedules. However, the arrival time needs to be fixed carefully, and there may be potentially long waits for patients under this policy. Actually the marginal reduction of patients' waiting time can become less significant if the specialist arrivals are more frequent. In contrast, a TL policy ensures that the maximal wait for patients is controlled. Yet the drawbacks of this policy lie on the fact that it exposes specialists to uncertain requests, and consequently uncertain schedules. Moreover, if there is a high volume of patients, specialists have to visit ED very frequently. This policy can be very inconvenient for specialists to follow in reality.

Table 4.3: Comparison of Specialists' Response Policies

	FT	TL
Specialist	More feasible	On call
Specialist arrivals	Fixed	can be many as amount of patient ↑
Waiting time	Marginal saving ↓ as daily arrivals ↑	Controlled

In this subsection, we figure out the optimal strategy for specialists' arrivals regarding the following criteria:

- Average per person waiting time;
- Frequency of specialist visits.

To be more specific, if both are similar under FT and TL policies, an FT policy is preferred as it is more practical and convenient for specialists to comply with. Otherwise, the policy resulting in shorter average waiting time per person and low frequency of specialists' arrivals is considered optimal. In general, the determination of the specialists' response policy depends on the volume of demands.

High Volume of Demand. Stay-in specialists should be hired to serve a high volume of patients' SCs. Because specialists may receive multiple requests before they finish a consultation in ED, therefore they tend to stay there for more SCs rather than leave for other tasks and come back to the same ED later.

Medium Volume of Demand. FT policies are proper in this case. *Twice a day* FT policy is optimal for a volume level of 9000, and *Three times a day* FT is preferred for a volume level of approximately 2000.

Low Volume of Demand. A TL policy is recommended in this case. Typically, if the demand for a certain specialist is less frequent than once a day, it is unnecessary to fix a time for that specialist to visit the ED every day. In the case of genitourinary consultation, specialists should adopt TL policies because an FT policy leads to either a long waiting time for patients or more frequent visits for specialists.

In Table 4.4, we showcase several scenarios with varied scales of patient volumes, and compare the performance of different policies in each scenario. Specifically, we consider three different scenarios: 1) A high volume of demand with 8,904 specialist requests annually. This includes all patients who are in demand of specialists in our ED records, regardless of the type of specialists. 2) A medium volume of demand with an annual amount of 2,190 specialists requests. This refers to the specialist requests in the category "other" in our ED dataset. 3) A low volume of demand with the number of 622 specialist requests in one year. This is typically the type of request for

genitourinary specialists in our dataset. In each scenario, we consider both FT and TL policies, and the expected length of each SC lasts either 30 or 20 minutes. The column *Request2Realisation* records the average waiting time per patient between sending out specialists' requests and the arrival of the specialist(s). The column *LOS* represents the average LOS in ED resulted from this policy. Under the TL policy, we record the total arrivals of specialists per year in the column *Notes*. Whereas under FT policies, we record the optimal arrival time(s) when specialists should arrive at the ED in the column *Notes*, it is because the specialists' optimal arrival time(s) are so crucial that they determine the average waiting time per patient under this policy. Moreover *FT1,FT2* and *FT3* refer to one, two and three arrivals for specialists to respond to ED SC requests per day, respectively. We can see that for a high volume of specialist demands, an FT policy with two visits a day can achieve a similar specialist delay (less than 6 hours on average) as the TL policy, which leads to almost two specialist visits per day to ED as well, if the expected specialist's treatment session lasts half an hour. In the case where expected specialist's session lasts 20 minutes, we need to set three visits per day for specialists under FT policies, in order to match the similar delay for patients under the TL policy. Furthermore, the total specialists' visits are similar under both policies as well. Similarly, an FT policy with three times per day for specialists to visit ED is optimal in the medium volume scenario. In contrast, a TL policy should be set for the low demand scenario. It is because patients' waiting time is much shorter under a TL policy than under an FT policy, and specialists visit the ED only once per day.

Table 4.4: Compare Different Optimal Policies

			30 min per session			20 min per session		
	Amount	Policy	Request2Realisation	LOS	Notes	Request2Realisation	LOS	Notes
All Consultation	8904	TL	5.1234	9.4633	614	3.711	8.0509	1045
		FT 1	7.893	12.2315	11 h	7.305	11.6435	12 h
		FT 2	5.8462	10.1847	8/16	4.7056	9.0441	11/23 h
		FT 3				3.3829	8.1679	8/16/24 h
Other	2190	TL	4.5993	9.4176	841	4.0067	8.825	911
		FT 1	9.6755	14.4903	14 h	9.8159	14.6307	14 h
		FT 2	6.1941	11.0069	8/16 h	5.8933	10.7081	8/16 h
		FT 3	4.7103	9.5231	8/16/24 h	4.2856	9.1003	8/16/24 h
Genitourinary	622	TL	5.0146	8.9606	419	4.5436	8.4957	433
		FT 1	9.93	13.882	17 h	9.6409	13.5929	17 h

The column *Request2Realisation* records the average waiting time per patient between sending out specialists' request and the arrival of the specialist(s). The column *LOS* represents the average LOS in ED resulted from this policy. Under the TL policy, we record the total arrivals of specialists per year in the column *Notes*. Whereas under FT policies, we record the optimal arrival time(s) when specialists should arrive at the ED in the column *Notes.FT1,FT2* and *FT3* refer to one, two and three arrivals for specialists to response to ED SC requests per day, respectively.

4.4 Prioritize Patients with Time-dependent Modified Triage Rule

In this section, we study the revised triage policy with the optimal specialists' response strategies. Hereafter, we use "patient" interchangeably with "job" or "customer", and "physician" with "server". Moreover, we only consider the case of no preemptions, i.e. servers cannot be interrupted once the service begins. In other words, the physician will not treat another patient before completing the treatment for the current one. In reality, the ED physicians may have to stop serving non-critical patients when a critical patient arrives, which only accounts for less than 10 % of all ED visits. ED physicians treat critically life-threatening patients whose triage codes are either 1 or 2 with highest priority. They even use preemptive policy if any of those critical patients arrives; that is, they have to start treating those critical patients immediately, and their current treatment is interrupted. However, we do not consider those critical patients here, as they are very few in ED.

Assumption 4.1 (Non-critical patients only) *We consider non-preemptive policy here, because an ED physician does not treat other non-critical patients before finishing one treatment.*

We will release this assumption and incorporate critical patients in the comprehensive simulation later.

In practice, patients in ED can have very complicated symptoms, and they may need multiple SCs. Moreover, ED physicians may not be able to figure out the proper specialist whom a patient needs to consult. However, the portion of complicated patients is not large, therefore we simplify the model and consider at most one round of SC and perfect judgement of consultation type at triage.

Assumption 4.2 *Each patient needs at most one round of SC.*

Assumption 4.3 *At triage, the nurse is able to match patients with the type of specialist perfectly. But the likelihood of actual request is not 100 %.*

4.4.1 Set up - Dynamic Programming

Suppose the ED physicians' treatment time is independent of patient's type and generally distributed with first and second moments m_j and σ_j.

Assumption 4.4 *ED physicians treat patients with a service time that follows the same distribution, regardless of the patient class.*

We consider an ED with a single physician. In the conventional heavy traffic framework, a multiple server system with N servers is asymptotically equivalent to the single-server system with a service rate N times faster than that of a single server. Hereafter we assume a single server system, yet we use time-varying service rate $u(t)$ with cumulative distribution function $G(t)$ to incorporate the possible different amount of physicians. In our modeling framework,

Decision epoch: t when any ED physician completes the service.

Observation Set: $\mathbf{S} = \{S_0(t), S_1(t), S_2(t), \cdots, S_N(t)\}$ is the set the amount of patients in class n, $n \in \mathscr{N}\{0, 1, 2, \cdots, N\}$ at decision point t. Suppose there are N different classes ($\{1, 2, \cdots, N\}$) of patients; and each class of patients requires different SCs. Let class 0 be the one of patients who are not predicted to have an SC in triage. Denote I the amount of classes with the optimal FT policies, and J the amount of classes with the optimal TL policies, and $N = I + J$. For the purpose of convenience, let the classes $\mathscr{I} = \{0, 1, 2, \cdots, I\}$ represent the classes of which FT is the optimal specialist response rule, and $\mathscr{J} = \{I + 1, I + 2, \cdots, N = I + J\}$ the classes of patients whose specialists should follow the TL policy.

The predicted probability of patient κ in class i who is going to require an SC is $p_{i\kappa}$ where $\forall \kappa \in \{1, 2, \cdots, S_i(t)\}$ and $\forall i \in \{1, 2, \cdots, N\}$. $\forall \kappa \in \{1, 2, \cdots, S_0\}$, the predicted probability of patient κ who does not require an SC is $p_{0\kappa}$. To avoid repetitiveness of classification, we set up a threshold of probability P such that

$$p_{i\kappa} \geq P, \quad \forall \kappa \in \{1, 2, \cdots, S_i(t)\}, \forall i \in \{1, 2, \cdots, N\}, \quad \forall t \geq 0. \tag{4.18}$$

Therefore, those patients who are more likely to require an SC are classified into corresponding class i, and $i \neq 0$, matching their symptoms with the type of specialist who is able to treat them; otherwise they are categorized into class 0 as ones who are unlikely to require an SC. In the case of $P = 0.5$, $p_{i\kappa} \geq 0.5$, $\forall \kappa \in \{1, 2, \cdots, S_i\}, \forall i \in \mathscr{N}, \forall t$.

Set of Admissible Actions: $A(\mathbf{S}(t))$ under statues $\mathbf{S}(t)$, choose a patient from any non-empty class.

$$A(\mathbf{S}(t)) = \{n, \kappa | S_n(t) \geq 1, \kappa \in [1, S_n(t)], n \in \mathscr{N}\}. \tag{4.19}$$

Patients arrive at ED following a time-varying Poisson process withl rate $\lambda_n(t)$, $\forall n \in \mathscr{N}$. Let

$$\Lambda_n(t) = \int_0^t \lambda_n(u) \mathrm{d}u \tag{4.20}$$

Transition Probability: Let t be a decision epoch. Consider *after action* state $\mathbf{S}(t)$ after a service with length of v, then the system at the next decision epoch $t + v$ will be $\mathbf{S}(t + v)$ with probability $\mathbf{P}(\mathbf{S}(t + v)|\mathbf{S}(t))$

$$\mathbf{P}(\mathbf{S}(\mathbf{t} + \mathbf{v})|\mathbf{S}(t)) = \prod_{n \in \mathscr{N}} \mathbb{P}_n(v, S_m(t + v) - S_m(t)) \tag{4.21}$$

where

$$\mathbb{P}_n(v, X) = e^{\Lambda_n(v)} \frac{(\Lambda_n(v))^X}{X!} \tag{4.22}$$

the probability that there are X arrivals in a time interval with length v for class n.

Value function minimize the total waiting time of existing patients from state $\mathbf{S}(t)$

$$V(\mathbf{S}(t)) = \min_{j,\kappa \in A((S)(t))} \quad \left\{ \int_0^\infty \mathcal{W}_j(v) + \sum_{\mathbf{S}'} \mathbf{P}(\mathbf{S}(t+v)|\mathbf{S}(t)) V(\mathbf{S}(t+v)) \mathrm{d}G(s) \right\}, \tag{4.23}$$

$$\mathbf{S}(t) \neq 0; \tag{4.24}$$

$$V(\mathbf{X}(0)) = 0. \tag{4.25}$$

where $\mathcal{W}_{j\kappa}(v)$ is the total known waiting time of all patients in the system at time $t+v$. Suppose the expected arrival time of next specialist is T,

$$\mathcal{W}_{j\kappa}(t) = \begin{cases} v(\sum_{n\in\mathcal{N}} S_n) + T - (t+v), & \text{with probability } p_{j\kappa}; \\ v(\sum_{n\in\mathcal{N}} S_n), & \text{with probability } 1 - p_{j\kappa}. \end{cases}$$

4.4.2 Stability Condition

The arrival rate function has a periodic pattern in the ED setting. Specifically, it follows the same intraday pattern, that is, $\lambda_n(t + T_c) = \lambda_n(t)$, where T_c denotes the cycle depending on the unit of time. Naturally, the scheduling of ED physicians should follow this periodic pattern, and service rate $u(t)$ is periodic as well.

We consider the condition of stability for this time-dependent system. In order to keep the system stable, the capacity of ED physicians should be larger than the demand of all visits. Specifically, this holds in each cycle of T_c. Let

$$\Lambda(t) = \sum_{n\in\mathcal{N}} \Lambda_n(u) \tag{4.26}$$

be the cumulative arrival rate function of all ED patients during a day, where Λ_n is calculated with Eq.4.20. Without loss of generality, we assume time zero as the beginning of each cycle, then $\Lambda(T_c)$ becomes the total arrival rate of a whole cycle.

Lemma 4.3 (Stability Condition) *The time-dependent system is stable if*

$$\int_0^{T_c} u(t)dt \geq \Lambda(t). \tag{4.27}$$

Therefore, all arrivals during one periodic cycle can be dealt with, and the waiting queues do not grow into infinity.

4.4.3 Structural Properties

Suppose $\forall t, \forall i \in \{1, 2, \cdots, I\}$, the upcoming due time when the specialist arrives is T_i, where $T_i \geq t$, and interval between the upcoming due time and the following one is Δt_i.

First, we show that non-idle policy is optimal. That is, the optimal scheduling policy should always assign an existing patient to the ED physician who completes a treatment, rather than let the physician be idle.

Proposition 4.2 (Existence of Optimal Policy) *There exists an optimal policy that does not allow servers to be idle except when the system is empty.*

Patients who are likely to require SCs following the FT rules actually have a deterministic due time for their service. If the due time is missed, they will have to wait for the next arrival of the type of specialist(s) they need. The following proposition describes the optimal scheduling for those patients whose required specialists following FT policies.

Proposition 4.3 (Optimal Scheduling for Patients under FT policies) *If* $T_1 = T_2 = \cdots T_I = T$, *let* t *be a decision point for the state* $\mathbf{S}$, *then*

1. *There exists a threshold* ΔTH *such that it is optimal to assign non-FT patients between time inter* $[\alpha, \beta]$, *where* $\alpha < T < \beta$.

2. *Within each class of patients, it is optimal to prioritize the one with largest probability;*

For those patients whose specialists on demand follow a TL rule, they are actually facing a stochastic due time, because the delay of the specialists' arrival is uncertain, and independent of the time when they send of request. Therefore, the scheduling policy below is different for these classes of patients.

Proposition 4.4 (Optimal Scheduling for Patients under the TL Policy) *Let t be a decision point for the state $\{S_0(t), S_1(t), S_2(t), \cdots, S_N(t)\}$, then*

1. *It is optimal to prioritize the class of patients who require specialist under the TL policy according to their arrival time.*

2. *Within each class of patients, it is optimal to prioritize the one with largest probability;*

4.5 Numerical Results

We use a database of all ED visits in the year of 2015 in St Mary's hospital, Montreal. There are 36,324 ED visits in total, among which 32,825 or (90.37 %) fall in to clinical non-critical categories (Triage Codes 3, 4 and 5). According to the expert opinions, duration of an SC session is 30 minutes on average, and can be as long as one hour. Thus we use triangle distribution $TR(0.25, 0.5, 1)$ (in the unit of hour) to simulate the length of an SC session. We consider other distributions for the duration of SC sessions in sensitivity analysis.

Our ED data shows that a certain amount of patients need more than one specialist, and there exists multiple rounds of specialist requests in ED for an individual patient. In our simulation model, we only consider the first specialist request for the sake of simplicity. We leave the streamlining multiple specialist requests to future research.

4.5.1 Patient Clusters and Their Clinical Trajectories

In the hospital where our data come from, the types of specialists are: 1) internal medicine; 2) oncology; 3) mental; 4) gynecology; 5) blood & immune; 6) heart disease;

7) digestion, tissue & skin; 8) genitourinary; 9) injury; 10) other. The trajectories from diagnosis code to the type of SC are summarized in Table C.2. Statistics of current LOS with specialist delays are summarized in Table 4.5, where the number of consultation sessions per year is recorded for each specialist type. While the numbers in the last three columns are the expected amount of hours, the standard deviations are reported in the bracket. The column *TTFT* refers to the Time To the First Treatment in ED, and it is similar among all the patients who are in demand of different SCs. The delay of SC is recorded in the column *R2R*. The delays between sending out a specialist request and the specialist's arrival are over five hours, except for the gynecology specialists, who arrive in about three hours on average. The longest average delay is for the mental specialists, with an average waiting time of over nine hours. The last column showcases the time from a patient's arrival to the time he or she sees the first specialist. This is due to the scope of our study, which focuses on the delay of the first specialist for each patient. So we present the status quo in Table 4.5 as the base case scenario for our simulation models.

4.5.2 Empirical Model on Estimation for the Probability of SC Request

To predict each patient's probability of SC request with available information in triage, we use logistic regression and regression tree (CART) for in-sample data first (first 2/3 of ED visits in 2015), and then compare Area under Curve (AUC) and Mean Squared Error (MSE) with out-of-sample data (the last 1/3 ED visits in 2015) in Table 4.6. Our prediction of SC requests is accurate with over 80 % AUC and less than 18 % MES for out-of-sample verification. Sensitivity and specificity are two important measures for prediction. We summarize sensitivity and specificity in Table 4.7 for each triage code. Because less critical patients tend not to need a specialist, the sensitivity is lower for patients with triage code 4 and 5, whereas their specificity is very high.

We also try other statistical supervised learning methods, such as a neural network

with different amounts of hidden levels, nearest neighbor with Gaussian kernel, kernel epsilon and support vector machine (SVM), to classify patients into categories with different likelihoods of SC demand. However, due to the limited amount of variables collected in triage, those methods do not improve the power prediction significantly (Table C.3) shown in the Appendix.

Moreover, our data shows the existence of an unbalance between patients who require consultation and those who do not, specifically, the ration is approximate $1:3$ (Table C.4). Therefore, we also try the method of balance. In fact, balance does not improve the power prediction in terms of AUC and MRE (Table C.5), rather it improves the sensitivity and specificity (Table C.6) as shown in the Appendix.

4.5.3 Optimal Specialist Arrival Time under FT Policy

Our empirical study shows that the patient arrivals to ED follows a non-homogeneous Poisson process. We show the time-varying arrival patterns of ED patients in Section C.2 in the Appendix. In order to verify the optimal arrival time for specialists under FT policies resulted from our analytical model, we use two types of functions to fit patients' arrival patterns - 1) piecewise linear function; 2) submodial *sin* function.

In Figure 4-5, we show the function fitting of SC demand patterns. The left is for the specialist demand of non-mental consultation and the right is for those patients who need to consult internal medicine specialists. The first row shows the *sin* function fitting, and the second row of graphs shows the linear function fitting. Apparently, linear functions tend to fit the actual arrival rates better than *sin* function. Actually R^2 of *sin* fit is 70.06 % for Internal Medicine, and 76.12 % for non-mental. Whereas R^2 of linear function fit is over 85 % for both specialist demands. The last row of plots show the numerical results of optimal arrival times for specialists. We calibrate the numerical results by enumeration and search for the optimal hour that can lead to the shortest average waiting time. The numerical results also indicate that the specialists' optimal arrival times vary during a week, resulted from the daily variation of patients' arrival flow during a week.

Table 4.5: Status Que LOS In Terms of SC

Specialist Type	Number of Sessions	TTFT (h)	R2R (h)	Time to first SC (h)
All Consultation	8904	1.4182 (1.0826)	6.6205 (7.9108)	12.3688 (10.9657)
Non-mental	7862	1.4126 (1.074)	6.2893 (7.2473)	12.0288 (9.8689)
Internal Medicine	1705	1.3547 (1.0392)	6.6844 (7.5998)	12.462 (10.3414)
Oncology	255	1.4025 (1.021)	7.4656 (7.849)	14.6156 (11.057)
Mental	1042	1.46 (1.1458)	9.1232 (11.4713)	14.9337 (16.8947)
Gynecology	347	1.5489 (1.2113)	3.1346 (2.2988)	6.095 (3.0677)
Blood & Immune	122	1.4424 (0.921)	9.62 (9.76)	14.8425 (13.098)
Heart Disease	714	1.3243 (0.922)	6.9442 (8.1975)	12.7876 (10.6612)
Digestion, Tissue & Skin	1072	1.4945 (1.165)	5.5162 (6.6403)	10.6637 (8.9971)
Genitourinary	622	1.5192 (1.1139)	5.2852 (5.9049)	10.7549 (8.4316)
Injury	835	1.3717 (1.0995)	5.7111 (6.6205)	11.2918 (9.7308)
Other	2190	1.4098 (1.0643)	7.0052 (7.4953)	13.238 (10.0084)

The standard deviation is in the bracket. The column *TTFT* refers to the Time To the First Treatment in ED, and it is similar among all the patients who are in demand of different SCs. The delay of SC is recorded in the column *R2R*, where R2R stands for Request to Realization. The last column is the time from a patient's arrival to the time he or she sees the first specialist. We consider only the time to the first SC due to the scope of our study.

Table 4.6: Results of Statistical Learning

Estimated probability of	AUC	MSE
CART (%)	81.60	17.67
Logit (%)	83.29	17.49

Table 4.7: Sensitivity and Specificity

Triage	3	4	5
Sensitivity (%) $(\mathbb{P}(Pred = 1 \mid Act = 1))$	79.34	67.88	42.18
Specificity (%) $(\mathbb{P}(Pred = 0 \mid Act = 0))$	78.06	89.48	97.5

Although, the arrival patterns of different patient types are similar under the FT policy, the optimal arrival time for the corresponding specialists are not the same due to the difference in patient volumes.

In Table 4.8 and 4.9, we present both analytical and numerical results under FT policies, and consider two values of expected duration of each SC session, namely 20 and 30 minutes. For each value of SC durations, we report the optimal hour when specialists should arrive at the ED, and the associated average per patient waiting. In the first section, we show numerically the optimal time when specialists should arrive once for consultation in the ED each day of a week. Because of the daily variation of demand volumes and patterns, the numerical results imply that specialists should visit ED at different times during a week. The second section of the table compares the optimal times of FT policies with one specialist visit per day, calculated from numerical and analytical models. The row *Daily* presents the optimal specialist response time calibrated by enumeration, regardless of the daily variation. The rows *sin* and *linear* present the optimal specialist response time under FT policies with one daily specialist visit, calculated from analytical models with *sin* and linear function fitting, respectively. Although analytical result with *sin* function has a smaller error of optimal timing, linear function fit tends to estimate the average

waiting time better, i.e. very close to the numerical results. However, the errors of analytical models come from 1) carryover errors from function fitting for actual demand patterns; and 2) daily variation of demand patterns, that is, the specialist demand patterns are not exactly cyclical on a daily basis. Moreover, we also consider a constrained scenario where specialists work only during business hours, so their consultation sessions in the ED take place only from 8 am to 6 pm every day. The last two sections of the table show the results under this scenario. The row *Constraint Actual* shows the optimal solution if the specialist has to finish the consultation session at 6 pm, and the row *Constraint linear* shows the analytical results with the same constraint. The last section is about the FT policy with two specialist visits per day, and it numerically compares the optimal timing for specialists to arrive at the ED with and without the constraint. Compared with the scenarios without the constraint, specialist are more convenient and avoid irregular working hours; yet patients who are in demand of SCs have to wait longer in the scenarios with the constraint.

Moreover, we illustrate the optimal arrival times for internal medicine, injury and non-mental specialists in Table 4.8, 4.9 and 4.10, respectively. We choose these three types of patients, because they represent medium, low and high volumes of associated specialist demands respectively. Under the FT policy with one daily specialist visit, the optimal arrival time for non-mental specialists is the earliest among the three types, and injury specialists should arrive at the ED in the evening. For the same type of specialists, they should start the SCs earlier if the SCs last longer. The three tables together show that specialists should arrive at the ED earlier if the patient volume is higher or the consultation session lasts longer, keeping the same demand patterns, according to Corollary 4.2.

Figure 4-5: Numerical Results of Optimal Specialist Arrival under FT Policy

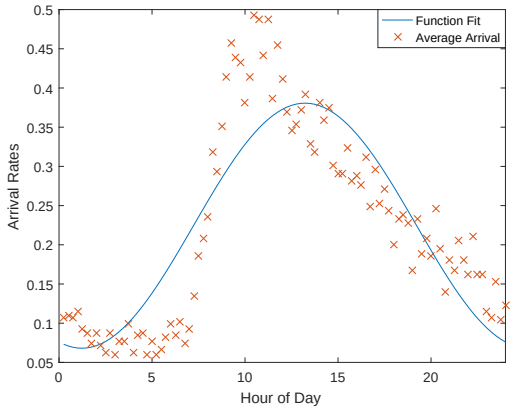
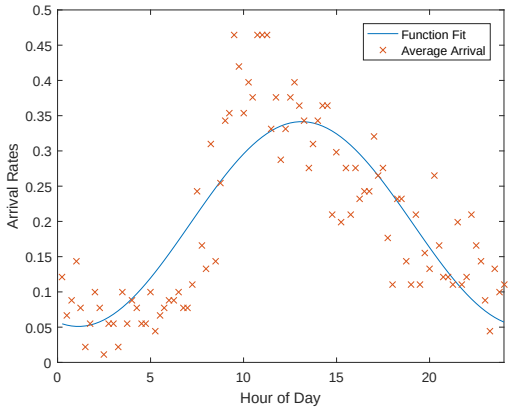
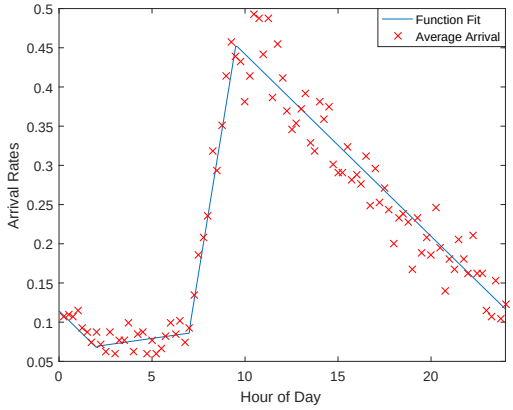
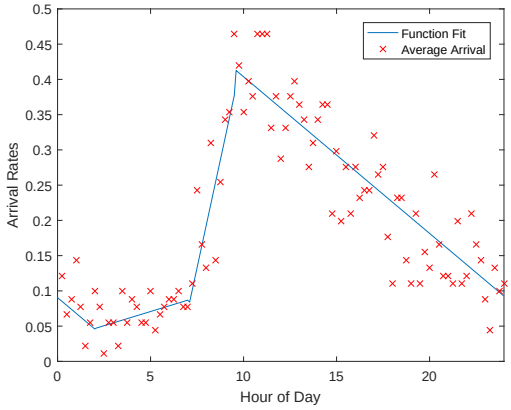
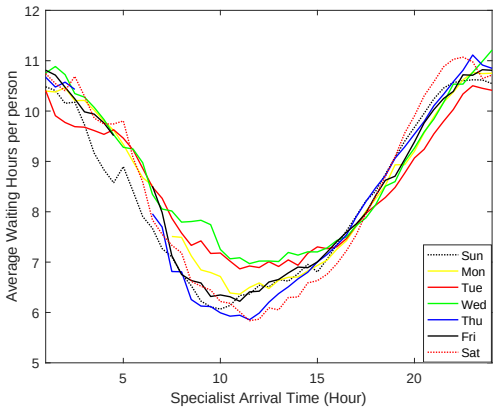
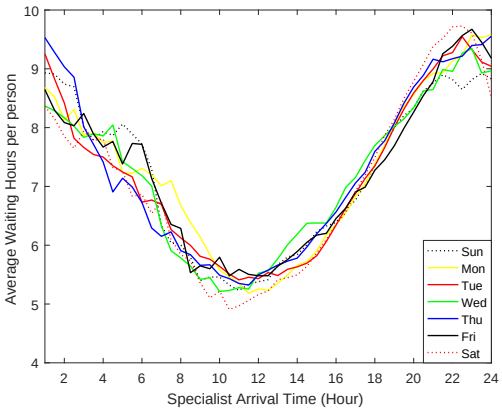
Non-mental	Internal Medicine
Function Fit - sin	
	
Function Fit - linear	
	
Optimal Arrival Time	
	

Table 4.8: Compare Optimal FT Policy - Internal Medicine Specialists

Internal Medicine Consultation 1705 sessions				
	30 min per session		20 min per session	
	Optimal Hour	Average waiting time (h)	Optimal Hour	Average waiting time (h)
Sun	15	8.6302	15	8.9845
Mon	16	8.5156	16.5	8.9442
Tue	15.5	8.7051	16.5	9.014
Wed	16	8.7259	17.5	8.8668
Thu	16	8.6604	18	8.6612
Fri	15.5	8.4916	17.5	8.8121
Sat	16	8.4394	16.5	8.9205
Daily	16	8.592	17.5	8.9122
sin	16.8915	8.1714	17.67	8.5607
linear	17.0642	8.5236	17.8428	8.9129
Constraint Actual	14.5	8.7804	15.5	9.0217
Constraint linear	14.66	8.8124	15.44	9.2017
Twice	14/21	7.6072	16/22	7.9605
Constrained Twice	14/16	8.6195	12/16	9.0745

All numerical results are based on 10 samples drawn from the arrivals that require internal medicine SCs.

Table 4.9: Compare Optimal FT Policy - Injury Specialists

Injury Consultation 835 sessions				
	30 min per session		20 min per session	
	Optimal Hour	Average waiting time (h)	Optimal Hour	Average waiting time (h)
Sun	18.5	9.0985	17.5	9.4662
Mon	16	9.3716	17	9.4034
Tue	16.5	8.9156	18	9.2894
Wed	16.5	9.3183	14.5	9.5999
Thu	18	9.0879	19	9.3026
Fri	16.5	9.1742	15.5	9.0945
Sat	17.5	9.1086	17.5	9.4355
Daily	16.5	9.208	17.5	9.4267
sin	18.0832	8.7673	18.4645	8.9579
linear	18.2561	9.1195	18.6374	9.3102
Constraint Actual	15.5	9.361	16	9.478
Constraint linear	15.85	9.4084	16.24	9.599
Twice	14/22	7.3633	15/23	7.615
Constrained Twice	12/16	8.854	11/16	9.3024

All numerical results are based on 10 samples drawn from patient arrivals with consultation request of injury specialists.

Table 4.10: Compare Optimal FT Policy - Non-mental Specialists

Non-mental Consultation 7862 sessions in total						
	30 min per session		20 min per session		10 min per session	
	Optimal Hour	Average waiting time (h)	Optimal Hour	Average waiting time (h)	Optimal Hour	Average waiting time (h)
Sun	10	5.3274	10	6.0669	11.5	7.8037
Mon	9.5	5.3384	11	6.364	15.5	7.9105
Tue	9	5.3366	11	6.8658	15	8.4628
Wed	9.5	5.6424	11.5	6.9701	16	8.3902
Thu	9	4.7021	11.5	5.8555	14.5	7.4261
Fri	9.5	4.9664	11	6.2213	14.5	7.8337
Sat	10	4.7968	11.5	5.8363	14.5	7.5141
Daily	10	5.2528	11	6.4034	14	8.0122
Sin	8.4572	3.9543	12.04	5.7493	15.6371	7.5442
Liner	8.6292	4.3061	12.2195	6.1012	15.8097	7.8964

All numerical results are based on 10 samples drawn from arrivals which require consultation of non-mental specialists.

4.5.4 Comprehensive Simulation

In reality, there are at least ten different types of SC demands, and each requires a consultation with the corresponding specialists, as we explained in subsection 4.5.1. This makes the integrated scheduling problems from triage to SC very complicated and analytically intractable. However, recently, simulation models have been popular to tackle these sort of complicated systematic scheduling problems. Therefore, we conduct a comprehensive simulation based on all out-of-sample ED visits (the last 1/3 of ED visits in 2015) with ARENA (Blackrock) software. We use the out-of-sample data to avoid the data overfitting, as we use the first 2/3 of data to predict probability of consultation demand.

We describe the scenarios to be tested with our simulation models below.

- **Base case.** It is the status quo where traditional triage rule is applied and yet no specialist arrival policies are applied.

- **Modified triage.** This scenario combines both modified triage rule and optimal specialist arrival policy. Per multiple patients with the same non-critical triage code (3, 4 and 5), the patient with a higher predicted probability of SCs get the priority.

- **Optimal specialist policy.** This scenario adopts the optimal specialist response policies for all types of specialists.

- **Combined.** This scenario combines both modified triage rule and optimal specialist arrival policy. In this scenario, urgent or life threatening patients (triage code of 1 or 2) always have the highest priority among the rest of the patients. Per multiple patients with the same non-critical triage code (3, 4 and 5), the patient with a higher predicted probability of SCs get the priority. Patients who are predicted to require a specialist whose arrival follows FT policies are prioritized within a certain period before the associated specialist's arrival time. The rest of the time, patients who are likely to require a specialist following a TL arrival rule are prioritized.

In all above scenarios, we use the actual scheduling of ED physicians: there are two ED physicians from 8am to 4pm on Monday to Friday, and one ED physician for the rest of time. Other assumptions in our simulation models include: ED physicians' service time of urgent patients, which follows triangular distribution $TR(0.1, 0.3, 0.8)$ in the unit of hours; ED physicians' service time of the other patients following $TR(0.05, 0.2, 0.4)$ in the unit of hours; and triage nurses' service time for all patients following $TR(0.05, 0.1, 0.2)$ with the unit in hours.

In the base case model, we use the true-to-life delay between sending out consultation requests and the arrival of the associated specialist, based on our dataset. Our empirical study shows that there is no statistically significant difference in the delay patterns between business and non-business hours. Both patterns are compared in Figure C-3 in the Appendix. In the base case, the LOS of patients who require SCs is calculated as the sum of the period from arrival time until the specialist's arrival and a specialist's service time. The LOS of patients who do not require SCs is calibrated as the sum of TTFT plus an ED physicians' service time. With these measures, we avoid the impact of multiple rounds of consultations and delay of admission. Table 4.11 reports the consequential waiting times under different policies.

In Table 4.11, we present the results of our simulation models. The results of base case model are presented in the unit of hour. We present the results of other scenarios with the percentage of changes from the corresponding base case results. For instance, the scenario of modified triage results in a 0.07 % shorter LOS but 7.05% longer R2R, no change on the amount of patients present in the ED compared with the base case model. R2R, or Request 2 Realization, refers to the delay between sending out specialist requests and the arrival of the associated specialist(s). In contrast to R2R, which is only among patients who require SCs, LOS is the average amount for all patients. Although modified triage incorporating the predicted demand of specialists slightly shortens overall LOS in ED, it increases the delay of specialists' response to SC requests. Optimal SC policies can shorten both LOS and R2R significantly. It dominates the performance of modified triage when combined together in the combined scenario. Moreover, Figure 4-6 shows the histogram of the

number of patients present in ED under the Optimal Specialist Policy. It shows a positive skewness of the distribution of the amount of patients present in the ED.

Table 4.11: Simulation Results

Scenario	LOS (%)	R2R (%)	Amount of Patients Present (%)
Base case	4.5374 h	7.1048 h	21.0611(8.9284) h
Modified Triage	-0.07	7.05	-
Optimal Specialist Policy	-16.81	-42.60	-16.77(-4.77)
Combined	-16.95	-42.04	-18.21(-6.12)

The standard deviations are in the bracket.

4.5.5 Sensitivity Analysis

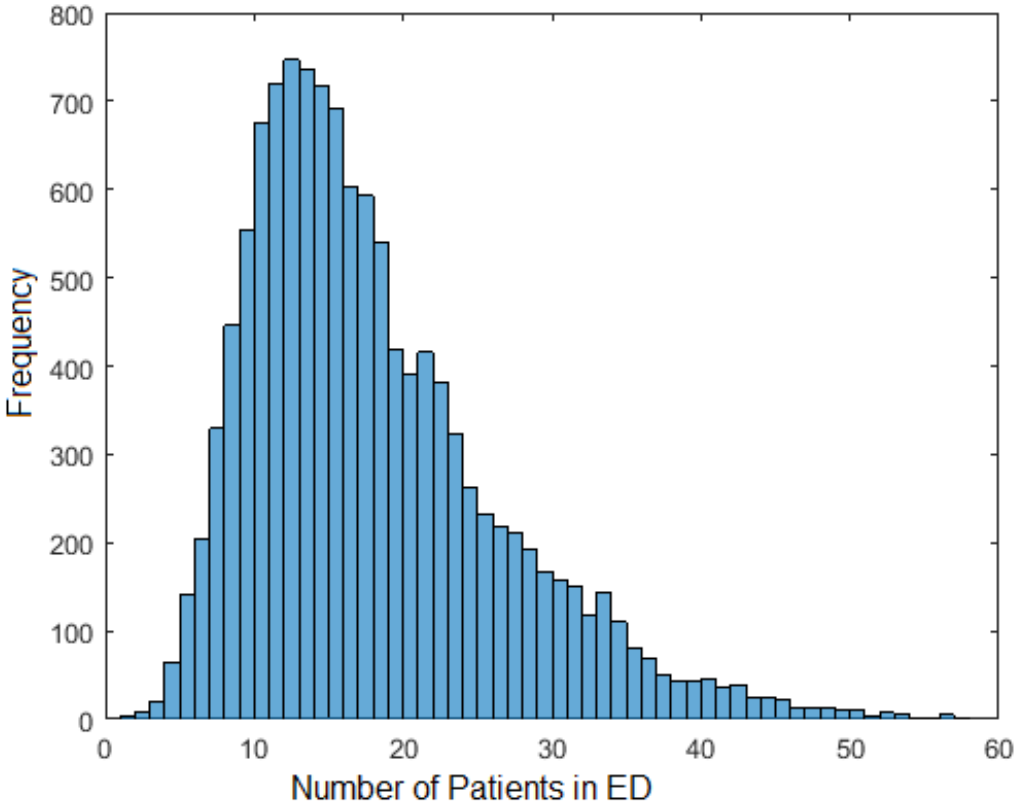
In this subsection, we conduct several scenarios for sensitivity analysis to examine the impact of several factors on the modified triage policy. We take the traditional triage with optimal specialist policy as the benchmark. We test the following factors in sensitivity analysis.

- **Impact of prediction accuracy.** Predicting each patient's probability of requiring specialists is not perfectly accurate due to limited information available in triage. The accuracy of the prediction can impact the performance of the modified triage rule.

- **Skewed distribution of SC duration.** We model the duration of an SC session with a triangle distribution, which features positive skewness. We use a normal distribution with the same first two moments to examine the impact of the skewness in this distribution.

- **Threshold of FT advancement.** We use a static threshold to switch priority between patients who are likely to require a specialist following an FT policy and those whose specialists follow a TL policy. The length of the period before the arrival time of a specialist is referred as FT advancement. For example, if

Figure 4-6: Histogram of Number of Patients in ED under Optimal Specialist Policy



a certain specialist is set to arrive at the ED at 11 am, an FT advancement of 2 hours means that starting from 9 am, the patients who are likely to require consultation from the specialist are prioritised among others.

We present the results from sensitivity analysis in Table 4.12. The first row is the benchmark with values of LOS and R2R in the unit of hours. For the following scenarios, we show the percentage of change compared to the benchmark. Although the accurate prediction and properly set static advancement can possibly contribute to the improved performance of modified triage rule, we see more significant improvements in the performance of modified triage rules with a symmetric distribution for the duration of SC sessions.

Table 4.12: Sensitivity Analysis

FT Advancement (h)	Actual Delay		$TR(0.5, 1, 11.5)$		$N(4.4, 2)$	
	LOS (%)	R2R (%)	LOS (%)	R2R (%)	LOS (%)	R2R (%)
Benchmark	3.7748	4.5785	3.87	4.5974	3.9861	4.7052
1	-0.62	-0.13			-2.19	-1.25
1.5	-0.15	0.08			-1.80	1.03
2	-1.76	0.98			-2.34	-1.08
2.5	-1.62	0.30	-0.48	0.91	-3.73	-0.73
3	-2.13	-0.86	-0.13	-0.94	-2.07	-1.72
3.5	-2.20	0.14	-0.72	0.80	-3.20	-0.12
4	-0.32	1.11	-0.67	0.10	-2.49	-0.23
4.5			0.18	0.02	-3.10	-0.22

The first row is the benchmark with actual length of LOS and R2R in the unit of hours. For the following scenarios, we show the percentage of change compared with the benchmark.

4.6 Conclusion and future research

This study is motivated by the prolonged delay of specialists' arrivals after the request of SC in a local community hospital in the city of Montreal. The limited amount of specialist demand makes it impossible to hire on-site specialists in the ED. The lack of systematic rules of specialists responses lead to the fact that patients can wait

for average 7 hours for one round of SC, and the total delay of several rounds of consultations can add up to multiple days. We study this problem based on the real data of all ED visits in the year of 2015.

First, we set up queueing models with non-homogeneous arrival rates to model the demand of SCs in the ED. Through our analytical model, we figure out the closed form of the expected average per person waiting time and optimal arrival time for a specialist based on demand patterns, volumes and duration of a consultation session, for an FT policy with one visit per periodic cycle of demand arrival flow.

Second, we provide a systematic method to determine the best response rule for different sort of specialists, based on the volume of demand, and features of varied proposed specialist response rule. These optimal policies can significantly shorten waiting time for patients; and moreover, they are convenient for specialists to implement in their busy schedules, thus are easy to implement and enforce. Hence the proposed guideline of determining an optimal arrival rules for specialists provides valuable managerial insights.

Then, we conduct patient classification in terms of their likelihood of requesting a certain SC with the information available at triage. Using multiple statistical learning methods, we are able to provide a moderately accurate prediction at triage. Balanced method is also used, and it may not improve prediction accuracy in terms of MSE, but can improve the performance of specificity and sensitivity.

Finally, we analytically measure the potential improvement on efficiency of the resource-based triage with the dynamic programming framework. The actual realized benefits may be offset by uncertainties and other delays in ED, as a result of our comprehensive simulation models.

The most straightforward future work should lay on a dynamic threshold to switch the prioritization among patients who are in demand of different specialists. Algorithms proposed for those fixed interval due-dates problems can be helpful. Other future research can focus on the following possible directions of time-varying queues. First, delay of test results should be studied as this is another factor that lead to the prolonged LOS in ED. A queue system or network with several tandem queues can be

applied in this case combining specialist and test delay together. Second, this study only considers at most one round of SC. Future work can extend to multiple rounds incorporating feedbacks in queueing system. Last, our analytical work on determining the optimal FT specialist arrival time can be easily extended to the case of optimal boarding time. The interface of ED and inpatient wards can be considered together in order to reduce LOS in ED. Due to the possible variation of specialists' flexibility and availability, we can also incorporate the capacity and workload of the specialist into our model in the future work.

This work attempts to shorten LOS in ED via improving an internal process (delay of SC). The following chapter also aims to reduce waits in ED, however, via designing an interface between ED and inpatient wards.

Chapter 5

Design of Observation Units (OU) for Acute Decompensated Heart Failure (ADHF) Patients

5.1 Introduction

Heart failure (HF) has been one of the growing epidemics in North America. Over 10 % of people suffer from HF in the U.S. (ACC, 2017), and currently around 20 % of Canadian population live with HF, and 50 thousand Canadians are diagnosed with HF each year (Heart & Stroke Foundation, 2017). With over 1 million HF patients hospitalized each year, HF is the single leading factor of hospitalization in U.S. (Nieminen and Harjola, 2005; Ross et al., 2006). Given the ongoing trend, by 2030 over 10 thousand more Canadians are projected to live with HF compared with 2013 (Tran et al., 2016), and 25% more HF patients in the U.S. (Heidenreich et al., 2011; Roger et al., 2011). HF has also exposed a heavy economic burden on healthcare system. Indeed, Heart & Stroke Foundation (2017) estimated that direct costs relevant with HF are over $2.8 billion annually in Canada, whereas in the U.S. the direct costs of treating HF are $34 billion per year, most of which is due to expensive hospitalization (Feng et al., 2008). Moreover, HF is expected to cost the U.S. health system $70 billion by 2030 (Collines et al., 2015).

Acute Decompensated Heart Failure (ADHF) is defined as "the sudden or gradual onset of the signs or symptoms of heart failure requiring unplanned office visits, emergency room visits, or hospitalization" (Joseph et al., 2009). It is among one of major factors of Emergency Department (ED) visits. According to clinical guidelines, ADHF patients with worsening clinical conditions are recommended to hospitalization. Currently, most of these patients are admitted due to the uncertainty of post-discharge events, including morbidity, mortality and re-admission (Collins et al., 2013). It is because early discharge of HF patients can result in a high chance of mortality and/or re-admission. In fact, 33% of these patients were dead or re-hospitalized within 60-90 days after early discharge from ED (Gheorghiade et al., 2006; Setoguchi et al., 2007). According to Weintraub et al. (2010) 10% to 20% of ED visits are discharged home directly, while they have 20% to 30% higher chance of post-discharge events. Ironically, hospitalization of HF patients is not proven to be the better way to reduce the likelihood of post-discharge events (Gheorghiade et al., 2005; Setoguchi et al., 2007).

A prospective cohort study by Smith et al. (2002) demonstrated that ED physicians tend to overestimate significantly the severe complication incident of ADHF patients, resulting in over-utilization of scarce healthcare resources. Although most patients have complex medical comorbidities, they do not demand an acute intervention beyond decongestion or intense monitoring as in hospital wards or ICU Collins et al. (2013). There are typically three types of ADHF patients.

- **Low risk patients** who respond to the initial therapy, and return to baseline. They can be discharged after a brief period of observation;
- **Intermediate-risk patients** who are partially responsive to the treatment with no high-risk features developed. They require continuous treatment and observation, consisting of inexpensive tests, acute therapy and an effective care transition; rather than inpatient admission(Peacock et al., 2010);
- **High-risk patients** who develop a worsening clinical feature, including continuous symptoms, worsening renal function, hypotension or an elevated troponin. They require prompt inpatient admission and/or further intensive care.

The purpose of this paper is to design an Observation Unit (OU) for ADHF patients in order to improve the quality of care without increasing relevant costs. OU, also called Short Stay Unit, Clinical Decision Unit, Chest Pain Unit, Rapid Diagnosis and Treatment Unit (Ross et al., 2012), was originated in 1960 (Gururaj et al., 1972). A dedicated OU has the following features.

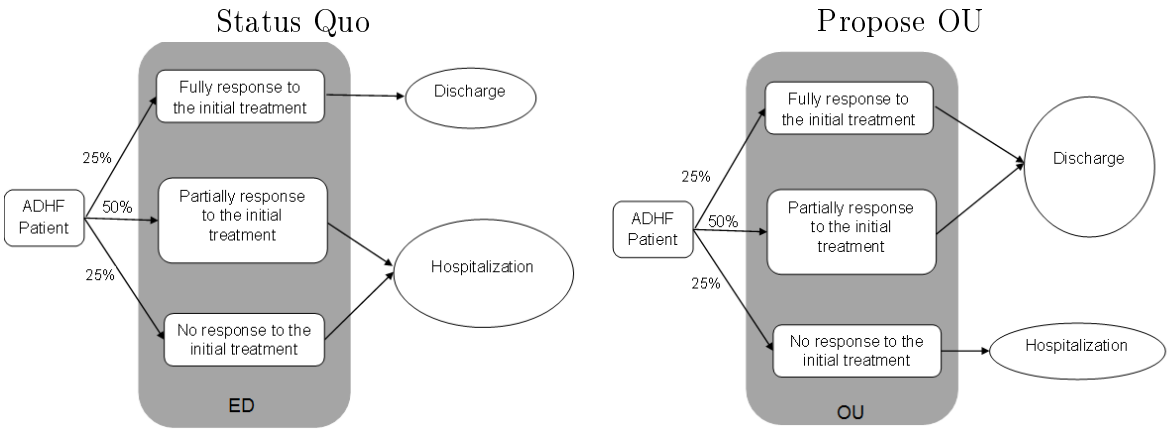
- A typical stay lasts less than 24 hours, no longer than 48 hours;
- A discharge rate is generally around 70-80 % (Hostetler et al., 2002; Mace et al., 2003; Ross et al., 2003);
- A better utilization of healthcare resources (McDermott et al., 1997; Mace, 2001; Goodacre et al., 2004).

Due to the direct access to proper treatment, clinical tests and education in an OU for ADHF patients, this dedicated OU is considered optimal (Ross et al., 2012),

without a doubt very promising. Indeed, a typical stay of less than 24 hours is sufficient to identify and treat low risk and intermediate risk patients with HF. They can be discharged home without being exposed to increasing post-discharge events. A preliminary study by Kosowsky et al. (2000) showed that most low-risk ADHF patients could see clinical improvements within 6 to 12 hours after their arrivals to ED. On the other hand, high-risk HF patients should be hospitalized. Moreover, 75% of OU admitted profiles can have response to treatment without development of worsening high-risk features, and can be discharged with a satisfactory follow-up plan (Collins et al., 2013).

In an ED without OU, low-risk patients are discharged early without sufficient observation; which leads to a high chance of post-discharge events. Due to the conservative perspective, intermediate-risk patients are admitted to hospital, similarly as high-risk profiles, resulting in a waste of inpatient wards, because these intermediate-risk patients do not need intensive inpatient care. If an OU is installed, all ADHF patients are observed in OU after initial treatment. Low and intermediate-risk patients can be discharged after all conditions become stable, reducing the incidence of possible post-discharge events. High-risk patients can be identified and admitted sequentially. Figure 5-1 compares ADHF patient flow between an ED without OU versus an ED with OU.

Figure 5-1: ADHF Patient Flow in ED



Collins et al. (2013) estimated that 50% of HF patients can be discharged after a short period in OU, which leads to decreased unnecessary admissions and reduced post-discharge events, and potentially leads to 1.2 million inpatient days and over $1.2 billion savings in U.S. per year. Controlled experiments conducted by Peacock et al. (2002) demonstrated that an effective OU management protocol of ADHF patients can reduce emergency department visits and re-admission rates in a 90 days post-discharge horizon. A sequential group design study conducted by Storrow et al. (2005) found out that ADHF patients in OU showed a decreased re-admission, fewer repeated ED visits and lower total costs compared with their peers admitted to hospital in a 30-day study window. Collins et al. (2009) conducted a cost-effective study of non-high-risk HF patients, and concluded that OU is more cost-effective than ED discharge regarding those with low or intermediate HF patients, taking into account the post-discharge events. Therefore OU of ADHF can be seen as the "safety net" of ED (Ross et al., 2012).

The goal of ADHF-dedicated OU is to combine treatment and risk-stratification simultaneously, after the initial evaluation and therapy in ED, which is the typical entry point for OU admission. From clinical or medicine perspective, an OU for ADHF patients is required to fulfill the following tasks summarized in Collins et al. (2013).

1. complete initial therapy or treatment for every patient, and allow them to have access to complete resolution within 24 hours;

2. facilitate monitoring of blood pressure, heart rate, urine output, body weight and other bio-chemical index;

3. provide patients with access to simple diagnostic testing, such as electrolyte testing, echocardiography, B-type natriuretic peptide (BNP) or N-terminal pro-B-type natriuretic peptide, and serial troponin measurement;

4. enable patient education and scheduling outpatient follow-up, which is believed crucial in avoiding re-admission by American College of Cardiology and American Heart Association.

All the above can be achieved with relatively less complex and more economic OU rather than resource intensive inpatient wards. However, in Canada, an ADHF dedicated OU has not been widely set up. The present work examines the optimal operational design of an ADHF dedicated OU, incorporating both quality and economic objectives.

Given a certain patient volumes and arrival patterns from a historical dataset, we provide theoretical quantitative decision support for the capacity of an ADHF dedicated OU. That is, we consider several stylized models to figure out the optimal amount of beds to install in a certain OU, satisfying a certain level of utilization rate (the fraction of time a bed is occupied) and loss rate (the proportion of patients being lost due to the full capacity of the OU).

Beyond the capacity design stage, we further consider possible admission and discharge policies for the OU with fixed capacity. We expect that the interactive admission and discharge policies could lead to both better healthcare outcomes of ADHF patients and simultaneously increase efficiencies and cost-effectiveness in the use of the limited OU resources. More formally, supported by a reliable clinical indicator, our goal is to admit patients with the highest uncertainties of risk levels to the OU, so that the limited resources can be optimally used for the purpose of risk stratification. Indeed, the relatively more apparent low-risk patients can be discharged from ED directly, and the more certain high-risk ones should be admitted to hospital wards without occupying OU beds. In the situation that a new patient requires admission to a fully occupied OU, we consider different sorts of possible early discharge criteria, and eventually figure out the optimal alternatives that results in possibly least likelihood of post-discharge events among those ADHF patients. ADHF patients arrive at the OU at random times; and each patient is assigned a risky score given all their physiological characteristics and demographic information. We assume the risky score of the patient population is uniformly distributed between 0 and 1. The patients with higher risky scores may need to stay longer in OU, and top 25% of higher risky patients need to be hospitalized, even after OU discharge; while 25 % lowest risky patients can possibly be discharged home directly or after a short stay in

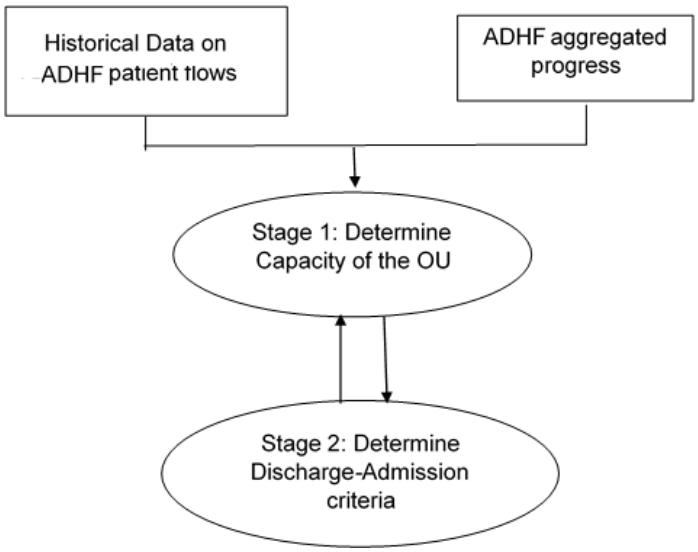
OU. The length of stay (LOS) in OU may follow a general probability distribution. Our data implies that the arrival distribution of ADHF may not follow a homogenous Poisson distribution. Thus our analysis will consider several stylized models to fix the possible range of the OU capacity given the over-dispersed arrivals of patients. This over-dispersion of patient arrivals and generosity of LOS also contribute to the infeasibility of analytically analysis of several admission and discharge policies. Thus we use simulation models to test different admission and discharge alternatives and verify the impacts of combined admission and discharges policies. Our goal is to design a comprehensive quality-guaranteed admission-discharge policy to minimize the chance of post-discharge events among all ADHF patients.

As such, this work provides a systematic framework for the operational design of a prospective ADHF dedicated OU. Specifically, we would like to address the following research questions (1) What is the optimal number of beds in the OU to balance the utilization and loss rate; (2) What is the optimal admission policy of the OU to take more effective use of the limited resources; (3) What is the optimal discharge policies in order to minimize the probability of post-discharge events; (4) What interactive admission-discharge policies work best to enhance the quality of care of HF patients and reduce the economic burden on the healthcare system. This is a data-driven work with the annual ADHF patients who visited the ED of St Mary's Hospital, a local community hospital in Montreal, Quebec. Although this work is designed for the case of ADHF, this systematic framework can be generalized to other applications in healthcare, such as the design of another specific OU or hospital ward.

As shown in Figure 5-2, there are two stages in this work.

The rest of the paper proceeds as follows. Section 2 conducts a literature review of relevant existing works in both clinical and operations management fields. In Section 3, we decide the capacity (i.e. the number of beds) of the specific OU based on historical patient arrival flows. We consider different stylized models and calibrate the possible rank for different levels of patient volumes. We explore various admission and discharge policies in Section 4. Section 5 compares the outcomes of different admission-discharge alternatives using simulation models. We further show that our

Figure 5-2: Road Map of the Proposed Study



proposed interactive admission-discharge policy outperforms a number of alternatives of interest. We conclude and show possible future research in Section 6.

5.2 Literature Review

Our study closely relates to literature on capacity and staff planning in operations management. Recently, more works have been focused on time-dependent arrival distribution, or non-homogenous Poisson arrivals, as homogenous Poisson arrivals rarely exist in reality. Interested readers can refer to Defraeye and Van Nieuwenhuyse (2016) for a literature review on staffing and scheduling problems under non-stationary demand over the period of 1991-2013. The main methods to determine capacity for time-dependent arrivals are Pointwise stationary approximation (PSA), effective arrival rate approximation (EAR), simple peak-hour approximation (SPHA), modified offered load (MOL), infinite server (IS), numerically integrate ODE, stationary backlog-carryover (SBC). One stream of studies propose efficient algorithms to calculate optimal dynamic staffing levels for time-varying arrivals of customers. For instance He et al. (2016) designed an innovative algorithm of staffing for non-Poisson non-stationary arrival process. They detailed the methods via composition

and then extend the algorithms to models with non-exponential service and abandonment where patience time follows non-exponential distribution. Liu and Whitt (2012) developed an algorithm to determine staffing level for time-dependent queues with nonhomogeneous Poisson arrival process and time-varying abandonment probability. Cheng and Huo (2013) conducted a numerical experiment of a time-varying staffing algorithm based on stationary independent period by period (SIPP) approach to set staffing requirement for time-varying cyclic queue $M_t/M/s_t + M$ with abandonment. Originally proposed in Stolletz (2008), and similarly as PSA and its extension like lag PSA, SBC requires to divide long time horizon into small time intervals in the first step, and then incorporates the carryover into the modified arrival rate (MAR) with Erlang-loss models. SBC outperformed PSA regarding the approximation of time-varying queue system. Furthermore, Stolletz and Lagershausen (2013) showed numerically the extension of SBC into more general arrival and processing distribution.

Our work is most relevant with the queueing models involving parameter uncertainty, where the mean and variance of arrival distributions are non-equal, and it can be considered as a special case of time-varying arrival rates. in the following we highlight most recent studies that are most relevant with our work. Bassamboo et al. (2010) found that uncertain parameters such as arrival rates leads to the invalid capacity forecast of traditional square-root safety staffing principle, while an adapted newsvendor model is proven accurate. Kocaga et al. (2015) studied staffing problem with uncertain arrival rates and outsourcing option.

Regarding the application of operations management concepts in the design of OU, the amount of existing study is scarce. Lovejoy and Desmond (2011) used Little's Law and average amount of patients in the system and average length of stay to estimate the capacity of an OU, providing a preliminary and pedagogic example in this context. However, operations management studies on admission and discharge policies in healthcare systems have provided significantly valuable insights. We highlight the most relevant work here. Shmueli et al. (2003) has been one of the most influential papers in terms of admission strategies in the healthcare domain. They proposed

to admit patients whose benefits from being admitted to Intensive Care Unit (ICU) exceed from a certain hurdle, in order to take better advantage of scarce health care resources. From the perspective of discharge policies, Chan et al. (2012) compared several major discharge strategies in the setting of ICU, and proposed a ratio-like discharge policy to minimize the readmission risk involved in early-discharging ICU patients. More recently, Mallor et al. (2015) studied comprehensively potential discharge strategies of ICU, aiming to minimize the rate of patient rejection, and thus improving the accessibility of ICU resources.

We derive the aggregated progress of ADHF from clinic and medical literature. It is feasible to identify high risk HF patients in OU (Collines et al., 2015). Actually, a study of Diercks et al. (2006) investigated the potential indicators that can help to define low-risk ADHF patients in OU. Moreover, evidence and consensus-based OU guidelines have been published by the Society for Cardiovascular Patient Care Peacock et al. (2009). ADHF patients require in-time assessment and proper therapy. Feasible and practical guidelines are discussed in Michota and Amin (2008). However, the OU management of ADHF patients largely depend on the development of newer treatment, innovative drugs and devices (Qureshi et al., 2015). Graff et al. (1999) studied the selective admission criteria for HF patients regarding mortality rate via a retrospective observational cohort study. Later on, Auble et al. (2004) conducted classification trees to identify low-risk patients with HF. The used variables include patients' demographic information, medical history, the most abnormal examination or diagnostic test values. The last two are measured either in ED (vital signs only) or on the first day of hospitalization. They also examined the death rates and readmission rates within 30 days of hospitalization for low-risk patients. The study of Fonarow et al. (2005) showed that routinely available vital signs and laboratory data obtained upon hospitalization can be reliable to identify low-, intermediate- and high-risk of ADHF patients in terms of mortality. Recently, Schrager et al. (2013) demonstrated that ADHF dedicated OU is favorable and proposed accelerated treatment protocols (ATP) driven guidelines. We derive the aggregated progress of ADHF with micro clinical indicators based on Young et al. (2002). Our admission and

discharge criteria largely base on Logeart et al. (2002), who use B-Type Natriuretic Peptide (BNP) as proxy of post discharge events, namely death and re-admission, and they show the association between serial BNP, BNP at discharge and the risk of death or re-admission via univariate Cox Analysis.

5.3 Capacity Design - Analytical Models

In this study, we refer capacity specifically to the number of beds in the prospective OU. We calibrate the possible range of the capacity given a certain level of patient flow in an OU, with various approaches in operations management. We first calculate the number of beds with the most common Square Root Principle. Then we consider Erlang loss model which captures the no-waiting feature of OU. Moreover, we estimate capacity with models incorporating overdispersion features of arrival patterns.

5.3.1 Square Root Principle

We use square root principle, the most common way of capacity decision in operations management, as a benchmark. For a queue with Poisson arrivals, traditional square root principle says that capacity C can be determined by

$$C = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} \tag{5.1}$$

where λ and μ are the average arrival rate and average service rate, respectively. Without a loss of generalization, Eq.5.1 can be written as Eq.5.2 below when service rate is set to unity.

$$C = \rho + \beta\sqrt{\rho} \tag{5.2}$$

where $\rho = \frac{\lambda}{\mu}$ is also called load rate in queueing theory.

5.3.2 Erlang-B type loss model

The Erlang-B type model allows a finite waiting cushion in the queue. Denoted by $M/G/n/n$, Erlang-B type model that depicts an OU in this context, provides a closed formula of resource utilization for a queueing model with Poisson arrivals, a general distribution for service time, a finite amount n of servers and no queue. That is, the amount of customers in the queue cannot exceed the number of servers. Indeed, if an OU is full and a new patient arrives, he or she gets diverted either home or hospital wards as a loss of customers. Later in the paper, we will discuss some alternatives where existing OU patients may possibly be discharged early to make room for new patients. In the context of OU, new patients do not wait in a queue for an available bed.

Bed utilization can be estimated by Eq.5.3.

$$U(n,\rho) = [1 - B(n,\rho)]\frac{\rho}{n}, \tag{5.3}$$

where $B(n,\rho)$ is called Erlang loss function (Erlang B function or blocking probability), is defined as

$$B(n,\rho) = \frac{\rho^n/n!}{\sum_{i=0}^{n} \frac{\rho^i}{i!}}. \tag{5.4}$$

The Erlang-B model features an insensitivity property, which says that the blocking probability is independent of the service-time distribution. It is applicable to the general service time distribution as long as it has a finite mean.

First we show the monotonicity of Erlang-loss function.

Lemma 5.1 (Monotonicity of Erlang-loss function) *Erlang loss function* $B(n,\rho)$

- *is decreasing in* n, $\forall n \in \mathbb{Z}^+$;
- *is increasing in* ρ.

Erlang-loss models explain that the blocking rate increases with smaller capacity,

and larger patient volumes. We then explain the monotonicity of utilization in Erlang loss cases.

Proposition 5.1 (Blocking Rate Determined Capacity) *$\forall \rho$, and a certain block rate b, the capacity n should be determined to satisfy*

$$B(n, \rho) \leq b. \tag{5.5}$$

In the instance we can set up a capacity for the OU to satisfy a certain threshold for patient loss, so that the accessibility of the OU resources can be ensured with a certain amount of OU beds.

Rather than the criteria of accessibility, health care providers also need to consider the expenses of healthcare service. Thus we next discuss **Cost-effective Capacity** from economic perspective, under the framework of Erlang loss models. Let Δ be the dollar amount health quality gain from OU, h dollar amount of losing a patient in the OU, c relevant expenses of an OU bed per unit of time, including nursing and facility costs.

To maximize the total economic gain with a certain capacity m

$$\max_m \{\Delta\lambda[1 - B(m, \rho)] - h\lambda B(m, \rho) - cm\} \tag{5.6}$$

which is equivalent to the minimal of total dollar amount expenses

$$\min_m \{\Delta\lambda B(m, \rho) + h\lambda B(m, \rho) + cm\} \tag{5.7}$$

The following property ensures the existence and uniqueness of the decision on capacity n in the optimization problem 5.7.

Proposition 5.2 (Cost-effective Capacity) *In an Erlang loss model, with Δ dollar amount health quality gain from OU, h dollar amount loss a patient for the OU, c relevant expenses of an OU bed per unit of time, the capacity m should be set to*

satisfy

$$(\Delta + h)\lambda(B(m, \rho) - B(m + 1, \rho)) = c. \tag{5.8}$$

Bassamboo et al. (2010) similarly proposed a newsvender form of an approximate capacity solution for an $M/M/b+M$ queueing model for a doubly stochastic Poisson process with an infinite capacity buffer and customer abandonment. Their proposed approximation is proven to outperform the standard square-root in the queueing model with customer abandonment. Collins et al. (2009) OU admission with a cost-effectiveness ratio of $ 44,249 per quality adjusted life year (QALY) versus $ 684,101 per QALY for hospitalization of non-high-risk HF patients. They consider a time horizon of 30 days. Though in the context of chest pain, Abbass et al. (2015) provided a more rational cost-ratio between hospitalization and OU. It shows that OU is 1.4 to 2 times less costly than inpatient care. The capacity decision based on ratios of OU and hospitalization and corresponding post-discharge event rates are reported in the numerical section 5.5.4.

5.3.3 Overdispersion of Arrival Distribution

Healthcare providers target both economic and quality of healthcare services. This requires the OU to display a Quality-and-Efficiency Driven (QED)regime. This means the queueing system has a large demand and a capacity with controllable idle rate between 0 and 1, and finally a negligible expected delay. Moreover, our empirical analysis on the arrival patterns of ADHF patients show that homogeneous Poisson distribution fails to fit the arrival distribution. Indeed, our patient arrival process incorporates overdispersion. This implies a significant larger variance than the mean of arrival rates.

The arrival process with overdispersion has been commonly treated as a doubly stochastic Poisson process (e.g. Maman, 2009; Mathijsen et al., 2017), which says that the arrival rate is non-homogeneous but follows a certain distribution. And the most popular parametric family of the Poisson rate is the Gamma distribution, re-

sulting in a mixed Poisson-Gamma distribution. Mixed Poisson-Gamma distribution is equivalent to a negative binomial distribution. The following algorithm explains the generation of a random variable of mixed Poisson-Gamma distribution.

Generating Mixed Poisson-Gamma variables

1. Normalized service rate as 1 via scaling arrival time points;
2. Discrete the whole arrival period into equal distance;
3. Estimate the rate Λ from gamma distribution with probability density function $G(a, b)$, where

$$G(a,b) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-\frac{x}{b}}, \quad \forall x \in (0, \infty); \tag{5.9}$$

4. Generate a Poisson variable with the rate Λ.

In terms of capacity decisions with overdispersed arrival distribution, Whitt (2006) is among the first one to propose a capacity decision to deal with overdispersion. It recommended that the capacity should incorporate variance

$$C = \mathbb{E}\Lambda + \beta\sqrt{\mathbb{V}\Lambda + \mathbb{E}\Lambda}. \tag{5.10}$$

Due to the fact that if a random variable X follows a mixed Poisson-Gamma distribution, its mean and variance are

$$\mathbb{E}(X) = \mathbb{E}\Lambda, \quad \mathbb{V}(X) = \mathbb{V}\Lambda + \mathbb{E}\Lambda. \tag{5.11}$$

Later on Maman (2009) proposed that under the assumption of a mixed Poisson-Gamma distributed arrival rate Λ with a mean λ and a standard deviation λ^c, where $0 < c \leq 1$, the capacity should be determined as

$$C = \lambda + \beta\lambda^c. \tag{5.12}$$

Noted that when $c = \frac{1}{2}$, the arrival rates are not overdispersed, then Eq.5.12 becomes the same form as the conventional square-root in Eq.5.2.

More recently, Mathijsen et al. (2017) proposed a capacity decision based on the Gamma distribution. Suppose the arrival rate follows a Gamma distribution with probability density function $G(a, b)$ as Eq.5.9, the capacity should become

$$C = ab + \beta\sqrt{ab(b+1)}. \tag{5.13}$$

We compare numerically the capacity decisions under all above different methods in Section 5.5.4. The above analytical models act as valuable starting point for our OU design, providing more specific ranges of capacity given a certain level of patient flow. Thus these analytical frameworks largely reduce the amount of simulation scenarios that we should conduct for this study.

5.4 Admission and Discharge Policies

After addressing the capacity decision, in this section, we discuss several potential admission and discharge strategies for the prospective OU. First we state the performance measures and specific proxies applied in this work.

5.4.1 Performance Measures

The goal of performance measures is to explicitly quantify the outcomes of different policies. For each policy we construct, we are typically interested in the following two characteristics in healthcare system.

- *Measures of Quality* includes mortality during and after treatment, readmission, and access to healthcare service.
- *Measures of Cost* includes all relevant expenses of treating ADHF patients.

Next, we describe proxies for each performance measure below.

- **Post-Discharge Event (PDE) Rate.** Though there are several possible clinical indicators implying chance of PDEs and risky level of ADHF patients, in terms of complication, re-admission or mortality (e.g.Graff et al., 1999; Auble et al., 2004). In this study, we choose BNP as the proxy of the likelihood of PDEs, including both mortality and re-admission rate. Indeed, Logeart et al. (2002) demonstrate the close correlation of BNP and the chance of PDEs. Specifically, higher BNP levels indicate a riskier case where the likelihood of mortality or re-admission can be higher than an individual with a lower BNP level. It also provides the aggregate time-dependent progress of ADHF patients' BNP levels in OU. So we choose BNP as a proxy due to its feasibility. However, as the clinical research goes on, there might be other indicators to apply in the future.

- **Block or Loss rate.** It measures accessibility of health care service, which aligns with the quality goal of health care service Chan et al. (2012). It is essential for an OU service to ensure equitable and maximal access for ADHF patients.

- **Hospitalization rate.** The economic feature of OU requires it to contribute to the reduction of overall ADHF patients' hospitalization rate given its diagnosis, treatment and risk-stratification features, resulting in the a lower cost for healthcare system.

- **Cost-Gain Ratio.** It measures the cost-effectiveness - the dollar amount of clinical benefits that one extra bed makes, while of course ensuring no sacrifice in terms of quality of care.

5.4.2 Admission Policies

It is possible to distinguish high-risk ADHF patients in ED from several clinical features, such as positive cardiac biomarkers, new ischemic electrocardiogram changes and certain ranges of systolic blood pressure, serum sodium, blood urea nitrogen (BUN) and creatinine (Collins et al., 2009). Those patients are required to have

inpatient care from clinically perspective, no matter through an OU or not. In the meanwhile, those low-risk patients can also be distinguished and may be discharged home without an OU admission as well. We investigate the possibility of setting a hurdle for OU admission so that only patients whose risk level is difficult to identify can be admitted to OU. We compare this hurdle strategy (FCFS-H) with the general admission strategy where every ADHF patient is admitted to OU as long as there is an available bed, similarly as in the ICU setting (Shmueli et al., 2003). These strategies are described below.

- **First Come First Serve (FCFS).** Under this policy, all ADHF patients are admitted to OU as long as a bed is available. New patients are rejected if all OU beds are occupied.

- **First Come First Serve with a Hurdle (FCFS-H).** Under this policy, only patients with moderate risks are admitted to OU if there is a bed available in OU. New patients are rejected if all OU beds are occupied. Low risk patients are discharged home directly without being admitted to OU, and patients with highest risk are admitted to inpatient ward directly without going through OU either.

5.4.3 Discharge Policies

When a new patient arrives at the OU with all beds occupied, discharge strategies provide guidelines to either early discharge an existing patient or reject the new patient. We consider the following potential discharge policies in the setting of an ADHF-dedicated OU.

- **Discharge before access.** New patients are always rejected if all OU beds are occupied under this policy.

- **At random.** Anyone of the new and all existing patients is selected at random to be discharged from the OU. This discharge rule has been considered as a benchmark in the setting of ICU (Mallor et al., 2015).

- **Longest time in service.** Existing patient who stays in the OU the longest will be discharged if a new patient arrives.
- **Shortest service time remaining.** Existing patient with the shortest remaining time in the OU will be discharged once a new patient arrives.
- **Likelihood of PDEs.** Existing patient whose chance of post-discharge events will be discharged once a new patient arrives. Similarly as Chan et al. (2012) who considers readmission risk with a crude metric of the likelihood of readmission, we measure the chance of post-discharge event with the BNP level, and discharge the existing patient with the lowest BNP level at the time when a new patient arrives. With the medical findings in Logeart et al. (2002), we linearly interpolate 30-day PDE chance between 10 % and 25 % to the BNP level between 350 to 700 ng/l.

Longest time in service and shortest service time remaining are both service time related metrics, similarly as Mallor et al. (2015) which considered the various form of service time related discharge rule in ICU setting.

5.5 Data Analysis and Parameter Estimation

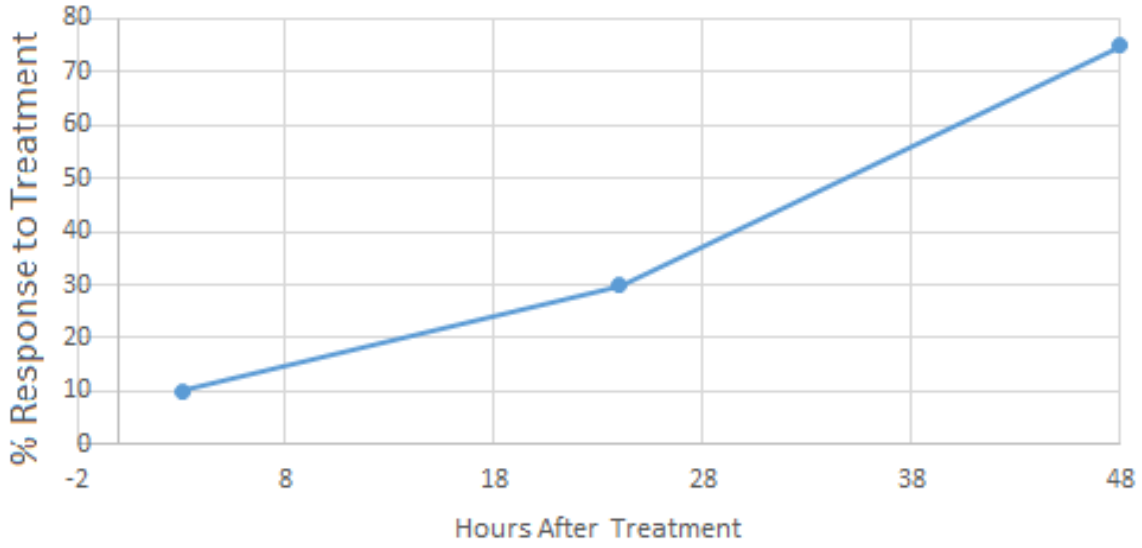
We compare the chance of post discharge events for existing patients and the probability of post discharge events for new patients if they are discharged early. All the probability can be estimated with sequential BNP and basic demographic information as in Logeart et al. (2002).

5.5.1 Cost Data

A recent research of Abbass et al. (2015) uses the claims data and shows that the inpatient admissions were between 1.4 to 2.2 times more costly than OU after adjusting for baseline characteristics, risk scores and diagnosis at discharge.

In order to compare the cost of OU and inpatient wards, we use the costs in Collins et al. (2009), where the hospital cost is $5,712 per patient whereas the cost per PDE

Figure 5-3: Proportion of Patients who Response to Treatment Overtime



is $ 4,588 as the weighted average of death and hospitalization cost. And the cost per OU bed is $ 381 per patient in the base case.The higher cost of hospitalization incorporate both longer length of stay and the higher intensity of care than an OU stay. All dollar amount were adjusted for inflation using the Medical Services Price Index for 2012 Abbass et al. (2015).

5.5.2 Service time or Length of Stay (LOS)

The average LOS is 29.8813 hours derived from Logeart et al. (2002). We randomly assign LOS of OU for each ADHF patient, without loss of generosity.

Logeart et al. (2002) provided an aggregate progress of BNP among OU patients. That is, in general, after t hours in OU, the BNP level of an OU patient becomes

$$BNP(t) = BNP_0 - 12t, \tag{5.14}$$

where BNP_0 is the individual's initial BNP level upon arrival in ED.

5.5.3 Arrival rate

Our data includes 1645 ADHF patients who visit ED of St Mary's Hospital from April 4, 2011 to December 16, 2015. 74.47% of these ADHF patients are admitted to hospital due to the absence of an OU in the hospital.

Having normalized average service rate into 1, the arrival rates have a mean $\mathbb{E}\Lambda = 1.1920$ and variance $\mathbb{V}\Lambda = 1.4141$. Arrival rates are estimated under different distributions in Table 5.1.

The estimation on varied distributions of arrival patterns confirms the existence of overdispersion, where the ratio of variance and expectation is no longer equal to 1. Thus the statistical significance of Poisson parameter is not as high as the Gamma parameters.

Table 5.1: Arrival Rates Parameters

Distribution	Parameter(s)	Estimation	Significance
Poisson $Poisson(\lambda)$	$\hat{\lambda}$	1.1920	.
Gamma $G(a, b)$	$(\hat{a}, \hat{b})$	(6.2770, 0.1899)	***
Mixed Poisson c	$\hat{c}$	0.9864	***

where . < 1 and *** < 0.001

5.5.4 Capacity Evaluation from Analytical Models

The number of OU beds under different analytical models are presented and compared in Table 5.2. Though the capacity decision does not vary much when the patient flow is low among different analytical models, Square Root approach (Eq. 5.2) and method of Maman (2009) (Eq.5.12) tend to conservatively estimate the amount of beds, resulting in more than 13 beds for a hospital with 5 times our original dataset. Erlang-loss model (Eq. 5.3) confirms that as 13 beds can possibly lead to 0 % blocking, which creates full access of OU for all the patients. While the method of Whitt (2006) (Eq. 5.11) and approach in Mathijsen et al. (2017) (Eq. 5.13) give a moderate estimation of around 10 beds, Erlang loss model (Eq. 5.3) provides a lower bound of 8

beds with a block rate not exceeding 10 % . However, the actual block rate should be higher with 10 beds due to the over-dispersion feature of the actual arrival patterns.

Given our analytical results, the last row of Table 5.2 shows the OU bed ranges to be tested in our simulation models . Our analytical models minimize the amount of simulation scenarios by providing the test range for each hospital scale. This largely improves the efficiency of the subsequent simulation.

Table 5.2: Numerical Results: Number of OU Beds

Patient Flow ρ	0.25	0.5	0.75	1	2	3	4	5
Square Root (Eq. 5.2)	0.68	1.36	2.04	2.72	5.44	8.15	10.87	13.59
10 % Block Rate	2	2	2	3	4	5	7	8
Exact block rate (%)	1.27	4.36	8.43	3.71	6.78	9.02	6.00	7.36
No Block	3	3	4	5	7	9	11	13
Cost Ratio 1.4	2	3	3	4	6	8	10	12
Cost Ratio 2	2	3	4	4	7	9	11	13
Maman (2009) (Eq. 5.12)	0.69	1.36	2.04	2.71	5.40	8.07	10.74	13.41
Whitt (2006) (Eq. 5.11)	1.33	2.06	2.68	3.26	5.31	7.16	8.90	10.58
Mathijsen et al. (2017) (Eq. 5.13)	1.01	1.63	2.19	2.72	4.71	6.61	8.48	10.32
Test Range	0-1	1-2	2-3	2-3	4-6	6-8	8-11	8-13

Service level $\beta = 1.28$. For a given patient flow ρ, the capacity can be calculated by plugging it into the equation of each analytical approach.

5.6 Simulation Models for Admission-Discharge Rules

In order to evaluate the performance of the proposed OU and compare the different admission-discharge policies, we develop different simulation models using ARENA software (Rockwell). Recently, simulation has been extensively used to understand complex systems and predict their behaviors. Furthermore, it is often used to provide decision support when designing new systems Sokolowski and Banks (2009).

In this study, we consider different models. First, a base case model that replicates the current ED without an OU in terms of patient handling policies and patient volumes. Second, we develop a core model for the ED with a proposed OU using

the original patient flow. Next, we extend the core model to incorporate several discharge policies. We test individual admission and discharge policies one-by-one in the different scenarios of varied OU capacity, i.e. the amount of beds. Finally, we test the combination of different admission-discharge policies. We test the interactive performance of combined admission and discharge policies. Detailed explanation will be given in the subsections 5.6.2, 5.6.3 and 5.6.4.

Moreover, we also consider other cases where patient volumes are different from our core model. Specifically, we test smaller scale hospitals where patient volumes are 25 %, 50 % or 75 % of the core model; and larger scale hospitals where patient volumes are two, three, four and five times the original database. In this model, in order to incorporate the overdispersion feature of arrival patterns, the ratio of mean and variance of arrival rates is kept constant. That is,

$$\frac{\mathbb{E}\Lambda}{\mathbb{V}\Lambda} \equiv \text{Constant} \tag{5.15}$$

where the ratio in Eq.5.15 is 1.1863 in the case of normalized arrival rate taking average service rate as unity.

We generate the equal amount of overdispersed arrival rates in all those cases with the Mixed Poisson-Gamma algorithm described in subsection 5.3.3 , and keep the risk score, service time and initial BNP levels the same for all patients as in the core model.

For our base and core model, we consider the following set up: a time horizon ranging from April 4, 2011 to December 16, 2015. In our base case, the average cost is $ 4456.08 per ADHF patient given 75 % of them are hospitalized from the ED without OU. In the core model, admission policy is used once a new patient arrives. Whereas discharge decisions are made once a new patient arrives at the OU if all beds are occupied. Table 5.3 summarizes different simulation models and scenarios. Figure 5-4 shows the screen shot of one of our simulating models.

Table 5.3: Summary of Simulation Models and Scenarios

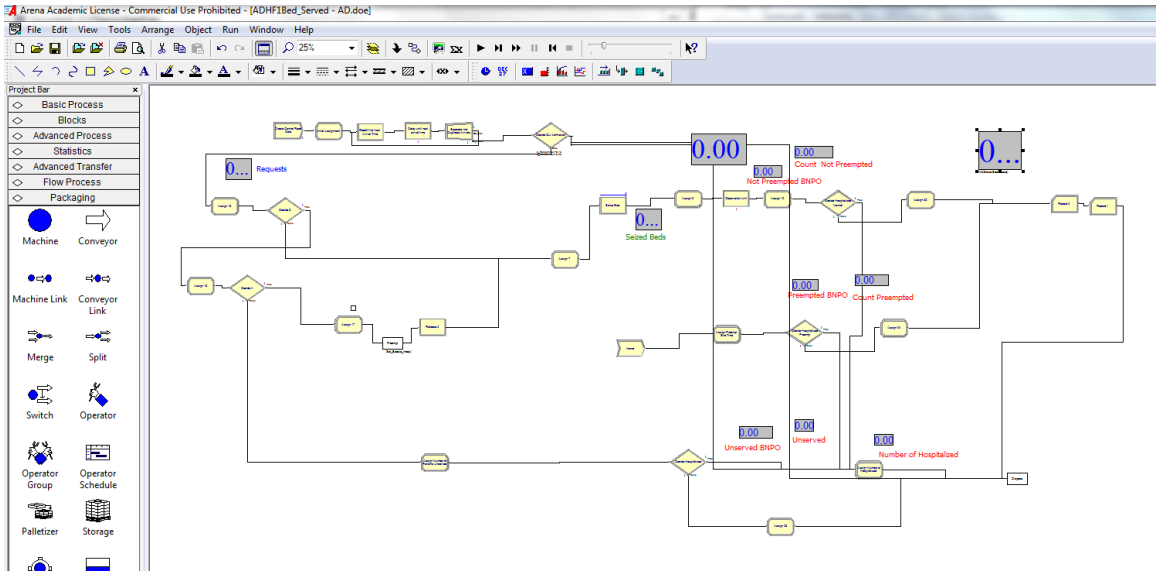
Model	Patient Flow	Structure	Results
Base case	Current $\rho = 1$	without OU	
Core	Current $\rho = 1$	with OU Scenarios: - Capacity design validation - Discharge policies - Admission-Discharge Policies	 Table 5.4 Table 5.6 Table 5.7
Smaller Hospital	$\rho < 1$	with OU Scenarios: - Capacity design validation	 Table 5.4
Larger Hospital	$\rho = 2, 3, 4, 5$	with OU Scenarios: - Capacity design validation - Admission-Discharge Policies	 Table 5.5 Table 5.9

5.6.1 Capacity Design Validation

In Table 5.4, we present the impacts of different capacity decisions on small hospitals, whose patient flows are lower than the hospital under analysis in the core model. Considering the results of our analysis, small hospitals may not need an OU. Because, in this case, more beds result in unnecessarily low utilization, while small amount of beds fail to accommodate majority of patients due to the highly dispersed arrival patterns. Moreover, from the cost perspective, the total cost in the small hospitals including hospitalization, treatment of PDEs and OU beds can be even more than the average per patient cost ($4456.08)in the base case.

In Table 5.5, we present the results of various capacity decisions on larger hospitals. The number of beds is positively associated with the accessibility; specifically an OU with more beds has a lower block rates, and consequently a higher proportion of patients are able to access the OU resources. As the accessibility is closely linked to quality of care; in the case of ADHF-dedicated OU, the hospitalization rates decrease as the accessibility of OU improves. This is because larger proportion of patients are able to be treated in OU and thus identified for hospitalization demand if needed. Only those high-risk patients who truly require further inpatient care are admitted.

Figure 5-4: Screen Shot of the Admission-Discharge Model



Whereas, those who are blocked from OU are more likely to be admitted even if they are not that risky enough to require hospitalization; because those patients fail to be discharged home directly for the purpose of high chance of post-discharge events. Given a full OU, they have to be hospitalized directly. Therefore, the high block rates result in high hospitalization rate. However, the utilization of each bed is negatively correlated with the amount of beds. Actually more beds may lead to lower mean utilization of each individual bed.

From conducted analysis, we observe that larger hospitals with higher patient volumes tend to have a higher accessibility than their smaller peers, with the same utilization rates. Table 5.5 summarizes the results from the larger hospital models. For example, this table shows that 2 beds for a hospital with patient volume of 2Λ leads to a utilization of over 58 % but more than 40 % patients are non-served; while in the case of a hospital with a 3Λ patient volume, it needs 4 beds to keep the same utilization rate. Only around one quarter of patients fail to access OU service. In a larger hospital of 4Λ patient volume, 7 beds result in less than 10 % of non-served patients and a slightly lower utilization rate of 56 %. When it comes to the case of 5Λ, the accessibility is around 94 % with 9 beds and similar utilization rates. This

is because of the "economy of scale" in the larger hospitals. Even with the presence of overdispersion in patients' arrival pattern, the larger amount of OU beds, i.e. the larger OU capacity, provides more flexibility to accommodate the highly uncertain patient arrivals. When the smaller amount of beds suffer from fully occupation in the smaller hospital, several extra beds in larger hospitals tend to be available upon the over-dispersed arrival of demand.

Table 5.4: Efficiency Comparison for Small Hospitals

Capacity of 1 OU bed						
Patient Flow	Utilization	Nonserved Proportion	Served PDE Rate	Nonserved PDE Rate	Hospitalized Proportion	Total Cost ($)
0.25	20.75%	23.83%	17.73%	19.38%	34.89%	4348.86
0.5	32.51%	39.39%	17.84%	19.34%	43.77%	4107.86
0.75	40.73%	49.79%	17.71%	19.51%	48.69%	4143.01
1	53.77%	57.26%	17.86%	19.36%	51.85%	4201.83
Capacity of 2 OU beds						
Patient Flow	Utilization	Nonserved Proportion	Served PDE Rate	Nonserved PDE Rate	Hospitalized Proportion	Total Cost ($)
0.25	13.21%	3.22%	17.73%	20.14%	24.62%	5271.61
0.5	23.68%	11.06%	17.75%	19.68%	28.94%	4000.95
0.75	32.76%	20.36%	17.88%	19.21%	32.95%	3731.05
1	44.66%	28.51%	17.56%	19.39%	37.51%	3734.33
Capacity of 3 OU beds						
Patient Flow	Utilization	Nonserved Proportion	Served PDE Rate	Nonserved PDE Rate	Hospitalized Proportion	Total Cost ($)
0.25	9.07%	0.24%	17.75%	2.50%	23.10%	6704.11
0.5	17.34%	2.13%	17.70%	19.91%	24.26%	4485.89
0.75	25.65%	6.75%	17.80%	18.40%	25.78%	3814.96
1	36.61%	12.46%	17.63%	18.68%	29.36%	3635.19

Original dataset. PDE stands for Post-Discharge Event, including complication, death and readmission.

5.6.2 Discharge Policies

Based on the original patient arrival data, we run several scenarios that incorporate different discharge policies as described in Table 5.3. Table 5.6 displays the outcomes

Table 5.5: Upper Bound Capacity Calibration

Patient Flow 2 Λ

Number of Beds	Utilization	Nonserved Proportion	Served PDE Rate	Nonserved PDE Rate	Hospitalized Proportion	Total Cost ($)
2	58.26%	43.83%	17.87%	19.41%	44.98%	3801.58
3	52.07%	24.86%	17.87%	19.66%	35.08%	3415.57
4	45.23%	12.77%	17.81%	19.11%	29.24%	3257.06
5	39.15%	5.47%	17.75%	18.92%	25.78%	3242.35

Patient Flow 3 Λ

Number of Beds	Utilization	Nonserved Proportion	Served PDE Rate	Nonserved PDE Rate	Hospitalized Proportion	Total Cost ($)
4	58.70%	26.69%	17.68%	18.72%	36.60%	3422.33
5	53.30%	16.78%	17.68%	18.42%	31.55%	3254.30
6	48.11%	9.73%	17.74%	18.56%	27.72%	3163.15
7	43.14%	5.29%	17.73%	18.56%	25.41%	3156.09
8	38.81%	2.37%	17.73%	20.12%	24.07%	3207.25

Patient Flow 4 Λ

Number of Beds	Utilization	Nonserved Proportion	Served PDE Rate	Nonserved PDE Rate	Hospitalized Proportion	Total Cost ($)
6	60.49%	15.50%	17.84%	19.16%	30.58%	3145.94
7	55.70%	9.24%	17.80%	19.06%	27.42%	3054.80
8	50.98%	4.92%	17.71%	19.30%	25.59%	3040.30
9	46.45%	2.55%	17.72%	19.93%	24.38%	3065.14
10	42.41%	1.16%	17.74%	22.51%	23.53%	3112.79

Patient Flow 5 Λ

Number of Beds	Utilization	Nonserved Proportion	Served PDE Rate	Nonserved PDE Rate	Hospitalized Proportion	Total Cost ($)
9	53.33%	5.90%	17.77%	19.35%	25.90%	2984.57
10	49.45%	2.92%	17.74%	19.02%	24.44%	2973.80
11	45.62%	1.40%	17.74%	18.70%	23.53%	2996.83
12	42.25%	0.36%	17.75%	18.94%	23.10%	3048.63
13	39.07%	0.12%	16.07%	2.50%	22.98%	3039.86

PDE stands for Post-Discharge Event, including complication, death and readmission.

of these scenarios and compares them.

The random policy acts as the benchmark in comparison, due to the non-existence of this type of OU yet. Basically, outputs under the Random policy provide the worst possible results, because patients are randomly picked to be prompted out without any clinical reason.

Compared to Random policy, while the minimized PDE policy increases the utilization rate of each OU bed, it does not significantly reduce hospitalization rate. Moreover, it has the highest proportion of non-served patients, i.e. the worst accessibility. As the accessibility is largely linked to the quality of care, the amount of PDE actually is high under this policy despite its name and original intention. The overall cost, including treatment of PDEs and hospitalization is highest under this policy.

Both policies based on service time outperform the other two in terms of high utilization of OU beds and low hospitalization rate. Yet these two policies feature differently. The strategy of discharging patients with shortest remaining service time results in highest utilization among all the policies. However, this policy results in a significantly higher chance of preempting patients compared with minimized PDE and Random policy. The high frequency of interrupting OU treatment can lead to potential hazard of patients' recovery, and it contributes to the relatively higher hospitalization rate than the longest service time policy. Another drawback of this policy lies on its feasibility. Because the OU operator makes the decision of selecting the patient with shortest remaining service time. The accuracy of this remaining service time forecast depends completely on clinical judgement of OU physicians and nurses. Therefore, the implementation may be difficult because the prediction of the patients' response and service time is still challenging so far.

The strategy of discharging patients who have already been treated in OU for the longest time is very straightforward to implement. Moreover, it results in the lowest hospitalization rate among all the policies, indicating the quality of care is ensured in this perspective, and largely alleviate the burden of hospital wards. Despite the fact that it can potentially save more expenses from hospitalization, it results in a moderately high utilization of OU resources. This policy features 100 % accessibility,

in the sense that all new patients are admitted to the OU, since obviously they have the shortest service time compared to any existing patient. In this strategy, the only chance that new patients are rejected occurs when more than one patients arrive at the same time and all OU beds are occupied, which is very rare - less than 6 % in the one-bed scenario, and almost none in the scenario of two and three beds. However, it expose the highest preempting chance for existing patients, specifically it preempts almost twice of patients in the 1 bed scenario, and 1.6 times in the scenarios of 2 and 3 beds, compared with the Random policy. Although this policy generates the highly frequent preempting events, the amount of patients who complete OU treatment is higher than shortest remaining service time, and still moderately comparative with minimized PDE and Random policies. Given the lowest overall expenses, this discharge policy is considered the optimal among all the discharge strategies.

5.6.3 Admission Policies

Based on the original patient arrival patterns in the core model, we test two alternative admission policies in separate scenarios. These policies are tested in four different configurations of discharge policies, leading to a total of 20 scenarios. Table 5.7 summarizes the results of these scenarios. For each couple of (admission, discharge) policy, we use a hurdle rate that indicates the percentage of lowest risk cohorts who are directly discharged home from ED, and the same percentage is used for highest risk patients who are admitted to hospital from ED. Take the first row in Table 5.7 for instance, under the Discharge before access policy, given a hurdle rate of 5 %, corresponding to 5 % highest risk patients that are directly admitted to hospital and 5 % lowest risk patients that are discharged home. In this case, the utilization rate is 51.75 %, the percentage of non-served patients is 50.21 % with an average PDE rate of 20.50 %. One can note that this policy does not preempt patients. The percentage of patients who complete their OU treatment is 40.73 % with an average PDE rate of 18.06 %. The overall hospitalization rate is 51.06 %. The total cost is 4142.09 dollars, including hospitalization and treatment of PDEs.

Table 5.6: OU Discharge Policies Comparison

Compare Discharge Policies under 1 Bed									
Policy	Utilization	Nonserved	PDE Rate	Preempt	PDE Rate	Normal	PDE Rate	Hospitalized	Cost ($)
Minimized PDE	65.28%	37.08%	19.10%	30.88%	28.73%	32.04%	19.19%	43.28%	3867.48
Longest Service	65.31%	5.78%	19.92%	62.25%	28.57%	31.98%	15.48%	25.71%	2945.85
Shortest Remain	68.68%	21.03%	18.26%	50.03%	28.81%	28.94%	17.59%	31.73%	3264.79
Random	60.59%	30.70%	19.52%	32.71%	29.14%	36.60%	16.46%	38.36%	3560.64
Compare Discharge Policies under 2 Beds									
Policy	Utilization	Nonserved	PDE Rate	Preempt	PDE Rate	Normal	PDE Rate	Hospitalized	Cost ($)
Minimized PDE	50.44%	12.28%	18.85%	23.34%	26.33%	64.38%	18.52%	27.36%	3259.95
Longest Service	50.56%	0.43%	17.00%	35.50%	27.86%	64.07%	16.57%	23.04%	3022.35
Shortest Remain	51.87%	4.62%	17.14%	33.13%	26.16%	62.25%	17.61%	23.83%	3060.16
Random	47.77%	10.15%	19.42%	21.88%	28.31%	67.96%	17.24%	27.72%	3257.71
Compare Discharge Policies under 3 Beds									
Policy	Utilization	Nonserved	PDE Rate	Preempt	PDE Rate	Normal	PDE Rate	Hospitalized	Cost ($)
Minimized PDE	38.75%	4.50%	18.46%	11.00%	24.91%	84.50%	18.07%	24.44%	3403.21
Longest Service	39.02%	0.00%	2.50%	15.68%	26.34%	84.32%	17.42%	22.92%	3315.53
Shortest Remain	39.37%	1.34%	16.34%	15.74%	24.56%	82.92%	17.74%	22.92%	3314.57
Random	37.83%	3.89%	19.46%	9.97%	27.67%	86.14%	17.55%	24.98%	3425.27

Original dataset. PDE stands for Post-Discharge Event, including complication, death and readmission. This is for Discharge policy only, the OU admits all ADHF patients. To illustrate the result, we take the 1st row for instance: under the minimized PDE policy with 1 bed capacity, the utilization rate is 65.28 % per bed. The proportion of non-served patients is 37.08 % with an average PDE rate of 19.10 %. The proportion of preempted patients is 30.88 % with an average PDE rate of 28.73 %. The proportion of patients who complete their OU stay is 32.04 % with an average PDE rate of 19.19 %. The overall hospitalization rate is 43.28 %. The total cost is 3867. 8 dollars, including hospitalization and treatment of PDEs.

For the illustrative purpose, we show different scenarios setting the same hurdle rate for both home discharge and hospitalization. In reality, we can definitely implement different hurdle rates according to clinical guidelines and economic constrains.

Opposed to the discharge before access policy where every ADHF patient is admitted in OU, FCFS-H strategy admits only those intermediate patients whose risk level are not that apparent to stratify based on the short period in ED. Assuming it is possible to identify those high risk patients who need to be hospitalized with their clinical information in ED, this cohort will be admitted to inpatient wards in ED without going through OU under FCFS-H admission strategy. The batch of patients with lowest risk are discharged home directly from ED assuming their conditions are already stable and PDEs are under control. More specifically, direct home discharge will not expose a higher chance of PDE than discharge after admitting in OU. Therefore, given a fixed amount of patient volume, the demand of OU beds decreases as the amount of patients admitted in OU is less. Moreover, as the hurdle increases, the amount of patients gets smaller. It is because a higher hurdle rate screens out higher percentage of lowest risky patients and the same amount of highest risk ones. As its original desire in Shmueli et al. (2003), this policy makes room of scarce resources for those who benefit most; specifically in the context of ADHF, patients with intermediate complications can have access to OU and one can identify their true risk level, which is impossible with the short stay in ED. As a result, those with identified need for hospitalization get admitted to inpatient wards after OU, and those with positive response to OU treatment are discharged home.

Consequently, the increasing hurdle rate reduces the utilization of the same amount of OU beds, resulting from the reduced demand of OU beds. Although the proportion of preempting and nonserved patients decreases, implying a higher accessibility, there is a higher chance of PDEs for those nonserved and preempted patients. This is because the nonserved and preempted patients tend to be more complicated as hurdle rate increases - the lowest risky patients are no longer included, and mitigate the risk-pooling effects, so the complications of OU patients are higher, resulting in higher PDE rates for both non-served and preempted patients.

Table 5.7: OU Admission Policies Comparison

Discharge Before Access Policy

Hurdle rate (%)	Utilization	Nonserved	PDE Rate	Preempt	PDE Rate	Normal	PDE Rate	Hospitalized	Cost ($)
5	51.75%	50.21%	20.50%			40.73%	18.06%	51.06%	4142.09
10	48.94%	42.67%	21.76%			37.75%	18.79%	49.48%	4028.01
15	46.05%	35.56%	22.87%			35.20%	19.48%	47.84%	3904.70
20	43.33%	28.57%	24.43%			32.89%	20.21%	46.57%	3803.81
25	39.48%	20.91%	25.80%			30.03%	20.88%	44.50%	3630.16

Minimal Post-discharge Event Policy

Hurdle rate (%)	Utilization	Nonserved	PDE Rate	Preempt	PDE Rate	Normal	PDE Rate	Hospitalized	Cost ($)
5	63.22%	30.15%	20.32%	29.54%	28.79%	31.25%	19.78%	40.73%	3697.01
10	59.56%	23.89%	21.64%	26.63%	29.43%	29.91%	20.52%	38.18%	3508.87
15	56.09%	17.69%	22.84%	24.68%	29.80%	28.39%	21.27%	35.14%	3291.19
20	52.36%	12.34%	24.45%	21.64%	29.78%	27.48%	21.80%	33.01%	3113.21
25	47.23%	7.36%	25.83%	17.51%	29.56%	26.08%	22.14%	30.88%	2906.62

Longest Service Time Discharge Policy

Hurdle rate (%)	Utilization	Nonserved	PDE Rate	Preempt	PDE Rate	Normal	PDE Rate	Hospitalized	Cost ($)
5	64.00%	5.05%	20.12%	55.56%	28.94%	30.33%	16.21%	25.65%	2890.71
10	61.20%	3.71%	21.56%	47.48%	29.62%	29.24%	17.15%	25.23%	2802.92
15	57.58%	2.92%	23.98%	40.06%	30.15%	27.78%	18.30%	25.05%	2734.57
20	53.72%	2.07%	24.09%	32.40%	30.49%	26.99%	19.77%	24.56%	2642.61
25	48.07%	1.16%	2.50%	24.13%	30.74%	25.65%	20.79%	25.05%	2570.16

Original dataset. PDE stands for Post-Discharge Event, including complication, death and readmission.

Table 5.8: OU Admission Policies Comparison Continued

Shortest Remaining Service Time Discharge Policy

Hurdle rate (%)	Utilization	Nonserved	PDE Rate	Preempt	PDE Rate	Normal	PDE Rate	Hospitalized	Cost ($)
5	66.14%	16.11%	19.57%	46.50%	28.89%	28.33%	18.05%	30.70%	3164.76
10	62.44%	11.91%	21.10%	40.55%	29.38%	27.96%	18.54%	29.73%	3047.68
15	58.46%	7.84%	22.35%	35.87%	29.91%	27.05%	19.31%	28.09%	2900.87
20	54.16%	4.80%	24.29%	30.09%	30.18%	26.57%	20.34%	26.99%	2778.61
25	48.27%	2.86%	25.77%	22.55%	30.23%	25.53%	21.27%	26.69%	2673.26

Original dataset. PDE stands for Post-Discharge Event, including complication, death and readmission.

The quality of health care is not compromised as the hurdle rates increase, because the overall costs are decreasing with the increase of hurdle rates. This ia an indication of the overall lower hospitalization and PDEs under FCFS-H strategy. We showcase the advantages of the FCFS-H admission policy from the perspective of streamlining and operations. However, the determination of the exact hurdle rate and hence the ultimate implementation of FCFS-H strategy should largely depend on the advancement of ADHF diagnosis and development of proper clinical guidelines, which ensure a highly accurate risk identification of ADHF patients in ED.

5.6.4 Interaction of Admission-Discharge Policies

We also compare possible combinations of varied discharge policies and different hurdle rates in Table 5.7. Performance of hurdle rates and discharge policies are independent of each other. Higher hurdle rates lead to more accessibility and fewer preempting, and consequently lower overall costs when combined with each discharge policy. Longest service time discharge policy outperforms the rest discharge policies with a higher utilization, higher access rate and lower overall costs in each level of hurdle rates.

With the optimal discharge policy, longest service time strategy, we apply different admission policies for larger hospitals with a patient volume of five times the original data set. The simulation results are presented in Table 5.9. The features of individual hurdle rates keep the same in this case. We can observe that higher hurdle rates enjoy higher accessibility and relatively higher quality of care indicated in the lower overall costs. We also see a much lower preempt rate and significant proportion of patients who complete their OU service in this large OU. Moreover, with the help of optimal discharge policy and the hurdle rates with its consequent lower demand, we figure out the capacity of OU can be smaller than the range calibrated from our analytical models in Table 5.2. In this case, we test the OU with 7 beds, one fewer than the lower bound in Table 5.2. This capacity results in lower overall costs than the ones with larger capacities.

Table 5.9: OU Admission-Discharge Policies Comparison

Number of beds	Hurdle rate (%)	Utilization	Preempt	PDE Rate	Normal	PDE Rate	Hospitalized	Cost ($)
8	0	60.48%	16.53%	25.83%	83.47%	17.15%	22.92%	2771.57
9	0	55.16%	9.06%	27.13%	90.94%	17.41%	22.92%	2833.97
7	0	66.13%	26.08%	26.25%	73.92%	16.71%	22.92%	2723.38
7	5	62.74%	19.70%	26.23%	71.25%	17.39%	22.92%	2682.38
7	10	58.01%	13.13%	26.32%	67.29%	18.28%	22.92%	2634.35
7	15	52.54%	7.17%	27.27%	63.59%	19.15%	22.92%	2594.14
7	20	46.44%	3.40%	27.72%	58.05%	20.21%	22.92%	2561.86
7	25	38.61%	1.03%	28.16%	49.91%	21.00%	22.92%	2508.82

Patient flow is 5 times of original dataset.

5.6.5 Sensitivity Analysis

We have already conducted sensitivity analysis on patient volume in the previous subsection. Here we conduct sensitivity analysis on the cost data. The core models is presented in section 5.5.1. Keeping the other costs constant, the long term PDE cost can increase by 19 times the core model without changing any capacity, discharge or admission suggestions. If the long term cost of treating PDE goes beyond further, FCFS discharge policy becomes the most cost-effectiveness among all discharge policies. With the rest costs fixed, the unit bed cost per patient can increase by 3 times without changing the advantage of OU in all sort of patient volume scenarios. If the OU bed cost increases further, an OU fails to save costs for the health care system by reducing the hospitalization rate. Due to the "economy of scale" in the larger hospital with 5 times of base-case patient volume, the OU cost can increase by 6 times to keep the cost-saving advantage.

5.7 Conclusion and Future Research

This study provides a comprehensive operational framework to install an ADHF dedicated OU, that reduces unnecessary inpatient admission and ensures low cost of ED patient triaging. As the current conservative norm, around 75 % of ADHF patients

are admitted to hospital due to the potentially high risk of complication, death and readmission of early discharge home after a short stay in ED. However, the proportion of ADHF patients who truly demand inpatient care is only 25 %, one third of current hospitalization rate. Therefore, the proposed OU is designed, so that ADHF patients can be treated for no longer than 48 hours in OU, and consequently the likelihood of post-discharge events is largely mitigated. Moreover, their hospitalization needs are identified in OU, ruling out the unnecessary inpatient admission afterwards. Given the fact that the treatment and nursing service are less intense in OU than those in hospital wards, the potential ADHF-dedicated OU features lower economic burden and higher quality of care, and thus is attractive for the care providers. We provide managerial insights of deciding capacity, specifying an optimal discharge policy and interpreting the features of a hurdle involved admission strategy. This is a data-driven work, as the motivation and simulation are rooted in over 1,500 ADHF patients' arrivals to a local community hospital in Montreal.

First, we use multiple models to facilitate capacity decision, trying to accommodate the overdispersed demand arrival pattern, and the type of service system with no waiting space. Although, arrival rates are overdispersed, traditional square-root principle can still be applicable for capacity sizing. However, the service level is no long guaranteed. Erlang-loss model with generalized service rate is still appealing, because it provides a lower bound with a given loss rate in the case of over-dispersed arrival pattern. Erlang-loss model provides an accurate capacity estimation for a larger hospital which has a higher patient volume. It is because larger hospitals with a larger OU capacity provides more flexibility, as more beds can accommodate more patients simultaneously, offsetting the negative impact of arrival over-dispersion. We also consider three innovative methods to handle specifically the over-dispersion of arrival rates, namely Whitt, 2006; Maman, 2009; Mathijsen et al., 2017 where the arrival pattern is modelled as a mixed Poisson process, and the Poisson arrival rate follows a Gamma distribution. These specific methods confirm the feasible range of the amount of beds that should be installed in varied OUs with different demand volumes. These analytical results provide a diminished possibility of capacity , sig-

nificantly enhancing the efficiency of sequential simulation.

Then we investigate the possible discharge policies to decide if and which patients to discharge and make room for new patients when all OU beds are occupied. We evaluate each capacity decision and discharge policy with ARENA simulation software from the perspectives of quality of care and economic burden. We figure out that the accessibility largely impacts the quality of care as even partial OU treatment can possibly reduce the chance of PDEs. While more beds in larger hospitals can provide more leeway to ease the negative impacts on accessibility of overdispersed demand arrivals, OU is not recommended to smaller hospitals as the uncertain patients may not get sufficient OU service and make the utilization of OU resources very limited. The simulation results demonstrate that the policy used to discharge an existing patient who has stayed in the OU for the longest time outperforms the other three discharge alternatives from the perspectives of feasibility, accessibility, quality of care and overall costs. Moreover, this strategy is proven robust in different scenarios with various patient volumes. Therefore it provides a valuable insight to operate the potential OU.

Moreover, we think one more step further ahead and ensure to accommodate the potential emergence of new medical technology. In fact, Abbass et al. (2015) cast doubt about the current appealing of OU to hospitals for reducing unnecessary admission and lowering cost of risk stratifying patients in ED. Indeed, in the future, new imaging tests or chemical indicators may distinguish low and high patients in early stages. We test the FCFS-H admission policy for this ADHF dedicated OU, and demonstrate that a slight downsized OU according to the specific hurdle rate determined by the future diagnosis capability is still appealing to the health care service providers, as FCFS-H strategy is flexible and feasible to accommodate technology advancement in the long run, without any compromise of quality standard or loss of initial investment. This systematic framework can be definitely generalized to other applications.

This work has several limitations. First, we make several assumptions of clinical measures and diagnosis due to the lack of reliable medical contribution. For instance,

we make linear interpolation on the chance of post-discharge events versus the BNP levels, and interpolate the progress of OU patients' average hourly BNP levels. Second, there is few economic data on the costs of ADHF-dedicated OU except Collins et al. (2009). We run sensitivity analysis to demonstrate the robust advantage of designing an OU. Yet a more accurate estimate of ADHF-dedicated OU cost is definitely helpful to gauge the benefits more precisely. Third, we have no individual level information on the progress of ADHF patients' response to treatment, or the individual PDE rate of different risk levels. This preliminary work relies on the aggregate behavior of ADHF patients. However, once those clinical and medical inputs become available, we can incorporate them in our framework and generate more realistic suggestions at ease.

Chapter 6

Conclusion and Future Research

In this dissertation, I conduct data-driven research to address problems in healthcare operations management from the strategic, operational and clinical perspectives and provide valuable managerial insights.

On the strategic level, we propose an incentive based payment scheme to encourage physicians to make decision for the maximal value of patients in Chapter 3. This study demonstrates that proper financial incentives in healthcare system are essential to ensure quality of care and control of expenses. Chapter 4 and 5 mainly contribute to the healthcare decision-making on the operational and clinical levels. In Chapter 4, we analyze and propose a systematic guideline for specialists' response to ED consulting requests with non-homogeneous queueing models; and propose an integrated decision-making linking triage to specialist consulting demands. Our empirical work contributes to clinical decision-making by identifying potential specialist consulting demands with limited information available at the triage stage. In Chapter 5, we propose to set up an ADHF dedicated OU in order to avoid unnecessary hospitalization and post-discharge events for ADHF patients. This potential OU, operated with our proposed capacity and admission-discharge policy, ensures the quality of ADHF treatment in ED without incurring extra costs for healthcare payers.

All these chapters are motived by empirical studies based on medium to large size datasets. Specifically, Chapter 3 is based on over 12 million U.S. individual live birth records from National Bureau of Economic Research; and Chapter 4 and Chapter 5 are based on 40 thousand individual patient visits to the Emergent Department of St Mary's Hospital in Montreal. We use extensive statistical methods to conduct patient clustering from the clinical perspective. Moreover, all our analytical models and proposed strategies are verified with these data sets, which include patient-level information.

There are numerous fields worthwhile for further investigation and exploration in the future. First, it would be of great value to study physicians' behavior, although we assume physicians' diverse behaviours and preferences offset each other with our census data in Chapter 3. However, future research should explore the impact of physicians' behavior on their clinical decision-making. It is also worthwhile to identify

certain measures to quantify their effort. This would contribute significantly to the design of accurate financial incentives to manage physicians' clinical decision-making. We expect that this stream of study largely relies on the decent data source and advanced analytical methods.

Second, to integrate healthcare system is promising in the future. As we demonstrated in Chapter 4, the problem of ED overcrowding actually involves multiple processes in a hospital, for instance, tests, specialists, interface with hospital wards and even community nursing houses. An integrated decision-making can help to improve efficiency in the entire healthcare system. Analytical models like a queue network consisting of multiple tandem queues are potential tools to effectively solve this type of problems. Moreover, advanced queue network models can deal with problems of complicated patient flows, which can involve abandonment, several rounds of tests or specialist requests and re-admission to inpatient wards.

Furthermore, all our chapters consider passive patients who are indifferent with their preference of physicians. However, it is worthwhile to investigate more realistic scenarios where patients are active. Patients can select their preferred physicians and leave the physicians they do not like. Moreover, future research should consider the more realistic case where patients do not perfectly conform to their physicians' decision. Advanced games and contract theory framework are expected to incorporate the interaction between healthcare providers and patients.

In addition, advanced dynamic programming based algorithms should be further developed in order to improve efficiency in complicated patient streaming problems. As the case in our Chapter 4 with multiple patient classes, a dynamic streaming policy is expected to outperform the current one.

Finally, if incorporating more developed clinical inputs, future research would provide more feasible and valuable insights. For example, information of individual records on the progress of ADHF treatment can contribute to the design of OU under the framework in Chapter 5. We expect that statistical learning methods can definitely contribute to this type of research with sufficient amount of reliable clinical data.

Appendix A

Literature Review on Design of Financial Incentives and Payment Schemes in Healthcare Systems

Settings	Payment Mechanisms	Literature
Physicians	Overview, Mixed	(Leger 2008); (Leger 2011); (Robinson 2001); (Lee and Zenios 2012)
	FFS	(Cutler 2002); (Adida et al. 2016)
	Capitation	(Ellis 1998)
	Bundle	(Adida et al. 2016); (Gupta and Mehrotra 2015)
	P4P and OAP	(Fuloria and Zenios 2001); Lee and Zenios (2012); (Shwartz et al. 2016)
	Blended payment	(Chu, Liu et al. 2003); (Sorensen and Grytten 2000); (Adida et al. 2016)
Hospitals	Overview	(McKillop, Pink et al. 2001); (Friesner and Rosenman 2004); (Rosenman and Li 2002); (Sutherland 2011); (Czypionka et al. 2014); (Hua et al. 2016)
	Retrospective payment	(Morey and Dittman 1996); (Nedelea and Fannin 2013)

	Prospective payment	(Ankjær-Jensen, Rosling et al. 2006); (Clement, Grosskopt et al. 1996); (Puenpatom and Rosenman 2008); (Ata, Killaly et al. 2013)
	DRG	(Fetter 1991); (Goldfield 2010); (Sutherland, Hamm et al. 2009); (Dismuke and Sena 1999); (Herwartz and Strumann 2012); (Sharma 2008); (Woodbury, Manton et al. 1993); (Gaal, Stefka et al. 2006); (Epstein and Mason 2006); (Fattore and Torbica 2006); (Bellanger and Tardif 2006); (Schreyögg, Tiemann et al. 2006); (Rouse and Swales 2006); (Shwartz and Lenard 1994)
	Global budget funding	(Peacock and Segal 2000)
	Activity based funding	(Biorn, Hagen et al. 2003); (Sommersguter-Reichmann 2000)
	Internal cost allocation	(Verheyen and Nederstigt 1992); (Verheyen 1998); (Morey and Dittman 1984)
Pharmaceuticals		(Kolassa 1997); (Song and Zipkin 2003); (Chick, Mamani et al. 2008); (Sun, Yang et al. 2009); (Mamani, Chick et al. 2013); (Malvankar-Mehta and Xie 2012); (Zhang, Zaric et al. 2011); Zaric et al. (2013); Mahjoub et al. (2014); Taylor and Xiao (2014); (Levi et al. 2016)

Table A.1: Literature under Category

Authors	Research Goals	Focus	Methodology	Model	Objective Function	Strength	Limitations & Possible extension
(Adida et al. 2016)	Analyze the performance of healthcare providers under FFS and bundle payment; Propose optimal mechanisms to improve their performance.	Physicians or other healthcare providers	Stochastic	Optimization, analytic approach or economic reasoning, principal & agent	Maximize risk-neutral insurers' utility and risk-averse providers' utility of payoff.	It is the first paper to study the financial incentives with analytic models, and most of their findings aligned with proceeding empirical results.	Some of their observations need confirmation from future empirial studies.
(Ankjær-Jensen, Rosling et al. 2006)	Describe and evaluate a case-mix system for Danish hospitals.	Hospital Inpatient & Outpatient	Methodology statement	NA	NA	The paper evaluated this new system in details.	The case-mix system did not work sufficiently well.
(Ata, Killaly et al. 2013)	Propose an optimal alternative policy to stabilize the admission flows and mitigate cream skimming in hospices.	Outpatient	Dynamic	Fluid	Maximize hospice manager's realized revenue.	The proposed mechanism contributed to stabilizing hospice's admission and discharge pattern over the year.	The admission and discharge rates might be different for various diseases, and a new policy should be designed to remedy it.
(Bellanger and Tardif 2006)	Assess the prospective payment system in France.	Hospital inpatient	Methodology statement	NA	NA	It introduced the methods to design a new French prospective payment system.	The impact on waiting time and other important operations may be of interest.
(Biorn, Hagen et al. 2009)	Investigate whether heterogeneity of hospitals affects the effectiveness of activity-based financing	Hospital	Econometrics	NA	Study heterogeneity with respect to hospitals' efficiency before and after ABF.	The study confirmed the activity-based funding had no negative impact for patients to access to different hospitals.	Mortality, equity and other quality factors of hospitals were not studied here, but might be worth investigating.

(Blake and Carter 2002)	Study how to strategically allocate resources in acute care hospitals	Hospital & physician	Deterministic	Linear goal programming	One model sets case mix and volume for physicians holding service costs fixed; the other translates case mix decisions into a commensurate set of practice changes for physicians.	The model allowed explicit allocation decision, and enabled the balance of constrained budgets and maintain normal operations for an institution.	The objective function could incorporate multiple goals for decision makers.
(Blake and Carter 2003)	Investigate how varied physician payment mechanisms interact with different hospital funding policies	Physician & hospital	Deterministic	Linear programming	Minimize weighted deviations from desired economic goals, and minimize the deviations from physician preferred income.	The model incorporated the interaction between hospitals and salaried physicians under multiple budget constrained scenarios.	It explained several payment mechanisms, without recommending an optimal one for each individual settings
(Chick, Mamani et al. 2008)	Study what contract can improve public health cost-effective outcome without sacrificing manufacturers' profits?	Pharmaceuticals	Stochastic	Games theory, supply chain contract	Both government and manufactures aim to minimize net costs.	Their proposed cost-sharing contract successfully motivated both payers and suppliers to achieve global optimization and guarantee the supply of vaccine.	Homogeneous population and epidemic model did not incorporate residual immunity of vaccine. Government may not be able to forecast the demand of specific types and quantity of vaccines. The case of multiple purchasers and suppliers may be of interest too.
(Chu, Liu et al. 2003)	Examine whether a Physician Compensation Program (PCP) can improve efficiency in a large Taiwan teaching hospital	Physician	Empirical	Data Envelopment Analysis, pobit model	Examine and explore the factors that impact hospitals' efficiency.	It confirmed the improvement of hospital efficiency after implementing PCP.	Benchmarking and other productivity measures may be of interest.

(Clement, Grosskopt et al. 1996)	Study what a shadow prices of hospital services are and How big the differences between reimbursement rates and shadow prices are	Inpatient & Outpatient	Deterministic	Nonlinear programming	Calculate shadow prices of hospital services by estimating Shephard-type distance function.	Distance function released the cost minimization assumption of cost function, and allowed retrieval, and can have varied applications.	The shadow prices of each DRG rates are of interest, but rely largely on available disaggregated data.
(Czypionka et al. 2014)	Examine the impact of ownership and financial incentives on hospital efficiency in Australia.	Inpatient	Empirical	Data Envelop Analysis	Calculate the efficiency index for different hospitals given certain indicators	They found the impact of ownership on Australia hospitals, and further found the impact of financial incentives by comparing their results with existing literature.	Quality indicators are worth considering, and relevant dataset is expected to be available.
(Dismuke and Sena 1999)	Examine whether DRG payment influenced the technical efficiency and productivity of diagnostic technologies in Portuguese public hospitals	Inpatient	Empirical	Parametric and non-parametric frontier model	Examine the impact of actual DRG payment on the productivity of diagnostic technology	The work confirmed the positive contribution of DRG to improving productivity and technical efficiency in Portuguese hospitals.	The length of stay, waiting time and other quality factors may be of interest.
(Epstein and Mason 2006)	Describe and evaluate the national cost-per-case tariff system for financing hospitals in England.	Hospital inpatient	Methodology statement	NA	NA	The work detailed the design of cost-per-case tariff system in England and included early assessment.	The variants of hospitals are worth considering. The comparison of different payment systems may be of interest.

(Fattore and Torbica 2006)	Analyze and assess rates of DRG derived from production costs in Italy.	Hospital inpatient	Methodology statement	NA	NA	The work explained how national and regional governments set up the rates, and pointed out several consequential problems.	Empirical study and better rate design should be developed. Comparison among other similar countries may be of interest.
(Fetter 1991)	Describe and evaluate the original DRG setup in the USA.	Inpatient	Methodology statement	NA	NA	Documented the history and evolutions of DRG.	Updated evolution and applications in different settings are of interest.
(Friesner and Rosenman 2004)	Investigate whether providers raise inpatient prices for non-government patients when faced with lower government reimbursement for outpatient services,	Hospital inpatient & outpatient	Empirical	NA	NA	The work concluded government owned hospitals tend not to have cost shifting behaviors.	The quality and efficiency may be of interest when evaluating changes of insurance plans.
(Fuloria and Zenios 2001)	Examine How purchasers can design a payment mechanism that motivates healthcare providers to choose treatment for the purpose of maximizing total social welfare	Physician	Stochastic	Dynamic programming	Maximize the purchaser's expected discounted payoff for each state.	It proposed an optimal payment scheme for providers to pursue long-term benefit for patients.	The patients are assumed passive, and providers are profit maximizers.
(Gaal, Stefka et al. 2006)	Describe and assess the cost methodology and price setting of a DRG system for Hungarian hospitals.	Hospital inpatient	Methodology statement	NA	NA	The work analyzed the issues of implementing DRG systems in Hungarian hospitals.	The impact on efficiency and comparison with other peer countries may be of interest.

(Gupta and Mehrotra 2015)	Evaluate the "bundle payments for care improvement" (BPCI) mechanism with normative models; design and analyze a constrained optimal mechanism.	Hospital inpatient & outpatient	Mathematical	Principal & agent framework; game theory	Payer aims to maximize total expected social benefit	They proposed an optimal uncertain selection mechanism to deal with the uncertain quantity of propsers submitted to the payer.	The framework is limited to the specific initiative and cannot generalize to other settings. They fail to incorporated trial and error in practice when implementing this mechanism.
(Herwartz and Strumann 2012)	Investigate whether prospective payment effectively increases local hospital competition in Germany	Inpatient	Empirical	Stochastic frontier analysis, DEA, spatial regression	Examine the spatial interdependence of hospital efficiency; whether or not the magnitude of negative spatial spillovers of hospital efficiency has increased after DRG reform.	The work concluded the increasing competition among German hospitals due to negative spatial spillovers.	Patient-level efficiency and impact of hospital specialty may be worth further investigating.
(Hua et al. 2016)	Analyze the competition between free public providers and toll private providers. Develpe a coordination mechanism with government policies for both types of providers to maximize social welfare.	Hospital inpatient & outpatient	Mathematic	Mixed duopoly game	Government aims to maximize the social welfare of the public service, while public players target to maximize their utility with capacity constrain, and private ones maximize their profits.	They found a unique Nash equilibrium for the competition of a two-tier system, though it may not outperform one-tier system.	They ignored the real time queue length information, but focused on the long time expected wait time.

(Hutchison, Hurley et al. 2000)	Develop and evaluate alternative methods of adjusting primary medical care capitation payments for variations in relative need for health care among enrolled practice populations.	Physician	Empirical	NA	NA	Needs-based capitation formulae based on socioeconomic and mortality data does not work well	Alternative formulae incorporating varied population should be developed.
(Jiang, Pang et al. 2012)	Study how to set up a unified performance-based contracting framework incorporating patient access-to-care requirements and complex outpatient care dynamics for an online appointment scheduling system	Outpatient	Dynamic	Queue	Minimize service provider's cost to meet certain waiting-time targets.	A linear performance-based contract worked, while a simplified threshold-penalty contract was optimal for dedicated-only patients,	Non-linear performance-based contract would be of interest, and more complex patient mix and dynamics of day-to-day appointment system are worth further investigation.
(Lee and Zenios 2012)	Study what the structural parameters of Medicare's dialysis payment system involving the risk adjustment and the transition toward a pay-for-compliance system are	Physician	Empirical	NA	NA	Pay for risk adjusted downstream would work better to ensure quality care.	It did not accommodate heterogeneous physicians and possible selective mechanism of providers. It assumed providers as profitable maximizers. It is a static model, while providers' dynamic response over the time might be of interest.

(Levi et al 2016)	Investigate the effectiveness of the uniform subsidy mechanism.	Pharmaceutical	Mathematic	Mathematic programming with equilibrium constrains	Central planner aim to maximize the consumption of the malaria drug or social welfare.	They confirmed the optimality and effectiveness of uniform subsidy, also found this mechanism can maximize social welfare as well.	The theoretical bounds on the effectiveness needs to be studied. Alternative policy is worth investigating in the setup with a fixed cost of market entry.
(Mahjoub et al. 2014)	Evaluate the impact of a risk-sharing featured pay-for-payment contract on the drug manufacturers	Pharmaceutical	Stochastic	Markov model, disease progression model	Drug manufacturers aim to maximize their profits	They found manufacturers' profit is not monotonic under this contract.	The perspective of healthcare payers is worth considered; and other risk rather than effectiveness should be studied.
(Malvankar-Mehta and Xie 2012)	Study what incentives should decision makers take in order to encourage optimal strategic allocation of HIV/AIDS prevention resources interactions in a multiple-level resource-allocation setting	Pharmaceutical	Deterministic	Linear programming	Maximize the number of infections averted with a lower level objective to maximize the utility of multiplicative function of equity, efficiency, and funds received.	It showed the potential of multiple level budget allocation taking consideration of lower-level decision makers' preferences.	It may be of interest to incorporate multiple upper-level decision makers. Sequential imperfect games are of interest too.

(Mamani, Chick et al. 2013)	Investigate why the allocation of influenza vaccines is inefficient across borders? What contractual mechanism can alleviate those inefficiencies?	Pharmacy	Stochastic	Game theory	Each government minimizes its perceived total cost of an outbreak; a central planner minimizes the overall financial and health costs of the system as a whole.	The proposed cost-sharing contract were proven to integrate multiple governments' decision making.	The work did not consider manufacturers, nor different social costs in different countries. The source countries can be uncertain and unknown in advance in practice. And the implementation of the contract relied on political ramification.
(Morey and Dittman 1984)	Study how hospital administrators can be helped to meet their profit maximizing and profit satisfying goals	Hospital	Deterministic	Nonlinear programming	Maximize a hospital's real profit by limiting total revenues to some preset targets.	The model helped to explain several tactical issues and departmental cross subsidies.	Multiple (more than two) patient classes could be extended.
(Morey and Dittman 1996)	Study whether an inverse relationship existed between the hospital's inefficiency rating and the extent of its cost pass-through reimbursement	Inpatient	Empirical	Hypothesis testing, DEA	Examine the relationship between hospital's inefficiency rating and its cost pass-through reimbursement.	It proposed a practical provisional rate to motivate hospitals to improve operational efficiency.	It did not consider the regional differences among hospitals. A cost-effective test could be more demonstrative.
(Nedelea and Fannin 2013)	Examine what the effects of environmental variables are on the technical efficiency of the Critical Access Hospital Program	Hospital	Empirical	Data Envelopment Analysis	Estimate technical efficiency scores of CAHs	The methods contributed to statistically efficiency of estimating parameters. And they concluded the CAH had no negative impacts on technical efficiency.	The benchmark and performance measures may be better defined.

(Oliveira and Bevan 2008)	Investigate what models can accurately estimate unavoidable hospital costs? How can a proposed model be used	Inpatient & Outpatient	Stochastic	Stochastic multilevel model	Explore log-linear or a semi-log relationship between the standardized cost and the covariates	This generalized multilevel model is useful for hospitals with different cost systems and geographic settings.	It involved several assumptions on distribution and independence. Panel data may be useful for robust estimation.
(Peacock and Segal 2000)	Analyze factors resulting in failure in Australian health system reform; and assess the methods used in building the national capitation model.	Hospital	Economic analysis	NA	NA	It detailed the specialty of Australian health systems and the possibility of customized capitation model.	An optimal payment scheme for Australia should be developed.
(Puenpatom and Rosenman 2008)	Examine how implementing capitation-based Universal Health Coverage (in Thailand impacts technical efficiency in larger public hospitals during the policy transition period	Inpatient	Empirical	Bootstrap DEA, truncated regression	Compare efficiency before and during the transition period	The work found the increase of efficiency varied depends on geography during transition period.	The benchmark and productivity measures may be worth further investigation.
(Rosenman and Li 2002)	Examine how donations, grants and contracts affect California healthcare clinics' average costs? What are the managerial insights	Outpatient	Empirical	N/A	Study whether or not grants and contracts are targeted money used to enhance quality.	The empirical study suggested the increase of quality of grants was due to grantors' motivation of quality creation.	More and further follow up empirical study may of interest.

(Rouse and Swales 2006)	Analyze model development and application of DEA to set up prices in DRG levels, from both theoretic and political perspectives.	Inpatient	Methodology statement	NA	NA	The paper showed the relative success of transferring theoretic DRG mechanism into practice in New Zealand.	The effective application relied largely on rigor governance and management as well as educating all participants.
(Sanchez-Martinez, Abellan-Perpiinan et al. 2006)	Analyze and evaluate the main costing and pricing (reimbursement) systems employed by hospitals in the Spanish National Health System (NHS).	Inpatient	Methodology statement	NA	NA	The work analyzed the cost accounting methods of allocating budget into cost centers among Spanish hospitals.	Unit cost allocation regarding specific disease treatment may be worth investigation.
(Schreyögg, Tiemann et al. 2006)	Describe the German DRG-system and the methodologies used to determine prices.	Inpatient	Methodology statement	NA	NA	The work analyzed the methods of determining DRG prices in the setting of German hospitals.	The accuracy and availability of reliable cost data need to improve.
(Sharma 2008)	Examine what the resource distribution dynamics are across Diagnosis Related Groups of elective surgery patients, in a continuing Prospective Payment System	Inpatient	Empirical	Stochastic kernel approach	Estimate empirical distributions of Length of Stay.	The work successfully modeled the distribution of length of stay among different DRGs.	The associated tariff calculation, resource allocation and scheduling problems may be worth investigating.

(Shwartz and Lenard 1994)	What are alternative methods for setting prices in hospital prospective payment systems, in order to enhance economic incentives?	Inpatient	Deterministic	Linear programming	Minimize the disincentive index with optimal equilibrium prices	Equilibrium pricing outperformed average cost pricing, contributing to allocate patients to hospitals, and incentive hospitals.	Potential impacts of this pricing mechanism may be of interest, such as waiting time, and other operational impacts.
(Shwartz et al. 2016)	The impact of DEA composite measures on the pay-for-performance scheme	Physicians & Hospitals	Mathematic	Data Envelopment Analysis, Monte Carlo, bootstrapping	Maximize efficiency index	They developed the innovative way of calculating composite measures with DEA.	Guidelines of performance measurement for pay-for-performance are not mature and worth further investigating.
(So and Tang 2000)	Study how the threshold policy would affect the clinic's prescription policy on the drug usage and how it would affect the clinic's profitability, patients' health, and pharmaceutical firm's profitability	Pharmacy & Physician	Stochastic	NA	Maximizes the expected profit for the clinic with an optimal prescription policy	The model was helpful to evaluate costs and benefits of several payment schemes.	Multi-dimension model should be investigated for patients' well-being; a non-linear function or patient-dependent response rate are of interest.
(Sommersguter-Reichmann 2000)	Study how the Austrian hospital financing reform impacts hospital productivity, efficiency, and technology changes	Inpatient & outpatient	Empirical	Data Envelopment Analysis	Calculate input-based Malmquist index, decomposed into different efficiency changes.	The study confirmed the shift of technology while no improvement of technical efficiency yet.	The benchmark and measures of efficiency may be worth further investigating.

(Sorensen and Grytten 2000)	Examine why and how different financing schemes have been used in different types of municipalities	Physician	Empirical	NA	Study physicians' utility that depends on income and leisure time.	The work suggested different optimal should be proposed for primary care physicians in different municipals.	Ethics and preference-based health measures may be worth further investigating.
(Sun, Yang et al. 2009)	Investigate how much of its limited supply of a drug would a country give up to contain a pandemic and how much of the drug would each country keep to guard against possible future infections	Pharmaceutical	Stochastic	Game theory	Minimize the average number of total infections; maximize the probability of no infections.	The work successfully modeled the uncertainties regarding onset and spread of influenza and efficacy of drugs. It proposed a robust agreement among different countries.	The distributions of disease within and between countries should be explicitly modeled. Multi-periods dynamic model of further spread of influenza over a longer time horizon need further study. The dynamic response curve of different population, imperfect information settings and coordinating contract may be worth further investigating.
(Sutherland, Hamm et al. 2009)	Examine how inaccurate comorbidity data affects cost weight values in designing DRGs and what the implications are for hospital payment	Hospital inpatient	Empirical	Bayesian framework, Markov chain Monte Carlo	Estimate probability of misclassification of the true comorbidity level given the reported comorbidity level.	The work showed the importance of incorporating disease severity and comorbidity adjustment to calculate cost weight of DRGs.	The application of the methods on electronic patient reports may be of interest. Interacting of different departments may be modeled with this method.

(Taylor and Xiao 2014)	Recommend donors what to subsize in order to improve the availability and affordability of malaria drugs	Pharmaceutical	Mathematic	Contract theory	Donors aim to maximize purchases, while retailers target to maximize profit.	They studied the subsidy mechanism on the long shelf life product, and explored the impact of shelf life length on the optimal subsidy. The results can be useful to improve customer access to other drugs.	They fail to incorporate any operation factors in the drug supply chain.
(Tiemann and Schreyogg 2012)	Investigate what the effects of privatization are on hospital efficiency in Germany	Inpatient	Empirical	Bootstrap DEA, panel regression	Estimate hospitals' efficiency score.	It found different factors impacting the effect of hospital privatization. It suggest privatization has positive impact on effectiveness of hospitals.	The benchmark and measures of hospital efficiency, productivity and performance may be better defined.
(Verheyen 1998)	Investigate how to resolve the tensions of multiple tasks a professional at the base of a hospital face in the internal budget system	Hospital & university	Economic analysis	NA	NA	The approach was able to integrate external and internal budgeting by aligning input-output decisions.	Auditing and governance might be important when implementing the theory in practice.
(Verheyen and Nederstigt 1992)	Fully analyze a patient-based cost-information system applied to Dutch hospitals.	Inpatient & outpatient	Economic analysis	NA	NA	The work was useful for internal resource allocation.	The data set under study might not have complete information. More accurate estimation relies on the availability of a reliable dataset.

(Woodbury, Manton et al. 1993)	Examine how to allocate global budget to individual hospitals in a DRG system.	Inpatient	Deterministic	Quadratic programming	Minimize square error between predicted and current budgets based on the variation of the frequency of each DRG over MTFs	The model was useful for the purpose of utilization control, specially under the setting of investor-owned hospitals.	Though reflecting marginal cost, zero weight for several DRGs may be difficult to implement in practice.
(Zhang, Zaric et al. 2011)	Study the optimal mechanism for price-volume agreement in the setting of asymmetric information?	Pharmaceuticals	contract theory	Principal-agent	Minimize payer's total cost subject to incentive constraint for suppliers.	They recommended rebate when negotiating price-volume agreement for the sake of compatible incentives	Market decisions of manufacturing and dynamics of contract are worth further investigating.

Table A.2: Taxonomy of the Papers

Appendix B

On Reducing Medically Unnecessary Cesarian Deliveries: The Design of Payment Models for Maternity Care

B.1 Extra Lemmas and Propositions

Proposition B.1 *All physicians will choose SB if overall utility of SB overcomes that of C-section; and they will prefer C-section if the overall utility of C-section overcome that of SB. That is, for a complexity level x_i, all physicians will choose*

- *SB, if $U_{SB}(\lambda, x) \geq U_{CS}(x)$*
- *C-section, if $U_{SB}(\lambda, x) \leq U_{CS}(x)$*

Lemma B.1 *If a reimbursement mechanism $(m_S^C(x), m_S^D(x))$ $\forall S \in \{SB, CS\}$ leads to a consequent threshold of planned CS s, quality of decision increases with respect to his benevolence. That is, the deviation from the clinical cutoff of planned CS $|s - x^*|$ is non-increasing as α increases.*

Lemma B.2 *Under bundled payment where a physician's facility costs dominate the monetary value of the physicians' effort invested in servicing a delivery, $\lambda = 1$ is optimal for physicians in delivery stage after the decision of spontaneous birth.*

B.2 Parameter Estimation

B.2.1 Successful rate of natural Birth $f(x)$

We describe the approach to estimate the rate of NB for those deliveries that SB is prescribed by the end of the prenatal care. For pregnancy complexities below the cutoff point, we calculate the probability of NB by simply dividing the number of NB cases to number of SBs for given pregnancy complexity level of x. Since SB was prescribed for a significantly low number of high risk women, completely different patterns exist for pregnant complexity below and upon the cutoff points. We extrapolate the pattern for clusters of low risks to those of high risks. In this context, we use a polynomial function with power 8 to fit the low risks with complexities below the cutoff point. This function fit has a R^2 of 99.87%, significantly well fitted to our data.

B.2.2 Cost of delivery and postpartum care

We incorporate two main sources of birth delivery for a payer: hospital cost right after birth, and quality of delivery in the long run. According to Canadian Institute for Health Information (2006), the non-complicated delivery and postpartum costs for CS and NB were CAD 4,200 and CAD 2,700 on average across Canada in 2002-2003, while the costs of delivery with complications were CAD 5,200 and CAD 3,200 for CS and NB, respectively. CS patients tend to have a longer stay after birth in hospital wards, and also require more intensive nursing care after the operation. Canadian Institute for Health Information (2006) tells 32% of NB were complicated and complicated CS delivery accounted for 34% in the same period. Because the complications of delivery is independent of pregnancy complexities, or appropriateness of planned CS, we take the weighted average costs for both NBs and CSs as their hospital costs.

In monetary means, quality of birth reflects in long-term health care requirement, such as re-admission, more nursing and community care, and medical and surgical demands, which leads to a greater use of health care resources and consequently higher

health care expenses. Here we use the incidence of postpartum complications as the proxy of birth quality based on the availability of our data set. Indeed, a mother or baby with any of those severe postpartum complications definitely requires more intensive care. We use the rate of CAD 9,700 per birth with postpartum complications, which is reported in Canadian Institute for Health Information (2006) as an average cost per baby admitted to neonatal intensive care unit (NICU) in 2002-2003.

We use Canada's historical annualized inflation rates of past 15 years `http://www.inflation.eu/inflation-rates/canada/historic-inflation/cpi-inflation-canada.aspx` to adjust the prices by increasing 25.13% to the consumption level in 2015 - 2016.

B.2.3 Physicians' Effort

Our empirical study confirmed the desire of leisure and comfort as a driver of emergent CS abuse. Due to the lack of literature on quantification of physicians' effort, we estimate physicians' efforts relevant with delivery by the duration and intensity of delivery modes.

We consider physicians' effort spent in a planned CS as the time and energy invested in an operation. We take the general obstetric consultation rate of CAD 100 in 2015-2016. Because the operation requires more intensive efforts, we double hourly rate for CS as CAD 200.

However, for the births with labor, physicians first spend time monitoring the progress of labor, and decide then a NB or an emergent CS. In general, the average time of labor is approximate 20 hours. We assume physicians spend efforts during certain time of this period, since the nurses and midwifes also take active roles for monitoring the labour. Hence the efforts on monitoring the labour are counted as CAD 400. Hence physicians' effort of serving emergent CS consist of two parts: efforts of monitoring labor and effort of implementing CS; the effort spent on emergent CS should be equivalent to that for planned CS, therefore, the effort for emergent CS is estimated as CAD 600. Similarly, their effort spent on a NB comprises of the part of monitoring labor and the part of assisting NB. Though the intensity of assisting NB is similar as that for CS, in general, NB takes around 0.5 hour, hence the total effort

for NB is estimated as CAD 500.

B.3 Sensitivity Analysis

B.3.1 Alternatives for Handling Missing Data

In addition to the multiple imputation of dealing with missing data, we also consider the following alternative methods:

Simple Delete. We simply keep the records with complete information and remove those with at least one missing data. This way may increase incidence of those risky factors and postpartum complications. It is because most missing data tend to happen in the normal cases, while people will keep records more likely when the abnormal factors or complications happens.

Replace with median. We replace the missing data with the median of existing of the same column. This way will more likely under-estimate the incidence of the risky factors or complications, especially in the case of binary of rare events; because the median is zero in that case.Due to the bias of variance and covariance of replacement with mean, we do not replaced missing data with mean; because we need to implement regression tree and logistic regression in the next step.

We show the estimation of $f(\lambda, x)$ under three imputation methods in Figure B-1.The logistic regression involves dummies of different payment resources. Our estimation is shown robust with respect to different imputation methods. However, due to the potential over-estimated bias of simple delete and undermined bias of median replacement, we believe multiple regression is the best alternative in handling missing data. Moreover, we report the results of different imputation methods with respect to clustering and corresponding incidence of postpartum complications for all the in-sample data in table B.1, B.2 and B.3.

Figure B-1: Successful Rates across Clusters under Different Imputation Methods in 2013

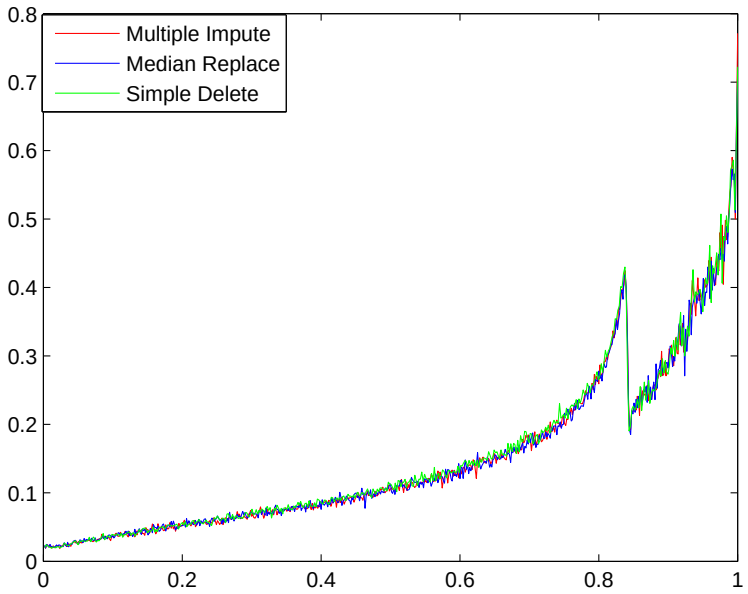


Table B.1: Estimated Incidence of Postpartum Complications 2013

Imputation	$I(CS,x)$	$I(SB,x)$	x^*	R^2
Multiple	8.6288	$4.3320+5.4032x$	79.5232	88.46
Median Replace	8.6290	$4.3269+5.3507x$	80.4026	89.09
Simple Delete	8.4714	$4.2026+5.9257x$	72.0387	90.60

All numbers are in percentage; R^2 is the R-square statistics of linear regression for $I(SB,x)$.

Table B.2: Estimated Incidence of Postpartum Complications 2012

Imputation	$I(CS,x)$	$I(SB,x)$	x^*	R^2
Multiple	8.3492	$4.4328+5.2500x$	74.5980	88.89
Median Replace	8.3376	$4.3289+5.1982x$	77.1170	89.43
Simple Delete	8.0245	$3.0429+7.3957x$	67.3580	93.19

All numbers are in percentage; R^2 is the R-square statistics of linear regression for $I(SB,x)$.

Table B.3: Estimated Incidence of Postpartum Complications 2011

Imputation	$I(CS, x)$	$I(SB, x)$	x^*	R^2
Multiple	8.3932	$4.6497+5.0772x$	73.7316	87.80
Median Replace	8.3728	$4.6540+4.9705x$	74.8174	88.91
Simple Delete	8.4253	$4.6261+5.3448x$	71.0822	88.98

All numbers are in percentage; R^2 is the R-square statistics of linear regression for $I(SB, x)$.

B.3.2 Different Number of Obstetricians in a Group

When the size of obstetrical physician group is very small, a given reimbursement policy tends to have worse outcomes, including extremely high CS rate and high expenses for payers, compared with a larger group (Lemma 3.7). Because physicians in a smaller group have to share more shifts and more workflows once recommending a spontaneous birth, they would prefer planning CS more than their colleagues in a larger group. We recommend to pool physicians into a larger group, not only can they pool their patients and coordinate to share work flows, but also enforce peer supervision under our proposed bonus mechanism.

B.3.3 Difference of Physicians' Effort

Whether the effort of serving NB is relative more effort-consuming than a CS impacts the incentive power of bonus. When NB requires more effort from physicians, the bonus amount has to serve as compensation of extra effort paid in serving a NB, and offset its incentive purpose. The bottom right plot of Figure B.4 shows the rate of planned CS rate increases as more effort demanded in NB than CS.

B.3.4 Physicians' Altruism α

Altruism implies that, as we found from our analytical models, higher value of benevolence determines more weight on considering patients' benefits and relatively less emphasis on their own net benefits. Physicians' net benefits include efforts and finan-

Table B.4: Sensitivity Analysis

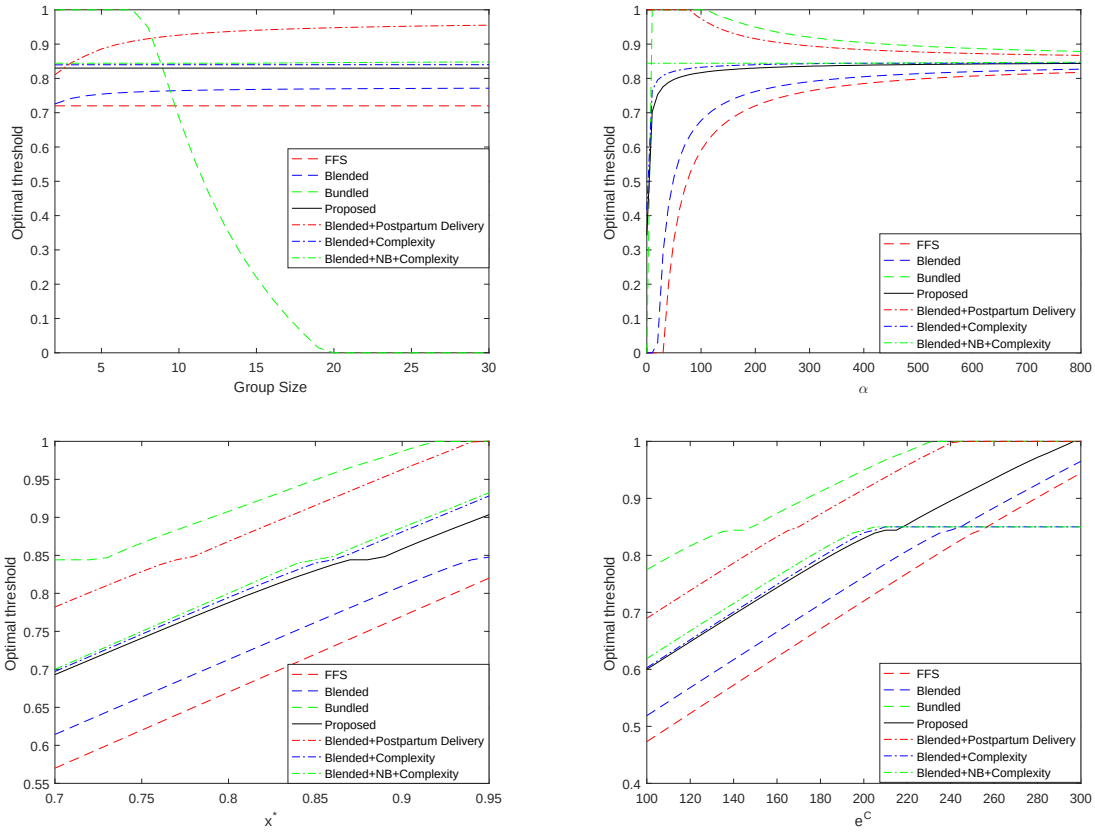


Table B.5: FFS

λ	s	Π_P	P^{PC}	P^N	P^{EC}	Δ
0.5	0.84	5157.8	204	200	300	0.73%
0.75	0.72	5389.3	252	250	350	4.70%
1	0.59	5684.1	304	300	400	9.87 %

Δ refers to the change from benchmark.

cial incomes. Given a blended reimbursement mechanism, CS rate decreases as the altruism increases, and CS rate converges to the ideal clinic cutoff point as altruism level is sufficiently high. In the case of bundled payment policy, CS rate increases as altruism increases, and converges to the ideal clinic cutoff point as well. As the deviation from ideal clinic point gets smaller, the total birth costs decrease and become stable at the lowest level when ideal clinic point is approached. As obstetricians' altruism increases, C-Section converges to the ideal clinic cutoff point, according to the top right plot of Figure B.4.

B.3.5 Clinical Optimal Threshold

The threshold between planned CS and SB considered by physicians, i.e. the value of x^* is possibly different from the ideal value 0.85. Actually, physicians' own perspective of x^* have a large impact on CS rates. The consequential CS rates increase along with the rates believed by physicians. We consider the range of potential beliefs from 0.7 to 0.95. As shown in the bottom left plot in Figure B.4, the deviation from clinical guide decreases as obstetricians believe higher CS threshold in their practice.

B.4 More Numerical Experiments

We present more numerical results with respect to different effort levels under FFS in Table B.5, under blended payment in Table B.6, under bundle payment in Table B.7, and proposed scheme in Table B.8. AUC of ROC for out-of-sample data of different classification methods are reported in Table B.9.

Table B.6: Blended Payment

λ	s	Π_P	P^{BP}	Δ
0.5	0.78	5260.8	228.2	2.74%
0.75	0.68	5473.3	268.2	6.33%
1	0.57	5731.1	307.5	10.77%

Δ refers to the change from benchmark.

Table B.7: Bundle Payment

λ	s	Π_P	P^{BL}	Δ
0.5	0.92	6729.9	5923.1	15.68%
0.5	0.85	6791.1	5986.6	16.24%
0.75	0.93	6752.6	5926.8	15.14%
0.75	0.85	6822.3	6001.6	15.75%
1	0.95	6767.1	5898.3	14.01%
1	0.85	6853.5	6016.5	15.25% %

Δ refers to the change from benchmark.

Table B.8: Proposed Mechanism

λ	s	Π_P	P^{BP}	B^{NB}	Δ
0.5	0.83	5206.4	208.0	53.9	1.68%
0.5	0.85	5352.3	200.0	288.1	3.93%
0.75	0.83	5253.4	208.0	122.9	2.06%
0.75	0.85	5402.2	200.0	359.0	4.19%
1	0.83	5295.6	208.0	183.9	2.36%
1	0.85	5434.8	200.0	405.2	4.11%

Δ refers to the change from benchmark.

Table B.9: AUC of ROC for out-of-samples across years

	2011	2012	2013
Logistic1	83.60	83.89	83.83
Logistic2	83.29	83.64	83.78
CRT1	77.70	77.98	78.03
CRT2	77.70	77.99	78.03

Logistic 1 refers to logistic regression with dummy variables of Payment resources; Logistic 2 represents logistic regression with clinical variables only. Similarly CRT 1 refers to classification and regression tree with dummy variables of Payment resources; CRT 2 represents classification and regression tree with clinical variables only.

B.5 Proofs

Proof of Lemma 3.1. Denote $G(s) = \int_0^s u_{SB}(\lambda, x)\mathrm{d}x + \int_s^1 u_{CS}(x)\mathrm{d}x$.

Sufficiency $\Rightarrow$. In order to achieve the maximum when $s^* \in (0, 1]$, first and second derivatives of $G(s)$ should satisfy

$$G'(s) = u_{SB}(\lambda, s) - u_{CS}(s) = 0$$
$$G''(s) = u'(SB, s) - u'(CS, s) \leq 0$$

That is, $G'(s)$ is monotonously decreasing, and there is only one zero point s^* in the interval $(0, 1]$. So $G'(s) < 0$ if $s < s^*$, and $G'(s) > 0$ when $s > s^*$. Thus leads to the conclusion.

If the maximum is achieved when $s^* = 0$, $G'(s)$ is unnecessarily monotonously decreasing, yet $u_{SB}(\lambda, x) < u_{CS}(x)$, $\forall x \in [0, 1]$, which leads to the conclusion.

Necessity $\Leftarrow$. Consider $\forall s_1 \in [0, s)$,

$$\begin{aligned} G(s_1) &= \int_0^{s_1} u_{SB}(\lambda, x)\mathrm{d}x + \int_{s_1}^{s} u_{CS}(x)\mathrm{d}x + \int_s^1 u_{CS}(x)\mathrm{d}x \\ &\leq \int_0^{s_1} u_{SB}(\lambda, x)\mathrm{d}x + \int_{s_1}^{s} u_{SB}(\lambda, x)\mathrm{d}x + \int_s^1 u_{CS}(x)\mathrm{d}x \qquad = G(s) \end{aligned}$$

Similarly, $\forall s_2 \in (s, 1]$,

$$\begin{aligned} G(s_2) = &\int_0^s u_{SB}(\lambda, x)\mathrm{d}x + \int_s^{s_2} u_{SB}(\lambda, x)\mathrm{d}x + \int_{s_2}^1 u_{CS}(x)\mathrm{d}x \\ &\int_0^s u_{SB}(\lambda, x)\mathrm{d}x + \int_s^{s_2} u_{CS}(x)\mathrm{d}x + \int_{s_2}^1 u_{CS}(x)\mathrm{d}x \qquad = G(s) \end{aligned}$$

That is, $G(s)$ achieves maximum at s. □

Proof of Lemma 3.2. Because every finite symmetric game has a symmetric Nash equilibrium, each physician has the same decision with respect to the certain complexity x.□

Proof of Lemma 3.3. Overall CS rate r can be derived as

$$\begin{aligned} r &= 1 - s + \int_0^s 1 - f(\lambda, x)\mathrm{d}x \\ &= 1 - \int_0^s f(\lambda, x)\mathrm{d}x \end{aligned}$$

due to $f(\lambda, x) > 0$ the overall CS rate is one-to-one mapping of the threshold of planned CS.

Proof of Lemma 3.4. First we prove $CH(\lambda, s)$ is a convex function of s, because

$$\begin{aligned} \frac{\partial CH(\lambda, s)}{\partial s} &= f(\lambda, s)c_H^N + (1 - f(\lambda, s)c_H^C) - c_H^C \\ &= f(\lambda, s)(c_H^N - c_H^C) \\ \frac{\partial^2 CH(\lambda, s)}{\partial s^2} &= (c_H^N - c_H^C)\frac{\partial f(\lambda, s)}{\partial s} > 0. \end{aligned}$$

Next $CI(\lambda, s)$ is a convex function of s due to

$$\begin{aligned} \frac{\partial CI(\lambda, s)}{\partial s} &= C(I_{SB}(\lambda, s) - I_{CS}) \\ \frac{\partial^2 CI(\lambda, s)}{\partial s^2} &= C\frac{\partial I_{SB}(\lambda, s)}{\partial s} > 0. \end{aligned}$$

Therefore, as a linear combination of three convex functions, Π^E is convex with respect to s.□

Proof of Proposition 3.1. First we show $\lambda = 1$ is optimal for payers in delivery stage. Because $u^D_{CS}(x) \geq 0$, $\forall x > s^E$ and $u^D_{SB}(\lambda, x) > 0$, $\forall x < s^E$,

$$
\begin{aligned}
&m^C_S(x) = 0, \quad \forall S \in \{CS, SB\}, \quad m^D_{CS} = e^C, \\
&m^D_{SB} = f(\lambda, s^E)e^N + (1 - f(\lambda, s^E))e^C + \lambda \bar{e}^{MN}
\end{aligned}
$$

So $M(\lambda, s)$ that minimizes total costs becomes

$$
\begin{aligned}
M(\lambda, s) &= \int_0^s f(\lambda, s)e^N + (1 - f(\lambda, s))e^C + \lambda \bar{e}^{MN} \mathrm{d}x + e^C(1 - s) \\
\frac{\partial M(\lambda, s)}{\partial s} &= f(\lambda, s)e^N + (1 - f(\lambda, s))e^C + \lambda \bar{e}^{MN} - e^C \\
&= f(\lambda, s)(e^N - e^C) + \lambda \bar{e}^{MN} \\
\frac{\partial^2 M(\lambda, s)}{\partial s^2} &= e^N - e^C < 0
\end{aligned}
$$

That is, $M(\lambda, s)$ is convex. Therefore the Π^E is convex according to Lemma 3.4. Therefore, s^E should satisfy

$$
\begin{aligned}
&\frac{\partial \Pi^E(\lambda, s)}{\partial s} \\
=&f(\lambda, s)(c^N_H + e^N) + (1 - f(\lambda, x)(c^C_H + c^C_e) + \lambda \bar{e}^{MN} - (c^C_H + e^C) + CI_{SB}(\lambda, x) - CI_C S \\
=&f(\lambda, s)(c^N_H + e^N - c^C_H - c^C_e) + \lambda \bar{e}^{MN} - (c^C_H + e^C) + C(I_{SB}(\lambda, x) - I_C S) = 0.
\end{aligned}
$$

Furthermore consider delivery stage and effort level λ,

$$
\begin{aligned}
&\frac{\partial \Pi^E(\lambda, s)}{\partial \lambda} \\
=&\int_0^s \frac{\partial f(\lambda, s)}{\partial \lambda}(e^N - e^C + c^N_H - c^C_H) + \bar{e}^{MN} \mathrm{d}x + \int_0^s C \frac{\partial I_{SB}(\lambda, s)}{\partial \lambda} \mathrm{d}x \\
<&0,
\end{aligned}
$$

$\lambda = 1$ for the optimal threshold s^E.□

Proof of Proposition 3.2. Suppose the optimum lies outside the interval between x^* and s^E.

Case (i): $s^E < x^*$.

$\forall x > x^*$, $\Pi^E(x) > \Pi^E(x^*)$ due to the convexity of $\Pi^E(\dot{})$ according to Lemma 3.4. Plus $\Pi^E_Q(x) > \Pi^E_Q(x^*)$, $\Pi^{VM}(x) > \Pi^{VM}_P(x^*)$.

$\forall x < s^E$, $\Pi^E(s) > \Pi^E(s^E)$, plus $\Pi^E_Q(s) > \Pi^E_Q(s^E)$ leads to $\Pi^{VM}(x) > \Pi^{VM}(s^E)$.

This contradicts the previous argument. Case (ii): $s^E > x^*$. We have the similar contradiction. Therefore the optimum lies in the interval between x^* and s^E.□

Proof of Lemma 3.7. Denote

$$\begin{aligned}\triangle u(x) &= u_{SB}(\lambda, x) - u_{CS}(x) \\ &\alpha b_{SB}(x) + m^D_{SB}(x) + \frac{u^D_{SB}(\lambda, x)}{J} - u_{CS}(x)\end{aligned}$$

. It is monotonic decreasing with respect to J given

$$\frac{\partial u(x)}{\partial J} = -\frac{u^D_{SB}}{J^2} < 0$$

Suppose under a certain $(m^C_S(x), m^D_S(x))$, $\exists s$ are chosen by a group of J physicians. That is, $\triangle u(x) < 0$ when $x < s$, $\triangle u(s) = 0$ and $\triangle u(x) > 0$ when $x > s$. The zero point s is decreasing when J increases.

Proof of Corollary 3.2. Suppose the lower and higher bounds of the feasible thresholds are $\underline{s}$ and $\bar{s}$.

Case (i) $\bar{s} < \min\{s^E, x^*\}$: $\bar{s}$ is optimal for both Π^E and Π^Q;

Case (ii) $\underline{s} > \max\{s^E, x^*\}$: $\underline{s}$ is optimal for both Π^E and Π^Q.□

Proof of Proposition 3.3. In the case of blended payment

$$\begin{aligned}u_{SB}(\lambda, x) &= \alpha(x^* - x) + \frac{P^{BP}}{J} - \frac{1}{J}(f(\lambda, x)e^N + (1 - f(\lambda, x))e^C + e^{MN}) \\ u_{CS}(x) &= \alpha(x - x^*) + P^{BP} - e^C\end{aligned}$$

Denote $\Delta u(x) = u_{SB}(\lambda, x) - u_{CS}(x)$, then it is continuous. Moreover

$$\frac{\partial \Delta u}{\partial x} = -2\alpha + \frac{1}{J}\frac{\partial f(\lambda, x)}{\partial x}(e^C - e^N) < 0$$

hence $\Delta u(s) = 0$ that is,

$$2\alpha(x^* - s) + \frac{1-J}{J}P^{BP} + e^C - \frac{1}{J}(f(\lambda, s)e^N + (1 - f(\lambda, s)e^C + e^{MN}) = 0$$
$$P^{BP} = \frac{J}{J-1}[2\alpha(x^* - s) + e^C - \frac{1}{J}(f(\lambda, s)e^N + (1 - f(\lambda, s)e^C + e^{MN})]$$

we can see P^{BP} decreases as s increases, due to

$$\frac{\partial P^{BP}}{\partial s} = \frac{J}{J-1}(-2\alpha + \frac{e^C - e^N}{J}\frac{\partial f(\lambda, x)}{\partial x}) < 0$$

Also due to PCC and PCN, $P^{BP} \geq \max(e^C, e^N f(\lambda, x) + e^C(1 - f(\lambda, x)) + e^{MN}) = e^C(1 - f(\lambda, x) + e^{MN})$, due to Assumption 3.2. So we have

$$\frac{J}{J-1}[2\alpha(x^* - s) + e^C - \frac{1}{J}(f(\lambda, x)e^N + (1 - f(\lambda, x)e^C + e^{MN})] \geq e^C(1 - f(\lambda, x)) + e^{MN}$$
$$2J\alpha(x^* - s) - (f(\lambda, x)(e^N - e^C) + e^{MN}) \geq 0$$

Denote $G(s) = 2J\alpha(x^* - s) - (f(\lambda, x)(e^N - e^C) + e^{MN})$, $G(s)$ is continuous and decreases as s increases

$$G'(s) = -2J\alpha + \frac{\partial f(\lambda, s)}{\partial s}(e^C - e^N) - \bar{e}^{MN} < 0$$
$$G(x^*) = -(f(x^*)(e^N - e^C) + e^{MN}) < 0,$$

followed by Eq. 3.2, therefore $s < x^*$.

Under FFS, $P^N \geq e^N + e^{MN}$, $P^{EC} \geq e^C + e^{MN}$, and $P^{PC} > P^N \geq e^N + e^{MN}$. Suppose the threshold $s \geq x^*$. Consider $\forall x \in (x^*, s)$,

$$u_{CS}(x) = \alpha(x - x^*) + P^{PC} - e^C$$
$$\geq \alpha(x - x^*) + e^{MN}$$

. While,

$$\begin{aligned} u_{SB}(x) &= \alpha(x^* - x) + \frac{1}{J}[f(\lambda, x)(P^N - e^N) + [(1 - f(\lambda, x))(P^{EC} - e^C) + e^{MN}] \\ &\leq \alpha(x^* - x) + \frac{1}{J}[f(\lambda, x)(P^{PC} - e^N) + [(1 - f(\lambda, x))(P^{EC} - e^C) + e^{MN}] \\ &\leq \alpha(x^* - x) + [f(\lambda, x)(P^{PC} - e^N) + [(1 - f(\lambda, x))(P^{PC} - e^C) + e^{MN}] \\ &\leq u_{CS}(x). \end{aligned}$$

It leads to a contradiction. Therefore $s < x^*$. □

Proof of Lemma 3.8. In delivery stage physicians' utility under a spontaneous birth is

- blended

$$u_{SB}^D(\lambda, x) = P^{BP} - [f(\lambda, x)e^N + (1 - f(\lambda, x))e^C - e^{MN}];$$

It is decreasing according to Eq. 3.3. Under FFS,

$$\begin{aligned} u_{SB}^D(\lambda, x) &= f(\lambda, x)P^N + (1 - f(\lambda, x))P^{EC} - [f(\lambda, x)e^N + (1 - f(\lambda, x))e^C - e^{MN}]; \\ \frac{\partial u_{SB}^D(\lambda, s)}{\partial \lambda} &= \frac{\partial f(\lambda, x)}{\partial \lambda}(P^N - e^N - P^{EC} + e^C) - \bar{e}^{MN} \\ &= \left[\frac{\partial f(\lambda, x)}{\partial \lambda}(e^C - e^N) - \bar{e}^{MN}\right] + \frac{\partial f(\lambda, x)}{\partial \lambda}(P^N - P^{EC}) \qquad < 0. \end{aligned}$$

Therefore, $\underline{\lambda}$ is optimal for physicians. □

Proof of Corollary 3.9. For a specific λ under each payment policy, we have

- FFS. $M(\lambda, s) = \int_0^s f(\lambda, x)P^N + (1 - f(\lambda, x))P^{EC}\mathrm{d}x + P^{PC}(1 - s)$;
- Blend. $M(\lambda, s) = P^{BP}$;

Apparently, $M(\lambda, s)$ is non-concave under blend payment.

Under FFS,

$$\frac{\partial M(\lambda, s)}{\partial s} = f(\lambda, s)P^N + (1 - f(\lambda, s))P^{EC} - P^{PC};$$
$$\frac{\partial^2 M(\lambda, s)}{\partial s^2} = \frac{\partial f(\lambda, s)}{\partial s}(P^N - P^{EC}) > 0.$$

□

Proof of Corollary 3.3. Suppose s^E is the optimum to the Problem 3.2. If $s^E \geq x^*$, feasible threshold is outside invertal (x^*, s^E) due to proposition 3.3. Consider $s^E < x^*$, and it is the solution to the Eq.3.8. We consider payer's amount of total economic cost, denoted as $\Pi(s)$, depending on the threshold s.

$$\Pi(s) = CH(s) + CI(s) + m^D_{CS}(s)(1 - s) + m^D_{SB}(s)s.$$

Under FFS, $m^D_{CS}(s) = e^C$, and $m^D_{SB}(s) = e^N f(\lambda, s^E) + e^C(1 - f(\lambda, s^E)) + e^{MN}$.

$$\begin{aligned}\frac{\partial \Pi(s)}{\partial s} &= f(\lambda, s)(c^N_H - c^C_H) + C(I_{SB}(\lambda, x) - I_{CS}) + f(\lambda, s)(e^N - e^C) + e^{MN} \\ &\geq f(1, s)(c^N_H - c^C_H) + C(I_{SB}(1, x) - I_{CS}) + f(1, s)(e^N - e^C) + \bar{e}^{MN}.\end{aligned}$$

It indicates a steeper decreasing slope of $\Pi(s)$ than that of Π^E_P in Eq.3.8. Therefore the optimum here $s^* \geq s^E$.

Under blended, $m^D_{CS}(s) = m^D_{SB}(s) = e^N f(\lambda, s^E) + e^C(1 - f(\lambda, s^E)) + e^{MN}$.

$$\begin{aligned}\frac{\partial \Pi(s)}{\partial s} &= f(\lambda, s)(c^N_H - c^C_H) + C(I_{SB}(\lambda, x) - I_{CS}) + \frac{\partial f(\lambda, s)}{\partial s}(e^N - e^C) + \bar{e}^{MN} \\ &\geq f(1, s)(c^N_H - c^C_H) + C(I_{SB}(1, x) - I_{CS}) + f(1, s)(e^N - e^C) + \bar{e}^{MN},\end{aligned}$$

Given non-positive $\frac{\partial f(\lambda, s)}{\partial s}$, and $\frac{\partial f(\lambda, s)}{\partial s}(e^N - e^C) > f(1, s)(e^N - e^C)$. It indicates a steeper decreasing slope of $\Pi(s)$ than that of Π^E in Eq.3.8. Therefore the optimum here $s^* \geq s^E$.

□

Proof of Lemma 3.10. Under bundled payment,

$$M(\lambda, s) = P^{BL} - \frac{1}{J}\left[\int_0^s f(\lambda, x)c_H^N + (1 - f(\lambda, x))c_H^C \mathrm{d}x + c_H^C(1 - s)\right];$$
$$\frac{\partial M(\lambda, s)}{\partial s} = \frac{1}{J}\left[-f(\lambda, s)c_H^N - (1 - f(\lambda, s))c_H^C + c_H^C\right];$$
$$\frac{\partial^2 M(\lambda, s)}{\partial s^2} = \frac{\partial f(\lambda, s)}{J\partial s}(c_H^C - c_H^N) < 0$$

□

Proof of Proposition 3.4. Under bundled payment,

$$u_{SB}(\lambda, x) = \alpha(x^* - x) + \frac{P^{BL}}{J} - \frac{1}{J^2}(f(\lambda, x)(c_H^N + e^N) + (1 - f(\lambda, x))(c_H^C + e^C) + e^{MN})$$
$$u_{CS}(x) = \alpha(x - x^*) + P^{BL} - \frac{c_H^C}{J} - e^C$$

Denote $\Delta u(x) = u_{SB}(\lambda, x) - u_{CS}(x)$, then it is continuous. Moreover

$$\begin{aligned}\frac{\partial \Delta u}{\partial x} &= -2\alpha + \frac{1}{J}\frac{\partial f(\lambda, x)}{\partial x}[c_H^C - c_H^N + J(e^C - e^N)] \\ &< -2\alpha + \frac{1}{J}\frac{\partial f(\lambda, x)}{\partial x}(\frac{c_H^C - c_H^N}{J} + e^C - e^N) < 0\end{aligned}$$

hence $\Delta u(s) = 0$ that is,

$$2\alpha(x^* - s) + \frac{1 - J}{J}P^{BL} + e^C - \frac{1}{J^2}(f(\lambda, s)(e^N * J + c_H^N) + (1 - f(\lambda, s))(c_H^C + e^C * J) + e^{MN}) = 0$$
$$P^{BL} = \frac{J}{J - 1}[2\alpha(x^* - s) + (e^C J + c_H^C) - \frac{1}{J^2}(f(\lambda, s)(e^N * J + c_H^N) + (1 - f(\lambda, s))(e^C J + c_H^C) + Je^{MN})]$$

we can see P^{BL} decreases as s increases, due to

$$\frac{\partial P^{BL}}{\partial s} = \frac{J}{J - 1}(-2\alpha + \frac{c_H^C - c_H^N + J(e^C - e^N)}{J}\frac{\partial f(\lambda, x)}{\partial x}) < 0$$

Also due to PCC and PCN, $P^{BL} \geq \max(\frac{c_H^C}{J} + e^C, (e^N + \frac{c_H^N}{J})f(\lambda, x) + (\frac{c_H^C}{J} + e^C)(1 -$

$f(\lambda,x))+e^{MN})=e^C+\frac{c_H^C}{J}$. So we have

$$\frac{J}{J-1}[2\alpha(x^*-s)+\frac{c_H^C}{J}+e^C-\frac{1}{J}(f(\lambda,x)(e^N+\frac{c_H^N}{J})+(1-f(\lambda,x))(\frac{c_H^C}{J}+e^C)+e^{MN})]$$
$$\geq e^C(1-f(\lambda,x))+e^{MN}$$
$$2J\alpha(x^*-s)-(f(\lambda,x)(\frac{c_H^N-c_H^C}{J}+e^N-e^C)+e^{MN})\geq 0$$

Denote $G(s)=2J\alpha(x^*-s)-(f(\lambda,x)(\frac{c_H^N-c_H^C}{J}+e^N-e^C)+e^{MN})$, $G(s)$ is continuous and decreases as s increases

$$G'(s)=-2J\alpha-\frac{\partial f(\lambda,s)}{\partial s}(\frac{c_H^N-c_H^C}{J}+e^N-e^C)<0$$
$$G(x^*)=(f(x^*)(\frac{c_H^N-c_H^C}{J}+e^N-e^C)+e^{MN})\geq 0,$$

therefore $s\geq x^*$.

□

Proof of Corollary 3.4. Suppose the optimum to Z_P is s^E. If $s^E=x^*$,

$$P^{BL}=\frac{J}{J-1}[(e^CJ+c_H^C)-\frac{1}{J^2}(f(\lambda,x^*)(e^N*J+c_H^N)+(1-f(\lambda,x^*))(e^CJ+c_H^C)+Je^{MN})]$$

due to

$$\frac{\partial P^{BL}}{\partial s}=\frac{J}{J-1}(-2\alpha+\frac{c_H^C-c_H^N+J(e^C-e^N)}{J}\frac{\partial f(\lambda,x)}{\partial x})<0$$

PCC and PCN hold.□

Proof of Proposition 3.5. Consider physicians' utility in delivery stage after the decision of spontaneous birth.

$$u_{SB}^D(1,x)=P^{BL}-e^{MN}-f(1,x)(e^N+\frac{c_H^N}{J})-(1-f(1,x))(e^C+\frac{c_H^C}{J})<0;$$
$$f(1,x)\left(e^C-e^N+\frac{c_H^C-c_H^N}{J}\right)<e^{MN}+e^C+\frac{c_H^C}{J}-P^{BL};$$
$$f(1,x)<\frac{J(e^C+e^{MN}-P^{BL})+c_H^C}{J(e^C-e^N)+c_H^C-c_H^N}.$$

□

Proof of Proposition 3.6. Suppose P_0 is the rate associated with $\bar{s}$ under original payment mechanism, that is, $\forall x \in (\bar{s}, x^*]$,

$$2\alpha(x^* - x) - \frac{1}{J}(f(\lambda, x)e^N + (1 - f(\lambda, x))e^C + e^{MN}) + P_0 + e^C \leq 0.$$

If the successful NB bonus B^{NB} is set up by

$$B^{NB} \triangleq -\frac{1}{f(s)}\min(2\alpha(x^* - s) - \frac{1}{J}(f(s)e^N + (1 - f(s))e^C + e^{MN}) - P + e^C,$$
$$\alpha(x^* - s) - \frac{1}{J}(f(s)e^N + (1 - f(s))e^C + e^{MN})).$$

therefore $\exists s > \bar{s}$, such that $\forall x \leq s$

$$u_{SB}(\lambda, x) - u_{CS}(x) \geq 0;$$
$$u_{SB}(\lambda, x) \geq 0.$$

That is, $\bar{s}$ can increase under NB rate.

For postpartum bonus B^{PO}, suppose sp is the intersection of $I(SB, x)$ and CS, i.e.

$$I(SB, x) \leq I(CS), \quad \forall x \leq sp;$$
$$I(SB, x) \geq I(CS), \quad \forall x > sp.$$

If $\bar{s} < sp$, then $\forall x \in (\bar{s}, sp]$,

$$B^{PO}(1 - I(SB, x)) > B^{PO}(1 - I(CS)).$$

therefore $\exists s > \bar{s}$, such that $\forall x \leq s$

$$u_{SB}(\lambda, x) - u_{CS}(x) \geq 0;$$
$$u_{SB}(\lambda, x) \geq 0.$$

That is, $\bar{s}$ can increase under postpartum outcome rate. □

Proof of Proposition 3.7. (i) When $\bar{s} < x^*$. Suppose P_0 is the rate associated with $\bar{s}$ under original payment mechanism, that is, $\forall x \in (\bar{s}, x^*]$,

$$2\alpha(x^* - x) - \frac{1}{J}(f(\lambda, x)e^N + (1 - f(\lambda, x))e^C + e^{MN}) + P_0 + e^C \leq 0.$$

For complexity premium B, if it is set up as

$$B \triangleq -\min(2\alpha(x^* - s) - \frac{1}{J}(f(s)e^N + (1 - f(s))e^C + e^{MN}) + P_0 + e^C,$$
$$\alpha(x^* - s) - \frac{1}{J}(f(s)e^N + (1 - f(s))e^C + e^{MN})),$$

therefore $\exists s > \bar{s}$, such that $\forall x \leq s$.

$$u_{SB}(\lambda, x) - u_{CS}(x) \geq 0;$$
$$u_{SB}(\lambda, x) \geq 0.$$

That is, $\bar{s}$ can increase under overall rate.

(ii) When $\underline{s} > x^*$, we can conclude that $\underline{s}$ can be reduced with similar way. □

Proof of Proposition 3.8. Overall rate add-on B with an original payment P_0 lead to a physician's utility function in serving a delivery in his hospital shift

$$u_{SB}^D = P_0 + Bf(\lambda, x) - f(\lambda, x)e^N - (1 - f(\lambda, x))e^C - \lambda\bar{e}^{MN};$$
$$\frac{\partial u_{SB}^D}{\partial \lambda} = \frac{\partial f(\lambda, x)}{\partial \lambda}(e^C - e^N + B) - \bar{e}^{MN};$$
$$B\frac{\partial f(\lambda, x)}{\partial \lambda} \geq \bar{e}^{MN} - \frac{\partial f(\lambda, x)}{\partial \lambda}(e^C - e^N),$$

which gives the lower bound of B leads to an increasing u_{SB}^D.

Per postpartum outcome-oriented bonus B, Postpartum outcome-oriented add-on B with an original payment P_0 lead to a physician's utility function in serving a delivery

in his hospital shift

$$\begin{aligned}
u_{SB}^{D} &= P_0 + B(1 - I_{SB}(\lambda, x)) - f(\lambda, x)e^N - (1 - f(\lambda, x))e^C - \lambda\bar{e}^{MN};\\
\frac{\partial u_{SB}^{D}}{\partial \lambda} &= \frac{\partial f(\lambda, x)}{\partial \lambda}(e^C - e^N) - \bar{e}^{MN} - B\frac{\partial I_{SB}(\lambda, x)}{\partial \lambda};\\
B\frac{\partial I_{SB}(\lambda, x)}{\partial \lambda} &\leq \frac{\partial f(\lambda, x)}{\partial \lambda}(e^C - e^N) - \bar{e}^{MN};\\
B\frac{-\partial I_{SB}(\lambda, x)}{\partial \lambda} &\geq \bar{e}^{MN} - \frac{\partial f(\lambda, x)}{\partial \lambda}(e^C - e^N)\\
&\geq \bar{e}^{MN} - \upsilon(e^C - e^N), \quad \upsilon = \max_{\lambda, x}\frac{\partial f(\lambda, x)}{\partial \lambda}
\end{aligned}$$

which gives the lower bound of B leads to an increasing u_{SB}^{D}. □

Proof of Proposition 3.9. Suppose $E(\lambda, x) = e^C f(\lambda, x) + e^N(1 - f(\lambda, x)) + e^{MN}$, under Complexity bonus scheme:

$$\begin{aligned}
U_{CS}^{CO}(x) &= \alpha(x - x^*) + P + B^{CO}\mathbb{1}_{x > x^*} - e^C;\\
U_{SB}^{CO}(x) &= \alpha(x^* - x) + \frac{P}{J} + B^{CO}\mathbb{I}_{x \leq x^*} - \frac{E(\lambda, x)}{J}.\\
\Delta U^{CO}(x) &= U_{SB}^{CO} - U_{CS}^{CO}\\
&= 2\alpha(x^* - x) + \frac{P(1 - J)}{J} + B^{CO}(\mathbb{I}_{x \leq x^*} - \mathbb{I}_{x > x^*})\mathbb{I}_{x \leq x^*} - \frac{E(\lambda, x)}{J} + e^C.
\end{aligned}$$

Whereas under NB bonus scheme,

$$\begin{aligned}
U_{CS}^{NB}(x) &= \alpha(x - x^*) + P - e^C;\\
U_{SB}^{NB}(x) &= \alpha(x^* - x) + \frac{P}{J} + B^{NB}f(\lambda, x) - \frac{E(\lambda, x)}{J}.\\
\Delta U^{NB}(x) &= U_{SB}^{NB} - U_{CS}^{NB}\\
&= 2\alpha(x^* - x) + \frac{P(1 - J)}{J} + B^{NB}f(\lambda, x) - \frac{E(\lambda, x)}{J} + e^C.
\end{aligned}$$

Let $\Gamma = 2\alpha(x - x^*) + \frac{P(J-1)}{J} + \frac{E(\lambda,x)}{J} - e^C$, then $\forall s$ and the same P,

$$B^{CO} = \Gamma;$$
$$B^{NB} = \frac{\Gamma}{f(\lambda, s)}.$$

Therefore the average bonus per cost is

$$\text{Complexity bonus} = \Gamma(1 - |s - x^*|) \leq B^{CO}$$
$$\leq \frac{\Gamma \int_0^s f(\lambda, x)\mathrm{d}x}{f(\lambda, s)} = \text{NB bonus}.$$

□

Proof of Proposition 3.10. Let (P^{PO}, B^{PO}) and (P^{CO}, B^{CO}) be the rates under Postpartum Outcome Bonus and Complexity bonus, resparately, to achieve the same s.

From PCC,

$$P^{CO} = e^C;$$
$$P^{PO} + B^{PO}(1 - I_{CS}) = e^C.$$

$$\text{Postpartum Outcome bonus} \leq P^{PO} + B^{PO}(1 - I_{CS})$$
$$\leq P^{CO} \leq \text{Complexity bonus}.$$

Therefore the average cost under Postpartum Outcome is less than that under Complexity bonus.□

Proof of Proposition 3.11. Suppose the same amount of P and B are reimbursed to each physician under *Relevaant party* and *Group* mechanism, they should lead to the same threshold of planned CS s by the end of antepartum stage, and same level of effort λ during intrapartum stage. According to expression of $M(\lambda, s)$ in Table 3.9, *Group* leads to higher $M(\lambda, s)$ than *Relevaant party*, due to $J \geq 2$.□

Proof of Proposition 3.12. In the case of high risk population, $\int_0^x g(u)\mathrm{d}u \leq x$, even all physicians are supposed to use full effort $\lambda = 1$, aiming to reduce CS rates for the group of physicians during antepartum stage. Suppose $sleqs^*$, the resulting overall CS rate r' becomes

$$\begin{aligned} r' &= 1 - \int_0^s g(u)f(1,u)\mathrm{d}u \\ &> 1 - \int_0^s g(u)f(1,u)\mathrm{d}u, \quad (*) \\ &= r; \end{aligned}$$

Therefore, s should be $s > s^*$ for the high risk population in order to get bonus B^{TH}. Denote $G(x) = \int_0^s g(u)\mathrm{d}u$, then $G(0) = 0$ and $G(1) = 1$. Eq.(*) is due to

$$\begin{aligned} &\int_0^s g(u)f(1,u)\mathrm{d}u \\ =&\int_0^s f(1,u)\mathrm{d}G(u) \\ =&f(1,x)G(x) - \int_0^s G(u)\frac{\partial f(1,u)}{\partial u}\mathrm{d}u \\ <&f(1,x)x - \int_0^s u\frac{\partial f(1,u)}{\partial u}\mathrm{d}u \\ =&f(1,x)x - f(1,x)x + \int_0^x f(1,u)\mathrm{u} \end{aligned}$$

Similarly, if $G(u) > u$ for the case of low risk population,

$$\begin{aligned}
&\int_0^s g(u)f(1,u)\mathrm{d}u \\
=&\int_0^s f(1,u)\mathrm{d}G(u) \\
=&f(1,x)G(x) - \int_0^s G(u)\frac{\partial f(1,u)}{\partial u}\mathrm{d}u \\
>&f(1,x)x - \int_0^s u\frac{\partial f(1,u)}{\partial u}\mathrm{d}u \\
=&f(1,x)x - f(1,x)x + \int_0^x f(1,u)\mathrm{u} \\
>&\int_0^x f(\lambda,u)\mathrm{u},
\end{aligned}$$

Therefore suppose $s \geq s^*$, the resulting overall CS rate r'

$$\begin{aligned}
r' &= 1 - \int_0^s g(u)f(1,u)\mathrm{d}u \\
&< 1 - \int_0^s g(u)f(\lambda,u)\mathrm{d}u, \\
&= r;
\end{aligned}$$

Therefore $s < s^*$ even not all physicians use full effort on their shifts.□

Proof of Lemma 3.11. Under the bundled payment with a rate add-on P^{BL}, a physician's expected compensation amount is

$$\begin{aligned}
\Pi_1 &= P^{BL} - \frac{1}{J}\left[\int_0^s f(\lambda,x)c_H^N + (1-f(\lambda,x))c_H^C\mathrm{d}x + c_H^C(1-s)\right] \\
&= P^{BL} - \frac{c_H^C}{J} + \frac{c_H^C - c_H^N}{J}\int_0^s f(\lambda,x)\mathrm{d}x,
\end{aligned}$$

which is the combination of blend payment with blend rate $P^{BL} - \frac{c_H^C}{J}$ and a NB bonus of $\frac{c_H^C - c_H^N}{J}$.□

Proof of Proposition 3.13. Under the blended payment with an NB rate add-on (P^{BP}, B^{NB}),physicians' expected compensation amount is $\Pi_1 = P^{BP} + Bf(\lambda,x)$ per delivery during their shift, $\forall\lambda$ and $\forall x$. Under the linear combination of blend payment

with blend rate $(1-\theta)P^{BP}$ and a bundled payment with rate θP^{BL}, physicians expected income is

$$\begin{aligned}\Pi_2 &= (1-\theta)P^{BP} + \theta\left[P^{BL} - \frac{1}{J}\left(f(\lambda,x)c_H^N + (1-f(\lambda,x))c_H^C\right)\right] \\ &= (1-\theta)P^{BP} + \theta(P^{BL} - \frac{c_H^C}{J}) + \frac{\theta}{J}f(\lambda,x)(c_H^C - c_H^N) \\ &= (1-\theta)P^{BP} + \theta P^{BP} + Bf(\lambda,x) \\ &= \Pi_1, \quad \forall x, \quad \forall \lambda.\end{aligned}$$

□

Proof of Lemma 3.12. Under NB bonus scheme,

$$\begin{aligned}U_{CS} &= \alpha(x-x^*) + P - e^C; \\ U_{SB} &= \alpha(x^*-x) + \frac{P}{J} + B^{NB}f(\lambda,x) - \frac{E(\lambda,x)}{J}. \\ \Delta U &= U_{SB}^{NB} - U_{CS}^{NB} \\ &= 2\alpha(x^*-x) + \frac{P(1-J)}{J} + B^{NB}f(\lambda,x) - \frac{E(\lambda,x)}{J} + e^C.\end{aligned}$$

Therefore

$$\begin{aligned}\frac{\partial \Delta U}{\partial P} &= \frac{1}{J} - 1 < 0; \\ \frac{\partial \Delta U}{\partial B} &= \frac{f}{(}\lambda, x) > 0;\end{aligned}$$

Therefore, the overall CS rate increases as B^{NB} decreases, or as P^{BP} increases.

Because $M(s) = P + B\int_0^s f(\lambda,x)\mathrm{d}x$,

$$\begin{aligned}\frac{\partial M(s)}{\partial s} &= f(\lambda,s); \\ \frac{\partial^2 M(s)}{\partial s^2} &= \frac{\partial f(\lambda,s)}{\partial s} < 0.\end{aligned}$$

Hence, $M(\lambda,s,m_D)$ is concave with respect to s. □

Proof of Proposition 3.14. Define $\Pi(s,P,B) = s\Pi(1,P,B) \forall s \geq 1$ and $B, P \geq 0$; define $\Pi(s,P,B) = s\Pi(|s|,|P|,|B|) \forall s, B, P \leq 0$. Therefore $Pi_P(s,B,P)$ is coersive. Because feasible domain is close, there exist at least a global minimizer for the objective function. □

Proof of Corollary 3.5. From Lemma 3.12, we can find the proper P and B for any threshold s for the problem Z_P, therefore the value maximization solution to Z_{BM} is feasible. □

Proof of Proposition 3.15. Delivery physicians get the same amount of money regardless of his own patient mix. The realized amount of successful NB is shared evenly by all. Therefore, they each should get the amount of money as long as they follow clinical guideline.□

Proof of Proposition 3.16. For those prescribed planned C-section, the probability of actually high risk pregnant women is

$$\Pr(H|CS) = \frac{\Pr(CS|H)\Pr(H)}{\Pr(CS|H)\Pr(H) + \Pr(CS|L)\Pr(L)} = \frac{a(1-s_0)}{a(1-s_0)+(1-a)s_0}$$

due to Bayes' rule, and the assumption of uniformly distributed patients with respect to the complexity, indicating the fraction of L type pregnant women is s_0. Similarly, the fraction of actual low risk patient among those having planned C-section becomes

$$\Pr(L|CS) = \frac{\Pr(CS|L)\Pr(L)}{\Pr(CS|L)\Pr(L) + \Pr(CS|H)\Pr(H)} = \frac{(1-a)s_0}{a(1-s_0)+(1-a)s_0}$$

For all planned C-sections, a fraction $\Pr(H|CS)$ has actual complexity $x > s_0$, and the physician can gain marginal utility $u_{CS}(\lambda, x)$ which is greater than $u_{SB}(\lambda, x)$ as $x > s_0$; whereas the rest $\Pr(L|CS)$ has actual complexity $x \leq s_0$, therefore the physician's gain of marginal utility $u_{CS}(\lambda, x)$ is less than $u_{SB}(\lambda, x$, suffering a loss of utility eventually. Similarly, the fraction of actual low risk cases among prescription of spontaneous birth $\Pr(L|SB)$ and the part of actual high risk but prescribed spontaneous

birth $\Pr(H|SB)$ can be expressed as

$$\Pr(H|SB) = \frac{(1-a)(1-s_0)}{(1-a)(1-s_0)+as_0}, \quad \Pr(L|SB) = \frac{as_0}{(1-a)(1-s_0)+as_0}$$

Noted that the percentage $\Pr(L|CS)$ of planned C-section and the percent $\Pr(H|SB)$ of prescription of spontaneous birth lead to loss of total utility.□

Proof of Proposition B.1. Let $U_j^k(S,x)$ be the total expected payoff / utility for physician j if he chooses a procedure S, when there are k physicians determine SB $\forall x \in [0,1]$ and $k \in \{0,1,2,\cdots,J-1\}$, then

$$U_j^k(S,x) = u(S,x) + \frac{k}{J}(u_{SB}^D(\lambda,x))$$

Because the best strategy for physician j is to maximize the total payoff with respect to each complexity x no matter what the other colleagues choose, and then he will prefer SB as long as

$$U_j^k(SB,x) \geq U_j^k(CS,x), \forall k \in \{0,1,2,\cdots,J-1\}$$

which is equivalent to

$$u_{SB}(\lambda,x) \geq u_{CS}(x)$$

The similar rational is for the preference of C-section.□

Proof of Lemma B.1. Suppose under a certain $(m_S^C(x), m_S^D(x))$, and α, $\exists s_0$ such that

$$u_{CS}^0(x) \geq u_{SB}^0(\lambda,x), \quad \forall x > s_0$$
$$u_{CS}^0(x) \leq u_{SB}^0(\lambda,x), \quad \forall x \leq s_0$$

. Consider $\forall \lambda' = \lambda + \triangle$, where $\triangle > 0$, physicians' modified utilities become

$$u_{CS}(x) = \triangle(x - x^*) + u^0_{CS}(x),$$
$$u_{SB}(x) = \triangle(x^* - x) + u^0_{SB}(x), \quad \forall x \in [0, 1].$$

Case (i): $s_0 < x^*$. Apparently $\forall x \in [0, s_0]$, $x < x^*$, $uCS(x) \leq u_{SB}(\lambda, x)$; $\forall x \in [x^*, 1]$, $uCS(x) \geq u_{SB}(\lambda, x)$. If $x \in (s_0, x^*)$, denote $\triangle u(x) = u_{SB}(\lambda, x) - u_{CS}(x)$, we have $\triangle u(s_0) > 0$ and $\triangle u(x^*) < 0$, so there must $\exists s \in (s_0, x^*)$ that satisfies $\triangle u(s) = 0$, and s is the new threshold, that leads to smaller deviation from x^*.

Case (ii): $s_0 > x^*$ is similar. $\forall x \in [0, x^*]$ still leads to $uCS(x) \leq u_{SB}(\lambda, x)$; $\forall x \in [s_0, 1]$ leads to $uCS(x) \geq u_{SB}(\lambda, x)$. And there a new threshold $s \in (x^*, s_0)$ determined by physicians with λ'. □

Proof of Lemma B.2. In delivery stage, any patient with complexity x leads to an expected utility

$$u^D_{SB}(\lambda, x) = P^{BL} - f(\lambda, x)\frac{c^N_H}{J} + (1 - f(\lambda, x))\frac{c^C_H}{J} - [f(\lambda, x)e^N + (1 - f(\lambda, x))e^C + e^{MN}];$$
$$\frac{\partial u^D_{SB}(\lambda, s)}{\partial \lambda} = \frac{\partial f(\lambda, x)}{\partial \lambda}(\frac{c^C_H - c^N_H}{J} + e^C - e^N) - \bar{e}^{MN}$$
$$> 0.$$

□

Appendix C

Design of Specialist Responsible Policies to Reduce Waiting Times in Emergency Departments

C.1 Notation

Table C.1: Notation

Specialist Response Policy	
T_j	the arrival time of the jth patient who requires a specialist consultation;
$N(t)$	the number of arrivals in $(0,t], \forall t > 0$;
$\lambda(t)$	the arrival rate at time t;
$L(t)$	the number of patients waiting in the system;
$W(t)$	total waiting time at time t of all patients arriving by t;
ST	generally distributed specialist's treatment time;
B	generally distributed the specialists' arrival time from now;
FT	a determinist period after which next specialist arrives.
Modified Triage	
$\mathbf{S}(t)$	Obsevable set, the number of patients of each class at time t;
$S_n(t)$	the number of patients of cluster n, $\forall n \in \mathscr{N}$;
$\mathscr{I}$	index of classes where specialists follow FT response rules;
$\mathscr{J}$	index of classes where specialists follow TL response rules;

C.2 Non-homogeneous Poisson Arrivals

We simulate two periodic functions

$$\lambda(t) = \max(1, 1 + b\sin\left(\frac{\pi}{12}t\right));$$
$$\lambda(t) = b\left[\sin\left(\frac{\pi}{12}t\right) + 1\right],$$

where b determines the magnitude of the peak arrivals.

Table C.2: Trajectory from diagnosis code to specialist type

Specialist Type	Possible Inpatient Ward
Internal Medicine	1, 5, 7, 8, 9, 11, 14, 22;
Oncology	2;
Mental Health	6;
Gynecology & Obstetrics	16.

C.3 Alternative results of Statistical Learning

Table C.3: Results of ALternative Statistical Learning

Estimated probability of	AUC	MSE
neural net hidden 1 (%)	50.00	25.14
neural net hidden 2 (%)	82.34	18.76
neural net hidden 3 (%)	82.29	18.78
nearest neighbor (%)	80.16	18.98
kernel epilson (%)	78.58	18.62
SVM (%)	79.23	18.67

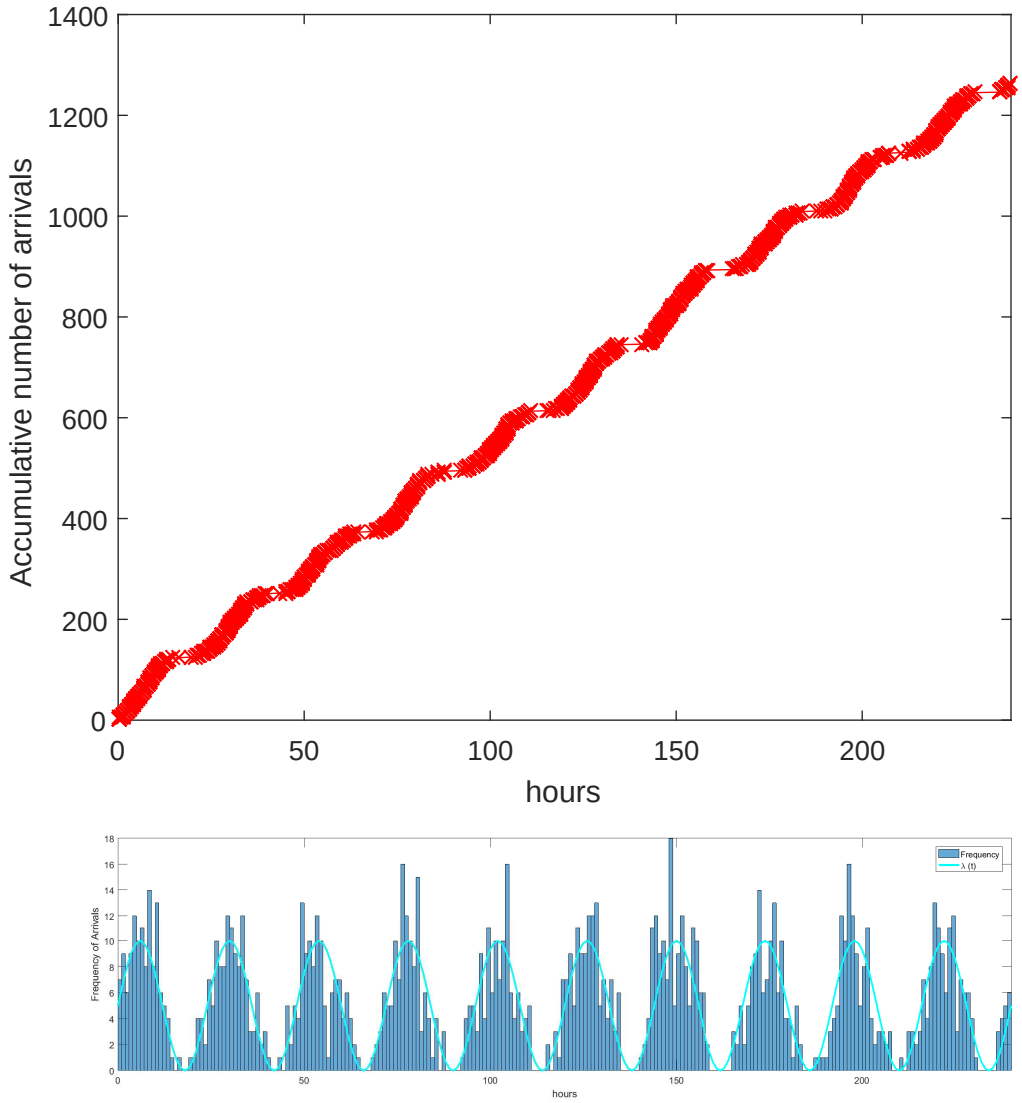


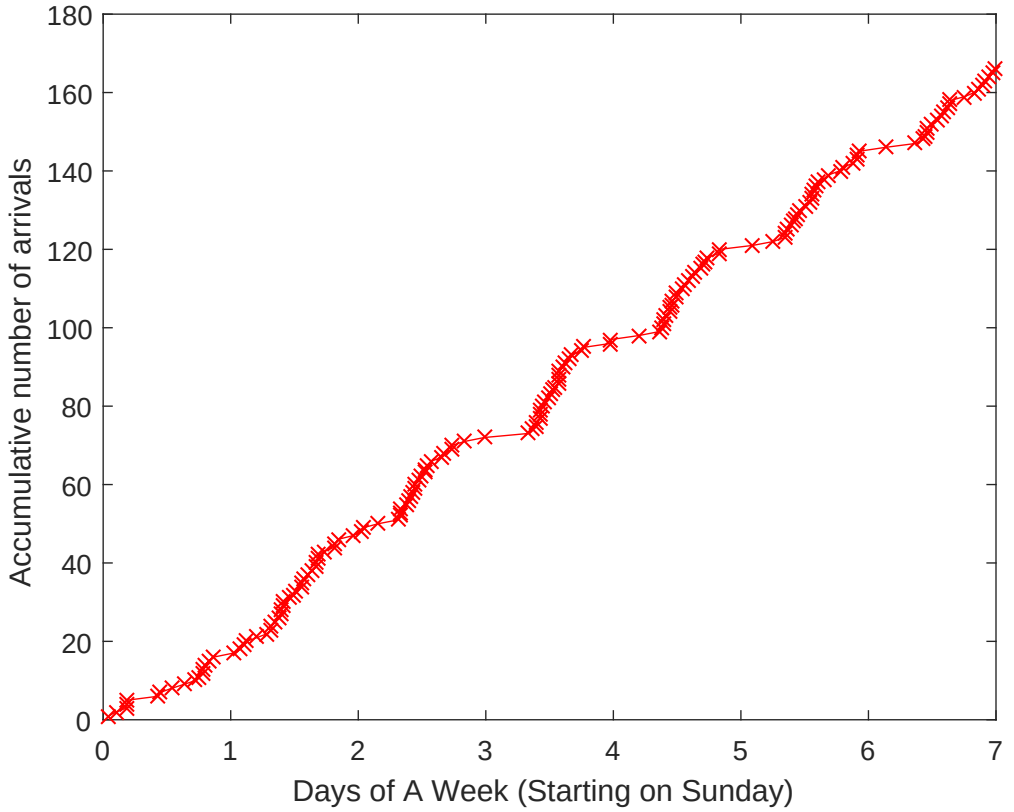
Figure C-1: Simulation of Arrivals with $\lambda(t) = 5\left[\sin\left(\frac{\pi}{12}t\right) + 1\right]$

Table C.4: Unbalance between request or non-request of specialist consultation

Consulting	3	4	5	Total
No	26.44	36.93	11.95	75.31
Yes	14.24	9.31	1.14	24.69
Total	40.68	46.24	13.08	100.00

All numbers are in %

Figure C-2: Daily Variations of Arrivals during a Week



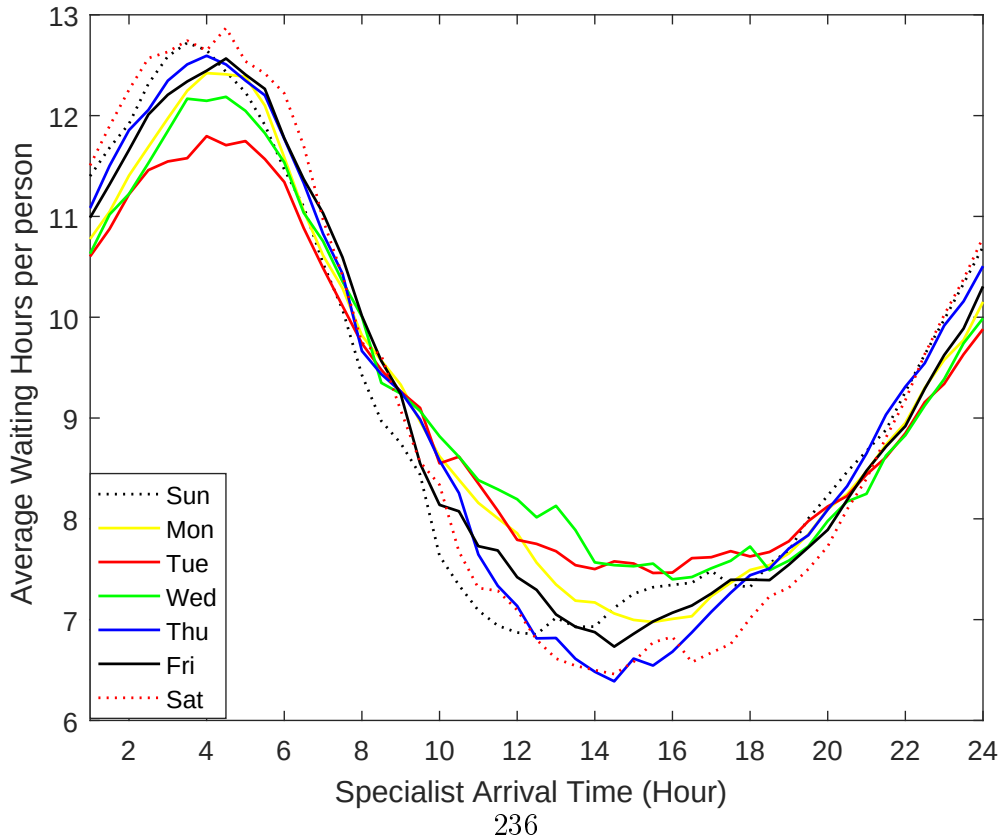
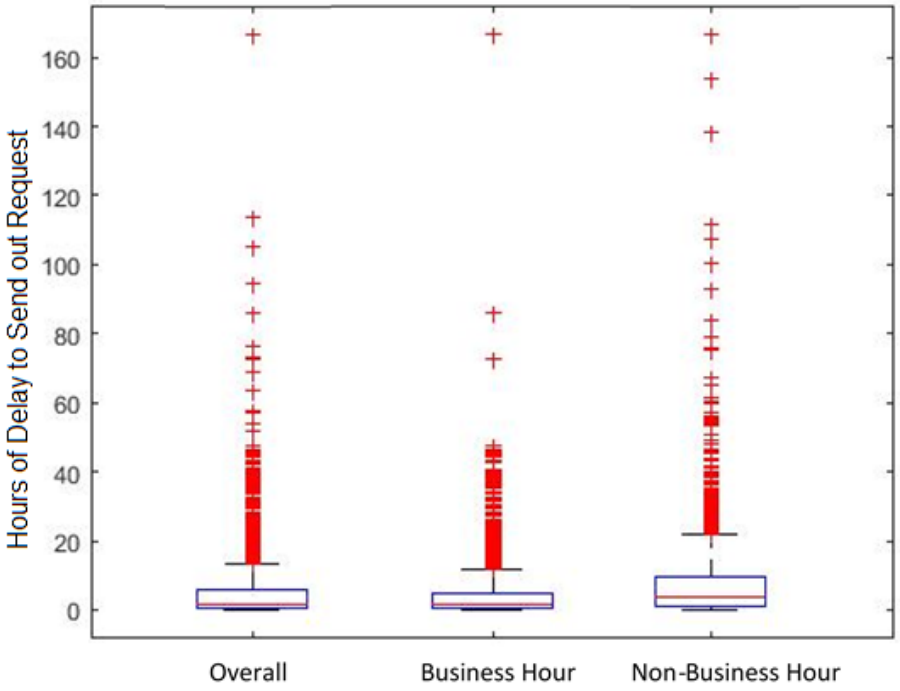


Figure C-3: Compare Delay of Sending out Consulting Request



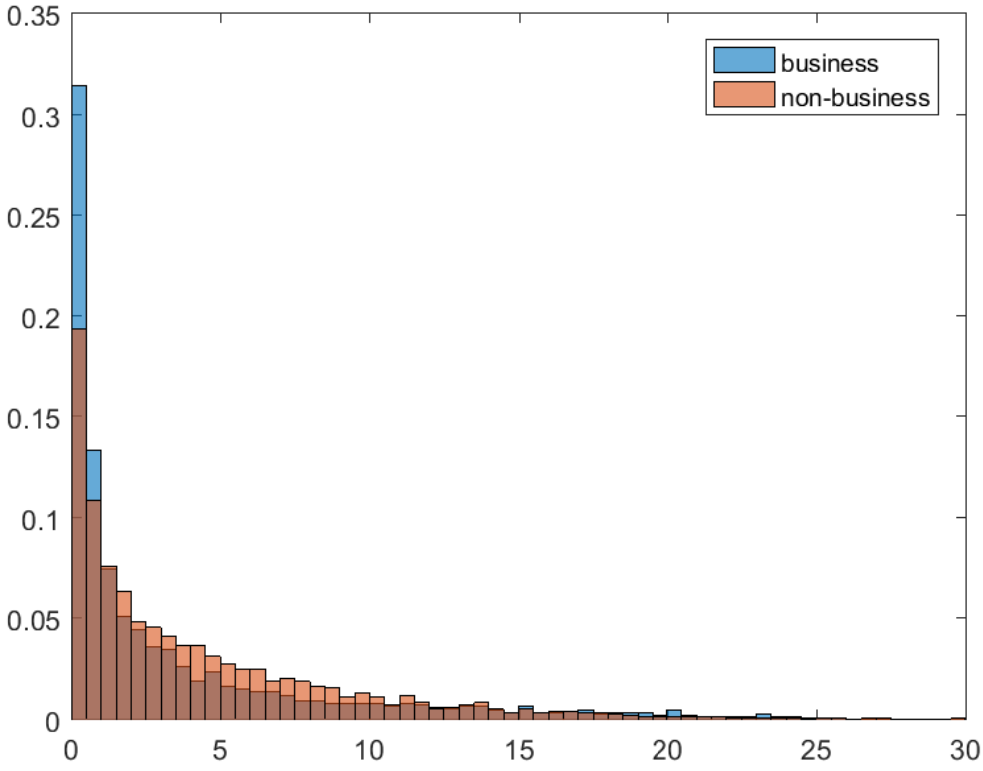


Table C.5: Results of Statistical Learning with Balance

Estimated probability of	AUC	MSE
CART (%)	80.49	17.36
Logit (%)	83.68	17.35

Table C.6: Sensitivity and Specificity with Balance

Triage	3	4	5
Sensitivity (%) ($\mathbb{P}(Pred = 1 \mid Act = 1)$)	79.34	73.13	59.18
Specificity (%) ($\mathbb{P}(Pred = 0 \mid Act = 0)$)	78.03	89.21	94.56

C.4 Delay of Specialist Requests

C.5 Proofs

Proof of Corollary 4.1.

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\left(\int_{a(t)}^{b(t)} f(x,t)\mathrm{d}x\right)\\ =&\int_{a(t)}^{b(t)} \frac{\partial f(x,t)}{\partial t}\mathrm{d}x + f(b(t),t)b'(t) - f(a(t),t)a'(t).\end{aligned}$$

□

Proof of Theorem 4.1. From the perspective of Lebesgue integral, the total waiting time is

$$W(t) = \int_0^t (t-\tau)\mathrm{d}N(\tau).$$

Because $M(t)$ is a martingale,

$$\mathbb{E}\left[\int_0^t (t-\tau)\mathrm{d}M(\tau)\right] = 0.$$

That is,

$$\mathbb{E}\left[\int_0^t (t-\tau)(\mathrm{d}N(\tau) - \lambda(\tau)\mathrm{d}\tau\right] = 0,$$
$$W(t) = \int_0^t (t-\tau)\lambda(\tau)\mathrm{d}\tau.$$

□

Proof of Proposition 4.1.

$$\mathbb{E}(L(24)) = \Lambda(24), \quad \mathcal{V}ar(L(24)) = \Lambda(24);$$
$$\mathbb{E}(L^2(24)) = \mathcal{V}ar(L(24)) + [\mathbb{E}(L(24))]^2.$$

□

Proof of Proposition 4.2 We show the result with a sample path argument. Let π be the optimal policy that always assign a customer if there exists to the server as long as the server completes a service; let π' be the alternative policy that does not assign a patient to an idle server. Obviously, π dominates π', because

$$\mathcal{W}_i^{\pi}(t) \leq \mathcal{W}_i^{\pi'}(t), \quad \forall t, \forall i \in \{0, 1, 2, \cdots, N\}$$

Therefore, $V^{\pi} \leq V^{\pi'}$. □

Proof of Proposition 4.3 Suppose all patients are ranked according to their probability such that $p_{i1} \geq p_{i2} \geq \cdots p_{iS_i(t)}$, $\forall i \in \{0, 1, 2, \cdots, N\}$. Consider a policy π that always assign the patient in class i with highest probability, and π assigns patients with due date T as a priority after threshold α and switch to prioritize the rest patients after β. Suppose other policies

- π_1 is the same with π except it priority random patients with each class;
- π_2 always prioritize non-FT patients, and prioritize patients with highest probability in each class;
- π_3 always prioritize FT patients, and prioritize patients with highest probability

in each class.

Denote $[vs, vf]$ is the duration of a service, such that $\alpha \leq vs \leq T \leq vf \leq \beta$ that is T happens during this service. Denote $v \triangleq vf - vs$, then

$$V^{\pi}(\{S_0(vs), S_1(vs), S_2(vs), \cdots, S_N(vs)\}) = v(\sum_{n=1}^{N} S_n(vs)) + p_{i1}(T - vs)$$
$$+ \sum_{\mathbf{S}(vf)} \mathbf{P}(\mathbf{S}(vf)|\mathbf{S}(vs)) V(\{S_0(vf), S_1(vf), S_2(vf), \cdots, S_i(vf) - 1, \cdots, S_N(vf)\})$$

$$V^{\pi_1}(\{S_0(vs), S_1(vs), S_2(vs), \cdots, S_N(vs)\}) = v(\sum_{n=1}^{N} S_n(vs)) + p_{i1}(T - vs)$$
$$+ \sum_{\mathbf{S}(vf)} \mathbf{P}(\mathbf{S}(vf)|\mathbf{S}(vs)) V(\{S_0(vf), S_1(vf), S_2(vf), \cdots, S_i(vf) - 1, \cdots, S_N(vf)\})$$
$$+(\Delta T - vf + vs)(p_{i1} - p_{i2})$$

So $V^{\pi_1} \geq V^{\pi}$.

$$V^{\pi_3}(\{S_0(vs), S_1(vs), S_2(vs), \cdots, S_N(vs)\}) = v(\sum_{n=1}^{N} S_n(vs)) + p_{i1}(T + \Delta T - vf)$$
$$+ \sum_{\mathbf{S}(vf)} \mathbf{P}(\mathbf{S}(vf)|\mathbf{S}(vs)) V(\{S_0(vf) - 1, S_1(vf), S_2(vf), \cdots, S_i(vf), \cdots, S_N(vf)\})$$

So $V^{\pi_3} \geq V^{\pi}$.

$V^{\pi_2} \geq V^{\pi}$ because the extra waiting time incur for non-FT patient when prioritizing FT patients who still wait in the system after the ED physician's treatment. $\square$

Appendix D

Design of Observation Units (OU) for Acute Decompensated Heart Failure (ADHF) Patients

D.1 Proofs

Proof of Lemma 5.1. Denote the reciprocal

$$R(n,\rho) \triangleq \frac{1}{B(n,\rho)}.$$

Let $S(n) = \sum_{i=0}^{n} \rho^i/i!$, then

$$\begin{aligned} R(n,\rho) &= \frac{S(n)}{\rho^n/n!} = \frac{S(n-1)+\rho^n/n!}{\rho^n/n!} \\ &= \frac{nR(n-1,\rho)}{\rho} + 1. \end{aligned}$$

Because

$$B(n,\rho) = \frac{\rho B(n-1,\rho)}{1+\rho B(n-1,\rho)}, \tag{D.1}$$

where $B(0,\rho) = 1$.

Using mathematical induction on n, because $B(0,\rho)$ is increasing in ρ, if $B(n,\rho)$ is also increasing in ρ, from Eq.D.1, $B(n+1,\rho)$ is increasing in ρ as well.

Moreover,

$$B(n-1,\rho) - B(n,\rho) = B(n-1,\rho)\left(1 - \frac{\rho}{1+\rho B(n-1,\rho)}\right) > 0,$$

this is because $B(n,\rho) \leq \max\{0, 1-\rho^{-1}\}$, which can be proved with Little's law as following.

Let N be the number of customers served in the system. The fraction of customers who are served is $\lambda(1-B(n,\rho))$, that is, the arrival rate excluding blocked customers. The expected waiting time is the mean service time in the loss model. So due to Little's Law $L = \lambda W$,

$$\mathbb{E}N = \frac{\lambda(1-B(n,\rho))}{\mu} = \rho(1-B(n,\rho)) < n,$$

Therefore $B(n,\rho) \leq \max\{0, 1-\rho^{-1}\}$. □

Proof of Proposition 5.1. The existence is trivial due to the monotonicity of $B(n,\rho)$ in n in Lemma 5.1. □

Proof of Proposition 5.2. With the monotonicity of $B(n,\rho)$ in n given a certain ρ in Lemma 5.1, we only need to prove the discrete convexity of $B(m,\rho)$ as a function of m, that is, $B(m,\rho) - B(m+1,\rho)$ is decreasing in m. The discrete convexity of $B(m,\rho)$ has first been proved in Messerli, 1972; Jagers and van Doorn, 1986 , and then (Wolff and Wang, 2002) extended the property to $G/GI/n/n$ models. □

Bibliography

American Academy of Physician Assistants (2010). Quality Incentive Program. Retrieved July 24, 2014,https://www.aapa.org/WorkArea/DownloadAsset.aspx?id=827

Abbass, Ibrahim, Trudy M. Krause, Salim S. Virani, J. Michael Swint, Wenyaw Chan, Luisa Franzini (2015) Revisiting the economic efficiencies of observations units *Managed Care* March 2015:46-52.

Adan, Ivo, Jacques Resing (2015) Queueing Systems. *Eindhoven University of Technology* Eindhoven, the Netherlands.

Adida E, Mamani H, Nassiri S (2016) Bundled Payment vs. Fee-for-Service: Impact of Payment Scheme on Performance. *Management Science* 63(5):1606–1624.

Afeche, P. (2013) Incentive-Compatible Revenue Management in Queueing Systems: Optimal Strategic Delay. *Manufacturing & Service Operations Management* 15(3): 423-443.

Affleck A., P. Parks, A.Drummond(2013) Canadian Association of Emergency Physicians Position Statement: Emergency department overcrowding and access block. *Canadian Journal of Emergency Medicine.* 15(6): 359-370.

Alberta Health Care Insurance Plan (AHCIP) (2016) Medical Procedure List. *Alberta Health Care Insurance Plan Schedule of Medical Benefits.* Accessed from http://www.health.alberta.ca/documents/SOMB-Medical-Procedures-2016-04.pdf. Accessed at November 30th, 2017.

Allard M, Jelovac I, Leger PT (2011) Treatment and referral decisions under different physician payment mechanisms. *Journal of Health Economics* 30(5):880–893.

American College of Cardiology (ACC). More hospitalizations, deaths for U. S. heart failure patients in winter. *ScienceDaily* ScienceDaily www.sciencedaily.com/releases/2017/03/170308150027.html (accessed on 8 March 2017).

Alberta Health Care Insurance Plan (AHCIP) (2016) Medical Procedure List. *Alberta Health Care Insurance Plan Schedule of Medical Benefits* http://www.health.alberta.ca/documents/SOMB-Medical-Procedures-2016-04.pdf.

American Medical Association (AMA) (2015) Evaluating pay-for-performance contracts. Accessed from https://www.acponline.org/system/files/documents/running_practice/payment_coding/commercial/ama-evaluating-p4p-contracs.pdf. Accessed at November 30th, 2017.

Ankjær-Jensen, A., P. Rosling and L. Bilde (2006). Variable prospective financing in the Danish hospital sector and the development of a Danish case-mix system. *Health Care Management Science* 9(3): 259-268

Arena, R., L.P. Whitsel, K. Berra, et al. (2015) Healthy lifestyle interventions to combat non-communicable disease: a novel non-hierarchical connectivity model for key stakeholders: a policy statement from the AHA, ESC, EACPR and ACPM. *Mayo Clin Proc.* 90:1082-1103.

Argon, Nilay Tanik, Serhan Ziya (2009) Priority Assignment Under Imperfect Information on Customer Type Identities. *Manufacturing & Service Operations Management* 11(4):674-693.

Ariadne Lab. (2017) Designing Capacity for High Value of Care: The Impact of Design in Clinical Care in Childbirth. Final Report, Available at https://massdesigngroup.org/sites/default/files/file/2017/170223_Ariadne\%20Report_Final.pdf. Accessed at November 30th, 2017.

Asplin B.R., D.J. Magid , K.V. Rhodes, et al. (2003) A conceptual model of emergency department overcrowding. *Annals of Emergency Medicine.* 42:181-184.

Asplin B.R., D.J. Magid(2007) If you want to fix crowding, start by fixing your hospital. Annals of Emergency Medicine; 49(3): 273-274.

Atar, Rami,Chanit Giat,Nahum Shimkin (2010) The $c\mu/\theta$ Rule for Many-Server Queues with Abandonment. *Operations Research* 58(5):1427-39.

Ata, B., B. L. Killaly, T. L. Olsen and R. P. Parker (2013). On Hospice Operations Under Medicare Reimbursement Policies. *Management Science* 59(5): 1027-1044.

Auble, Thomas E., Margaret Hsieh, William Gardner, Gregory F. Cooper, Roslyn A. Stone, Julie B. McCausland, Donald M. Yealy (2004) A Prediction Rule to Identify Low-risk Patients with Heart Failure. *ACAD EMERG MED* 12(6)514-521.

Baron, O., O. Berman, D. Krass and J. Wang (2014) Using Strategic Idleness to Improve Customer Service Experience in Service Networks. *Operations Research* 62(1): 123-140.

Bassamboo, Achal, Ramandeep S. Randhawa, Assaf Zeevi (2010) Capacity Sizing Under Parameter Uncertainty: Safety Staffing Principles Revisited. *Management Science* 56(10):1668-1686.

BC Health Authority (2016). BC MSC payment schedule index: BC Fee for Services for Obstetric and Gynaecology. Accessed from https://www2.gov.bc.ca/assets/gov/health/practitioner-pro/medical-services-plan/msc_payment_schedule.pdf. Accessed at November 30th, 2017.

Bellanger, M. M. and L. Tardif (2006). Accounting and reimbursement schemes for inpatient care in France. *Health Care Management Science* 9(3): 295-305.

Bertsimas, D., G. Mourtzinou (1997) Trasient laws of non-stationary queueing systems and their applications. *Queueing System* 25:115-155.

Beveridge, Robert, Barbara Clarke, Laurie Janes, Nancy Savage, Jim Thompson, Graham Dodd, Michael Murray, Cheri Nijssen-Jordan, David Warren, Alain Vadeboncoeur (2016). Implementation guidelines. *Canadian Association of Emergency Physicians* Available at http://caep.ca/resources/ctas/implementation-guidelines.

Biglaiser G, Ma CA (2007) Moonlighting: Public service and private practice. *Rand Journal of Economics* 38:1113–1133.

Biorn, E., T. P. Hagen, T. Iversen and J. Magnussen (2009) How different are hospitals' responses to a financial reform? The impact on efficiency of activity-based financing. *Health Care Management Science* 13(1): 1-16.

Biorn, E., T. P. Hagen, T. Iversen and J. Magnussen (2003). The Effect of Activity based financing on hospital efficiency: A Panel data analysis of DEA Efficiency Scores 1992-2000. *Health Care Management Science* 6: 271-283.

Biorn, E., T. P. Hagen, T. Iversen and J. Magnussen (2009). How different are hospitals' responses to a financial reform? The impact on efficiency of activity-based financing. *Health Care Management Science* 13(1): 1-16.

Birnbaum, H.G., R.C. Kessler, D. Kelley, R. Ben-Hamadi, V.N. Joish, P.E. Greenberg (2010) Employer burden of mild, moderate, and severe major depressive disorder: mental health services utilization and costs, and work performance *Depress Anxiety* 27(1):78-89.

Blake, J. T. and M. W. Carter (2002). A goal programming approach to strategic resource allocation in acute care hospitals. *European Journal of Operational Research* 140(3): 541-561.

Blake, J. T. and M. W. Carter (2003). Physician and hospital funding options in a public system with decreasing resources. *Socio-Economic Planning Sciences* 37(1): 45-68.

Bond K., Ospina M., Blitz S., et al. (2007) Frequency, determinants, and impact of overcrowding in emergency departments in Canada: a national survey of emergency department directors. *Healthcare Quarterly.* 10:32-40.

Brandeau, Margaret L., Francois Sainfort, William P. Pierskalla (2004) Health care delivery: current problems and future challenges. *Operations Research and Health Care: A Handbook of Methods and Applications.* Spinger. 1-14.

Brekke KR, Nuscheler R, Straume OR(2005) Gatekeeping in health care. *Journal of Health Economics* 26:149–170.

Bremaud, P.(1981) Point Processes and Queues: Martingale Dyanmics. *Springer-Verlag.*

Brick A, Layte R (2011) Exploring trends in the rate of caesarean section in ireland 1999-2007. *The Economic and Social Review* 42(4):383–406.

Brown Iii HS (1996) Physician demand for leisure: implications for cesarean section rates. *Journal of Health Economics* 15(2):233–242.

Bryant J, Porter M, Tracy SK, Sullivan EA (2007) Caesarean birth: consumption, safety, order, and good mothering. *Soc Sci Med* 65(6):1192–201.

Burns LR, Geller SE, Wholey DR (1995) The effect of physician factors on the cesarean section decision. *Medical Care* 33(4):365–382.

Cachon GP, Lariviere MA (2001) Contracting to assure supply: how to share demand forecasts in a supply chain. *Management Science* 47(5):629–646.

Canadian Association of Emergency Physicians and National Emergency Nurses Affiliation (CAEP & NENA) (2001) Joint position statement on emergency department overcrowding. *Canadian Journal of Emergency Medicine.* 3 (2): 82-84.

Chan L., K.M. Reilly , R.F.Salluzzo (1997) Variables that affect patient throughput times in an academic emergency department. *American Journal of Medicine Quarterly.* 12:183-186.

Chan, Carri W, Vivek F. Farias, Nicholas Bambos, Gabriel J. Escobar (2012) Optimizing Intensive Care Unit Discharge Decisions with Patient Readmissions. *Operations Research* 60(6): 1323-41.

Chan, Carri W, Jing Dong, Linda V. Green (2016) Queues with Time-Varying Arrivals and Inspections with Applications to Hospital Discharge Policies. *Operations Research* Forthcoming.

Chaix-Couturier C, Durand-Zaleski I, Jolle D, Durieux P (2000) Effects of financial incentives on medical practice: results from a systematic review of the literature and methodological issues. *International Journal for Quality in Health Care* 12(2):133–142.

Carter E.J.,S.M. Pouch ,E.L. Larson (2014) The Relationship Between Emergency Department Crowding and Patient Outcomes: A Systematic Review. *Journal of Nursing Scholarship.* 46(2): 106-115.

Catalyst for Payment Reform (CPR) (2012) Maternity care payment action brief. Available at http://www.pbgh.org/storage/documents/Issue_Brief.pdf. Accessed at November 30th, 2017.

Center of Disease Control (2014) Trends in Low-risk Cesarean Delivery in the United States, 1990?2013. *National Vital Statistics Reports* 63(6).

Cheng, Zhi-long (2010) Integrated production and outbound distribution scheduling: review and extensions. *Operations Research* 58(1):130-148.

Cheng, Feng, Jiazhen Huo (2013) The staffing requirement with time-varying demand and customer abandonment in call centers. *Innovation and Supply Chain Management* 7(1):019-024.

Cheng, T.C.E., M.Y. Kovalyov (2001) Single machine batch scheduling with sequential job processing. *IIE Transactions* 33:413-420.

Chhajed, Dilip (1995) A fixed interval due-date scheduling problem with earliness and due-date costs. *European Journal of Operations Research* 84:385-401.

Chick, S. E., H. Mamani and D. Simchi-Levi (2008). Supply chain coordination and influenza vaccination. *Operations Research* 56(6): 1493-1506.

Child Birth Connection (2011) Transforming maternity care blue print for action: major recommendations at a glance. Available at http://transform.childbirthconnection.org/wp-content/uploads/2011/03/Major-Recommendations-at-a-Glance.pdf.

Chu, H.-L., S.-Z. Liu, J. C. Romeis and C.-L. Yaung (2003). The Initial Effects of physician compensation programs in Taiwan hospitals: Implications for staff model HMOs. *Health Care Management Science.* 6: 17-26.

Canadian Institute for Health Information (CIHI) (2006) Giving birth in Canada: the costs. Available at https://secure.cihi.ca/free_products/Costs_Report_06_Eng.pdf. Accessed at November 30th, 2017.

Canadian Institute for Health Information (CIHI) (2012) National Health Expenditure Trends, 1975 to 2012. *Canadian Institute for Health Information.* 178.

Canadian Institute for Health Information (CIHI)(2014) Public Release on Emergency Care. *http://www.cihi.ca/cihi-extportal/internet/en/document/types+of+care/hospital+care/emergency+care/release*$_0$7102014.

Canadian Institute for Health Information (2016) Giving Birth in Canada: Regional Trends From 2001-2002 to 2013-2014. *Canadian Institute for Health Information* Ottawa, ON.

Clement, J., S. Grosskopt and V. G. Valdmanis (1996). A Comparison of Shadow prices and reimbursement ratea of hospital services. *Annals of Operations Research* 67: 163-182.

Center for Medicare & Medicaid Services. (2014, June 19,2014). EHR Incentive Programs . Retrieved July 24, 2014, http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html?redirect=/ehrincentiveprograms/

Collins, Sean P, Daniel P. Schauer, Amit Gupta, Hermine Brunner, Alan B. Storrow, Mark H. Eckman (2009) Cost-effectiveness analysis of ED decision making in patients with non-high-risk heart failure. *The American Journal of Emergency Medicine* 27:293-302.

Collins, Sean P., Peter S. Pang, Gregg C. Fonarow, Clyde W. Yancy, Robert O. Bonow, Mihai Gheorghiade (2013) Is Hospital Admission for Heart Failure Really Necessary? The Role of the Emergency Department and Observation Unit in Preventing Hospitalization and Rehospitalization *Journal of the American College of Cardiology* 61:121-126.

Collins, Sean P., Alan B. Storrow, Phillip D. Levy, Nancy Albert, Javed Butler, Justin A. Ezekowitz, G. Michael Felker, Gregory J. Fermann, Gregg C. Fonarow, Michael M. Givertz, Brian Hiestand, Judd E. Hollander, David E. Lanfear, Peter S. Pang, W. Frank Peacock, Douglas B. Sawyer, John R. Teerlink, and Daniel J. Lenihan (2015) Early Management of Patients With Acute Heart Failure: State of the Art and Future Directions - A Consensus Document from the SAEM/HFSA Acute Heart Failure Working Group *Academic Emergency Medicine* 22:94-112.

CPR (2012) Maternity care payment action brief.

Cunningham F, Leveno K, Bloom S, Hauth J, Rouse D, Spong C (2010) *Williams Obstetrics* (New York: McGraw-Hill), 23 edition.

Currie J, MacLeod WB (2008) First do no harm? tort reform and birth outcomes. *QUarterly Journal of Economics* 123(2):795–830.

Diagnosis and unnecesary procedure use: evidence from C-Section. *Working Paper National Bureau of Economic Research.*

Cutler, D. M. (2002). Equality, Efficiency, and Market Fundamentals: The Dynamics of International Medical-Care Reform. *Journal of Economic Literature* 40(3): 881-906.

Czypionka, T., M. Kraus, S. Mayer, G. Rö hrling (2014). Efficiency, ownership, and financing of hospitals: The case of Austria. *Health care management science*, 17(4), 331-347.

Dai, J G, Pengyi Shi(2017) A two-time-scale approach to time-varying queues in hospital inpatient flow management. *Operations Research* Forthcoming.

Das A, Gopalan S., Chandramohan D. (2016) Effect of pay for performance to improve quality of maternal and child care in low- and middle-income countries: a systematic review. *BMC Public Health* 16:321-327.

Davis, D. R. (2009) Declining Fruit and Vegetable Nutrient Composition. *Hort Science* 44:15.

Decady ,Yves, Lawson Greenberg (2014) Health at glace: ninety years of change in life expectancy. *Statistics Canada.* Catalogue no.82-624-X. ISSN 1925-6493. `http://www.statcan.gc.ca/pub/82-624-x/2014001/article/14009-eng.pdf` (accessed on March 25, 2017.)

Defraeye, Mieke, Inneke Van Nieuwenhuyse (2016) Staffing and scheduling under nonstationary demand for service: a literature review. *Omega* 58:4-25.

Diercks DB, Peacock WF, Kirk JD, Weber JE. (2006) ED patients with heart failure: identification of an observational unit-appropriate cohort. *Am J Emerg Med.* 24(3):319-324.

Dismuke, C. E. and V. Sena (1999). Has DRG payment influcenced technical efficiency and productivity of diagnostic technologies in Portuguese public hospitals? An empirical analysis using parametric and non-parametric methods. *Health Care Management Science* 2: 107-116.

Down, Douglas G.,Ger Koole, Mark E. Lewis (2011) Dynamic Control of a Single-server System with Abandonments. *Queueing System* 67:63-90.

EBioMedicine (2015) Increasing Healthspan: Prosper and Live Long EBioMedicine. *EBioMedicine* 2:1559 `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4740330/pdf/main.pdf` (access on 26 March 2017).

Eggleston K. (2005). Multitasking and mixed systems for provider payment.*Journal of Health Economics* 24(1):211-23.

Eijkenaar F. (2013). Key issues in the design of pay for performance programs.*European Journal of Health Economics* 14(1):117-131.

Ellis, R. P. (1998). Creaming, skimping and dumping: provider competition on the intensive and extensive margins. *Journal of Health Economics* 17(5): 537-555.

Epstein, D. and A. Mason (2006). Costs and prices for inpatient care in England: Mirror twins or distant cousins? . *Health Care Management Science* 9(3): 233-242.

Epstein AJ, Nicholson S (2009) The formation and evolution of physician treatment styles: An application to cesarean sections. *Journal of Health Economics* 28:1126–1140.

Fabbri D, Monfardini C (2008) Style of practice and assortative mating: a recursive probit analysis of caesarean section scheduling in italy. *Applied Economics* 40(11):1411–1423.

Faloon T (2012) Physician remuneration options. *Canadian Medical Association.* Available at `https://www.cma.ca/Assets/assets-library/document/en/practice-management-and-wellness/module-8-physician-remuneration-options-e.pdf`. Accessed at November 30th, 2017.

Fattore, G. and A. Torbica (2006). Inpatient reimbursement system in Italy: How do tariffs relate to costs?. *Health Care Management Science* 9(3): 251-258.

Feasby, T. E. and C. Gerdes (2006). Pay-for-performance: Can it work in Canada?. *Healthcare Papers* 6(4): 47-50.

Fang J, Mensah GA, Croft JB, Keenan NL (2008) Heart failure-related hospitalization in the U.S., 1979 to 2004. *J Am Coll Cardiol.* 52(6):428-34.

Feinstein AJ, Soulos PR, Long JB, Herrin J, Roberts BK, Yu JB, Gross CP (2013) Variation in receipt of radiation thepary after breast-conserving surgery: Assessing the impact of physicians and geographic regions. *Medical Care* 51:330–338.

Fetter, R. B. (1991). Diagnosis Related Groups: Understanding Hospital Performance. *Interfaces* 21(1): 6-26.

Feder, J. (2013) Bundle with care- Rethinking Medicare incentives for post-acute care services. *New England Medicine* 369:400-401.

Fonarow, Gregg C., Kirkwood F. Adams Jr, William T. Abraham, Clyde W. Yancy, W. John Boscardin (2005) Risk Stratification for In-Hospital Mortality in Acutely Decompensated Heart Failure: Classification and Regression Tree Analysis *JAMA: American Medical Association* 293(5):572-580.

Foo PK, Lee RS, Fong K (2013) Hospital and physician prices and treatment choice in labor and delivery .

Forster A.J., I. Stiell , G. Wells, A.J. Lee, C.V. Walraven (2003) The Effect of Hospital Occupancy on Emergency Department Length of Stay and Patient Disposition. *Acedemic Emergency Medicine.* 10(2):127-133.

Fuloria PC, Zenios SA (2001) Outcomes-adjusted reimbursement in a health-care delivery system. *Management Science* 47(6):735–751.

Font JCi (2009) Do incentives, complexity and the demand for leisure explain caesarean-section deliveries? *International Journal of Social Economics* 36(9):906–915.

Friesner, D. L. and R. Rosenman (2004). Inpatient-outpatient cost shifting in Washington hospitals. *Health Care Management Science* 7: 17-26.

Fuloria, P. C. and S. A. Zenios (2001). Outcomes-Adjusted Reimbursement in a Health-Care Delivery System. *Management Science* 47(6): 735-751.

Gaal, P., N. Stefka and J. Nagy (2006). Cost accounting methodologies in price setting of acute inpatient services in Hungary. *Health Care Management Science* 9(3): 243-250.

Geelhoed G.H., N.H. Klerk (2012) Emergency department overcrowding, mortality and the 4-hour rule in Western Australia. *Medical Journal of Australia.* 196:122?126.

Ghaffarzadegan N, Epstein AJ, Martin EG (2013) Practice variation, bias, and experiential learning in cesarean delivery: A data-based system dynamics approach. *Health Services Research* 2013(48):713–734.

Gheorghiade M, De Luca L, Fonarow GC, Filippatos G, Metra M, Francis GS.(2005) Pathophysiologic targets in the early phase of acute heart failure syndromes. *Am J Cardiol* 96(suppl):11G-17G.

Gheorghiade M, Abraham WT, Albert NM, et al.(2006) Systolic blood pressure at admission, clinical characteristics, and outcomes in patients hospitalized with acute heart failure. *JAMA* 296:2217-26.

Gilboy, Nicki, Paula Tanabe, Debbie Travers, Alexander M. Rosenau (2011) Emergency Severity Index (ESI) A Triage Tool for Emergency Department Care (Version 4) Implementation Handbook 2012 Edition. *Agency for Healthcre Research and Quality (AHRQ)* 12-0014.

Goer, H., A. Romano and C. Sakala (2012) Vaginal or Cesarean birth? A Systematic Review to Determine What is at Stake for Mothers and Babies. *New York: Childbirth Connection.*

Goh, J., J. Pfeffer, S. Zenios (2015) Workplace stressors and health outcomes: health policy for the workplace *Behav Sci Policy* 1(1):43-52.

Goldfield, N. (2010). The evolution of diagnosis-related groups (DRGs): From its beginnings in case mix and resource use theory, to its implementation for payment and now for its current utilization for quality within and outside the hospital. *Quality Management in Health Care* 19(1): 3-16.

Gonzalez P (2004) Should physicians' dual practice be limited? An incentive approach. *Health Economics* 13:505–524.

Goodacre S, Nicholl J, Dixon S, et al.(2004) Randomised controlled trial and economic evaluation of a chest pain observation unit compared with routine care. *BMJ.* 328:254.

Gosden T, Forland F, Kristiansen I, Sutton M, Leese B, Giuffrida A, Sergison M, Pedersen L (2000) Capitation, salary, fee-forservice and mixed systems of payment: effects on the behaviour of primary care physicians. *Cochrane Database of Systematic Reviews* (3).

Goyert G, Bottoms S, Treadwell M, Nehra P (1989) The physician factor in cesarean birth rates. *New England Journal of Medicine* 320:706–709.

Graff L, Orledge J, Radford MJ, Wang Y, Petrillo M, Maag R.(1982) Correlation of the Agency for Health Care Policy and Research congestive heart failure admission guideline with mortality: peer review organization voluntary hospital association initiative to decrease events (PROVIDE) for congestive heart failure. *Ann Emerg Med.* 34(4 Pt 1):429-437.

Greene WH, Hensher DA (2010) Modeling ordered choices: a primer. *Cambridge University Press.*

Gregory KD, Curtin SC, Taffel SM, Notzon FC (1998) Changes in indications for cesarean delivery: United States, 1985 and 1994. *American Journal of Public Health* 88(9):1384-1387.

Gruber J, Kim J, Mayzlin D (1998) Physician fees and procedure intensity: The case of cesarean delivery. *National Bureau of Economic Research Working Paper Series* No. 6744.

Gruber J, Owings M (1996) Physician financial incentives and cesarean section delivery. *Rand Journal of Economics* 27(1):99–123.

Grytten J, Sorensen R (2003) Practice variation and physician-specific effects. *Journal of Health Economics* 22(3):403–418.

Grytten J, Skau I, Sorensen R (2013) Do mothers decide?:the impact of preferences in healthcare. *Journal of Human Resources*48(1):142-168.

Gupta, D., M. Mehrotra (2015). Bundled Payments For Healthcare Services: Proposer Selection and Information Sharing. *Operations Research* 63(4): 772-788.

Gururaj VJ, Allen JE, Russo RM.(1972) Short stay in an outpatient department. An alternative to hospitalization. *Am J Dis Child.* 123:128-132.

Hall, Nicholas G., 'Maseka Lesaoana, Chris N. Potts (2001) Scheduling with fixed delivery dates. *Operational Research* 49(1):134-144.

Haughom, John, Paul Horstmeier, John Wadsworth, Russ Steheli, Leslie Hough Falk (2014) The changing role of healthcare data analysts - how our most successful clients are embracing healthcare transformation. *Health Catalyst.* `https://www.healthcatalyst.com/wp-content/uploads/2014/12/whitepaper-Changing-Role-Healthcare-Data-Analysts.pdf`(accessed on 27 March 2017).

Harper, PR, Shahani AK (2002) Modelling for the planning and management of bed capacities in hospitals. *The Journal of the Operational Research Society* 53(1):11-18.

Harper PR, Winslett DJ (2006) Classification trees: A possible method for maternity risk grouping. *European Journal of Operational Research* 169(1):146–156.

Hatem M, Sandall J, Devane D, Soltani H, Gates S (2008) Midwife-led versus other models of care for childbearing women. *The Cochrane Database of Systematic Reviews* 4.

Health Care Financial Review (1996). Medicare program payments: 1967 and 1994. (1996 Supplement): 28-29.

He, Beixiang, Yunan Liu, Ward Whitt (2016) Staffing a service system with non-Poisson non-stationary arrivals. *Probability in the Engineering and Informational Sciences* 30:593-621.

Health Analytics (2015) Healthcare Big Data Analytics: From Description to Prescription. `http://healthitanalytics.com/news/healthcare-big-data-analytics-from-description-to-prescription` (accessed on April 2, 2017).

Health Canada (2012). Health Care System Patient Wait Times Guarantee (PWTG) Pilot Project Fund. Retrieved July 24, 2014, `http://www.hc-sc.gc.ca/hcs-sss/finance/hcpcp-pcpss/pwgt-gtap-eng.php`

Health Canada (2012).Health Care System Patient Wait Times Guarantee (PWTG) Pilot Project Fund.

Heart & Stroke Foundation (1982) The burden of heart failure. *2016 Report on the Health of Canadians* heartandstroke.ca.

Heidenreich PA, Trogdon JG, Khavjou OA, et al.(2011) Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. *Circulation* 123:933-44.

Halweil, Brian (2007) Still no free lunch: Nutrient levels in U.S. food supply eroded by pursuit of high yields. *Critical Issue Report. The Organic Center* `https://www.organic-center.org/reportfiles/Yield_Nutrient_Density_Final.pdf` (access on March 26, 2017).

Herwartz, H. and C. Strumann (2012). On the effect of prospective payment on local hospital competition in Germany. *Health Care Management Science* 15(1): 48-62.

Hillman A, Welch W, Pauly M (1992) Contractual arrangements between hmos and primary care physicians: three tiered hmos and risk pools. *Medical Care* 30:136–148.

Holt N.R. D. Aronsky (2008) Systematic Review of Emergency Department Crowding: Causes, Effects, and Solutions. *Annals of Emergency Medicine.* 52:126-136.

Honaker J, King G, Blackwell M (2015) A Program for Missing Data. `https://cran.r-project.org/web/packages/Amelia/Amelia.pdf`

Hostetler B, Leikin JB, Timmons JA, Hanashiro PK, Kissane K. (2002) Patterns of use of an emergency department-based observation unit. *Am J Ther.* 9:499-502.

Hu, B. and S. Benjaafar (2009) Partitioning of Servers in Queueing Systems During Rush Hour. *Manufacturing & Service Operations Management* 11(3): 416-428.

Hua, Z., W. Chen, Z. Zhang (2016). Competition and Coordination in Two-Tier Public Service Systems under Government Fiscal Policy. *Production and Operations Management Forthcoming*

Huang,Junfei, Boaz Carmeli, Avishai Mandelbaum (2015) Control of Patient Flow in Emergency Department, or Multiclass Queues with Deadlines and Feedback. *Operations Research* 63(4):892-908.

Hueston W (1994) Development of a cesarean delivery risk score. *Obstetrics & Gynecology* 84(6):965-968.

Hutchison, B., J. Hurley, S. Birch, J. Lomas, S. D. Walter, J. Eyles and F. Stratford-Devai (2000). Needs-based primary medical care capitation: Development and evaluation of alternative approaches. *Health Care Management Science* 3: 89-99.

Ibrahim, R., B. Kucukyazici, V. Verter, M. Gendreau, and M. Blostein (2015) Designing Personalized Treatment: An Application to Anticoagulation Therap. *Production and Operations Management.* 25(5):902-918.

Institute of Medicine (2001). Crossing the Quality Chasm: A New Health Systems for the 21st Centuary. Washington DC, National Academy Press.

Institute of Medicine (2007). Rewarding Provider Performance: Aligning Incentives in Medicalcare. Washington DC, National Academy Press.

Jacobson, Evin Uzun, Nilay T. Argon, Serhan Ziya (2012) Due-date Scheduling: Asymptotic Optimality of Generalized Longest Queue and Generalized Largest Delay Rules. *Operations Research* 60(4):813-832.

Jagers, A.A and E. A. van Doorn (1986) On the continued Erlang loss function. *Operations Research Letters* 5:43-46.

Janiak, Adam, Tomasz Krysiak (2007) Single processor scheduling with job values depending on their completion times. *J. Sched.* 10:129-138.

Jiang H, Pang Z, Savin S (2012) Performance-based contracts for outpatient medical services. *Manufacturing & Service Operations Management* 14(4):654–669.

Jiang H, Pang Z, Savin S (2017) Improving patient access to care: performance incentives and competition in healthcare markets. Working paper, available at https://www.jbs.cam.ac.uk/fileadmin/user_upload/research/workingpapers/wp1701.pdf. Accessed at November 30th, 2017.

Joseph, Susan M., Ari M. Cedars, Gregory A. Ewald, Edward M. Geltman, and Douglas L. Mann (2009) Acute Decompensated Heart Failure: Contemporary Medical Management. *Tex Heart Inst J.* 36(6): 510âĂŞ520.

Johnson EM, Rehavi MM (2016) Physicians treating physicians: information and incentives in Childbirth. *American Economic Journal: Economic Policy* 8(1):115-141.

Joustra, P., E. van der Sluis and N. M. van Dijk (2009) To pool or not to pool in hospitals: a theoretical and practical comparison for a radiotherapy outpatient department. *Annals of Operations Research* 178(1): 77-89.

Keeler E, Fok T (1996) Equalizing physician fees had little effect on cesarean rates. *Medical Care Research and Review* 53:465–471.

Kessler, R.C., H.S. Akiskal, M. Ames, et al.(2006) Prevalence and effects of mood disorders on work performance in a nationally representative sample of U.S. workers *Am J Psychiatry* 163(9):1561-1568.

Kim B (2010) Do doctors induce demand? *Pacific Economic Review* 15(4):554–575.

Kwapien, Agata (2016) Top 5 examples of big data in healthcare that can save peopleâĂŹs lives. *Datapine.* http://www.datapine.com/blog/big-data-examples-in-healthcare/# (accessed on 27 March 2017).

Kocaga, Yasar Levent, Mor Amony, Amy R. Ward (2015) Staffing call centers with uncertain arrival rates and co-sourcing. *Production and Operations Management* 24(7):1101-17.

Kolassa, E. M. (1997). Elements of Pharmaceutical Pricing. NY, USA, the Pharmaceutical Products Press.

Kosowsky JM, Gasaway MD, Hamilton CA, Storrow AB. (2000) Preliminary experience with an emergency department observation unit protocol for heart failure. *Acad Emerg Med.* 7(10):1171.

Knight, M., J. J. Kurinczuk, P. Spark, P. Brocklehurst et al.(2008) Cesarean Delivery and Peripartum Hysterectomy. *Obstetrics & Gynecology* 111(1): 97-105.

Knight M, Kurinczuk JJ, Spark P, Brocklehurst P, Committee ftUKOSSS (2008) Cesarean delivery and peripartum hysterectomy. *Obstetrics & Gynecology* 111(1):97–105 10.1097/01.AOG.0000296658.83240.6d.

Kucukyazici B., Zhu C. (2017) Design of Financial Incentives and Payment Schemes in Healthcare Systems. *Wiley Encyclopedia of Operations Research and Management Science* Under review.

Lally, Sarah (2013) Transforming Maternity Care: A Bundled Payment Approach. *Integrated Healthcare Association* No.10, September 2013.

Lee, Chung-Yee, Chung-Lun Li (1996) On the fixed interval due-date scheduling problem. *Discrete Applied Mathematics* 68:101-117.

Lee, DKK, Zenios SA (2012) An evidence-based incentive system for medicare's end-stage renal disease program. *Management Science* 58(6):1092–1105.

Lee P.A., B.H. Rowe, G. Innes, et al. (2013) Assessment of consultation impact on emergency department operations through novel metrics of responsiveness and decision making efficiency. *Canadian Journal of Emergency Medicine.* 1-8.

Leger, P. T. (2011). Physician payment mechanisms: an overview of policy options for Canada. *CHSRF Series on cost drives and health system efficiency* 3.

Legér, P. T. (2008). Physician Payment Mechanisms. Financing Health Care. L. Mingshan and E. Jonsson, Weinheim: Wiley - VCH: 149 - 172.

Levi, R., G. Perakis, G. Romero (2016) On the Effectiveness of Uniform Subsidies in Increasing Market Consumption. *Management Science.* 63(1):40-57.

Li, Dong, Kevin D. Glazebrook (2010) An Approximate Dynamic Programing Approach to the Development of Heuristics for the Scheduling of Impatient Jobs in a Clearing System. *Naval Research Logistics* 57(3):225-236.

Liu, Yunan, Ward Whitt(2012) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations Research* 60(6):1551-1564.

Liu, Cheng-Hsiang, Cheng-I Hsu (2015) Dynamic job shop scheduling with fixed interval deliveries. *Prod. Eng. Res. Devel.* 9:377-391.

Logeart, Damien, Gabriel Thabut, Patrick Jourdain, Christophe Chavelas, Pascale Beyne, Florence Beauvais, Erik Bouvier, Alain Cohen Solal (2004). Predischarge B-type natriuretic peptide assay for identifying patients at high risk of re-admission after decompensated heart failure. *Journal of the American College of Cardiology.* 43(4): 635-41.

Lovejoy, William S., Jeffrey S. Desmond (2011) Little's law flow analysis of observation unit impact and sizing. *Academic Emergency Medicine* 18:183-189.

Lucas R., H. Farley, J. Twanmoh, et al. (2009) Emergency department patient flow: the influence of hospital census variables on emergency department length of stay. *Academic Emergency Medicine.* 16:597-602.

Mace SE. (2001) Asthma therapy in the observation unit. *Emerg Med Clin North Am.* 19:169-185.

Mace SE, Graff L, Mikhail M, Ross M. (2003) A national survey of observation units in the United States. *Am J Emerg Med.* 21:529-533.

Mallor, Fermin, Cristina Azcarate, Julio Barado (2015) Optimal control of ICU patient discharge: from theory to implementation. *Health Care Management Science* 18:234-250.

Main EK, Morton CH, Hopkins D, Giuliani G, Melsop K and Gould JB (2011) Cesarean deliveries, outcomes, and opportunities for change: toward a public agenda for maternity care safety and quality. *California Maternal Quality Care Collaborative* White Paper, Palo Alto, CA.

Mahjoub, R., F. Odegaard, G. S. Zaric (2014). Health-based pharmaceutical pay-for-performance risk-sharing agreements. *Journal of the Operational Research Society.* 65(4), 588-604.

Mallor, Fermin, Cristina Axcarate, Julio Barado (2015). Optimal control of ICU patient discharge: from theory to implementation *Health Care Management Science* 18:234-250.

Malvankar-Mehta, M. S. and B. Xie (2012). Optimal incentives for allocating HIV/AIDS prevention resources among multiple populations. *Health Care Management Science* 15(4): 327-338.

Maman, Shimrit (2009) Uncertainty in the demand for service: the case of call centers and emergency department. *Master's thesis, Technion - Israel Institute of Technology,* Haifa, Israel.

Mamani H, Chick SE, Simchi-Levi D (2013) A game-theoretic model of international influenza vaccination coordination. *Management Science* 59(7):1650–70.

Marr, Bernard (2015) How big data is changing healthcare. *Forbes.* `https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/#4afaec3e2873` (accessed on 27 March 2017).

Mathijsen, B.W.J., A.J.E.M. Janssen, J.S.H. van Leeuwaarden, A.P. Zwart (2017) Robust heavy-traffic approximations for service systems facing overdispersed demand. Working paper.

McCarthy M.L., S.L. Zeger, R. Ding, et al. (2009) Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients. *Annals of Emergency Medicine.* 54(4): 492-503.

McCusker J., A. Vadeboncoeur, J.F. Lévesque , et al. (2014) Increases in Emergency Department Occupancy Are Associated With Adverse 30-day Outcomes. *Academic Emergency Medicine.* 21(10): 1092-1100.

McDermott MF, Murphy DG, Zalenski RJ, et al.(1997) A comparison between emergency diagnostic and treatment unit and inpatient care in the management of acute asthma. *Arch Intern Med.* 157:2055-2062.

Messerli, E.J. (1972). Proof of a convexity property of the Erlang B formula. *Bell System Tech. J.*. 51:951-953.

McKillop, I., G. H. Pink and L. M. Johnson (2001). Acute care in Canada: A Review of Funding, Performance, Monitoring and Reporting Practices. *Canadian Institute of Health Information (CIHI)*. Ottawa, Ontario.

Michota, Franklin A. Jr, Alpesh Amin (2008) Bridging the gap between evidence and practice in acute decompensated heart failure management *Journal of Hospital Medicine* 3(S6) S7-S15.

Ministère de la Santé et des Services Sociaux du Québec (MSSS) (2010) Plan strategique 2010-2015.

Ministère de la Santé et des Services Sociaux du Québec (MSSS) (2011) Rapport Annuel De Gestion 2010-2011. Québec: Gouvernement du Québec. `http://publications.msss.gouv.qc.ca/acrobat/f/documentation/2011/11-102-01F.pdf`.

Health Services and Medical Management Division (2010) Perinatal Practices. *Minnesota Department of Human Services Legislative Report* Available at `https://www.leg.state.mn.us/docs/2010/mandated/100219.pdf`.

Morey, R. C. and D. A. Dittman (1984). Hospital Profit Planning under Medicare Reimbursement. *Operations Research* 32(2): 250-269.

Morey, R. C. and D. A. Dittman (1996). Cost Pass through reimbursement to hospitals and their impacts on operating efficiencies. *Annals of Operations Research* 67: 117-139.

Medical Services Commission Payment Schedule (2006) Obstretrics and Gynecology. *British Columbia Medical Services Commission Payment Schedule* `http://www2.gov.bc.ca/assets/gov/health/practitioner-pro/medical-services-plan/msc-payment-schedule-june-2016.pdf`.

Nedelea, I. C. and J. M. Fannin (2013). Technical efficiency of Critical Access Hospitals: an application of the two-stage approach with double bootstrap. *Health Care Management Science* 16(1): 27-36.

Nieminen MS, Harjola VP (2005) Definition and epidemiology of acute heart failure syndromes. *Am J Cardiol.* 96(6A):5G-10G.

OECD Health Statistics (Database) (2014). OECD Health Data: Health expenditure and financing.

Ontario Health Insurance Plan (2016) Physician Services Under the Health Insurance Act. *Ontario Ministry of Health and Long-Term Care* http://www.health.gov.on.ca/en/pro/programs/ohip/sob/.

Oliveira, M. D. and G. Bevan (2008). Modelling hospital costs to produce evidence for policies that promote equity and efficiency. *European Journal of Operational Research* 185(3): 933-947.

Omar, M. Ashour, G. E. Okundan Kermer (2016) Dynamic patient grouping and prioritization: a new approach to emergency department flow improvement. *Health Care Management Science* 19:192-205.

O'Neill L, Kuder J (2005) Explaining variation in physician practice patterns and their propensities to recommend services. *Medical Care Research and Review* 62:339–357.

Ontario Health Insurance Plan (OHIP) (2016) Physician Services Under the Health Insurance Act. *Ontario Ministry of Health and Long-Term Care* Accessed from http://www.health.gov.on.ca/en/pro/programs/ohip/sob/. Accessed at November 30th, 2017.

Optum (2013) Can value-based reimbursement models transform health care? White Paper. Available at https://www.optum.com/content/dam/optum/resources/whitePapers/can-value-base-reimburesment-models-transform.pdf. Accessed at November 30th, 2017.

Ospina M.B., Bond K., Schull M., et al. (2006) Measuring overcrowding in emergency departments: a call for standardization . *Ottawa: Canadian Agency for Drugs and Technologies in Health.* [Technology report no 67.1]

Peacock, WF 4th, Remer EE, Aponte J, Moffa DA, Emerman CE, Albert NM.(2002) Effective observation unit treatment of decompensated heart failure. *Congest Heart Fail* 8(2):68-73.

Peacock, S. and L. Segal (2000). Capitation funding in Australia: Imperatives and impediments. *Health Care Management Science* 3: 77-88.

Peacock WF, Fonarow GC, Ander DS, et al.(2009) Society of Chest Pain Centers recommendations for the evaluation and management of the observation stay acute heart failure patient-parts 1-6. *Acute Card Care* 11:3-42.

Peacock WF, Braunwald E, Abraham W, et al.(2010) National Heart, Lung, and Blood Institute working group on emergency department management of acute heart failure: research challenges and opportunities. *J Am Coll Cardiol* 56:343-51.

Peacock, S. and L. Segal (2000) Capitation funding in Australia: Imperatives and impediments. *Health Care Management Science* 3: 77-88.

Porter, M.E., T.H. Lee (2013) The Strategy That Will Fix Health Care. *Harvard Business Review* October Issue.

Potts, Chris N., Mikhail Y. Kovalyov (2000) Scheduling with batching: a review. *European Journal of Operations Research* 120:228-249.

Pruss-Ustun, A., J Wolf, C CorvalÃąn, R Bos and M Neira (2016) Preventing disease through healthy environments: a global assessment of the burden of disease from environmental risks. *World Health Organization (WHO) publication.* `http://apps.who.int/iris/bitstream/10665/204585/1/9789241565196_eng.pdf?ua=1` (accessed on March 25, 2017.)

Puenpatom, R. A. and R. Rosenman (2008). Efficiency of Thai provincial public hospitals during the introduction of universal health coverage using capitation. *Health Care Management Science* 11(4): 319-338.

Qureshi, Waqas Tariq, aved Butler, Sean P. Collins, Alec J. Moorman, Mihai Gheorghiade (2015) Early Medical Management of Hospitalization for Heart Failure (HHF) *Management of Heart Failure: Volume 1: Medical* edited by Baliga, Ragavendra R. and Haas, Garrie J. Springer London. 113-149. ISBN 978-1-4471-6657-3.

Raghupathi, W., Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems* 2-3.

Ransom S, McNeeley S, Kruger M, Doot G, Cotton D (1996) The effect of capitation and fee-for-service remuneration on physician decision making in gynecology. *Obstetrics & Gynecology* 87:707–710.

Roberge D., Pineault R., Larouche D. et al. (2010) The Continuing Saga of Emergency Room Overcrowding: Are We Aiming at the Right Target? *Healthcare Policy.* 5(3): 27?39.

Robinson, J. C. (2001). Theory and Practice in the Design of Physician Payment Incentives. *The Milbank Quarterly* 79(2): 149-177.

Roger VL, Go AS, Lloyd-Jones DM, et al.(2011) Heart disease and stroke statistics 2011 update: a report from the American Heart Association. *Circulation* 123:e18-e209.

Rodwin M (1993) *Medicine, Money and Morals: Physicians' Conflicts of Interest* (Oxford: Oxford University Press).

Rosenman, R. and T. Li (2002). Buying healthcare Quality with Grants and Donations. *Health Care Management Science* 5: 25-31.

Ross MA, Compton S, Richardson D, Jones R, Nittis T, Wilson A. (2003) The use and effectiveness of an emergency department observation unit for elderly patients. *Ann Emerg Med.* 41:668-677.

Ross H, Howlett J, Arnold JM, et al.(2006) Treating the right patient at the right time: access to heart failure care. *Can J Cardiol* 22:749-54.

Ross, Michael A., Taruna Aurora, Louis Graff, Pawan Suri, Rachel O'Malley, Aderonke Ojo, Steve Bohan, and Carol Clark (2012) State of the Art: Emergency Department Observation Units. *Critical Pathways in Cardiology* 11(3):128-138.

Rouse, P. and R. Swales (2006). Pricing public health care services using DEA: Methodology versus politics. *Annals of Operations Research* 145(1): 265-280.

Sagner, M., A. McNeil, P. Puska, et al.(2017) The P4 health Spectrum - a predictive, preventive, personalized and participatory continuum for promoting healthspan. *Prog Cardiovasc Dis* 59:506-520.

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond and S. L. Kronick (2012) Patient Streaming as a Mechanism for Improving Responsiveness in Emergency Departments. *Operations Research* 60(5):1080-97.

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond and S. L. Kronick (2014) Complexity-Augmented Triage: A Tool for Improving Patient Safety and Operational Efficiency. *Manufacturing & Service Operations Management* 16(3): 329-345.

Sakala C, Corry M (2008) Evidence-based maternity care: what it is and what it can achieve. *Childbirth Connection, Reforming States Group, Milbank Memorial Fund* 2008 October. Available at `www.milbank.org/reports/0809MaternityCare/0809MaternityCare.html`.

Sánchez-Martínez, F., J.-M. Abellán-Perpiñàn, J.-E. Martínez-Pérez and J. Puig-Junoy (2006). Cost accounting and public reimbursement schemes in Spanish hospitals. *Health Care Management Science* 9(3): 225-232.

Schrager, Justin, Matthew Wheatley, Vasiliki Georgiopoulou, Anwar Osborne, Andreas Kalogeropoulos, Olivia Hung, Javed Butler, Michael Ross (2013) Favorable Bed Utilization and Readmission Rates for Emergency Department Observation Unit Heart Failure Patients *Academic Emergency Medicine* 20(6)554-561.

Schreyögg, J., O. Tiemann and R. Busse (2006). Cost accounting to determine prices: How well do prices reflect costs in the German DRG-system?. *Health Care Management Science* 9(3): 269-279.

Schull M.J., P.M. Slaughter , D.A. Redelmeier(2002) Urban emergency department overcrowding: defining the problem and eliminating misconceptions. *Canadian Journal of Emergency Medicine.* 4:76-83.

Schull M.J., K. Lazier, M. Vermeulen, et al. (2003) Emergency department contributors to ambulance diversion: a quantitative analysis. *Annals of Emergency Medicine.* 41:467-476.

Seddik, Yasmina, Christophe Gonales, Safia Kedad-Sildhoum (2013) Single machine scheduling with delivery dates and cumulative payoffs. *Journal of Scheduling* 16:313-329.

Seddik, Yasmina, Christophe Gonales, Safia Kedad-Sildhoum (2015) Performance guarantees for a scheduling problem with common stepwise job payoffs. *Theoretical Computer Science* 562:377-394.

Setoguchi S, Stevenson LW, Schneeweiss S. (2007) Repeated hospitalizations predict mortality in the community population with heart failure. *Am Heart J* 154:260-6.

Sharma, A. (2008). Inter-DRG resource dynamics in a prospective payment system: a stochastic kernel approach. *Health Care Management Science* 12(1): 38-55.

Shmueli, Amir, Charles L. Sprung, Edward H. Kaplan (2003). Optimizing admissions to an Intensive Care Unit. *Health Care Management Science* 6:131-136.

Shumsky RA, Pinker EJ (2003) Gatekeepers and referrals in services. *Management Science* 49(7):839–856.

Shwartz, M. and M. L. Lenard (1994). Improving Economic Incentives in Hospital Prospective Payment Systems Through Equilibrium Pricing. *Management Science* 40(6): 774-787.

Shwartz, M., J. F. Burgess Jr, J. Zhu (2016). A DEA based composite measure of quality and its associated data uncertainty interval for health care provider profiling and pay-for-performance. *European Journal of Operational Research* 253(2): 489-502.

Smith WR, Poses RM, McClish DK, Huber EC, Clemo FL, Alexander D, Schmitt BP.(2002) Prognostic judgments and triage decisions for patients with acute congestive heart failure. *Chest* 121:1610-1617.

Smith, M., C.Baker, L.Branch, R.Walls, R.Grimes, J.Karklins, M.Kashner, R.Burrage, A.Parks and P.Rogers(1992)Case-mix groups for VA hospital-based home care. *Medical Care* 30(1):1-16.

So KC, Tang CS (2000) Modeling the impact of an outcome-oriented reimbursement policy on clinic, patients, and pharmaceutical firms. *Management Science* 46(7):875–892.

Sokolowski, John A., Catherine M. Banks (2009). Principles of modeling and simulation: a multidisciplinary approach. *John Wiley & Sons.* Hoboken, NJ.

Solberg L.I., B.R. Asplin, R.M. Weinick, et al. (2003) Emergency department crowding: consensus development of potential measures. *Annals of Emergency Medicine.* 42:824-834.

Sommersguter-Reichmann, M. (2000). The Impact of the Austrian Hospital Financing Reform on hospital productivity: empirical evidence on efficiency and technology changes using a non-parametric input-based Malmquist approach. *Health Care Management Science* 3: 309-321.

Song, J.-S. and P. Zipkin (2003). Supply Chain Operations: Assemble-to-Order Systems. *Handbooks in Operations Research and Management Science: Supply Chain Management* 11.

Sorensen, R. J. and J. Grytten (2000). Contract design for primary care physicians: Physician location and practice behaviour in small communities. *Health Care Management Science* 3: 151-157

Spong C, Berghella V, Wenstrom KD, Mercer BM, Saade GR (2012) Preventing the first cesarean delivery. summary of a joint eunie kennedy shriver national institute of child health and human development, society for maternal-fetal medicine, and american college of obstetricians and gynecologist workshop. *American journal of obstetrics and gynecology* 120(5):1181–1193.

Staheli, Russ (2014) 4 ways healthcare data analysts can provide their full value. *Health Catalyst.* `https://www.healthcatalyst.com/healthcare-analytics-best-practices` (accessed on 27 March 2017).

Stolletz, Raik (2008) Approximation of the non-stationary M(t)/M(t)/c(t) queue using stationary queueing models: the stationary backlog-carryover approach. *European Journal of Operational Research* 190(2):478-493.

Stolletz, Raik, Svenja Lagershausen (2013) Time-dependent performance evaluation for loss-waiting queues with arbitrary distribution. *International Journal of Production Research* 51(5):1366-78.

Storrow AB, Collins SP, Lyons MS, Wagoner LE, Gibler WB, Lindsell CJ. (2005) Emergency department observation of heart failure: preliminary analysis of safety and cost. *Congest Heart Fail.* 11(2):68-72.

Sun P, Yang L, de Vericourt F (2009) Selfish drug allocation for containing an international influenza pandemic at the onset. *Operations Research* 57(6):1320–1332.

Sun B.C., R.Y.Hsia, R.E. Weiss , et al. (2013) Effect of Emergency Department Crowding on Outcomes of Patients. *Annals of Emergency Medicine.* 61(6): 605-611.

Sutherland, J. M. (2011). Hospital Payment Mechanisms: An Overview of Policy Options for Canada. Canadian Health Services Research Foundation (CHSRF) Series on Cost Drivers and Health System Effiency. *Canadian Health Services Research Foundation.* Ottawa, Ontario.

Sutherland, J. M. and R. T. Crump (2011). Exploring Alternative Level of Care (ALC) and the Role of Funding Policies: An Evolving Evidence Base for Canada. CHSRF Series of Reports on Cost Drivers and Health System Efficiency. *Canadian Health Service Research Foundation.* Ottawa, Ontario.

Sutherland, J. M., J. Hamm and J. Hatcher (2009). Adjusting case mix payment amounts for inaccurately reported comorbidity data. *Health Care Management Science* 13(1): 65-73.

Taljaard M, Donner A, Villar J, Wojdyla D, Faundes A, Zavaleta N, Acosta A, World Health Organization Global Survey on M, Perinatal Health Research (2009) Understanding the factors associated with differences in caesareansection rates at hospital level: the case of latin america. *Paediatr Perinat Epidemiol* 23(6):574–81.

Taylor, A., and W. Xiao (2014). Subsidizing the Distribution Channel: Donor Funding to Improve the Availability of Malaria Drugs. *Management Science* 60(10): 2461-2477.

Tor, I. and L. Hilde (2000). The effect of capitation on GPs' referral decisions. *Health Economics* 9: 199-210.

Tiemann, O. and J. Schreyogg (2012). Changes in hospital efficiency after privatization. *Health Care Manag Science* 15(4): 310-326.

Thomson Healthcare (2007) The healthcare costs of having a baby. Available at `www.marchofdimes.com/downloads/The_Healthcare_Costs_of_Having_a_Baby.pdf`. Accessed at November 30th, 2017.

Tran, Dat T., Arto Ohinmaa, Nguyen X. Thanh, Jonathan G. Howlett, Justin A. Ezekowitz, Finlay A. McAlister, Padma Kaul (2016) The current and future financial burden of hospital admissions for heart failure in Canada: a cost analysis. *CMAJ Open* 4(3):E365-E370.

Truven Health Analytics (2013) Cost of Having a Baby. *Truven Health Analytics.* Available at `http://transform.childbirthconnection.org/wp-content/uploads/2013/01/Cost-of-Having-a-Baby1.pdf`. Accessed at November 30th, 2017.

U.S. Department of Health and Human Services (2015). Office of Disease Prevention and Health Promotion. *ODPHP Publication.* No: B0132. Accessed from `https://www.healthypeople.gov/2020/topics-objectives/topic/maternal-infant-and-child-health/objectives`. Accessed at November 30th, 2017.

Ulu C, Honhon D, Alptekinoğlu A (2012) Learning consumer tastes through dynamic assortments. *Operations Research* 60(4):833–849.

Van Mieghem, Jan A. (2003) Due-date Scheduling: Asymptotic Optimality of Generalized Longest Queue and Generalized Largest Delay Rules. *Operations Research* 51(1):113-122.

Verheyen, P. (1998). The Missing Link in Budget Models of Nonprofit Institutions: Two Practical Dutch Applications. *Management Science* 44(6): 787-800.

Verheyen, P. A. and P. F. P. M. Nederstigt (1992). A cost-allocation system applied to Dutch hospitals. *European Journal of Operational Research* 58(3): 393-403.

Villar J, Carroli G, Zavaleta N, Donner A, Wojdyla D, Faundes A, Velazco A, Bataglia V, Langer A, Narvez A, Valladares E, Shah A, Campodnico L, Romero M, Reynoso S, Pdua KSd, Giordano D, Kublickas M, Acosta A (2007) Maternal and neonatal individual risks and benefits associated with caesarean delivery: multicentre prospective study. *BMJ* 335(7628):1025.

Wagner, K.H., H. Brath (2012) A global view on the development of non communicable diseases. *Prev Med* 54 Suppl: S38-S41.

Warrington T.A., J. Brunkow (2011) To bundle or not to bundle. *Healthcare Financial Management Association* White Paper.

Watanabe, S. (1964) On discontinous additive functionals and Levy measures of Markov processes. *Japanese J. Maths* 34:53 - 70.

Weintraub NL, Collins SP, Pang PS, et al.(2010) Acute heart failure syndromes: emergency department presentation, treatment, and disposition: current approaches and future aims: a scientific statement from the American Heart Association. *Circulation* 122:1975-96.

Whitt, W. (2006) Staffing a call center with uncertain arrival rate and absenteeism. *Production Operations Management* 15(1):88-102.

Whitt, W. (2016) Queues with time-varying arrival rates: a bibliography. `www.columbia.edu/\~ww2040/TV_bibliography_091016.pdf`.

World Health Organization (2010). Health systems financing: the path to universal coverage. *The World Health Report.*

World Health Organization (2015) Statement on caesarean section rates. *WHO Report* WHO/RHR/15.02.

WHO (2016) Global Health Observatory (GHO) data: life expectancy. *World Health Organization (WHO).* `http://www.who.int/gho/mortality_burden_disease/life_tables/situation_trends_text/en/` (accessed on March 25, 2017.)

WHO (2017) Don't pollute my future! The impact of the environment on children's health. *Word Health Organization (WHO) report.* `http://apps.who.int/iris/bitstream/10665/254678/1/WHO-FWC-IHE-17.01-eng.pdf?ua=1` (accessed on March 25, 2017.)

Wolff, R.W. (1982) Poisson Arrivals See Time Average. *Operations Research* 30(2):223-231.

Wolff, R.W., C.L. Wang (2002) On the convexity of loss probabilities. *J. Appl. Prob.* 39:402-406.

Woodbury, M. A., K. G. Manton and J. C. Vertrees (1993). A Model for Allocating Budgets in a Closed System Which Simultaneously Computes DRG Allocation Weights. *Operations Research* 41(2): 298-309.

World Bank (2014). Country and lending groups. 2013. Available: `http://data.worldbank.org/about/country-classifications/country-and-lendinggroups` (accessed 26 March 2017).

Yan, C., J. Kingston-Riechers and A. Chuck (2009). Financial Incentives to Physician Practices: A literature review of evaluations of physician remuneration models. www.ihe.ca. *Institute of Health Economics.* Edmonton, Alberta.

Young, James B., William T. Abraham, Darlene P. Horton, Lynne Warner Stevenson, Charles L. Emerman et al.(2002) Intravenous nesiritide vs nitroglycerin for treatment of decompensated congestive heart failure: a randomized controlled trial. *JAMA: American Medical Association* 287(12)1531-1541.

Zaric, George S. (2013) Operations research and health care policy. *Spinger.*

Zaric, G. S., H. Zhang, R. Mahjoub (2013). Modeling Risk Sharing Agreements and Patient Access Schemes. *In Operations Research and Health Care Policy* (pp. 295-310). Springer New York.

Zhang, H., G. S. Zaric and T. Huang (2011). Optimal Design of a Pharmaceutical Price-Volume Agreement Under Asymmetric Information About Expected Market Size. *Production and Operations Management.* 20(3):334-346.

Zorc, Sasa, Chick Stephen E., Hasija Sameer, Outcomes-Based Reimbursement Policies for Chronic Care Pathways (2017). INSEAD Working Paper No. 2017/35/DSC/TOM. Available at `SSRN:https://ssrn.com/abstract=2973048` or `http://dx.doi.org/10.2139/ssrn.2973048`