Topics in Service Operations Management

A dissertation presented by

Evrim Didem Güneş

to INSEAD faculty
in partial fulfillment of the requirements for the degree of
Ph.D in Management

August, 2004

Dissertation Committee:
Stephen E. Chick (Co-Chair)
O. Zeynep Akşin (Co-Chair)
Luk N. Van Wassenhove
Thomas D'Aunno

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

UMI Number: 3160455

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI®

UMI Microform 3160455

Copyright 2005 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Abstract

This thesis studies problems in the management of service processes in which the front line employee has an important role in matching the offered service level with customer needs. In doing so, the employee should take the operational implications of different service offerings into account. Two conditions are necessary for employees to accomplish this role: The ability to diagnose customer characteristics and incentives to provide the appropriate service. The thesis consists of three parts, each concerned with specific aspects of this issue in different service contexts.

The first part assumes perfect diagnosis capability and focuses on providing incentives to employees, in a setting where firms try to generate additional revenue by cross selling to a certain customer segment. It provides a modeling framework to analyze the ties between market segmentation decisions, incentives, and process performance in such service delivery systems. Additional revenues can be generated by providing an extended service (as opposed to a standard service), and the front line employee chooses between these two service levels after observing the revenue generation potential of a customer. We characterize the optimal market segmentation decision, optimal service level choice and a set of optimal linear incentive contracts that enable their implementation. It is shown that a market segmentation scheme combining revenue generation concerns with their process implications is essential for success. Characteristics of appropriate incentive schemes are identified.

The second part considers the ability to make the diagnosis in a health care setting. It combines and extends previous work on breast cancer screening models. We model a service system, which explicitly incorporates aspects of the dynamics of health care states, program outreach, and the screening volume-quality relationship. Using simulations, the impact of increasing standards for minimum reading volume to improve quality, expanding outreach with or without decentralization of service facilities, and the potential of queueing due to stochastic effects and limited capacity are analyzed. The results indicate a strong relation between screening quality and the cost of screening and treatment, and emphasize the importance of accounting for service dynamics when assessing the performance of health care interventions. For breast cancer screening, increasing outreach, without improving quality and maintaining capacity, results in less benefit than predicted by standard models.

The third part models the incentives of health care providers to invest in the ability to make a correct diagnosis in disease screening services. Screening is organized centrally in some countries (e.g. UK), while it is provided by private clinics or local initiations in others (e.g. US). This essay investigates potential implications of competition on screening test quality in a duopoly competition setting. The model accounts for disease prevalence in population, and patient utility for test quality and transportation to the facility. A preliminary analysis shows that competition may result in higher or lower quality than the socially optimal level. Policy implications and further research directions are discussed.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Acknowledgements

This thesis is the outcome of my five years at INSEAD. It took a lot of sweat and tears, sleepless nights, hopeless moments, much struggling and fighting to get here. But it could not have been possible without the enormous support and love, which I was so fortunate to have around all these years.

I would like to express my deepest gratitude to my co-chairs, Zeynep Akşin and Steve Chick for their continuous support and help for my dissertation. Zeynep has been a teacher, supervisor, co-author, and a good friend. I am grateful to her for introducing me to the problem when I was looking for a thesis subject and being a guide in the service operations field with her experience. Her serenity, her enthusiasm and dedication for the project were invaluable help to finish this work. I am most grateful to Steve for having accepted to supervise me at a difficult time. I learned much from him about writing concisely, and always remembering the reality checks. He had been an assuring support with his knowledge of the health care field. His attention to details and his endless patience, his accessibility and willingness to help whenever I needed was essential.

I am grateful to my committee members, Luk Van Wassenhove and Tom D'Aunno, for their support and valuable feedback they provided. Luk has been a great source of wisdom for research in general and also helped me to find financial support in my last year, providing me a conference organization experience, which I enjoyed a lot. Tom has kindly accepted to read my very "operations" papers with a genuine interest and provided a sanity check.

I would like to thank all INSEAD TM faculty for their help and friendliness. My special thanks are to Beril Toktay, who had been my academic advisor in the first year and an excellent teacher for many of our core courses. But her support continued until the end, and was no less than a thesis supervisor's and a close friend's, who was always ready to give feedback for papers, for presentations (sometimes even without me asking!) and for research and life problems in general. Ayse Öncüler, sitting in the office next-door, always welcomed me for complaining, crying or sharing the joy of an achievement. I not only enjoyed the simulation class but also playing basketball together with Enver Yucesan, thanks to his enthusiasm. Cristoph Loch was an excellent teacher, both in technology management and in soccer. Ludo Van der Heyden always amazed me with his sharp comments and it was great to have his feedback.

Department secretaries Delphine Delafontaine, Claire Derouin and François-Xavier Priour were so nice office-neighbors. I would run to them asking for an emergency help with some administrative work, the French grammar, or an urgent chocolate need, and was never disappointed. I'd like to thank the library staff, the bar staff, and the security guards, whom I have seen more often than my family in last few years. I felt at home here thanks to their kindness.

The financial support from the GE fund for two years of my studies is gratefully acknowledged.

As I said at the beginning, this thesis is an output of my 5 years at INSEAD. It is not the only output though, nor the most important one. I lived the most beautiful 5 years of my life here in Fontainebleau (our "village"), enjoying the forest, singing birds in the quiet streets, the autumn colors, and Paris with all its charm (although not as much as I had wished for). I met so many nice people, I learned so much about different cultures, and their music and food, I learned how to say hi in many languages, and I got the ability to recognize hundred accents in English. The most important things I am taking with me back home, apart from this humble thesis are the friendships, and the new person in me, with all these mostly-nice-sometimes-bitter memories of five years. I would like to express my sincere gratitude to all those people who provided me moral support and enriched my otherwise ordinary life.

I think the completion of this thesis was a relief for many people at INSEAD, one could feel the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

deep breaths taken after my defense, since I had made quite some noise to get there! In the first years, I had great classmates who eased the course-work burden: Many thanks to Jonghoon Bae, Zhixing Xiao, Ilker Yoney, Odisseas Triakaliotis, Konstantin Korotov, Kostas Lioukas, José Luis Peydro. José Miguel Gaspar was a big support during my first years. Sumitro Banerjee has always been ready to help and advise, or to have fun together. Stelios Kavadias always encouraged me and it was great fun to share the office with him. Many people comforted me with their existence, with their intelligence and particularities, with their smile, kindness and encouragement, which made the school a pleasant place: Atalay Atasu, Roxana Barbulescu, Onur Boyabatli (my wise student, dinner-fellow), Olivier Chatain (the reliable one), Alessandra Cillo, Raj Iyer, Nishant Dass (recent night-shift-mate), Vassilis Dimitrakas (gracias!), Maria Galli, Sofia Gueorgieva (long nice letters), Mumin Kurtulus, Pedro Matos, Ana Cavalheiro, Valéria Noguti, Selçuk Onay (great company), Christine Gilles (the normal person), Vini Onyemah (he is crazy), Ludovic Phalippou (warmth from the south, the south!), Andreas Robotis, Lambra Sainz, Metin Şengül (best party-mate), Dabu Sengupta (good things happen to good people), Tomoaki Shimada, Julie Urda (big smile calling me little Evrim!), Michael Yaziji (just so good) and Anne-Catherine Delmelle (a touch of art); Thank you! Moreover, I want to thank Ayşe Kocabiyikoglu, my last office-mate, also a great complaining and gossiping-mate, who has always provided relaxation to me. Govert Vroom was like a personal coach to me; I am grateful to him for valuable advice and long discussions, for sharing nice trips and the family atmosphere as well as ridiculous moments. I am indebted to Laurens Debo (thanks x36!) who was always on my side discussing research, partying in the office, doing sports, and pushing me to continue whenever I was down, even after he had moved away. Svenja Sommer has been one of my angels that I could rely on whenever I needed; studying for courses or rehearsing presentations, doing proofs for my papers, preparing for the job market or for a dinner party, she was always there, always perfect.

It is difficult to thank enough to Deniz Özdemir and Arzu Ozoguz for their place in my life here, without which I could not have survived easily. I enjoyed many political discussions, dinners, Paris trips with them. Arzu has showed me a new world (the chic one!), and shared my worst moments of pain with an open heart and wisdom as well as the best ones. Deniz is my friend for 10 years now and she has been my first-aid in the PhD from day one, sharing every good and bad since then. She was my memory here, and sometimes one really needs good memories!

I would like to thank two best friends from my pre-PhD life, Hande Yaman and Ozlem Ozhan, who are like sisters to me by now. I owe much to their on-line, on-phone, or on-visit, but always-loving, always-caring presence in my life.

Finally, I know I could not have finished this without the support of my family. My dearest aunt, Meral teyze, has been like a second mother to me in Paris (also feeding my friends with her delicious Turkish food). My beloved sister, Elif always cheered me up with surprises; my dear brother Evren, was a calming voice at frustration moments, telling me to stop working too hard. My mom Münevver and my dad Kamber always believed in me, they did not doubt even a second that I would succeed. At those times I thought it was because of ignorance or overconfidence, and the blind-love they had for their daughter. But now I am happy to see that they were right once again; I took another big step in my life, but only with the strength I got from their loving support, and only because they taught me to try my best in the pursuit of the good. Teşekkürler!

Fontainebleau, August 2004

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

To my family

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Contents

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

List of Figures

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

List of Tables

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1 Introduction

The service sector is the largest and the fastest growing sector of the world economy, representing more than half of GDP in developing economies and 71% in high income economies [10]. In the US, more than 80% of jobs in the private sector are in services [67]. Although some argue that there is no real difference between service and manufacturing operations [81], service operations have unique characteristics not found in manufacturing, notably customer participation in the service process, intangibility, simultaneity (inseperability of production and consumption), heterogeneity, and perishability [48]. Among those, customer participation and heterogeneity are deemed as the ones that are the most important [85]. This thesis investigates topics in service operations mainly dealing with heterogeneity of services, i.e. variability in the service delivered to the customer.

There are four factors contributing to the variation in service delivery, as determined by [50]: heterogeneous customers with different service expectations, lack of rigorous policies and processes, high employee turnover and the nature of customization. Variation in service delivery may be introduced deliberately in order to meet heterogeneous customer expectations, with the aim of increasing satisfaction, hence profitability, as argued by [57]. This can be done by customizing the service for each customer's expectations. Alternatively, different service types can be determined in advance and the one that suits most to a customer's characteristics can be offered for each customer.

In doing so, a service type is decided for each customer segment, i.e. a "match" of service type with customer characteristics is defined. A good match can be defined by scientific evidence, like a match of treatment with diagnosis groups in health care services. Or, it can be

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

defined by management according to a certain criteria, like a match of different loan options with customers having different income levels.

In ensuring the correct match, the front line employee has an important role, being the contact person during the service encounter [19]. The dynamic nature of customer characteristics (for e.g. changing health states in the health services context) increases the importance of the front line employee and makes delegation of service type decision to front line employees more desirable. In this dissertation, the main issue we focus on is service operations design and control for service processes in which the front line employee has an important role in matching the service offered with the customer needs.

Two necessary conditions for front line employees to accomplish this role properly are *abilities* and *incentives*. First, a server should be able to understand customer characteristics and needs, and to choose the right service level that matches. Second, the server should have the incentive to indeed offer the right service level. The manager should define what is a good match, considering the implications of different service levels on costs and revenues. Then the manager must ensure those two conditions are satisfied to achieve a good "match" in order for the service to be successful.

The issues that this thesis aims to understand better are: How best to match the service levels with customer characteristics; how best to design incentive systems that help autonomous service employees choose desired decisions; the dynamics of customer characteristics and needs, and the effect of errors in finding a good match between needs and service offerings; the effect of competition on the outcomes for disease screening services where a good match is important. This dissertation consists of three essays on service operations management, that investigate the above issues in different service contexts.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1.1 Summary of Chapter 2: Value Creation in Service Delivery: Relating Market Segmentation, Incentives and Operational Performance

In Chapter 2, the service context considered is call centers, which represents a huge part of the service sector in the US and Europe, a key customer contact point. In 2001 there were 55,800 call centers in North America, of which 90% are located in the US [9]. The Tower Group estimates that nearly 34 billion retail banking transactions were conducted in the US during 1998, growing to 44 billion in 2003, representing 25% of transactions [7].

The chapter focuses on value creation strategies like cross-selling (i.e. to sell a new product or service to an existing customer) or add-on sales. In mature markets where market growth is slow, cross-selling or add-on sales are used as strategies to increase the profitability of existing customers by increasing the revenues generated from existing customers, as opposed to increasing the market share. These initiatives are part of a firm's Customer Relationship Management (CRM) strategy.

We use cross selling as a representative of additional value creating activity, while the question presented in this chapter is valid more generally. Additional value can be created in a service interaction not only by cross selling but also by customizing the service, spending more time and effort to serve a customer and increase customer satisfaction. Cross-selling represents a higher service level that takes a longer time, but provides an opportunity of increasing revenues.

The basic idea of a value creation strategy we consider is to provide different service types to customers according to their needs and convince some subsegments to spend more money with the firm. Customers are heterogeneous in their willingness to pay for the additional

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

product or service, i.e. they have different revenue generation potentials. Since service time is longer with a cross-selling attempt, which increases congestion, only customers with a high enough revenue generation potential should be targeted for cross-sell.

Thus, the first task of a manager is to decide which market segment to cross-sell, which is the strategy that defines the "match". If only a part of the market is targeted for revenue generation, the strategy is called "service level differentiation". A market segmentation divides the customer pool into two or more subgroups, according to their revenue generation potentials. In the marketing literature market segmentation is commonly done without taking its operational implications into account, and the operational models commonly assume a given market segmentation scheme to decide on service type. In this chapter, market segmentation decision is taken jointly with the service type decision, accounting for the revenue generation potentials as well as increased congestion caused by longer service time of a cross sell attempt.

In implementation of the defined strategy, customer service representatives have to decide between a basic service (standard service) or a cross sell attempt in addition to the basic service (extended service), after they see the customer characteristics, i.e. revenue generation potentials. Chapter 2 assumes that servers have the ability to observe customer characteristics perfectly. The main concern is whether they have the incentives to implement the desired strategy, since they need to put extra effort for an extended service. Then the next task of the manager is to give the servers the right incentives so that they make the correct match in order to avoid unnecessary congestion by trying to cross sell to customers who would most likely not accept the offer. Similarly, they should not miss an opportunity to cross sell to a customer with high potential of accepting the offer. In this setting, appropriate incentive design is essential in ensuring a match between servers' performance and the service provider's desires, in terms of

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

service types for each customer segment. Thus we investigate how incentives affect operational performance and success in implementing policies on market segmentation and service level differentiation.

In chapter 2 we explore: When is it worth undertaking a value creation initiative like cross-selling? What are customer characteristics that induce a desire to spend effort on value generation? Is it better to increase the profitability of all customers by uniformly targeting all customers of a firm, or should a company segment the market and pursue a service level differentiation strategy? What server incentive schemes should be used to implement the desired actions of each strategy? How should these incentive schemes address the tension between creating value and providing fast and efficient service? Are only value generation related incentives (like sales incentives) enough to achieve the desired strategies?

These questions are answered using a principal-agent model. The firm is modeled as a queue, which allows incorporating effects of value creation activities on waiting, an undesired outcome. The principal (the manager), has a choice between the strategies labeled: remain a cost center, target all customers, or pursue service level differentiation. The key trade-off that the manager faces in making this decision is between revenue generated and costs. Revenues are determined by the value creation effort and customer profitability characteristics. Costs are in the form of incentive payments necessary to induce the desired actions by the servers, as well as the system-wide cost due to the congestion effect of the additional effort expended for value creation. The agent observes the customer characteristics and decides on a service level, among two service levels, standard or extended service.

The model is unique in that it combines value creation and process related issues, providing a coherent framework to analyze sales initiatives like cross-selling or service level

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

differentiation strategies. The analysis first shows the conditions for different strategies to be optimal. Value creation strategies are favored for moderately loaded systems where the time spent for an extended service does not push capacity utilization to a maximum and where the expected revenue from the extended service is high enough to compensate for the extra cost of congestion. When the unit congestion cost is relatively high compared to revenues, service level differentiation is preferred to targeting all customers. Firms that can operate in an integrated mode, e.g. where marketing and operations jointly optimize the market segmentation decision, are clearly better off in terms of achieving profitability. The optimal market segmentation decision depends on the distribution of revenue generation in the customer base.

Once the decision for the best match is made, we illustrate how the desired strategy can be implemented through the design of appropriate incentive contracts. We show that when the two tasks being considered have opposite performance effects (like the extended service is assumed to have in this analysis, by increasing revenues but increasing congestion), then optimal linear contracts may involve punishments. Furthermore, we find that incentive payments for these two tasks (for example service and sales) depend on each other, and providing incentives for only one dimension, as is frequently observed with sales based incentives, can lead to undesirable behavior. In the setting where we have a sophisticated manager and server who optimize the market segmentation decision, we show that there is a unique optimal linear contract.

This chapter appears to be the first model of cross-selling that looks at design issues at marketing-operations interface. It illustrates the trade-offs between value creation and operational measures like congestion. It also shows the superiority of market segmentation schemes that take operational implications into account. The role of front-line employee in ensuring a

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

good match between customer characteristics and service level is explicitly modeled and the structure of incentive contracts are discussed.[1]

1.2 Summary of Chapter 3: Breast Cancer Screening Services: Trade-offs in Quality, Capacity, Outreach, and Centralization

Chapter 3 is in the context of health care services, which represents an important part of the world economy (average global spending on health is 9.1% of GDP [8]). It models a specific health service: breast cancer screening. Breast cancer is the most common cancer among women, and the second leading cause of cancer related deaths after lung cancer. Most developed countries have organized screening programs to pro-actively detect breast cancer [69], as early diagnosis of most types of breast cancer is very effective. This chapter develops and analyzes a model on breast cancer screening aiming to contribute to the understanding of factors that influence breast cancer screening effectiveness.

In a screening program, there are two service levels: screening test or screening test followed by a diagnostic test. The front line employees are radiologists who read the screening tests (mammograms) and decide whether there is a suspicion for cancer that would necessitate a diagnostic test. A "good match" refers to trying to detect the disease at an early stage and avoid offering diagnostic test when there is no disease (a form of waste). In the screening services context, there is a clear definition of "good match" because all detected patients are treated. There is no incentive issue because of the hypocratic oath, a radiologist has incentives to detect the tumors and to avoid unnecessary diagnostic tests. The more critical issue is the

[1]The paper version of Chapter 2 has been accepted to appear in *Manufacturing and Service Operations Management*.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ability of the radiologist to recognize the customer characteristics correctly and make the right decision.

This chapter models mammogram as a part of the broader screening service delivery model. Mammogram reading is one of the most difficult tasks in radiology; missing existing tumors (false negatives) or asking for diagnostic test when there is no cancer (false positives) are common. Radiologists need to stay sharp in order to keep these errors at a minimum, i.e. have high screening quality. There are standards set by public health authorities on minimum reading volume of a radiologist in the interest of keeping screening quality high. However, these standards differ between countries. We include a model of quality as a function of reading volume in the system dynamics model of a breast cancer screening program, and use the standards as a control of the quality level that can be achieved. We model two types of service (screening tests and follow-up diagnostic tests), a finite service capacity, and the potential of waits due to finite service capacity and randomness associated with patient scheduling. To this end, we use a stochastic compartmental model with 21 compartments representing the states on two dimensions: health status and the state in the health system. We assume a three-stage health status model, (1) healthy, (2) preclinical, or early stage, breast cancer (3) clinical, or late stage, breast cancer.

The main contribution of this chapter is that it explicitly incorporates for the first time, aspects of the dynamics of health care states, program outreach, and the screening volume-quality relationship in an integrated way. The analysis is concerned with the three most important factors for screening program effectiveness: test quality, program outreach and service capacity. The simulation experiments analyze the impact of all these factors on health outcomes and system costs: specifically the impact of increasing standards for minimum reading volume

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

to improve quality, expanding outreach with or without decentralization of service facilities, and the potential of queueing due to stochastic effects and limited capacity in the simulation analysis.

Concerns have been raised in public [82] and by screening program organizers [94] about the waiting times to get a mammogram. This chapter shows that waiting should be a less significant concern than the quality of screens, unless there is a severe capacity problem. The analysis also illustrates that screening quality can affect the system load indirectly, by changing the number of unnecessary diagnostic tests. A comparison of the costs of diagnosis and treatment after improvements in screening program outreach versus screening test quality shows that improving screening quality has significant advantage over improving screening outreach. The results indicate a strong relation between screening quality and the cost of screening and treatment, and emphasize the importance of accounting for service dynamics when assessing the performance of health care interventions.

One important element of the model is that it accounts for the dynamics of disease progression. This allows us to model the impact of operations on the future demand characteristics, the health state in Chapter 3. A similar model could be used in other service contexts, to analyze customer profitability over time, for example in a Customer Relationship Management related problem like the one in Chapter 2.

As in the case of call centers, waiting times are not desirable in the case of breast cancer screening. In Chapter 2 the congestion cost was a loss in customer goodwill. For breast cancer screening, the cost of waiting is reflected by the lost opportunity of early detection, which occurs if the disease progresses to a late stage during a wait. If the quality of screening test is low, unnecessary diagnostic tests increase the load on the system and may increase the waits if

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the capacity is not sufficient. The interactions between system capacity, screening test quality and screening program outreach are modeled in this chapter.[2]

1.3 Summary of Chapter 4: A Model for Disease Screening: Quality Decision under Competition

Chapter 4 presents a model to analyze potential effects of competition on disease screening services delivery. Health services are financed and its delivery is organized using different systems in different countries. In some countries the health care system is funded primarily by taxes (e.g. Greece, United Kingdom), while in others private sources and insurance provide funds (e.g. US, Germany). Screening services can be centrally organized in some countries like in the UK or Canada, or they can be recommended by the public health authorities and promoted, but not centrally organized, like in the US, where competition plays an increasingly important role.

The analysis in Chapter 3 highlighted the importance of screening quality. Motivated by that, and by anecdotal evidence from the US about private mammogram clinics' complaints on reimbursement rates, this chapter focuses on the investment in screening quality in a competitive environment. The main focus is the investment on improving the ability to make a better screening test, i.e. the screening test quality. Unlike Chapter 3, Chapter 4 does not model the individual radiologist activities and the evolution of quality as a function of reading volume. Here we look at quality as an attribute which can be improved through investments like training or equipment. In Chapter 3, the customer utility was not modeled as a function of quality.

[2]The paper version of Chapter 3 has been accepted to appear in *Health Care Management Science*.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Here we model customer behavior and the customer utility for getting screened as a function of quality.

The model is simplistic relative to the complexity of the problem by choice in order to keep the model of competition tractable. The main contribution of this essay is to present and model the disease screening problem in a competitive setting, and set the ground for further work for a better understanding of preventive health care delivery system design. To this end, optimal quality level is analyzed in a duopoly competition and the result is compared with the quality level that maximizes social welfare.

The preliminary analysis shows that if everyone in the target population were willing to get screened, and the only differentiation between two facilities from the target population's perspective were the screening quality, then a social planner would open only one screening facility and make all the quality investment there. This is a reasonable policy, consistent with intuition and considering economies of scale and learning effects that were not considered in the model. The analysis of the competitive case for two facilities in the same location shows that there is no equilibrium with two competing facilities. When the two facilities are located apart from each other, an equilibrium can be achieved. It is shown that if the cost of improving screening quality is very low, the equilibrium can be unstable. For the stable case, the investment in quality of profit maximizing providers increases as the demand becomes more elastic with quality. The implications of these results are discussed.

To summarize, the services in this model are assumed to be either a screening test versus a screening test plus treatment at an early stage (the diagnostic test is modeled implicitly). The demand is assumed to be determined by customer utility from screening, a function of the valuation of screening quality and the distance to the facility, which determines the willingness

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

of a patient to get screened from one facility. Like Chapter 3, the main concern here is the ability to have the right match of the service with the demand characteristics, and to detect the disease at an early stage. However, Chapter 4 looks at a higher level problem, considering the industrial organization of the screening service, while Chapter 3 had the perspective of a social planner, and considers the incentives of managers (i.e. profit maximizing providers) to invest on the ability to do this match.

Chapter 5 summarizes the conclusions of the dissertation and discusses the limitations that lead to directions for further research.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2 Value Creation in Service Delivery: Relating Market Segmentation, Incentives and Operational Performance

2.1 Introduction

In a wide range of industries today, market share growth is no longer significant and growth is mainly driven by increasing the profitability of existing customers. This increase in profitability is not only pursued by improving efficiency measures but also by persuading existing customers to spend more money with the firm [47]. Value creation initiatives like cross-selling are one form of achieving this aim. Today, these initiatives are part of a firm's Customer Relationship Management (CRM) strategy and are supported by a panopoly of IT systems. The global market for CRM systems, service and technology is estimated to be around $25 billion [16].

Increasingly, however, companies report failures of CRM related initiatives. Among some of the most cited reasons for these failures are failure to integrate it to the back-office operations and failure to train and motivate the staff. Indeed, as noted in the Economist (July 2001) [37]; "durable customer relations are partly about clever technology. Mainly, they require relentless attention to detail: good products, prompt service, well-trained staff with the power to do a little extra when they judge it right to do so". Service employees have an important role in determining customer needs and acting accordingly, since they are the ones who interact with the customer during a service encounter. Even in environments where good customer information exists, and automated prompts guide servers' effort to enhance customer profitability, the ultimate decision of how to deal with the customer rests with the server. In assessing the customer profitability, the server uses public information such as past buying

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

behaviour, as well as private information such as what the customer says during the phone call. In this type of a setting, appropriate incentive design is essential in ensuring a match between servers' performance and the service provider's desires.

This chapter focuses on value creation strategies like cross-selling or add-on sales. Our model captures the deepening of an existing relationship with a customer, where depth is characterized by additional revenue per transaction and not by additional transactions. In this setting, we explore the following questions: When is it worth undertaking a value creation initiative like cross-selling? What are customer characteristics that induce a desire to spend effort on value generation? Is it better to increase the profitability of all customers by uniformly targeting all customers of a firm, or should a company pursue a service level differentiation strategy? What server incentive schemes should be used to implement the desired actions of each strategy? How should these incentive schemes address the tension between creating value and providing fast and efficient service? Are only value generation related incentives (like sales incentives) enough to achieve the desired strategies?

These questions are answered using a principal-agent model. Profitability characteristics of the customers are assumed to be given. The firm is modeled as a queue. The principal, or the manager, has a choice between the strategies labeled as: remaining a cost center, targeting all customers, or pursuing service level differentiation. While remaining a cost center is the status quo option, i.e., not pursuing any additional value creation activity, targeting all customers requires expending additional effort on all customers, and service level differentiation requires additional effort for a segment of the customer base. The key trade-off that the manager is facing in making this decision is between value or revenue generated and costs. Revenues are determined by the value creation effort and known customer profitability characteristics. Costs

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

are in the form of incentive payments necessary to induce the desired actions by the servers, as well as the system-wide cost due to the congestion effect of the additional effort expended for value creation. Using this framework, we show under what conditions each strategy is preferred, and what type of incentive scheme is necessary to ensure its implementation. In the first part of the analysis, it is assumed that the segments (high and low) for the service level differentiation strategy are given exogenously, and both the manager and the server take these as given. The sensitivity of the results to this market segmentation decision are explored subsequently, which leads to the second part of the analysis, where the segmentation choice is a decision variable.

The remaining parts of the chapter are organized as follows. Relevant literature is reviewed in Section 2.2. The model is introduced in Section 2.3. We analyze the resulting principal-agent problem in Section 2.4, and characterize optimal strategies and contracts for given customer profitability and customer segmentation choice. The sensitivity of these results to the customer segmentation decision is explored in Section 2.8.1. The optimal market segmentation choice and a contract that allows its implementation are characterized in Section 2.8.2. The chapter concludes with a discussion of the main results in Section 2.9.

2.2 Literature Review

Any value creation strategy requires an understanding of the relationship between customer needs and service offerings, and how these generate value. The fact that different needs may require different offerings and thus generate different profits for the firm is the basic premise that motivates a vast literature in marketing on market segmentation. Proliferation of direct and interactive forms of communication in recent years have brought concepts like one-to-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

one marketing or relationship marketing to the forefront, leading to a stream of literature that focuses on estimating customer profitability. These papers typically focus on value estimation and ignore costs, despite the need to the contrary [49]. The papers that do consider costs typically include only the marketing costs incurred for a customer in their profitability estimates [83], or they assume a fixed service cost element ignoring the interaction between service level and operational costs [18]. For example, [86] explicitly include the supply chain costs in their model of customer profitability analysis. However they have an activity-based cost accounting model, which allocates the costs after they are incurred rather than considering the operational costs explicitly before making the service level decision. This type of analysis is classified as *retrospective* by [102], as it is based on historical data. In contrast to this approach, the *prospective* approach considers the fact that customer profitability can be changed or influenced through the service provider's actions.

The approach in this chapter can be viewed as being closer to the *prospective* analysis described in [102]. We assume that the likelihood of generating revenue from a customer depends on the level of service provided. Thus, customer profitability is determined by the likelihood of generating revenue from high level service and the associated congestion cost of offering such high level service to a particular customer. While it may be possible to estimate profitability for individual customers, typically service levels are determined for a segment of customers rather than individuals. Thus, we consider the case where a market segmentation decision separates customers into groups, and customers in a group are assumed to have an average revenue generation potential, which can be derived from the prior on the distribution of the revenue for an individual customer. The manager determines the optimal level of service that should be provided to customers in each segment, given revenue generation probabilities

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

and cost parameters. The simplest case with two segments is considered for the analysis in this chapter. The choice between service levels is represented as a choice between performing a basic service task or a combined basic and extension task, where the latter represents a higher service level.

The impact of combining tasks in processes, in terms of its effect on congestion has been extensively studied in the operations management literature, mainly considering the systems as cost minimizing units. An important finding is the pooling result, which says that combining tasks decreases congestion. The importance of various human resource issues in assessing the performance of a pooled system have been discussed and incorporated in different settings ([72, 24, 92, 95, 25]). The interaction between combining tasks, incentives, and value generation, that we consider herein, have not been addressed before.

Our model lies at the interface of the problems dealt with in marketing and operations management. Marketing research focuses on value generation, but since there is no explicit modeling of the operational side, cannot take this value data to generate action plans in terms of appropriate service levels. The operations management literature that deals with process design, on the other hand, focuses on costs, and does not consider the value implications of various process designs. [4] analyze the congestion effect of a particular value creation initiative in call centers. The revenue generation from a customer is not explicitly modeled. [51] model a marketing effort decision analytically in a queueing setting. While their analysis models value creation from a customer, it only considers direct marketing cost associated with this value creation effort.

There is a huge literature that deals with incentive contracts and agency problems in economics, marketing and more recently in operations management. Among the classical

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

papers on agency theory, [54] and [59] assume a generic function for the output rather than using the models for the underlying operational system through which the effort leads to outcomes. In marketing, a stream of literature on salesforce compensation has started from models with deterministic output functions [43] and evolved into agency theoretic models. [13] present various salesforce compensation plans in a principal agent framework. The assumption of constant marginal cost is common in this literature (see for example [71]. For a thorough review of the salesforce compensation literature the reader is referred to [32].

In the salesforce compensation context, our study provides a link between the incentive and operations problems by explicitly modeling the operational costs of pursuing this additional value as opposed to assuming a constant marginal cost of production. In the typical setting considered by the salesforce compensation literature, the server is a salesperson whose job description is selling. In our model, we consider service settings where the primary role of the server is to provide service and the additional extension task can be considered as a sales activity. As such, the sales activity constitutes an additional component of the server's job description.

In the operations management literature, there are some studies considering incentive effects in different operational settings. A good review of this literature can be found in [93]. Studies on queueing systems have more often focused on pricing issues and related customer incentives as in the articles of [78, 23] and [79]. Examples of papers which consider server incentives in congestion prone settings are [52] and [100]. The latter considers incentive issues in service contexts such as medical services or call centers, where there is a gatekeeper who makes an initial diagnosis of a customer's problem, and then either solves it or refers it to a specialist. The effect of different contracts on the referral rate the gatekeeper chooses are

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

investigated in an environment where the gatekeeper has an ability (unknown to the firm) to deal with problems of varying difficulty. The incentive side of our model is similar in structure to the gatekeeping problem. However, we consider incentive problems stemming from the variance in processing times and customer identities, as opposed to the server's identity.

2.3 The Model

We model the provision of a service that can be offered at two different levels. The standard level requires no effort from the server and generates no revenue. This represents the prevailing level of service if no value creation is sought by the server. On the other hand, if the server opts for the high level of service, this requires effort and results in the possibility of generating revenues. Using this model, we analyze the service level decision, which determines the customer segment for which the revenue generating high level of service is optimal. Corresponding incentive contracts are characterized. The firm is modeled as a profit maximizing, single server Markovian queue with unlimited waiting space.

Customer Base and Value Generation: Customers arrive according to a Poisson Process of rate λ. There are two customer types, high and low, which we label as H and L respectively. The server can observe the customer type at the start of service, and incurs no cost for diagnosing a customer's type. For any customer, the probability of being a high type is q, and the probability of being a low type is $(1-q)$. Thus, the parameter q determines the size of the high type segment. The revenue generation potentials depend on the type of the customer and the service level offered as will be explained later in more detail. The probability of generating revenue R by offering a high level of service is p_H for the high type and p_L for the low type customers. We also make the assumption that $p_H > p_L$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

We illustrate in Section 2.8.1 how these parameters relate to the market segmentation decision of the firm. Until then, these parameters are taken to be given. The basic model, where these parameters are taken as given, can be seen as representing the case of a functional organization, where customer related information and any segmentation decision is taken by the marketing function and not questioned elsewhere. Thus, both the manager and server in the operations function take these as given. This assumption is relaxed in Section 2.8.2, where both the manager and the servers are more sophisticated. The manager determines an optimal market segmentation scheme, which in turn determines the parameters q, p_H, and p_L. The server may not accept this segmentation scheme, unless he is offered the appropriate incentives to do so. This latter setting represents the case of a more integrated organization, where both manager and server have an understanding of the entire process rather than just a functional view.

Service Process and Costs: There are two service levels that can be offered to the customer: 'standard' or 'extended' service. Server effort is represented by the binary variable denoted by $e_H \in \{0, 1\}$ and $e_L \in \{0, 1\}$ for high type and low type customers respectively. $e_H = 0$ or $e_L = 0$ represents the case with no effort and $e_H = 1$ or $e_L = 1$ the case where the server exerts effort. Standard service does not require any effort from the server, so the effort is 0. It generates no extra revenue, hence we normalize the revenue in this case to 0. The service time for this type of service is exponentially distributed with rate μ. The second level, 'extended' service, where a service extension is provided, can be interpreted as additional personalized attention, or a cross-sell attempt. This extension requires effort on the part of the server, so the effort is 1. As a result of the server's effort, a revenue of R is generated with a fixed probability that depends on the type of customer being served as explained before. This effort is also

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

reflected in the time spent for the service. The service time is exponential with rate $\mu - k$ ($\mu > k > 0$) in this case, where k represents the content or complexity of the extension task. Spending effort is unpleasant for the server, so he has a disutility of C_S whenever $e_H = 1$ or $e_L = 1$. Given the nature of the extension task, this implies that the server does not enjoy the sales activity. This represents the direct cost of providing high level service to a customer.

In addition, there are indirect costs due to the congestion experienced by customers in the system. For any customer, the time spent in queue costs c per unit time to the firm. This parameter can be interpreted as the loss of goodwill of the customers, or the cost of keeping the waiting space (for example phone line) busy, and represents the importance of congestion for management.

Note that in this model, increasing effort results in a *decrease* in service rate, contrary to the more common assumption in the literature that increasing effort increases service rate (for e.g., [52, 64]. An important implication related to this is that high effort *might not* be desirable, because of this consequent decrease in service rate, which would decrease the profitability of the customer (i.e., revenues net of costs of serving that customer) due to the increase in costs.

Information and Decision Structure: We assume there is a manager (she) who wants to implement a policy π, which is defined as the effort levels provided for each customer type, i.e., $\pi = (e_H, e_L)$. There is one server (he) who serves each arriving customer after observing their type. The manager only observes the time spent for the customer and the revenue generated. She cannot observe the realization of customer type, nor the distribution from which the service time is drawn. Hence, she does not know if a certain outcome is the result of the server's effort choice or of chance. On the other hand, the server does not incur a cost for the waiting time of customers, and furthermore, he does not like expending effort. As a result, his decision

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

may not be optimal for the firm if he is not compensated appropriately. The manager wants to ensure that the desired service levels are offered to each customer, which might depend on the customer type. So she needs to find an incentive scheme that would induce the best decision by the server in the presence of moral hazard (the effort is not observed) and private information (customer type for a given realization is not observed).

Performance Measures: There are two outcomes that are the results of a server's effort decision, which contribute to the overall system performance: service time, x_1 and the revenue generated, x_2. The manager decides on a compensation scheme and declares a policy, $\pi = (e_H, e_L)$ that she wants the server to implement. Then the agent (the server) decides on the effort levels (e_H, e_L), that maximize his utility (compensation less disutility for effort). The effort decision is taken once and applied to all customers in a particular segment, i.e., the decision is not taken dynamically. We assume that the contract is linear in the two outcomes x_1 and x_2, and that both the principal and the agent are risk neutral, maximizing their expected linear utilities.

The manager's objective is to maximize profits, i.e., revenues minus the costs as payments to the server and the cost associated with congestion in the system. The first cost component for the manager is the compensation of the server, w. We define the payment scheme as

$$w = \alpha_1 x_1 + \alpha_2 x_2 \tag{1}$$

for any customer served, where x_1 is the service time and x_2 is the revenue generated for that particular customer.

We assume that the principal measures performance on a customer basis. For each customer served, the outcome measures (service time and revenue generated) are determined and

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the corresponding bonus amount is added to the server's account. We explore how measuring performance on average outcomes rather than single customer realizations can change these results in Section 2.7.

The optimal policy for the firm is determined by taking into account the revenue generation potentials of the two customer types, and the additional costs for extended service. These costs include the direct cost of effort by the server, and the indirect cost of extra congestion in the system. A customer is said to be profitable if the revenue generated from him or her exceeds these costs. In order to avoid trivial cases, we assume that the direct costs of providing extension to the low types is less than the expected revenues, i.e., $Rp_L > C_S$, so that when only the direct costs are considered it is profitable to provide the high level of service to the low type customers. This makes the problem more interesting, and also allows us to illustrate the effect of the indirect cost of service extension.

2.4 Model Analysis

The optimal policy analysis is first done for a given market segmentation scheme, i.e., considering the values q, p_H and p_L as parameters. Recall that this represents the case of a functional organization. In the subsequent section, we discuss the consequences of changing these parameters when customer segments are redefined.

To analyze the optimal contracts, we will use the two-stage procedure suggested by [54]. This approach is simply to break-up the principal's problem into a computation of costs and benefits for different actions taken by the agent. For each policy π, we consider the incentive scheme that minimizes the expected cost of getting the agent to choose effort levels stipulated by that policy, and then select the policy with maximum profit for the manager.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The first cost component for the manager's objective function, compensation, is defined by

$$E[x_1] = \frac{1}{\mu} + (q\frac{e_H k}{\mu(\mu - k)} + (1 - q)\frac{e_L k}{\mu(\mu - k)}) \tag{2}$$

and

$$x_2 = \begin{cases} Re_H & \text{with probability } p_H \text{ , if customer is high type} \\ Re_L & \text{with probability } p_L \text{ , if customer is low type} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The second cost component is the congestion cost, measured by the average waiting time in the queue. Note that the waiting time for any customer type depends not only on the service rate chosen for that type, but also on the choice for the other type. This is because of the effect of service time mean and variance on the waiting time of a given customer. More precisely, when the same service rate is chosen for all customers, the service time is drawn from an exponential distribution and the queue is an $M|M|1$ queue, whereas if different rates are chosen for the two types, the queue is an $M|G|1$ queue with hyper-exponential service times.

The expected queueing time for an average customer can be found using the Pollaczek-Khintchine formula,

$$W(\pi) = (\frac{\lambda E[x_1^2]}{2(1 - \lambda E[x_1])})$$

where $\pi = (e_H, e_L)$, x_1 is the service time and λ is the arrival rate. We can write the expected cost of waiting in line for an average customer as $T(\pi) = c.W(\pi)$.

The objective function for the manager, for given effort levels and compensation rates

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

is the long run average profit rate, which can be written as:

$$E[\Pi^P(\alpha_1,\alpha_2)] = \lambda q[(1-\alpha_2)Rp_He_H - \alpha_1\frac{e_Hk}{\mu(\mu-k)}] + \lambda(1-q)[(1-\alpha_2)Rp_Le_L - \alpha_1\frac{e_Lk}{\mu(\mu-k)}] - \alpha_1\lambda\frac{1}{\mu} - \lambda T(\pi).$$

The agent (the server) observes the customer type at each service start, and decides on an effort level which maximizes his expected utility, which is the expected wage minus the cost of effort. That is, he solves two separate problems for the two customer types:

$$E[\Pi^A(\alpha_1,\alpha_2)|H] = \alpha_2Rp_He_H + \alpha_1(\frac{1}{\mu} + \frac{e_Hk}{\mu(\mu-k)}) - C_Se_H \quad (4)$$

$$E[\Pi^A(\alpha_1,\alpha_2)|L] = \alpha_2Rp_Le_L + \alpha_1(\frac{1}{\mu} + \frac{e_Lk}{\mu(\mu-k)}) - C_Se_L. \quad (5)$$

Equation (4) is the expected utility when the customer is of high type and (5) is the expected utility when the customer is of low type. Since a policy is defined as the effort levels chosen for both customer types, there are four possible policies that the principal and the agent can choose:

1 : (0, 0) : Standard: no effort for any customer;

2 : (1, 0) : Differentiation: effort only for high type customers;

3 : (1, 1) : Extension: effort for all customers;

4 : (0, 1) : Reverse differentiation: effort only for low type customers.

Analyzing the incentives of the agent, we can show that policy (0, 1) can be dropped from the analysis.

Proposition 1 *Offering extended service only for the low type customers (policy (0,1)) is never optimal for the agent.*

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Proof. It is easy to see that the agent never prefers this policy if $p_H > p_L$. If it is optimal for the agent to put effort for low type customers, it must be optimal to do so for high type as well, since the expected revenue from high type is higher. Hence policy (0,1) can never be optimal for the agent. ■

To find the optimal policy for the manager, we maximize the long run average profits for the three alternative policies. These can be written as follows, conditioned on the policy chosen:

$$E[\Pi^P(\alpha_1, \alpha_2)|(0,0)] = -\lambda T(0,0) - \alpha_1 \frac{\lambda}{\mu} \quad (6)$$

$$E[\Pi^P(\alpha_1, \alpha_2)|(1,0)] = \lambda q[(1-\alpha_2)Rp_H - \alpha_1 \frac{1}{(\mu - k)}] - (1-q)\alpha_1 \frac{\lambda}{\mu} - \lambda T(1,0) \quad (7)$$

$$E[\Pi^P(\alpha_1, \alpha_2)|(1,1)] = \lambda(1-\alpha_2)[qRp_H + (1-q)Rp_L] - \alpha_1 \frac{\lambda}{(\mu - k)} - \lambda T(1,1) \quad (8)$$

The **first-best solution** refers to the optimal solution, which can be achieved when there is no information asymmetry, i.e., when the customer types and the effort can be observed by the principal. In this case, since the customer types and effort levels are observed, the agent will be compensated for the effort he expends. Formally, the first best contract is defined as a payment for each customer type, different from the contract definition given in (1), and it can be found solving the following program. Given a policy $\pi = (e_H, e_L)$, this program finds the first best contract $w^{FB} = (w_H, w_L)$ that maximizes the expected profits subject to the participation constraint of the agent, assuming the reservation utility for the agent is zero. Note that since the efforts are observed, there is no need for incentive compatibility constraints.

$$\max_{w_H, w_L} E[\Pi^{FB}|(e_H, e_L)] = \lambda R(qp_H e_H + (1-q)p_L e_L) - \lambda T(\pi)$$

$$s.t.$$

$$q(w_H - C_S e_H) + (1-q)(w_L - C_S e_L) \geq 0$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

In the optimal contract, the constraint inequality will be an equality. The first best contract compensates the agent as much as his effort cost, C_S, whenever $e_H = 1$ or $e_L = 1$, and pays 0 (the reservation wage) otherwise:

$$w^{FB} = (w_H, w_L) = (C_S e_H, C_S e_L) \tag{9}$$

2.5 Stage 1: Optimal Contracts for Each Policy

We can now solve for the optimal contract under each alternative policy, which constitutes the first stage of the analysis in the two-stage solution methodology. The complete optimization program for the differentiation policy, (1, 0), is shown below. For (0, 0) and (1, 1) the results are presented in the Appendix, given the similarity of the analysis to the case for (1, 0). We use superscripts FB and $*$ to refer to the first-best solution and the optimal that a manager can achieve, respectively. In this optimization program, the set of (α_1, α_2) values that generates a set of feasible effort levels are defined by the agent's incentive compatibility constraints ICH and ICL. In addition, the expected utility that the agent gets from this contract should be at least as much as the reservation utility, which defines the outside option for the server. This is the individual rationality constraint (IR2), which defines the feasible set of contracts together with the incentive compatibility constraints. Finally, we have the constraint $\alpha_2 \leq 1$ to ensure that the compensation for the revenue generated is not more than the revenue itself. Note that for any policy π, the congestion cost in the objective function, $T(\pi)$, is a constant. That is, the incentive contract design problem is not affected by the congestion measure, once a policy is given. The term $T(\pi)$ plays a role only in the second stage problem, where the optimal policy will be chosen.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Policy 2: Differentiation

$$\max_{\alpha_1,\alpha_2} E[\Pi^P(\alpha_1,\alpha_2)] = \lambda q[(1-\alpha_2)Rp_H - \alpha_1\frac{1}{(\mu-k)}] - (1-q)\alpha_1\frac{\lambda}{\mu} - \lambda T(1,0)$$

s.t.

$$q(\alpha_2 Rp_H + \alpha_1\frac{1}{\mu-k} - C_S) + (1-q)\alpha_1\frac{1}{\mu} \geq 0 \qquad \text{(IR2)}$$

$$\frac{\alpha_1}{\mu-k} + \alpha_2 Rp_H - C_S \geq \frac{\alpha_1}{\mu} \qquad \text{(ICH)}$$

$$\frac{\alpha_1}{\mu-k} + \alpha_2 Rp_L - C_S \leq \frac{\alpha_1}{\mu} \qquad \text{(ICL)}$$

$$\alpha_2 \leq 1$$

$$\begin{aligned} \alpha_1^* &\in [-qC_S\frac{p_H - p_L}{\mu p_L + kq(p_H - p_L)}, 0] \\ \alpha_2^* &= \frac{C_S}{Rp_H} - \alpha_1^*\frac{\mu-(1-q)k}{\mu(\mu-k)qRp_H} \\ E[\Pi^*(\alpha_1,\alpha_2)|(1,0)] &= \lambda qRp_H - \lambda qC_S - \lambda T(1,0) \end{aligned}$$

First-best solution:

$$\begin{aligned} w^{FB} &= (w_H, w_L) = (C_S, 0) \\ E[\Pi^{FB}|(1,0)] &= \lambda qRp_H - \lambda qC_S - \lambda T(1,0) \end{aligned}$$

The optimal contract rates for each policy can be seen graphically in Figure 1.

The numbered regions are the feasible regions for each policy considering the agent's incentive compatibility constraints; 1 2 and 3 stands for standard, differentiation and extension policy respectively. The first best solutions are achieved by the contracts that bind the individual rationality constraints for the policy, which are labelled by IR1, IR2 and IR3 for standard,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

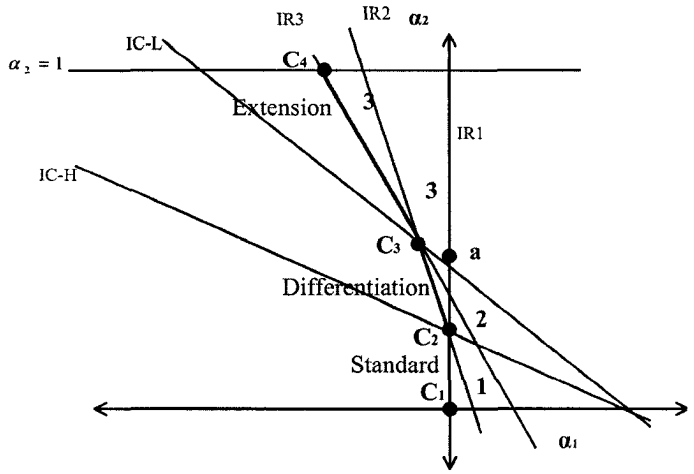

Figure 1: Optimal contracts and sensitivity

differentiation and extension policies respectively. There are infinitely many contracts for each policy, in the feasible range determined by the incentive compatibility constraints of the agents and the constraint $\alpha_2 \leq 1$. Optimal contracts for each policy are designated by the thick lines, labeled by the policy name.

The contract rates at the end points seen in Figure 1 are as follows:

$$
\begin{aligned}
C_1 &= (0, 0) \\
C_2 &= (0, \frac{C_S}{Rp_H}) \\
C_3 &= (-qC_S \frac{(p_H - p_L)(\mu - k)\mu}{\mu p_L - k(qp_H + (1-q)p_L)}, \frac{C_S}{R} \frac{(\mu - k)}{\mu p_L - k(qp_H + (1-q)p_L)}) \\
C_4 &= (-(R(qp_H + (1-q)p_L) - C_S)(\mu - k), 1)
\end{aligned}
$$

In all cases, the manager can achieve the first best solution with the incentive contracts (since the agent is risk neutral). As a result, the optimal profits for the principal are the same as the first-best case.

Note that the compensation rates (α_1, α_2) are not independent. The rate offered for service determines the rate that should be offered for revenue generated. Thus we find that incentive schemes which have only value generation (for example sales) related incentives,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

commonly found in practice, are not always optimal. The trade-offs between the two performance metrics need to be taken into account explicitly for incentive contract designs.

For the differentiation policy, α_1 should be non-positive, and for the extension policy it is entirely in the negative region. The negative compensation rate implies a "punishment" for long service. This is used as a means to balance the commission paid for the revenue generated and to incentivize the server for fast service. For the differentiation policy, the punishment for service time prevents the server from doing extended service to low class customers. This is so because in the optimal contract region (line segment $[C_2, C_3]$) the expected commission rate, α_2, is not high enough to compensate for the punishment associated with providing long service for low class customers, which have expected revenue of Rp_L. An increase in R and p_L would further reduce the commission rate to maintain the right incentives in place for differentiation. On the other hand, a high effort cost for the server (C_S) increases the commission rate required to do high level service. In addition, for higher values of q and p_H, both the commission rate and the punishment rate increases. This implies for everything else fixed, a firm facing a high end market will tend to offer stronger service time and sales incentives to implement the differentiation policy.

2.6 Stage Two: Policy Selection

In the second stage, the optimal policy is determined, i.e., the strategic level decisions are taken, given the best performance with each policy. This choice will be made (by the principal) to maximize net profits, found as revenues net of the cost of congestion and the compensation paid. In general, we can say that as long as the expected revenue generated is greater than

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the cost of effort, there will be a range of congestion cost values where the value generation policies, extension or differentiation are optimal. Since there is a trade-off between the value generated and the congestion in the system, as the unit congestion cost increases, there will be a switch to the policies which offer high level service for a smaller portion of the customer base.

Even when the effort is costless for the server, we obtain results which confirm the above intuition. If $C_S = 0$, i.e., there is no cost associated with effort for service extension, then

$$\begin{aligned} E[\Pi^P(\alpha_1, \alpha_2)|(1,0)] &\geq E[\Pi^P(\alpha_1, \alpha_2)|(1,1)] \\ \text{iff } c &\geq Rp_L\mu[\mu(\mu - k - \lambda) + k\lambda(1-q)]\frac{(\mu - k - \lambda)}{\lambda k\,(2\mu - k - \lambda)}. \end{aligned}$$

In other words, if the unit congestion cost is higher than a critical value, differentiation is preferred to offering high level service to everybody, regardless of the fact that it is costless for the server to offer high level service. As a result, given a fixed market segmentation scheme, the optimal policy choice is characterized by the critical value of congestion cost which trades off the extra revenue from extension to a customer segment with the extra load it brings to the system. The derivation of all the results are provided in the Appendix.

Proposition 2 *The critical value of unit congestion cost such that for $c \geq c^*$ differentiation is preferred to extension is given by*

$$c^* = \frac{\mu}{\lambda}\frac{(\mu - k - \lambda)}{(2\mu - k - \lambda)}(Rp_L - C_S)[\frac{(\mu - \lambda)(\mu - k)}{k} - \lambda q].$$

Similarly, the critical value of unit congestion cost such that for $c \geq c^{**}$ the standard policy is

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

preferred to differentiation is

$$c^{**} = \frac{(\mu - k)}{\lambda} \frac{(\mu - \lambda)}{(2\mu - \lambda - k)} (Rp_H - C_S)[\frac{(\mu - \lambda)(\mu - k)}{k} - \lambda q].$$

When there is only one possible market segmentation scheme, i.e., parameters q, p_H, and p_L are fixed, the optimal strategy the firm follows is found by comparing the unit congestion cost c with the critical values c^* and c^{**}. There are three cases.

Case I ($c > c^{**}$): the firm is not in a value creation environment, so extension is not suitable for any market segment. The firm prefers that the operation remains a cost center.

Case II ($c^{**} > c > c^*$): the firm can apply relationship management or value creation strategies, but not to its entire customer base. The only customers that are worth spending effort on are the high type customers. The firm opts for service level differentiation.

Case III ($c \leq c^*$): all customer segments are profitable and extension is worthwhile for all. The firm chooses to target all of its customers for additional value creation.

Given capacity (μ), profiles of the segments (q, p_H), complexity of the extension of service and the abilities of the servers (k), the revenues vs. direct costs of extension (R, C_S), and the congestion averseness of the firm (c), the firm will be in one of the above regions, which dictates the best policy to implement. The value of $c/(Rp - C_S)$, i.e., the ratio of the cost of a customer waiting in line one unit of time to the expected revenue generated from a customer affects the policy regions' relative size. The comparative statics results stated below show the effect of the parameters on the policy choice. By taking derivatives with respect to appropriate problem parameters, we have the following.

Proposition 3

$$\frac{\partial c^*}{\partial q} < 0 \ , \quad \frac{\partial c^*}{\partial C_S} < 0 \quad \textit{and} \quad \frac{\partial c^*}{\partial p_L} > 0;$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

also

$$\frac{\partial c^{**}}{\partial q} < 0 \quad , \quad \frac{\partial c^{**}}{\partial C_S} < 0 \ \textit{and} \ \ \frac{\partial c^{**}}{\partial p_H} > 0.$$

The range between the two critical values, (c^*, c^{**}), determines the attractiveness of the differentiation policy for the firm. Differentiation becomes more attractive as c^* decreases (by increasing the proportion of high type customers and the cost of effort, or decreasing the revenue potential from low types), or c^{**} increases (by decreasing the proportion of high type customers and the cost of effort, or increasing the success probability for high type customers). Note that both threshold values, c^* and c^{**} decrease with the size of the high class segment (q). With a bigger high class segment, the differentiation policy may increase the congestion too much compared to standard policy, so c^{**} decreases. Similarly, a big high class segment implies a small low class segment, in which case the additional revenue generation potential of the extension policy is little compared to the differentiation policy, therefore c^* decreases as well, favoring the differentiation policy. This implies that the differentiation policy will be favored by a bigger high class segment size in low congestion cost/high capacity environments, whereas it will be hindered in high congestion cost/low capacity environments.

We next analyze incentive contract sensitivity to assumptions made in the analysis. In Section 2.8.1, we explore the implications of changing the market segmentation decision on the policy choice.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2.7 Alternative Compensation Schemes and Incentive Contract Sensitivity

In this section we explore the possibility of control mechanisms other than linear contracts, like full monitoring via payment over average outcome values and discuss their implementation.

In the model considered in this chapter, the incentive contract is designed such that payments are made after each customer is served, according to the outcomes observed for that single customer. The payment per customer scheme is motivated by what is observed in practice for sales related incentives. Note however that the model also assumes that the principal can observe average waiting times, and in theory she can use this information to monitor the actions of the agent indirectly. In that case, knowing the service and demand parameters, and observing the average waiting times, the principal can calculate the realized value of q, i.e., the proportion of customers that the server has spent effort on, and pay exactly $\lambda q C_S$ to compensate for the effort cost of the server. This can be seen as a monitoring plan as studied by [63]. Given a risk neutral agent, both this monitoring scheme and the earlier proposed linear contract result in no loss of efficiency. Thus monitoring does not improve the profits for the principal.

A monitoring contract like the one explained above is an alternative solution to the linear contract, but there are some practical concerns that need to be addressed in its implementation. For the monitoring contract to be applicable, the payments need to be made after a sufficient number of interactions so that long-run averages can be observed. Furthermore it should be appropriate to compensate the server on the basis of averages as opposed to a per customer basis. Depending on the scale of the demand process and the payment intervals, this might not

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

be possible in all settings.

When some of the assumptions made in the analysis are relaxed, monitoring may become even more difficult to implement. Consider the case when there is a diagnosis cost for the server for each decision he makes about the customer type. Or a case where the service extension for the low type customer takes a longer time than for the high type. In both of these settings, monitoring only the average waiting time will not be enough and information on average revenue generation will also be required to ensure the proper actions by the server. Convincing a server that basing their compensation on expected successful cross-sells is fair, may be more difficult to do than doing so for the waiting time performance, given the possible volatility in customer revenue generation potentials and arrival patterns in shorter time frames. To account for the diagnosis cost, the linear incentive contract can be modified such that α_2 is adjusted up for the differentiation policy and down for the extension policy. When each customer type has a different extension task time, the linear incentive contract for the differentiation policy would be the same, while for the extension policy it can be used with modified (α_1, α_2).

Finally, although we consider a single server system, in practice where there are several servers, it would not be appropriate to attribute the average waiting time to the performance of a single server. The server, being punished or rewarded for a performance measure that he does not perceive to be totally under his control, would not be able to get good guidance and motivation from such a contract.

A second alternative compensation plan is a profit sharing contract, which is a classical solution suggested by the principal agent literature for the risk neutral agent case. This type of contract would be possible if the principal could charge the waiting time cost to the agent. In

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

our analysis we assume this is not an option because of practical concerns as explained in the previous paragraph.

This discussion illustrates that the linear contracts are not the unique solution to our model, though they provide a feasible and robust compensation mechanism under different assumptions. Next we investigate these contracts in more detail.

Linear Contracts We briefly discuss the robustness (i.e., the ability to induce the desired action when some conditions change) of the linear incentive contracts to errors in parameter estimation and contract design, and the impact of deviating from the optimal contracts as a result of these errors. The capacity of the system μ and the revenues R are relatively easy to assess. On the other hand, the disutility of the agent for expending effort for the service extension, C_S, can be difficult to quantify exactly. The server himself may have an assessment of this cost, but can miscommunicate this information to the manager. Moreover, evidence from call centers shows that this parameter can have a wide range. Servers differ in their preferences between generating extra revenue and providing fast and efficient service [19]. Similarly, there can be errors in assessing the parameters q, p_H and p_L.

There are two types of deviations from an optimal contract. The first is failing to satisfy the individual rationality constraint as an equality, thereby paying the agent more than his reservation utility. The second, is failing to satisfy the incentive compatibility constraints of the agent for the optimal policy, hence getting another policy implemented. The cost of this error would be either lost revenues (when the policy implemented is differentiation instead of extension) or increased congestion in the system (in case the policy implemented is extension instead of differentiation).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

These potential deviations can be visualized in Figure 1. Assume the optimal policy is differentiation, and the contract chosen is the point $C_2 = (0, \frac{C_S}{Rp_H})$. Let the estimated effort cost be δC_S, $\delta > 1$. Then the contract offered moves up from the optimal point C_2. For $\delta \leq \frac{p_H}{p_L}$, the contract offered is in the region designated by number 2, and this move would be a deviation of the first type as explained above. On the other hand, for $\delta > \frac{p_H}{p_L}$, the move would be to a point like 'a', and would constitute the second type of deviation. The latter would result in operational costs due to the increase in system congestion, in addition to the efficiency loss.

The ratio $\frac{p_H}{p_L}$ is a measure of robustness of the contracts that induce the differentiation policy, (1, 0). This fraction determines the allowable range for estimation error δ of C_S. We see that as p_H and p_L get closer, the feasible region for (1, 0) becomes smaller and the robustness of the contracts deteriorates. Intuitively, as the two market segment's characteristics differ more from each other, it becomes easier to differentiate between them and provide distinctly different incentives for the treatment of the two types of customers. In that case, small estimation errors in the contract parameters would not cause dramatic differences in the servers' behavior.

On the other hand, when the optimal policy is extension, if the condition $\frac{k}{\mu} < \frac{p_L}{(qp_H+(1-q)p_L)}$ does not hold, than the linear contract suggests a punishment for the revenue generated and a positive compensation on the service time outcome. We consider a negative α_2 value to be infeasible in practice. Further intuition can be obtained by re-writing the condition as

$$(\frac{k}{\mu(\mu-k)})/\frac{1}{\mu} < \frac{p_L}{q(p_H-p_L)} \tag{10}$$

This condition is comparing cost and benefits for the extension policy: the ratio of "increase in service time by extension" and "service time for standard service" should be less than the ratio of "probability of revenue from low types" and "contribution of high types to success

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

probability". I.e., if the service time difference is huge (k is big) or if the low types' revenue generation potential is very small compared to high types (not high enough to compensate the service time increase on the left hand side) then the extension policy cannot be implemented via this linear contract.

In conclusion, environments with high $\frac{p_H}{p_L}$ make the linear contracts more robust for implementation of the differentiation policy while the extension policy becomes more difficult to implement by linear contracts in these environments. This suggests that the contract design should take into account the policy choice of the firm. For standard policy, no incentive is actually needed so a flat salary would be enough. For the differentiation policy, a linear contract that gives incentives on only the revenue dimension would be appropriate. For the extension policy, the contract should include both of the two outcome dimensions. Moreover, it requires a punishment on service time. For some extreme cases, where k is too high relative to the capacity and the revenue generation potential, the linear contract would suggest a punishment on revenue, which may not be practical. In those cases, if possible, a monitoring contract could be used instead.

2.8 Market Segmentation Problem

In the previous sections, we analyzed the policy problem given two customer types and characterized the policy choice depending on the unit congestion cost. However, in reality the customer types are the result of the market segmentation decision of the firm.

In order to incorporate the market segmentation decision, the model is further developed as follows. We assume that for any customer, the probability of generating revenue R by offering high level service is a realization $\widehat{p}$ of a random variable P, which we call the probability of

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

success. Management knows the density function $f(p)$ of this random variable, but cannot observe the realized value for each customer. Given the density of success probabilities $f(p)$, a segmentation scheme is determined first. This is done by dividing the customer base into two segments, using a critical probability value θ. Namely, the customers with $p > \theta$ are defined as the High-type, and the remaining, with $p \leq \theta$ are the Low-type customers. Then the average representative success probabilities can be assessed for each segment, where

$$p_H = E[P|p > \theta] \tag{11}$$

$$p_L = E[P|p \leq \theta]. \tag{12}$$

Note that for any density $f(.)$, $p_H > p_L$ holds. Similarly, for each arriving customer, the probability of being a high-type, q, is found as

$$q = P(customer\ type = H) = P(p > \theta) = 1 - F(\theta). \tag{13}$$

This parameter determines the size of the high-type segment. In this section, we first demonstrate the sensitivity of the policy choice and the profits to the market segmentation decision and then solve the optimal market segmentation problem.

2.8.1 Sensitivity to Market Segmentation

Determining a market segmentation scheme corresponds to the selection of a value for θ. Each customer has a potential for generating a revenue R. However, unless the server chooses to undertake an extension task, this potential cannot be realized. The parameter k represents the content of this extension task, and can be seen as a measure of customer needs. Thus, revenues are not generated unless customer needs are met. The higher these customer needs, the higher will be the k parameter, and as a result the higher the impact of extension on congestion and

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

costs. In other words, the parameter k characterizes the operational impact of the value creation activity on the system. Hence, the optimal choice of θ can be seen as a market segmentation decision that takes into account both customer revenue generation potential and customer service needs. In this section, gains from optimizing θ are illustrated by analyzing the case where θ is fixed first and then the policy decision is made, taking this θ value as given.

We can characterize the policy choice as a function of these two key parameters: the additional load that service extension brings, k, and the threshold for the minimum probability of success for a high type customer, θ. Using the results of the analysis in Section 2.6, for any given value of θ policy choice can be defined by critical values of k as follows:

Case I ($k \leq k^*(\theta)$): extension policy is optimal

Case II ($k^*(\theta) < k < k^{**}(\theta)$): differentiation policy is optimal

Case III ($k^{**}(\theta) \leq k$): standard policy is optimal

The following result provides some structural properties of the curves $k^*(\theta)$ and $k^{**}(\theta)$. All proofs are in the Appendix.

Proposition 4 *$k^{**}(\theta) > k^*(\theta)$ for all $\theta \in (0, 1)$, and $k^*(\theta)$ and $k^{**}(\theta)$ are non-decreasing with θ:*

$$\frac{\partial k^*}{\partial \theta} \geq 0 \qquad \frac{\partial k^{**}}{\partial \theta} \geq 0.$$

The result states that for any given market segmentation decision, the maximum affordable load for the extension task is lower for the extension policy than it is for the differentiation policy. Moreover, as the high type segment size decreases, i.e., θ is increased, extension tasks with higher complexity (i.e., higher k) can be supported by the value creation strategies, extension and differentiation. When the policy is standard, the profits are the same for all values of θ

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

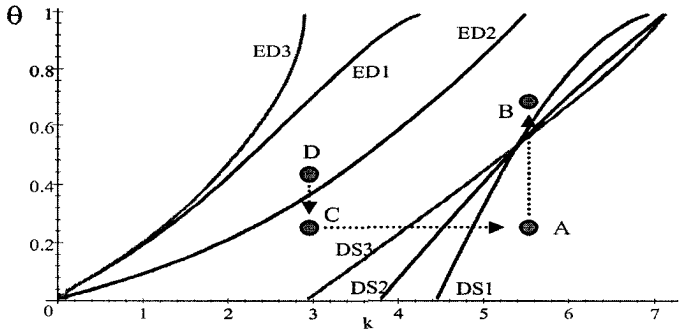

Figure 2: Policy choice for different customer profiles

(equivalently (q, p_H, p_L)). Similarly, the profits are the same for all extension policies, since there is no market segmentation in practice and all customers receive the same service. This observation drives the following result, which points out the importance of optimally selecting θ.

Remark 5 *For a given customer profile, a segmentation scheme θ that leads to the selection of the differentiation policy as optimal ($k^*(\theta) < k < k^{**}(\theta)$) by definition yields higher expected profits compared to a segmentation scheme θ' that leads to the policies 'extension' ($k < k^*(\theta')$) or 'standard' ($k^{**}(\theta') < k$) as optimal.*

This observation shows that making the segmentation decision cleverly would pay off, leading to higher profits. In other words, if we can make differentiation the optimal policy with a given customer profile through an appropriate choice of θ, we are better off than targeting all customers or remaining a cost center.

We illustrate the gains from optimizing the market segmentation decision θ with a numerical example shown in Figure 2, where the curves $k^*(\theta)$ and $k^{**}(\theta)$ are plotted for three

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

success probability density functions. Note that all curves have a positive slope as indicated by Proposition 4. In this figure, the optimal policy is extension in the regions to the left of the lines EDi, standard on the right of the lines DSi and differentiation in between these two lines for each density function i.

The result stated in Remark 5 is illustrated by the move from points A to B, and from D to C. With these moves into the differentiation region, the profits are increased only by changing θ. In addition, the effect of task complexity is illustrated by the horizontal moves. For example, while point C is in the differentiation region, when k increases we may move to a point such as A, where the standard policy is optimal and no value is generated. Then an optimal decision would be a move to a point such as B, by increasing θ.

To sum up, we have made three observations: First, the differentiation policy is potentially the most profitable policy. Second, the policy choice hinges on the market segmentation decision, so in order to achieve the maximum profits (using a differentiation policy if feasible), θ should be chosen optimally. Finally, the optimal θ choice is a function of k, representing the content or complexity of the extension task, and the success probability density function. Given these observations, the next question we address is: what is the optimal value for θ, and how do we enforce it given the private information of the server?

2.8.2 Market Segmentation Decision

Up to this point, the analysis is done taking the market segmentation variable θ as a parameter. In this section, θ is a decision variable, both for the principal (the manager) and the agent (the server). Recall that this represents the case of an integrated firm with a sophisticated manager and server. In this setting, $\theta = 0$ indicates choice of the extension policy, $\theta = 1$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the standard policy and $0 < \theta < 1$ implies the differentiation policy. The manager declares her θ decision. However, the server can choose another θ without the manager observing it, given the private information he has about the customers and his objective of maximizing his own utility. Therefore the contract (α_1, α_2) should be incentive compatible for the marginal customer (who has success probability θ) rather than the average customer for the particular segment, unlike the previous analysis of Section 2.4. There is only one incentive compatibility constraint that imposes indifference between offering standard and extended service for this marginal customer. The reservation utility for the agent is taken to be zero.

First let us define:

$$\begin{aligned} p_H(\theta) &= \frac{\int_\theta^1 x f(x)dx}{\int_\theta^1 f(x)dx} \quad , \quad q(\theta) = \int_\theta^1 f(x)dx \\ W(\theta) &= \frac{\lambda}{\mu}\left(\frac{((\mu-k)^2 + k(2\mu-k)\int_\theta^1 f(x)dx)}{((\mu-k)(\mu-\lambda) - \lambda k \int_\theta^1 f(x)dx)\,(\mu-k)}\right) \end{aligned}$$

Then the optimization problem that the principal solves is as follows:

$$\max_{\theta,\alpha_1,\alpha_2} E[\Pi] = \lambda(1-\alpha_2)Rp_H(\theta)q(\theta) - \lambda\alpha_1\left(\frac{q(\theta)}{\mu-k} + \frac{(1-q(\theta))}{\mu}\right) - \lambda c W(\theta)$$

$s.t.$

$$\begin{aligned} \alpha_2 Rq(\theta)p_H(\theta) + \alpha_1\left(\frac{q(\theta)}{\mu-k} + \frac{(1-q(\theta))}{\mu}\right) - C_S q(\theta) &\geq 0 \qquad \text{(IR)} \\ \alpha_2 R\theta + \alpha_1 \frac{1}{\mu-k} - C_S &= \alpha_1 \frac{1}{\mu} \qquad \text{(IC)} \end{aligned}$$

Since the efficient solution satisfies the individual rationality constraint as an equality, for any given value of θ, the optimal contract (α_1, α_2) is at the intersection of the two constraints, (IR) and (IC), which is found as:

$$\begin{aligned} \alpha_1 &= -C_S q \frac{(p_H - \theta)(\mu-k)\mu}{-kq(p_H-\theta) + \theta(\mu-k)} \\ \alpha_2 &= C_S \frac{\mu-k}{R(-kq(p_H-\theta) + \theta(\mu-k))} \end{aligned}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

We see that there is a unique optimal (linear) contract for the optimal market segmentation scheme as opposed to the infinitely many contracts in the case where θ was not a decision variable. Moreover, this contract requires a punishment for the service time component.

This optimal contract is a function of the variables determining the system capacity and the complexity of the value creation task, μ and k, as well as the variables defining the server disutility and customer characteristics, C_S and $f(\cdot)$. The comparative statics analysis shows the following result.

Proposition 6 *Assume a market segmentation decision θ, and a segment size q. For two customer pools X, Y with densities of success probabilities $f(p)$, $g(p)$, if $X \geq_{st} Y$ for $p \geq \omega \geq \theta$ (i.e $1 - F(p) \geq 1 - G(p)\ \forall p \in [\omega, 1]$) then the magnitude of the contract rates for both task dimensions is higher for X than it is for Y. i.e., $\alpha_1^{(X)} < \alpha_1^{(Y)}$ and $\alpha_2^{(X)} > \alpha_2^{(Y)}$*

This result compares two customer pools, one with more concentration on the high-end compared to the other, and shows that firms operating in a high-end market should give more commission rate for the revenue and more punishment for the service time. In a high-end market, in comparison to another, low-end market, a higher average success probability imposes high punishment in service time to balance for it. This in turn implies a higher commission rate for the revenue, since for the marginal customer, the success probability is the same for both markets (which is θ).

Having found the optimal (α_1, α_2), the program reduces to finding the θ that maximizes the objective value, given this contract. The following result characterizes the unique optimum for the market segmentation problem:

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Proposition 7 *Let* $f(x) > 0$ *for all* $x \in [0, 1]$ *and* θ' *be the solution to*

$$(R\theta - C_S)\left((\mu - k)(\mu - \lambda) - \lambda k q(\theta)\right)^2 \frac{1}{(2\mu - k - \lambda)\lambda k} - c = 0 .$$

Then the optimal $\theta = \theta^*$ *is found as:* $\theta^* = \theta'$ *if* $0 \leq \theta' \leq 1$, $\theta^* = 1$ *if* $\theta' > 1$.

The optimal cut-off point for the market segmentation problem, θ^* equates the expected marginal revenues from offering extended service to a customer with success probability θ, to the cost of expected congestion with that definition of the high-type segment. So for a customer with success probability $p = \theta^*$, we are indifferent between offering standard service and extended service.

Then the long run average profit rate would be as follows:

$$E[\Pi|\theta^*] = \lambda q(\theta^*)(p_H(\theta^*)R - C_S) - \lambda c W(\theta^*) \tag{14}$$

The optimality condition implies that for $C_S > 0$, $\theta = 0$ would never be optimal. In other words, it is never optimal to offer extended service to *all* customers. There is always some portion of customers, for whom even the direct cost of effort would not be paid off if the extended service is offered. This implies that the strategy of targeting all customers is only possible in settings where you have the type of server that has $C_S = 0$. In a setting where the value creation activity being pursued is cross-selling, this implies that an optimized cross-selling initiative has to be targeted. The common approach of attempting a cross-sell on all customers is clearly not supported when servers show the slightest disutility with respect to the sales activity. On the other hand, a very large unit congestion cost c or a very high utilization rate λ/μ satisfies the following condition and makes $\theta^* = 1$.

$$(R - C_S)\frac{((\mu - k)(\mu - \lambda))^2}{(2\mu - k - \lambda)\lambda k} \leq c . \tag{15}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

If the condition given in (15) is satisfied then the standard policy becomes optimal.

Next, we provide some comparative statics results for the optimum.

Proposition 8 *a)Optimal* θ *increases with k, and* $\frac{\lambda}{\mu}$. *i.e.,* $\frac{\partial \theta^*}{\partial k} > 0$ *and* $\frac{\partial \theta^*}{\partial (\lambda/\mu)} > 0$

b) Given two customer pools X, Y with densities of success probabilities $f(p)$, $g(p)$, *if* $X \geq_{st} Y$ *(i.e.,* $1 - F(p) \geq 1 - G(p)\ \forall p \in [0, 1]$ *) then* $\theta^*_X \geq \theta^*_Y$.

This proposition shows the environmental conditions that would affect the optimal market segmentation decision θ^*. In the first part, we show that an increase in the complexity of the extension task, and an increase in system utilization would increase θ^*. This is an intuitive result given that both factors would increase the load on the system and thus would make the extended service more difficult to offer for the customers with low revenue generation potentials.

The second part of the proposition shows the effect of the customer profile, represented by the success probability distributions. We compare two customer pools, one with more concentration on the high-end compared to the other. This result shows that for the high-end markets, the definition of a high-type customer will be up-graded, i.e., one would be more 'picky' selecting the customers to offer service extension. The intuition behind this result comes from the issue of resource allocation among the customers. The customers with high success probabilities would be given the priority when allocating the scarce service time. When there are many customers at the high end, the targeted high-type market size is filled with high-end customers and this results in selectiveness in defining customer types.

Another implication of this result is that if we consider a firm with constant capacity that could operate in two markets under the conditions stated in proposition 8b, the congestion level

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

in the high-end market is expected to be higher compared to the low-end market. Increasing selectiveness increases the marginal revenues expected from a high type customer, so the manager can afford to face the costs of higher congestion levels in the system. This result has implications on the customer experiences in different markets. For any customer with a given value of success probability p, being in a low-end market would be preferable to being in a high-end market for two reasons. First, the expected waiting time is lower. Second, the cut-off level of success probability, θ, to receive a higher service level is lower so that there is a higher chance to receive high level service in a low-end market. Hence we show the importance of one's relative position in the population with respect to the service received.

2.9 Summary of Model Findings-Concluding Remarks

A stylized model that captures the key levers for value creation strategy choice is presented and analyzed. The model is unique in that it combines value creation and process related issues, thereby providing a coherent framework to analyze sales initiatives like cross-selling or service level differentiation strategies. The analysis characterizes under what conditions a firm would choose to remain as it is or to attempt value creation. The latter choice is further elaborated in terms of a choice between a differentiation strategy and one of "targeting all customers". Value creation strategies are favored by moderately loaded systems where the extension task's complexity does not push capacity utilization to a maximum and where the expected revenue from the extension task is high enough to compensate for the extra costs of congestion. When unit congestion cost is relatively high compared to revenues, service level differentiation is preferred to targeting all customers.

For integrated firms that optimize θ, service level differentiation is the optimal choice.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

High congestion costs, high task complexity for the extension task, or high system utilization lead to the status quo choice where no value creation is pursued. Unless servers experience no disutility for the extension task, "targeting all customers" is never optimal, even if congestion cost is low. Firms that would like to pursue this strategy would need to select employees who enjoy the extension task, or provide additional automated support to their employees in an effort to reduce this disutility.

The analysis of market segmentation makes two important points. The first one is that a change in market segmentation (θ) can imply a change in policy choice. Thus we can talk about better θ choices that lead to higher profitability. More specifically we find that choosing and sustaining a service level differentiation strategy may hinge on the market segmentation decision. Firms that can operate in the integrated mode, where for example marketing and operations jointly optimize θ, are shown to be clearly better off in terms of achieving this profitability. The second point is that the optimal market segmentation decision depends on the distribution of revenue generation in the customer base. This in turn implies that actions that change the underlying customer profitability distribution shape, like better targeted sales for example, can instigate a shift in policy. Thus, emphasizing the need for a prospective type of analysis as opposed to the currently prevailing retrospective analysis in customer value estimation.

The analysis herein uses the probability of revenue generation from a specific extension task as the measure of customer profitability. The parameter k, on the other hand, captures the difference in needs (in terms of service extension to create value) between the H and L-type customers. As a result, the θ choice discussed in Section 2.8.1 can be seen as a segmentation decision that combines value and customer needs concerns. Our analysis demonstrates how

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

one can perform value coupled with needs based segmentation, where the θ choice allows one to make a trade-off between customer profitability and needs. This type of segmentation decision is in line with the recommendations in [28] and [53], who critique pure customer profitability based segmentation schemes.

The market segmentation analysis clearly illustrates that even if firms have very reliable individual customer data and can effectively estimate individual customer value, unless the aggregation decision which determines the segmentation scheme is done correctly, value cannot be maximized. As shown by the analysis, a correct aggregation decision needs to make the trade-offs between operational performance, the breadth of the value creation activity (i.e., segment size), and the depth (i.e., profitability potential) of these types of activities. An organization, that acts as what we labeled as the functional organization, will not be able to reap all the benefits of a value creation initiative. In other words, functioning as an integrated firm is essential to success.

Once a strategy choice is made, we illustrate how the desired strategy can be implemented through the design of appropriate incentive contracts. For the setting where θ is not optimized, a set of optimal linear incentive contracts are explicitly characterized for each strategy. We show that when the two tasks being considered have opposite performance effects (like the standard service and extension activities are assumed to have in this analysis), then optimal linear contracts may involve punishments, i.e., disincentives for one of the dimensions. Furthermore, we find that incentive payments for these two tasks (for example service and sales) depend on each other, and providing incentives for only one dimension, as is frequently observed with sales based incentives, can lead to undesirable behavior. In the setting where we have a sophisticated manager and server who optimize the market segmentation decision,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

we show that there is a unique optimal linear contract. This contract is clearly a function of θ, illustrating the close ties between market segmentation, process design, and incentives in value creation initiatives.

The model illustrates that success in implementation of a service level differentiation program depends on proper incentive contract design, which requires good parameter estimation. The policy implementation is particularly susceptible to misunderstanding employee preferences (characterized here by effort cost) and are not robust in settings where distinctly different customer segments cannot be formed. While our analysis assumes that there are only two customer segments (H and L), this last result suggests that as companies increase the number of segments that they define for their service level differentiation strategies (for example cases with three and five segments can be found in retail banking), implementations will become less robust. For integrated firms, the uniqueness of the optimal incentive contract illustrates the high sensitivity of these types of value creation initiatives to incentive design.

A final remark can be made for firms that plan to initiate cross-sell type of value creation programs without re-configuring their capacity. Considering the fact that a customer's experience (in terms of waiting time and service level in our setting) will impact their satisfaction, one can expect firms that have targeted a higher end customer pool to experience more customer dissatisfaction associated with their value creation initiatives if the service capacity is kept the same. This suggests that firms with relatively higher end customer pools should be more careful in implementing programs like cross-selling or add-on sales, and points out the importance of considering the capacity implications of these programs. Future research should focus on explicitly modeling the customer experience in value creation initiatives.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2.10 Appendix to Chapter 2

Problem Formulations and Solutions for Incentive Contract Problems of Policies 1 (Standard) and 3 (Extension):

Policy 1: Standard

$$\max_{(\alpha_1,\alpha_2)} E[\Pi^P(\alpha_1,\alpha_2)] = -\alpha_1\frac{\lambda}{\mu} - \lambda T(0,0)$$

s.t.

$$\begin{aligned} \frac{\alpha_1}{\mu} &\geq 0 \quad \text{(IR)} \\ \frac{\alpha_1}{\mu-k} + \alpha_2 R p_H - C_S &\leq \tfrac{\alpha_1}{\mu} \quad \text{(ICH)} \\ \frac{\alpha_1}{\mu-k} + \alpha_2 R p_L - C_S &\leq \tfrac{\alpha_1}{\mu} \quad \text{(ICL)} \\ \alpha_2 &\leq 1 \end{aligned}$$

$\alpha_1^* = 0 \qquad \alpha_2^* \in [0, \frac{C_S}{Rp_H}]$
$w^{FB} = (w_H, w_L) = (0,0)$
$E[\Pi^*(\alpha_1,\alpha_2)\|(0,0)] = -\lambda T(0,0) = E[\Pi^{FB}\|(0,0)]$

Policy 3: Extension

$$\underset{(\alpha_1,\alpha_2)}{Max}\, E[\Pi^P(\alpha_1,\alpha_2)] = \lambda(1-\alpha_2)R(qp_H + (1-q)p_L) - \alpha_1\frac{\lambda}{\mu-k} - \lambda T(1,1)$$

s.t.

$$\begin{aligned} \frac{\alpha_1}{\mu-k} + \alpha_2 R(qp_H + (1-q)p_L) - C_S &\geq 0 \quad \text{(IR)} \\ \frac{\alpha_1}{\mu-k} + \alpha_2 R p_H - C_S &\geq \tfrac{\alpha_1}{\mu} \quad \text{(ICH)} \\ \frac{\alpha_1}{\mu-k} + \alpha_2 R p_L - C_S &\geq \tfrac{\alpha_1}{\mu} \quad \text{(ICL)} \\ \alpha_2 &\leq 1 \end{aligned}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$\alpha_1^* \in (-(R(qp_H+(1-q)p_L)-C_S)(\mu-k), -qC_S\frac{(p_H-p_L)(\mu-k)\mu}{\mu p_L-k(qp_H+(1-q)p_L)})$
$\alpha_2^* = \frac{C_S}{R(qp_H+(1-q)p_L)} - \alpha_1^*\frac{1}{(\mu-k)R(qp_H+(1-q)p_L)}$;
$w^{FB} = (w_H, w_L) = (C_S, C_S)$
$E[\Pi^{*M}(\alpha_1,\alpha_2)\|(1,1)] = qRp_H + (1-q)Rp_L - C_S - T(1,1) = E[\Pi^{FB}\|(1,1)]$

Proof of Proposition 2:

Differentiation is preferred to extension only if:

$$qRp_H - qC_S - T(1,0) \geq qRp_H + (1-q)Rp_L - C_S - T(1,1),$$

which is equivalent to

$$T(1,1) - T(1,0) \geq (1-q)(Rp_L - C_S) \ . \tag{16}$$

This condition compares the costs and profits obtained by offering extended service to low type customers. Below, we re-write condition (16) with some new notation. We also use $T(.) = cW(.)$ as defined in the model. Let

$\Delta W_L = W(1,1) - W(1,0)$, $\Delta T_L = T(1,1) - T(1,0)$, and $\Delta R_L = (1-q)(Rp_L - C_S)$.

Then we can rewrite (16) as $\Delta T_L = c\Delta W_L \geq \Delta R_L$, which is equivalent to:

$$c \geq \frac{\Delta R_L}{\Delta W_L} = c^* = \frac{(1-q)(Rp_L - C_S)}{-(-2\mu + 2q\mu + k + \lambda - q\lambda - qk)\lambda\frac{k}{(\mu-\lambda-k)\mu(\mu^2-\mu k-\lambda\mu+\lambda k-\lambda qk)}}$$

After simplification, we get

$$c^* = \frac{\mu}{\lambda}\frac{(\mu-k-\lambda)}{(2\mu-k-\lambda)}(Rp_L - C_S)(\frac{(\mu-\lambda)(\mu-k)}{k} - \lambda q) \ .$$

Selection between differentiation and standard policies is done in a similar way, this time comparing the costs and profits of offering extended service to the high type customers.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The condition to determine the critical unit congestion cost c^{**} is $qRp_H - qC_S - T(1,0) \geq -T(0,0)$, or equivalently $qRp_H - qC_S \geq T(1,0) - T(0,0)$. This inequality can be represented as a comparison of marginal gains (ΔR_H) and losses (ΔT_H), i.e $\Delta R_H \geq \Delta T_H = c\Delta W_H$.

Thus the breakeven value of unit congestion cost c^{**} that makes us indifferent between the two policies differentiation and standard is found as follows:

$$c^{**} = \frac{\Delta R_H}{\Delta T_H} = \frac{q(Rp_H - C_S)}{\lambda q k \frac{2\mu-\lambda-k}{(\mu^2-\mu k-\lambda\mu+\lambda k-\lambda q k)(\mu-k)(\mu-\lambda)}}$$

$$c^{**} = \frac{(\mu-k)}{\lambda}\frac{(\mu-\lambda)}{(2\mu-\lambda-k)}(Rp_H - C_S)[\frac{(\mu-\lambda)(\mu-k)}{k} - \lambda q].$$

Proof of Proposition 4

In this proof, we use the two basic assumptions stated below, which imply profitability of all customers considering direct costs only and stability for all policies.

$$i) \quad Rp_L - C_S > 0 \tag{17}$$

$$ii) \quad \frac{\lambda}{\mu-k} < 1 \tag{18}$$

Define k^* and k^{**} implicitly as follows:

$$\psi(\theta,k) = c^*(\theta,k) - c = 0 \quad \text{for } k = k^*$$

$$\phi(\theta,k) = c^*(\theta,k^{**}) - c = 0 \quad \text{for } k = k^{**}$$

First we show that $\psi(\theta,k)$ and $\phi(\theta,k)$ are strictly decreasing in k:

$$\frac{\partial}{\partial k}\psi(\theta,k) = \mu^2(Rp_L - C_S)\frac{2\lambda^2 k - 4\mu\lambda k - 2\mu^3 + 2\mu^2 k + 5\lambda\mu^2 - 4\lambda^2\mu + \lambda^3 + \lambda q k^2}{\lambda k^2(-2\mu + k + \lambda)^2}$$

$$= -(Rp_L - C_S)\{\frac{\mu^2}{\lambda(2\mu-k-\lambda)^2}(\frac{(\mu-\lambda)(\mu-k)}{k} - \lambda q)\}$$

$$-(Rp_L - C_S)\{\frac{\mu^2(\mu-\lambda)}{\lambda k^2}\frac{(\mu-k-\lambda)}{(2\mu-k-\lambda)})\} < 0 \tag{19}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

It is easy to see that the above expression is negative since $(Rp_L - C_S) > 0$ by (17), and also for both terms, $\frac{\mu^2}{\lambda(2\mu-k-\lambda)^2}(\frac{(\mu-\lambda)(\mu-k)}{k} - \lambda q) > 0$ and $\frac{\mu^2}{\lambda(2\mu-k-\lambda)^2}(\frac{(\mu-\lambda)(\mu-k)}{k} - \lambda q) > 0$ hold by (18).

Similarly $\frac{\partial\phi(\theta,k)}{\partial k} < 0$ as the expression found below is negative by our assumptions.

$$\frac{\partial\phi(\theta,k)}{\partial k} = \frac{(-\mu+\lambda)\left(Rp_H - C_S\right)(\mu\lambda k^2+\lambda^2 qk^2+2\mu^2 k\lambda + 2\mu^4-2\mu^3 k - 3\mu^3\lambda + \lambda^2\mu^2-\lambda^2k^2-\mu\lambda qk^2}{\lambda k^2(-2\mu+k+\lambda)^2}$$
$$= -(\mu-\lambda)(Rp_H - C_S)\frac{\mu^2(2(\mu-k)-\lambda)(\mu-\lambda)+\lambda k^2(\mu-\lambda)(1-q)}{\lambda k^2(-2\mu+k+\lambda)^2} < 0 \tag{20}$$

Now we can show the results in the Proposition:

Part I: $k^{**} > k^*$ for any given θ.

This follows from the fact that $\phi(\theta,k) > \psi(\theta,k)$ and both $\phi(\theta,k)$ and $\psi(\theta,k)$ are decreasing in k (as shown in (19) and (20)). Then the solution of $\phi(\theta,k) = c$ must be greater than the solution of $\psi(\theta,k) = c$.

Part II: $k^{**}$and k^* are increasing in θ.

Recall that k^* is implicitly defined by the equation $\psi(\theta,k) - c = 0$. So we use implicit differentiation to find $\frac{\partial k^*}{\partial\theta}$:

$$\frac{\partial k^*}{\partial\theta} = -\frac{\partial(\psi(\theta,k)-c)}{\partial\theta}[\frac{\partial(\psi(\theta,k)-c)}{k}]^{-1}|_{k=k^*} \tag{21}$$

The first part of (21) is found to have a $(-)$ sign as shown below.

$$-\frac{\partial(\psi(\theta,k)-c)}{\partial\theta} = -\frac{d}{d\theta}\psi(\theta,k) = -(\frac{\partial\psi(\theta,k)}{\partial q}\frac{\partial q}{\partial\theta} + \frac{\partial\psi(\theta,k)}{\partial p_L}\frac{\partial p_L}{\partial\theta}) < 0 \tag{22}$$

To see that (22) holds, we check all the terms as follows:

1-First term of (22) is $(+)$ since $\frac{\partial\psi(\theta,k)}{\partial q} < 0$ (by Prop. 3) and $\frac{\partial q}{\partial\theta} \leq 0$ by definition of q.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2-Second term of (22) is (+) since $\frac{\partial\psi(\theta,k)}{\partial p_L} > 0$ (by Prop. 3) and $\frac{\partial p_L}{\partial\theta} \geq 0$ by definition of p_L.

The second part of (21) has a (−) sign also as shown in equation (19). So, $\frac{\partial(\psi(\theta,k)-c)}{k} < 0$. Therefore,

$$\frac{\partial k^*}{\partial\theta} = -\frac{\partial(\psi(\theta,k)-c)}{\partial\theta}[\frac{\partial(\psi(\theta,k)-c)}{k}]^{-1}|_{k=k^*} > 0 \ \textit{by eqn.(19) and (22)} \tag{23}$$

The same analysis is repeated to show that $\frac{\partial k^{**}}{\partial\theta} > 0$:

$$\frac{d}{d\theta}\phi(\theta,k) = \frac{\partial\phi(\theta,k)}{\partial q}\frac{\partial q}{\partial\theta} + \frac{\partial\psi(\theta,k)}{\partial p_H}\frac{\partial p_H}{\partial\theta} > 0 \tag{24}$$

$$\frac{\partial k^{**}}{\partial\theta} = -\frac{\partial(\phi(\theta,k)-C_S)}{\partial\theta}[\frac{\partial(\phi(\theta,k)-C_S)}{k}]^{-1}|_{k=k^{**}} > 0 \text{ by eqn. (20) and (24)} \tag{25}$$

Proof of Proposition 6

Given $f(p), g(p), \theta$ such that

$$q = \int_\theta^1 f(p)dp = \int_\theta^1 g(p)dp \quad \text{and } \exists \text{ a } \omega \text{ s.t. } \int_\omega^1 vf(p)dp \geq \int_\omega^1 g(p)dp \quad , \forall\omega \geq \theta,$$

$$(p_H(\theta))^X \geq (p_H(\theta))^Y \tag{26}$$

is true. Then since

$$\begin{aligned}\frac{\partial\alpha_1}{\partial p_H} &= -C_S q(\mu-k)^2\mu\frac{\theta}{(-kqp_H+kq\theta+\theta\mu-\theta k)^2} < 0\\ \frac{\partial\alpha_2}{\partial p_H} &= C_S(\mu-k)k\frac{q}{R(-kqp_H+kq\theta+\theta\mu-\theta k)^2} > 0\end{aligned}$$

$\alpha_1^{(X)} < \alpha_1^{(Y)}$ and $\alpha_2^{(X)} > \alpha_2^{(Y)}$ follows from 26.

Proof of Proposition 7

The necessary condition for optimality is found as follows:

$$\frac{\partial}{\partial\theta}(E[\Pi]) = f(\theta)(-(R\theta-C_S) + \frac{(2\mu-k-\lambda)c\lambda k}{\left((\mu-k)(\mu-\lambda)-\lambda k\int_\theta^1 f(x)\,dx\right)^2}) = 0$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

For $f(\theta) > 0$, it reduces to the condition stated in proposition. So θ' is a local optimum. Moreover,

$$\frac{\partial}{\partial\theta}(E[\Pi]) \geq 0 \text{ for } \theta \leq \theta' \text{, and } \frac{\partial}{\partial\theta}(E[\Pi]) \leq 0 \text{ for } \theta \geq \theta' .$$

Thus we conclude that θ' is a global maximum. If $\theta' > 1$, then $\frac{\partial}{\partial\theta}(E[\Pi])|_{\theta=1} > 0$ and the optimum is at the boundary, i.e., $\theta^* = 1$.

Proof of Proposition 8

For a given density $f(.)$ of success probabilities, define

$$\xi_f(\theta) = (R\theta - C_S)((\mu-k)(\mu-\lambda) - \lambda k \int_\theta^1 f(x)dx)^2 \frac{1}{(2\mu - k - \lambda)\lambda k} - c = 0 \text{at optimum.} \tag{27}$$

We use the following four results in the proof:

$$\frac{\partial \xi_f()}{\partial\theta} = \frac{((\mu-k)(\mu-\lambda) - \lambda k(1-F(\theta)))(R((\mu-k)(\mu-\lambda) - \lambda k(1-F(\theta)) + 2\lambda k f(\theta)(R\theta - C_S))}{(2\mu - k - \lambda)\lambda k} > (\tag{28}$$

$$1 - F(\theta) \geq 1 - G(\theta) \Rightarrow \xi_f(\theta) \leq \xi_g(\theta) \tag{29}$$

$$\frac{\partial \xi_f(.)}{\partial k} < 0 \text{ as shown in proof of Proposition 4} \tag{30}$$

$$\frac{\partial \xi_f(.)}{\partial(\lambda/\mu)} < 0 \tag{31}$$

a) The results follow easily taking the derivatives using the implicit function theorem as follows:

$$\frac{\partial\theta^*}{\partial k} = -\frac{\partial \xi_f(.)}{\partial k}\left[\frac{\partial \xi_f(.)}{\partial\theta}\right]^{-1}_{\theta=\theta^*} \geq 0 \quad \text{by (28) and (30)}$$

$$\frac{\partial\theta^*}{\partial(\lambda/\mu)} = -\frac{\partial \xi_f(.)}{\partial(\lambda/\mu)}\left[\frac{\partial \xi_f(.)}{\partial\theta}\right]^{-1}_{\theta=\theta^*} \geq 0 \text{ by (28) and (31)}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

b) Given X, Y with densities $f(x)$, $g(x)$, and $X \geq_{st} Y$ defined as $1 - F(\theta) \geq 1 - G(\theta)$ for all $\theta \in [0, 1]$, *i.e.*, $\overline{F(\theta)} \geq \overline{G(\theta)}$ $\forall\, \theta \in [0, 1]$. We can re-write the optimality condition for any customer pool as follows:

$$\xi(\theta) = (R\theta - C_S)\left((\mu - k)(\mu - \lambda) - \lambda k\overline{F(\theta)}\right)^2 \frac{1}{(2\mu - k - \lambda)\,\lambda k} - c = 0 \qquad (32)$$

If θ_Y^* is optimal for Y, $\xi_g(\theta_Y^*) = 0$ by the optimality condition, stated in (32). We know that $\xi_f(\theta) \leq \xi_g(\theta)$ (by 29). Equivalently,

$$\xi_f(\theta_Y^*) < \xi_g(\theta_Y^*) = 0. \qquad (33)$$

Then, $\theta_X^* \geq \theta_Y^*$ since θ should increase to make $\xi_f(\theta_X^*) = 0$, given the results $\xi_f(\theta_Y^*) < 0$ (by (33)) and $\frac{\partial \xi()}{\partial \theta} > 0$ (by(28)).

The objective function for any θ is then found as:(let $p_H(\theta) = p_H$, $q(\theta) = q$ for ease in exposition)

$$E[\Pi|\theta] = (1 - \alpha_2)Rp_H(\theta)q(\theta) - \alpha_1(\mu - kq(\theta)) - cW(\theta)$$

$$= (1 - (\frac{C_S}{R}\frac{\mu}{\mu\theta + kq(p_H - \theta)}))Rp_Hq - (-qC_S\frac{p_H - \theta}{\theta\mu + kq(p_H - \theta)})(\mu - kq) - cW(\theta)$$

$$= \frac{q\mu C_S p_H - q\theta\mu C_S + kq^2\theta C_S - kq^2 C_S p_H}{\theta\mu - kq\theta + kqp_H} + \frac{Rq\theta\mu p_H - q\mu C_S p_H - Rkq^2\theta p_H + Rkq^2 p_H^2}{\theta\mu - kq\theta + kqp_H} - cW(\theta)$$

$$= \frac{1}{\theta\mu - kq\theta + kqp_H}(\theta\mu - kq\theta + kqp_H)q(Rp_H - C_S) - cW(\theta) = q(Rp_H - C_S) - cW(\theta)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3 Breast Cancer Screening Services: Trade-offs in Quality, Capacity, Outreach, and Centralization

3.1 Introduction

Breast cancer is the most common cancer among women, and the second leading cause of cancer related deaths after lung cancer. The World Health Organization (WHO) estimates that more than 1.2 million people were diagnosed with breast cancer worldwide in 2001 [60]. The American Cancer Society (ACS) estimated that breast cancer would be diagnosed in 211,300 women, and 39,800 would die from the disease in 2003 in the U.S. [3]. The value of mass screening and the ages at which it is appropriate are currently in dispute [105]. However most developed countries have organized screening programs to pro-actively detect breast cancer [69], as early diagnosis of most types of breast cancer is very effective. The 5 year survival rate is 98% with early stage breast cancer treatment [3]. This chapter contributes to the understanding of factors that influence breast cancer screening effectiveness.

While many have examined breast cancer screening in operations [75, 90, 73, 111], statistics [33, 115, 36, 114], and health [69, 88, 22, 31, 55, 17], this paper appears to be the first to include aspects of all of the following interacting effects in the same model: (i) disease progression, (ii) the link between service quality and the volume of mammogram screens provided by the health care provider, (iii) participation levels in the health care program, (iv) factors influencing the participation in a mammogram screening program, and (v) limited service capacity and the potential effect of the utilization of service resources on patient waiting and an increased potential for poor health outcomes due to late diagnosis. Accounting for

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

system dynamics can strongly influence model-based health policy decisions ([29, 30], others below), so we do so here. The first of four model-based experiments in this chapter assesses the cost implications for two approaches to improving early detection: outreach and quality improvements. The second studies interactions between participation levels and the potentially deleterious health effects of waiting due to stochastic effects and highly utilized capacity. The third and fourth examine interactions between service decentralization, access, and screening quality. The analysis shows that increasing outreach without improving quality and maintaining capacity may result in less beneficial results than predicted by standard models due to the interactions of these effects.

Critical factors for breast cancer screening program success include two quality measures, *sensitivity* (the probability of detecting cancer in a patient with the disease) and *specificity* (the probability of a negative result in a patient without the disease), as well as *acceptability*, the extent to which those for whom the test is designed have access to and agree to participate in testing [89]. Figure 3 summarizes some interactions between these factors: quality, access, system capacity, and health outcomes. *Screening quality* is influenced by radiologist experience, the annual volume of readings, and film quality standards, among other factors. The popular press [82] recently highlighted problems with screening quality in the U.S. and the importance of the experience of radiologists who read mammograms. One potential cause identified is low minimum accreditation standards: 480 mammograms readings per year [6] compared with 3,000/year in British Columbia, Canada and 5,000/year in the U.K. [65]. Some argue that radiologists should read a minimum of 2,500 mammograms per year to stay sharp [65]. While imperfect reading quality is perhaps inevitable, the lack of quality incurs system costs. False positives add extra cost and consume service capacity for follow-up tests, incur

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

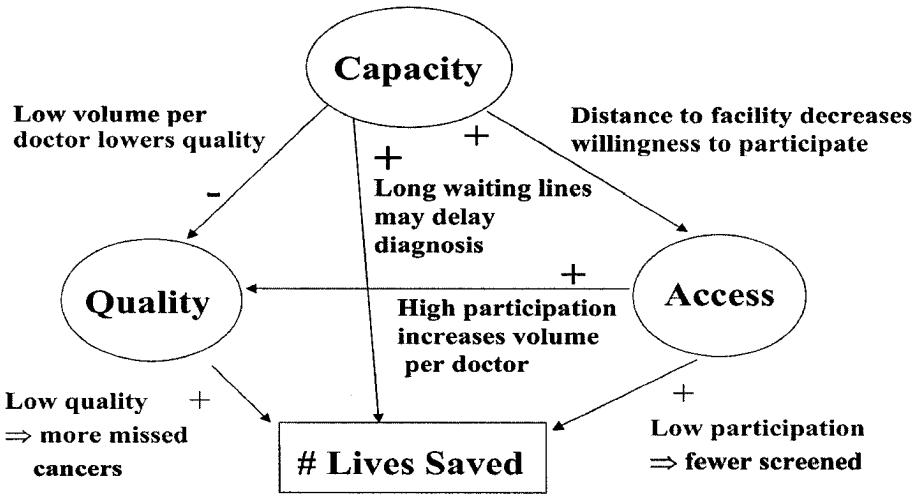

Figure 3: Mammogram Screening System Performance Is Influenced by Several Interacting Effects

the potential for unnecessary treatment, and can burden patients [40]. False negative results decrease the chances of survival by missing the opportunity for early detection and treatment.

For *acceptability/access*, the WHO [89] recommends that mammography should not be introduced for breast cancer screening unless the resources are available to ensure effective and reliable screening of at least 70% of the target age group, women over the age of 50 years. Factors that influence participation include the availability of local health care services, trust in health care providers, and the level of governmental or private health care coverage. This chapter does not examine the human and political factors that may significantly affect acceptability, but does model operational factors such as the dynamics of capacity and waiting, as well as the relationship between the proximity of service facilities and the likelihood of participating in a screening program [41]. While not all regions experience problems with waits, some do, and limited service capacity or scheduling are operational issues that can cause waits of even 3-6 months [46, 1].

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Delays in the screening and diagnosis system [104, 80, 26, 27] can reduce survival rates by delaying the stage of disease at diagnosis [96]. Increasing the minimum annual screening volume for accreditation may aggravate the capacity problem by reducing the number of radiologists willing or eligible to provide the service, thereby reducing accessibility. On the other hand, increasing the number of readings would increase the accuracy for communities that are still served. The dynamic interactions of reading volume and quality, access, delays in service and health outcomes, among other complications, present a challenge for health care service system design.

This chapter presents a system dynamics model of screening services that combines the interactions described above, and uses the model to analyze the impact of different interventions on the system performance in terms of health outcomes and costs. The general modeling framework is applicable to preventive health care services in general and is aimed at contributing to a better understanding of health care system design. Section 3.2 identifies papers that have studied some aspects of these important determinants of mammogram screening program success, but no paper seems to account for all of these potential interactions at once. It then presents a mathematical model of these interactions. The model is validated with data from published studies where possible, and simulation experiments in Section 3.3 are motivated by policy issues that arise in the U.S. and French health care contexts. A system dynamics model with deterministic differential equations might seem appropriate at first glance, but the model employs stochastic dynamics so that waiting can be more adequately described. Section 3.4 discusses implications and limitations of the model, as well as further research directions.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3.2 Problem Formulation

We view breast cancer screening provision as a problem of matching the supply (screening service) and demand (participation in the program, screening frequency) while ensuring sufficient quality (high sensitivity and specificity of the test). The objective is to reduce the breast cancer deaths, keeping system costs in mind.

There are several related health care service delivery papers in the operations management literature. Location models [73, 111] help situate screening facilities, but model neither the waiting list and health outcomes nor the documented relationship between screening volume and quality. [66] is unique in that it considers the deleterious health effects of waiting due to constrained health care service capacity. Their deterministic differential equation model is appropriate for the short-term transient dynamics of their application (smallpox control) but is less appropriate for assessing the long run performance of a continuously operating health care service. [116] models the interrelationship of service levels and system capacity, costs, and health outcomes (for renal disease), and obtains structural results with a fluid model. Those papers do not include the stochastic effects and waiting that we model. [103] is related in that it models the deleterious health effects of delayed health service provision in the context of kidney transplant. That work applies more to sequential stochastic assignment problems, but does not model programs with repeated screening.

Several studies of breast cancer screening in the operations research literature focus on modeling the disease progression to optimize the screening schedules for an individual [115, 68, 99, 11]. Their goal is similar to one of ours: to evaluate performance of different screening policies to find one maximizing the health outcomes and/or minimizing the costs. Among

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

different objectives of these models are minimizing costs associated with screening and the disease [99, 11, 112], minimizing detection delay given a number of screens [68] or more general utility functions related to detection time [115]. A limitation of these studies is that they do not consider the system-level constraints such as limited service capacity or delayed arrivals of women with a demand for screening. Moreover, they model the quality of screening test as a constant or as a function of tumour size, but do not incorporate the effect of radiologist experience. Our model addresses these points, while it takes the screening frequency as a given parameter.

Section 3.2.1 describes our choice for breast cancer disease progression and mammogram program service delivery structure models. Section 3.2.2 does the same for the relationship between screening volume and quality, and describes assumptions about acceptability. Section 3.2.3 gives cost assumptions.

3.2.1 Disease Progression and Service System Structure

We model two types of service (both screening tests and follow-up diagnostic tests), a finite service capacity, and the potential of waits due to finite service capacity and randomness associated with patient scheduling. There are n parallel servers (radiologists) that serve c queues (facilities).

Figure 4 depicts a system with $c = 1$ facility. A fraction h of individuals that reach the target screening program age (say, 50 year old women) join the screening program, the remainder are considered to be unenrolled. Enrolling individuals have an early stage cancer with probability p. With frequency f, enrollees attempt to obtain a screening mammogram (a type $t = s$ job) but wait in queue if all servers are busy. The service time for screening

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

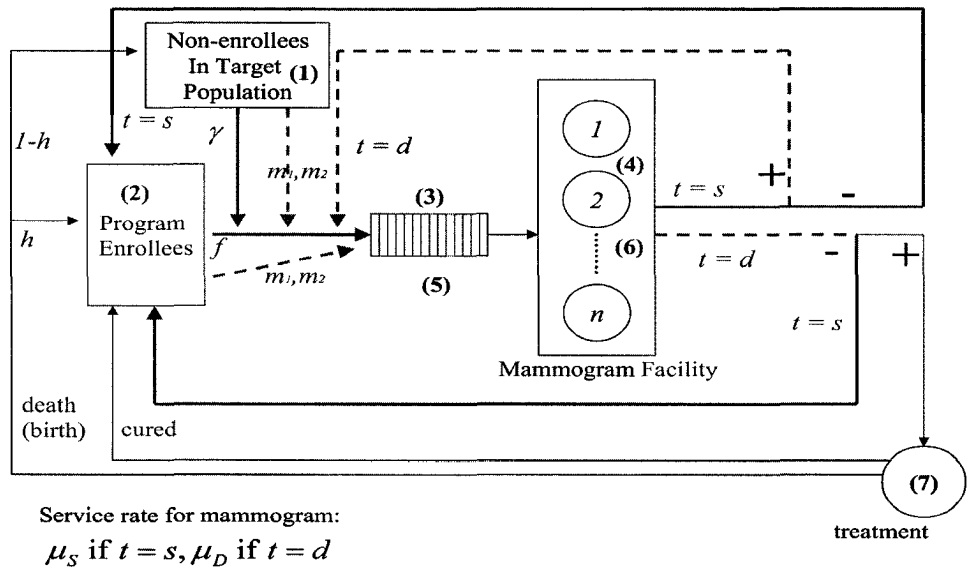

Figure 4: Operational Service System View for Screening

mammogram is an exponential random variable with rate μ_S. The service rate represents the bottleneck resource in the facility. If the screening mammogram result is positive, a diagnostic test is required. This means a $t = s$ job will return to the queue as a type $t = d$ job with a probability that depends on the health state of the patient and sensitivity and specificity of the screening. Diagnostic test flows, represented with dashed lines, have a service rate $\mu_D = \mu_S/a$, a higher sensitivity than screening mammograms, and incur greater costs than screening mammograms. If cancer is detected after completion of diagnosis, the woman goes under treatment. Otherwise she goes back to the target population. To reflect exogenous sources of detection such as self-exam and general practitioner referral, diagnostic tests may also be requested directly without a screening mammogram (with rate m_1 and m_2 respectively for women with early and late stage cancer). Unenrolled individuals may enroll in the program later, with rate γ.

The numbers in Figure 4 identify the state of progress through the health service system, (1) unenrolled, (2) enrolled but not yet scheduled for a screening, (3) waiting for a screening

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

mammogram, (4) getting a screening mammogram, (5) waiting for a diagnostic test, (6) getting a diagnostic test, or (7) undergoing treatment for cancer. We do not model treatment capacity explicitly, so a woman diagnosed with breast cancer begins treatment immediately.

We assume a three-stage health model like many others [75, 115, 68, 112]: (1) healthy, (2) preclinical, or early stage, breast cancer (3) clinical, or late stage, breast cancer. The number of individuals in health status $j = 1, 2, 3$ and service system state $i = 1, 2, \ldots, 7$ is denoted $X_{i,j}$. The service capacity constraint limits the number being served at any given time, $\sum_{j=1}^{3} X_{4,j} + X_{6,j} \leq n$.

Patient flows through the health system are represented vertically in Figure 5, and changes in health status are illustrated with horizontal flows from compartment to compartment. Table 1 gives the default values of the parameters that determine the flows, including parameters introduced above as well those that describe screening quality introduced below. A description of how their values were validated relative to published data and statistical information from national agencies can be found in the Appendix. We assumed a fixed accuracy level for diagnostic follow-up tests, while the accuracy of screening mammograms is calculated as a part of the model, as described in the following section. The parameter values give good fit with incidence and breast cancer death data from Statistics Canada [2].

We model two causes of death, disease specific mortality and all-cause mortality [21]. Noncancer related deaths occur at rate g and may affect women in all compartments (arcs for these transitions are not shown in the figures for clarity). Cancer-related deaths are assumed to affect only women with late stage cancer ($X_{i,3}$) or women undergoing treatment ($X_{7,j}$). A constant population size is assumed (a new individual enters the target population as a death occurs). Conclusions from results below are robust over a reasonable range of g.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 1: Summary of Default Values for Parameter Estimates

parameter	description
$g = 0.04$	death rate/person/year for reasons other than breast cancer (implied conditional life expectancy is 75 years: 50 years upon entry to target population, plus $1/g = 25$ years)
$\gamma = 0.01$	screening program enrollment rate/person/year for unenrolled
$p = 0.0055$	probability of having cancer at the entry to the target population
$b_3 = 0.0915$	probability of death from late stage cancer when not having treatment/person/ year
$b_{73} = 0.0915$	death rate/person/year from late stage cancer during treatment
$b_{72} = 0.0081$	death rate/person/year from early stage cancer during treatment
$s_{12} = 3.0122 \times 10^{-3}$	rate/person/year of acquiring preclinical cancer
$s_{23} = 0.585$	rate/person/year of cancer advancing from preclinical to clinical stage ([113] also shows fit with exponential distribution)
$m_1 = 0.01$	rate/person/year for self-referral for diagnosis, from preclinical stage
0.95	specificity of diagnostic test
0.90	sensitivity of diagnostic test for preclinical stage
0.99	sensitivity of diagnostic test for clinical stage
$m_2 = 4.13$	rate/person/year for self-referral for diagnosis, from clinical stage [96]
$r_0 = 100$	treatment completion rate/person/year after a false diagnosis
$r_1 = 0.2$	treatment completion rate/person/year after early diagnosis
$r_2 = 0.2$	treatment completion rate/person/year after late diagnosis
$\delta = 1$	sensitivity of screening for late cancer
$a = 1.5$	service effort for diagnostic test / screening mammogram
$p_1 = 0.05$	probability of recurrence after treatment of an early stage cancer
$p_2 = 0.15$	probability of recurrence after treatment of a late stage cancer

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

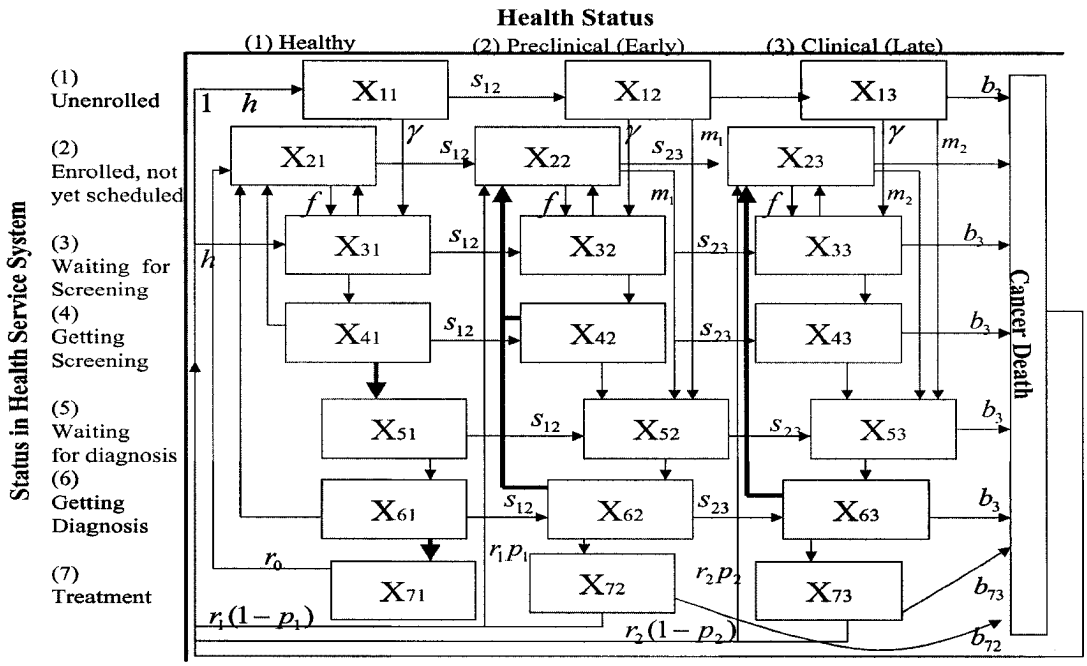

Figure 5: Stochastic Compartmental Model View for Service System and Health Status

For a given sensitivity and specificity, the dynamics of disease progression, program enrollment, and screening outcomes are assumed to be Markovian. In addition to the $X_{i,j}$, the state includes information about the volume of screening mammogram readings of the servers, which influences the reading quality, as described in Section 3.2.2. Quality influences the system dynamics model in Figure 5 via the flows on the thick arrows that are associated with false positive ($X_{4,1}$ to $X_{5,1}$ and $X_{6,1}$ to $X_{7,1}$) and false negative ($X_{4,2}$ and $X_{6,2}$ to $X_{2,2}$, and $X_{6,3}$ to $X_{2,3}$) test results.

3.2.2 Volume-Quality Relationship and Acceptability

Many factors influence the quality of readings [17, 40, 87], and acceptability (measured here by the fraction of women that enroll in a breast cancer screening program). Here we focus on the relationship between volume and quality. The quality-volume relationship is uncertain and complex, and its accuracy is questioned by some articles [14, 39]. We rely on the literature

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

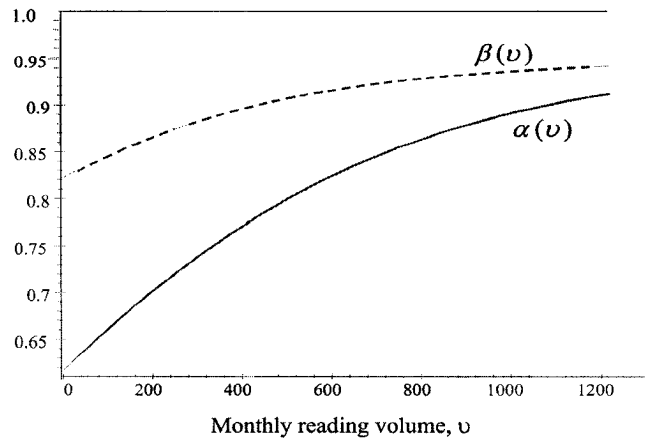

Figure 6: Sensitivity $\alpha(v)$ and Specificity $\beta(v)$ as a Function of Monthly Reading Volume, v

that supports this relationship [65, 42] to provide a model for quality of mammogram reading. We then describe a simple model for acceptability.

Sensitivity and specificity increase as the average reading volume increases [65, 42]. We assume a logistic relationship between volume and quality that embodies qualitative features from empirical observations [42].

$$\begin{aligned} \text{sensitivity} &= \alpha(v) = \frac{0.95}{1 + 0.5393e^{-0.0021v}} \\ \text{specificity} &= \beta(v) = \frac{0.95}{1 + 0.158e^{-0.0024v}} \end{aligned}$$

where v is a monthly reading volume. These functions are plotted in Figure 6.

There are several choices for modeling the volume of readings per server. We chose to measure the screening volume for fixed-length time periods and base quality during one time on the volume in the preceding time period. Specifically, if $vol(t)$ is the total number of completed mammogram readings up to time t years, then the volume of readings during the previous month is

$$v(t) = vol\left(\frac{\lfloor 12t \rfloor}{12}\right) - vol\left(\frac{\lfloor 12t \rfloor}{12} - 1\right).$$

An alternative approach to model quality-volume relationship could be a discrete state Marko-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

vian model (with knowledge measured by the number of readings recalled and that are effective for quality output, $v(t) \in \{0, 1, 2, \ldots\}$, with transition $v \rightarrow v + 1$ upon service completion, and $v \rightarrow v - 1$ with forgetting rate αv, where $1/\alpha$ is the 'half life' of recalling a screen). This approach leaves a memory of readings that is highly volatile. A similar continuous state, continuous time model for $v(t)$ might also be easier to study analytically, but also neglects the issue of regulatory checks which count readings in specific fixed length time periods. Formally, our choice to measure volume in fixed-length time periods changes the stochastic process in Section 3.2.1 from a Markov chain to a generalized semi-Markov process. We use simulation for the analysis, a standard tool for studying this class of processes.

Acceptability, measured by the probability h that a woman initially enrolls in a screening program, is simplified here to depend upon the distance from the nearest facility. We set h to match national enrollment statistics in some experiments. In others that test the effect of distributed facilities versus a centralized facility (requiring longer travel), we assume the odds ratio of enrolling drops by 3% for each additional 8km traveled, as in an empirical study [41]. This distance-enrollment relation oversimplifies a complex set of effects, but like the quality-volume relationship, it is probably the best we can use on the basis of research now available. Other factors that involve screening quality and acceptability can be modeled similarly.

3.2.3 Cost Assessment

The economic costs of screening, diagnostic follow-up tests and treatment in Table 2 are based on values taken from the literature and converted to 2003 U.S. dollars using the consumer price index for medical services where necessary. The cost of follow-up diagnosis tests is based on a weighted average of the costs of diagnostic mammogram, sonography, fine needle aspiration

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 2: Assumed Cost Structure (all in 2003 US$, [97])

Screening mammogram	$145
Diagnostic test	$471

Estimated Treatment Cost per Case of Preclinical and Clinical Stage Breast Cancer

Stage	% at Diagnosis	Est. Discounted Long-term Cost [45]
Preclinical: local	65%	$ 54,013
Clinical: regional	30%	$ 70,066
Clinical: distant	5%	$ 59,463

and biopsy reported in [97]. Long term discounted costs of treatment and continuing care are based on a three stage model for breast cancer classification [45]: local, regional or distant cancer. Local cancer corresponds to our preclinical stage. Both regional and distant stages correspond to our clinical stage. We therefore use a weighted average late treatment cost for the regional and distant stages, $\$68,551 = (\$70,066 \times 0.3 + \$59,463 \times 0.05)/0.35$. This lets us use incidence data to calculate the expected costs of breast cancer cases. While the cost conclusions below are based on data available from one HMO, some generality is preserved because the relative costs of different tests and treatment at different stages are more relevant for our purposes than the absolute figures, and cost comparisons with other previous studies do not indicate a sizable difference [45].

Like [97, 22, 55], we do not explicitly account for the cost of increasing program enrollment, service capacity, or changing screening standards. Those costs are likely to be highly dependent upon the target population. [5, 98] illustrate how to include such costs. We also vary the program enrollment probability over a wide range for sensitivity analysis. Realistically it may be expensive or even impossible to achieve very high, or even very low, enrollment levels.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3.3 Analysis

This section presents four analyses motivated by issues in the U.S. and French health systems in three subsections. The first experiment assesses the cost implications for two approaches to improving early detection, either through outreach or through quality increases due to increasing the minimum screening volume standards. The second experiment examines interactions between quality and the potentially deleterious health effects of waiting in the presence of insufficient capacity. The last two examine the interactions of service decentralization, access, and screening quality. Since the model is not easy to analyze in closed form, simulation experiments were used to estimate long run averages with batch means [74] for health outcomes and annual costs. Tests for stationarity with the Heidelberger and Welch [56] test led us to remove 20 years of 'warm-up' from the beginning of each simulation of 400-520 years for each parameter setting.

3.3.1 Increasing standards or expanding outreach

The National Cancer Institute (NCI) recommends mammography screening every one to two years for American women over 40 [61]. The General Accounting Office [108] estimates that 2/3 of the mammography machine capacity is utilized and that 64% of the target population had a screening mammogram in 2000, less than the 70% recommended by the WHO. Waiting is not significant on the whole, although waits of several months occur in some metropolitan and rural areas. The U.S. FDA currently requires radiologists to interpret a minimum of 480 screenings per year [6]. If participation increases to 70%, more cancers will be detected early both because more women are being screened and because screening quality improves with an increased volume per radiologist. On the other hand, an increase in demand along with

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 3: Parameter Values for Numerical Experiments in Section 3.3.1

Parameter	Values Set for Experiments
Probability of enrollment (h)	0.55, 0.60, 0.65, 0.70, 0.75
Volume standard (std)	480, 2500
Screening service rate (μ_S)	1000, 5000
Number of servers (n)	30, 6

recent decreasing trends in capacity [108] may exacerbate waiting and lessen the ability to detect cancers early. An alternative to increasing participation directly is to improve screening quality, and therefore health outcomes, by increasing the minimum annual screening standard from 480 to 2, 500 (the figure recommended by [65]).

This section examines the following questions. What health benefits can be gained by increasing outreach and what would be the impact for the capacity requirements? What are the benefits of increasing the minimum screening volume to 2, 500 per year? What are the implications on capacity requirements?

We simulated a target population of 25, 000 women. Since readings are typically not the only service provided [14], we simulated both the 480 base-level screening standard and increased reading standard of 2, 500/year by presuming that the maximum rate of readings would be about twice the standard level (so $\mu_S = 1,000$ for base-level screening, and $\mu_S = 5,000$ for the increased level). Initially, the fraction enrolling immediately is $h = 0.55$, so that the long run participation in the screening program roughly matches the empirical 64% value [108] (some women enroll later spontaneously or upon noticing symptoms). To evaluate the effect of increasing outreach and acceptability, we checked multiple h from 0.55 to 0.75 for both scenarios. The number of radiologists is set so that 30, 000 mammogram screenings per year can be done. Table 3 summarizes the parameters used in the experiments.

Health Outcomes. Figure 7 shows that the average number of early diagnoses per

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

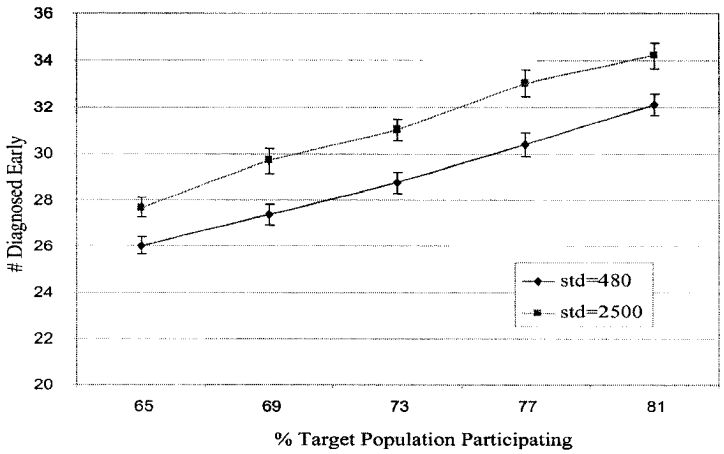

Figure 7: Annual Cases Diagnosed Early as Function of Participation for Two Levels of Reading Volume Standards (480 and 2500/year). Error Bars Show 95% Confidence Intervals

year increased when the reading standards were increased to 2, 500 from 480, regardless of the level of participation in the screening program. This was a direct result of the improved sensitivity and specificity of readings at higher volumes. Figure 7 also shows that for a given volume standard, increased participation levels resulted in increased early detection. This was a compound effect due to more women being screened and higher quality of readings due to a higher volume per radiologist.

An increase in the number of early diagnoses was reflected in a decrease in the breast cancer mortality results. Increasing the reading volume standard to 2, 500 at 65% participation level had approximately the same effect on the breast cancer death rate as increasing the participation to 69% (a 2-3 % decrease in the number of breast cancer deaths). In order to understand the relative costs and benefits of increasing outreach versus increasing quality, we compared these two specific options. Table 4 summarizes the parameters for the numerical experiments, and Table 5 reports the results (intervals represent 90% confidence intervals). With both options, an equivalent improvement in health outcomes was achieved.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 4: Current Situation and Two Improvement Options

Option	Description
0	Current situation: participation=65%, $std = 480$
1	Increase participation: participation=69% $std = 480$
2	Increase minimum accreditation standards: participation=65% $std = 2500$

Table 5: Comparison of Health Outcomes for Current Situation and Improvement Options from Table 4

Option	# Breast Cancer Deaths	# Early Diagnoses	# Late Diagnoses	# Screening Mammograms	# Diagnostic Mammograms	# False Treatments
0	17.0 ± 0.3	25.9 ± 0.25	54.9 ± 0.59	$16,182 \pm 10$	$2,765 \pm 9.84$	133 ± 0.44
1	16.6 ± 0.3	27.4 ± 0.25	53.3 ± 0.59	$17,132 \pm 10$	$2,909 \pm 9.82$	141 ± 0.44
2	16.6 ± 0.16	27.8 ± 0.26	53.5 ± 0.19	$16,176 \pm 12$	$2,107 \pm 5$	101 ± 0.75

Cost of Screening and Treatment. Table 6 summarizes the costs of each program, combining the screening and diagnostic tests and treatment costs from Section 3.2.3 and the outcomes in Table 5, assuming the cost of treating after a false positive is the same as the cost of treatment after early diagnosis. For example, the estimated total annual cost of screening and treatment for option 1 included the costs of screening mammograms, diagnostic follow-up tests, and treatment for early and late stage cancers, and false positive diagnosis: $17,132 \times \$145 + 2,909 \times \$471 + 27.4 \times \$54,013 + 53.3 \times \$68,551 + 133 \times \$54,013 = \$16.0 \times 10^6$.

Increasing the reading volume standards (Option 2) resulted in costs $2,600,000 less than option 1, while providing equivalent health outcome benefits because of two effects. An improvement in specificity decreases the unnecessary diagnostic procedures. An improvement in sensitivity increased the chances of detecting an actual tumor. These quality improvements are desirable for both costs and health outcomes. On the other hand, increasing outreach while

Table 6: Cost Summary (US$) for Scenarios in Table 4

Option	Estimated Total Annual Cost of Screening and Treatment
0	$16.0 \times 10^6 \pm 0.1 \times 10^6$
1	$16.6 \times 10^6 \pm 0.1 \times 10^6$
2	$14.0 \times 10^6 \pm 0.14 \times 10^6$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

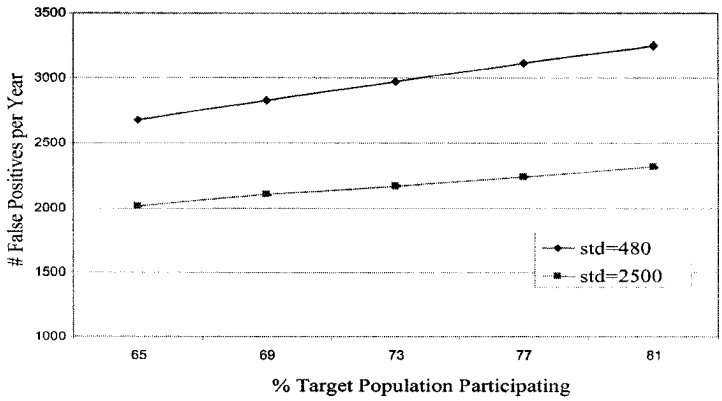

Figure 8: Average Number of False Positive Test Results per Year

keeping the standards the same (Option 1) increased costs by increasing the total screening costs and the number of unnecessary diagnostic mammograms in order to achieve comparable health benefits (Figure 8 compares the number of false positive test results). The combined impact of these options 1 and 2 would be a further reduction in breast cancer deaths (4.7%), with an estimated total cost of $\$14.3 \times 10^6$. The marginal cost increase due to increased outreach is therefore lower when quality standards are higher ($\$16.6\text{-}16.0 = \$0.6 > \$14.3\text{-}\$14.0 = \$0.3$) because there are fewer unneeded diagnostic tests.

The costs of achieving these improvements are *not* included in calculations, since they depend on the specific health care context and require capacity investment. Increasing participation to $69 - 70\%$ may be expensive. Small improvements on both participation and quality will be preferable to increasing one or the other if improvement costs are convex.

This result is not a call for not increasing the outreach of screening programs, but a warning for the costs of low quality screening. Increasing outreach provided substantial health outcome benefits and is desirable in order to provide an egalitarian public health service. If the quality of the screening test were low, by expanding outreach there would be excess waste

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

in the system and the costs would increase unproportionally with the health outcome benefits. Further, the benefits from screening more women were not fully realized when the standard (that is, the quality) was low, since a high percentage of the early stage cancers were missed among the ones screened.

Capacity Requirements. The simulation results did not indicate a problem with insufficient capacity and waiting up to a participation rate of 81%. Waiting times were not significant and did not affect health outcomes. Comparing the capacity requirements of the two options reinforced the benefits of increasing reading volume standards (Option 2). The higher quality due to the higher standards level reduced the load on the system that resulted from diagnostics required to resolve false positive results (Figure 8). Consequently, when the quality was low, the utilization level was always higher due to that indirect effect on the system load, so increased capacity requirements are a more serious problem with lower quality readings. The degree to which increased waits may negatively affect health outcomes is explored in Section 3.3.2. The greater the resource needed for diagnostic tests (larger a), the greater the capacity constraint effect caused by false positives. Decoupling screening from diagnostic mammogram capacity would reduce that effect.

The above comparisons are based on the costs of screening, diagnosis and treatment. They do not include patient-related costs like anxieties associated with false positives, or the effects of false positive results and long waiting lists on the willingness of women to request screening. If those additional factors were considered, the advantage of the option of improving quality over the option of increasing outreach would be even more significant.

These observations require some caveats. We focused on the effect of increasing the standards for reading volume on quality in our experiments. There can be some other conse-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 7: Parameters for the Numerical Experiments in Section 3.3.2

Parameter	Value
Screening interval $(1/f)$	2 years
Number of trained radiologists (n)	4
Max #readings/year/radiologist (μ_S)	2, 500
Initial enrollment probability (h)	(0.35, 0.4, 0.45, 0.5, 0.55, 0.65, 0.75, 0.85, 0.95)
Resource need for diagnostic test (a)	1.5
Target population size	25, 000

quences of increasing the standards. Fewer doctors may be willing to dedicate a significant proportion of their time to mammogram reading with higher demand levels. As the number of eligible doctors decreases, participation may decrease too since the transportation times will increase. Section 3.3.3 explicitly accounts for the participation and distance effect in a separate experiment. Finally, increasing outreach improves the chance of early detection to a broader cross-section of women, and may influence program design decisions on ethical grounds.

3.3.2 Limited Capacity, Waits, and Delayed Detection

Capacity crises may occur if demand increases and/or capacity decreases. While this may not be the case globally, waits occur in some areas [108], and many countries plan to increase participation. In the UK, women aged between 50-64 are screened, but work is being carried out to extend invitations to women up to age 70 by 2004 [107]. France intends to improve breast cancer screening participation to 80% of the target population by 2007 [62]. While these extension plans are implemented, capacity implications should be considered carefully, since capacity may be slower to influence because of extensive training required.

We ran simulations with the input parameter values in Table 7 to explore the relationship between capacity, utilization, waiting, and health outcomes. The recommended screening interval differs from country to country. Here we set it to 2 years.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

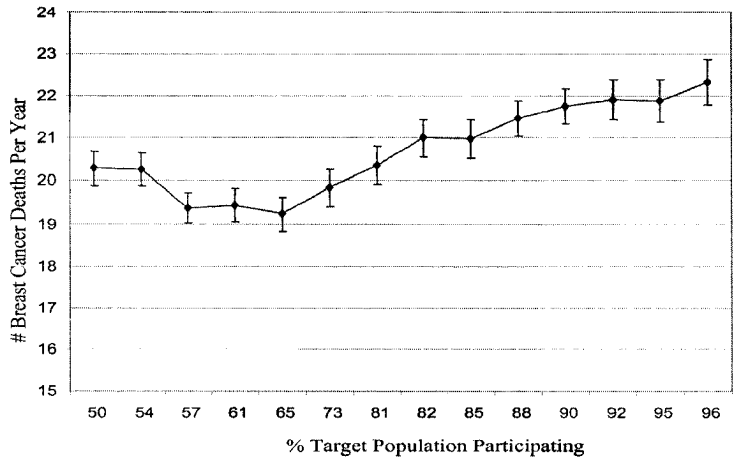

Figure 9: Breast Cancer Deaths First Decrease, then Increase with Increasing Participation Rates when Waits Become Significant

Figure 9 shows how insufficient capacity counteracts the benefits expected from increasing participation. This happened when the additional demand was not met and long waiting lines were observed. Figure 9 shows that there was a decrease in the number of cancer deaths as participation increased to 65% (corresponding to a utilization of 99%). Additional demand increased congestion and women had to wait to get regular screening mammograms. The output is given in Table 8. In this experiment, when participation is 96%, the average waiting time is 8.5 months (average waits by Little's Law are $6873/9670 = 0.71$ year). Capacity constraints or other causes for several months delay beyond a two-year screening interval can lead to poorer health outcomes due to fewer early detections. These deleterious health effects can be mitigated by improving quality for two reasons. First, each test is more accurate, improving detection. Second, a reduced burden due to less frequent follow-up diagnostic tests can free up capacity to further reduce waiting.

Additional runs with no service burden due to diagnostic tests ($a = 0.001$) indicated qualitatively the same result, with a small twist. A decrease in mammogram resource require-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 8: Simulation Results for Limited Capacity Scenario

h	% Participating	% Screened	Total Demand	# Waiting	Util.	Breast Cancer Deaths
0.35	50	45 ± 0.8	7222 ± 20	2 ± 0.0	0.77	20.3 ± 0.4
0.4	54	53 ± 0.8	7446 ± 18	3 ± 0.1	0.83	20.3 ± 0.4
0.45	57	57 ± 0.8	8263 ± 20	6 ± 0.1	0.88	19.4 ± 0.4
0.5	61	61 ± 0.8	8773 ± 15	13 ± 0.4	0.94	19.4 ± 0.4
0.55	65	64 ± 0.8	9280 ± 10	76 ± 6.7	0.99	19.2 ± 0.4
0.65	73	65 ± 0.7	9450 ± 10	1629 ± 20.6	1.00	19.8 ± 0.4
0.75	81	66 ± 0.9	9530 ± 11	3385 ± 21.5	1.00	20.4 ± 0.5
0.85	88	66 ± 0.8	9600 ± 78	5128 ± 23.5	1.00	21.5 ± 0.4
0.95	96	66 ± 0.9	9670 ± 98	6873 ± 16.6	1.00	22.3 ± 0.5

ments for diagnostic tests increases the optimal participation rate. The minimum breast cancer death rate was obtained at 81% utilization.

Our results suggest that waiting would affect health outcomes only when there is a severe capacity problem. Although screening mammograms are planned and scheduled, there is a tendency to stretch out the screening intervals, a phenomenon called "slippage" ([44] reported in [1]). Waiting cannot be avoided completely even when there is organized screening and scheduling in place. It is therefore important to consider the stochastic aspects of the demand for screening and the fact that schedules may not be implemented as intended. When capacity is insufficient, the problem will be aggravated and will have adverse effects on health outcomes.

3.3.3 Decentralization Decision / Learning From Peers

This section models one factor that influences participation: the use of decentralized facilities to reduce the distance traveled to the nearest facility [41], operationalized by mobile clinics or putting equipment in the facilities of more primary care providers, and modeled by the distance/access relationship in Section 3.2.2. Decentralization may have a positive effect in that improved participation offers more chances of early detection, and increases volume

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

and quality. On the other hand, more facilities implies lower volume per facility. If quality is improved in centralized facilities because outlier results can be shared with peers, this effectively improves the volume that each colleague sees. Reading quality in a centralized facility will be somewhere in between the quality that corresponds to the volume seen working alone, and the total volume of a centralized facility. As a result, decentralization may have mixed effects on reading quality while increasing the participation rates. The net effect is unknown [109].

We consider four cases with respect to two factors: (1) the effect of decentralization on quality (with learning from peers in a centralized facility, or without learning) and (2) capacity (sufficient capacity exists or not). If there is learning with centralization, we assume the best possible case, that the quality of each individual radiologist is based on the total volume of readings at the facility. Without learning, quality is modeled as before, as a function of the individual reading volume.

We assume that 60,000 women are evenly distributed over 100km, and go to the nearest of $c = 1, 2, 4$ or 8 facilities, assumed to be evenly distributed. The facilities house a total of 8 radiologists, each of whom serves at a rate of $\mu_S = 5000$/year. To model the sufficient and insufficient capacity cases, we set the maximum enrollment probability (with no travel) to $h_0 = 0.35$ and $h_0 = 0.75$ respectively. Participation rates ranged from $44-49\%$ for $h_0 = 0.35$ (sufficient capacity) and from $77-80\%$ for $h_0 = 0.75$ (insufficient capacity). With learning from peers, the volume associated with a fully centralized facility ($c = 1$) corresponds here to a sensitivity and specificity of about 0.95, which represents an upper bound quality level. When there is no learning, average sensitivity ranges in 0.78-0.83 while specificity is about 0.89.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Learning With Centralization. When quality is attenuated because centralization is associated with better reading performance, then Figure 10 indicates that the value of decentralization depends upon whether there is sufficient capacity to meet demand or not. If the system is already under-capacitated, then the fraction of the population actually screened may decrease, in spite of the fact that more people seek screening. Decreased reading quality in a decentralized setting increased demand for follow up tests due to false positives, reducing the effective capacity for screening mammograms. On the other hand, if there was sufficient capacity, then decentralization increased the ability to screen more women. The right panel of Figure 10 indicates that the net effect on annual breast cancer deaths was more complicated. Initially, decentralization reduced cancer deaths, due to early detection for more women. The benefits of increasing participation outweighed the losses in quality and pooling efficiency. But too much decentralization decreased reading quality and missed early stage cancers. Moreover, a loss of pooling advantage further increased waiting times and decreased the chance of early detection, so breast cancer deaths started to increase again. For the fully centralized (1 facility) and fully decentralized (8 facilities) cases, the number of breast cancer deaths were at about the same level. This suggested that learning in a centralized facility (which increased the sensitivity from 0.77 to 0.94) could provide the same benefits as decentralization, which increased participation from 44% to 49%. When there was insufficient capacity, the results do not suggest that decentralization decreased breast cancer deaths in the same way.

No Learning with Centralization. If quality is not affected by decentralization, we observed less impact of decentralization on the percent population screened, because the 'false positives effect' was weaker. By increasing the number of facilities, the percent of the population screened remained constant when there was insufficient capacity. The fraction screened

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

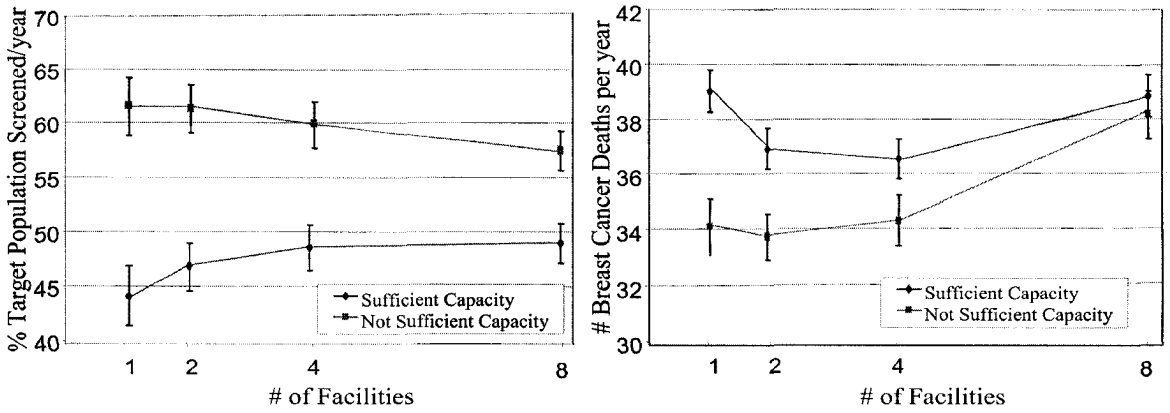

Figure 10: Learning with Centralization: Percentage of Target Population Screened (left) and Annual Breast Cancer Deaths per 60,000 (right)

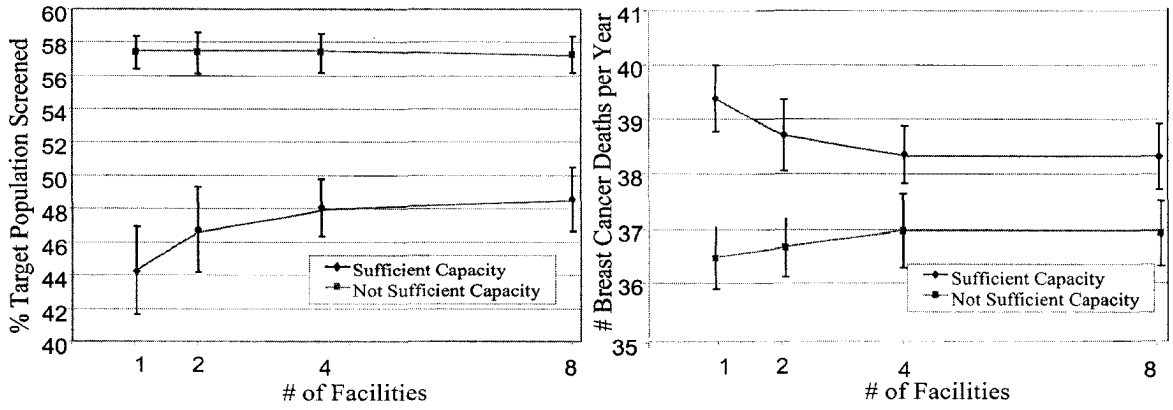

Figure 11: No Learning with Centralization: Percentage of Target Population Screened (left) and Annual Breast Cancer Deaths per 60,000 (right)

increased when there was sufficient capacity (left panel of Figure 11). The effect on annual breast cancer deaths followed a similar pattern: Since there is no loss in quality, when there is sufficient capacity the death rate decreased. Decentralization had no statistically significant effect on death rates when there was insufficient capacity because resources were already fully utilized (right panel of Figure 11).

The model suggests that fixed costs aside, decentralization is advantageous up to the point where screening quality drops significantly. If quality can be maintained in decentralized facilities, decentralization is beneficial as long as there is enough capacity to meet the increased

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

demand. If decentralization is not an option for other reasons, then instituting practices that enhance learning in centralized facilities can provide most of the reduction in cancer deaths that decentralization can provide.

3.4 Discussion

Our stochastic system dynamics model includes several factors that have not yet been considered all at once in the mammogram screening literature. Simulations here illustrated the system behavior, health outcomes and costs for some aspects of breast cancer screening programs due to public policy actions like improving enrollment rates or quality standards for radiologist certification.

A similar approach can be useful in other applications like colorectal cancer screening, where volume and quality; demand and the degree of facility decentralization; or capacity, service delays and outcome quality are interrelated. Colonoscopy is widely viewed as the most accurate screening test for colon cancer, and demand for colonoscopy has surged so much in recent years that patients may wait for months or be turned away [70]. Service design issues for colonoscopy also include the use of multiple screening policies with different costs, sensitivities and specificities.

The experiments here highlight the importance of the sensitivity and specificity dimensions of screening quality. Low quality results in additional follow-up tests that waste system capacity. The U.S., France and other countries have plans to increase adherence to regular screening by decentralization or other means, and many regions are experiencing a decline in service capacity. Any increase in participation should be accompanied both by an assurance that sufficient capacity will be established, and a maintenance or increase in screening quality

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

to insure that delays due to system dynamics do not decrease or reverse the anticipated public health benefit. Low reading volume standards reduce the quality of readings and increase screening costs by increasing the workload due to follow-up tests. Health outcomes could deteriorate because of a decreased effectiveness of screening and potential delays that might result in a late diagnosis. Decentralization of screening service to increase participation in screening is found beneficial only if the quality of screening tests can be maintained. These interactions between volume, quality, capacity and waiting influence health outcomes and system costs in ways that have not fully been accounted for in previous studies.

These aggregate conclusions should be understood relative to the limitations of the model. The homogeneous population assumption ignores risk factors involving age, genetic disposition, and environmental effects. Scheduling can ideally reduce waiting times but cannot prevent waiting completely because of the compliance issues discussed in Section 3.3.2, so we did not consider it here. Since health effects are primarily deleteriously affected by waiting times when capacity is insufficient, the value of scheduling would appear to be a second-order effect. The three-stage health model does not focus on tumor growth dynamics and patient-to-patient variability, but is consistent with a number of other papers. Quality was assumed to be a function of screening volume alone here. Adjustments can be made to handle other features [40, 101], like age and variability in reading quality between doctors, other skill factors, film quality, and controllable trade-offs between specificity and sensitivity in reading assessments, but we did not do so here.

Screening and treatment costs are included, but not the cost of the improvement options. Those must be added based on the specific health care environment to obtain a full cost-benefit analysis and to better inform controversy over the real value of breast cancer screening. Our

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

aggregate level model did not focus on the incentives of each actor in the health care system. The incentives of patients, providers and payers also play a role in determining service capacity and participation rate figures. Poor insurance coverage decreases the willingness of women to participate. Low reimbursement rates and high certification standards may decrease the willingness of the radiologists to provide service, in favor of other more profitable tasks. These issues could be explored with suitable data.

3.5 Appendix to Chapter 3: Parameter Estimates, Model Validation and Transition Rates

3.5.1 Parameter Estimates and Model Validation

Table 1 summarizes the default values for parameters. They were estimated from medical journal articles or national statistical publications wherever possible to improve model validity. Where that was not possible, we made reasonable assumptions (b_3, b_{72}, b_{73}) or fit parameters (g, r_1, p_1, r_2, p_2, γ, m_1, δ, a) so that the simulation output was of the same magnitude as corresponding country statistics taken from Canadian Cancer Surveillance On-Line [2] (Table 9). [96] reports delay data for the time to apply for diagnosis after developing symptoms. We used the aggregate data to obtain an average delay of 2.9 months, or $m_2 = 4.13$/year. The probability of developing cancer per year is 2.5% for ages $50 - 59$, 3.1% for ages $60 - 69$, and 3.3% for ages $70 - 79$ (NCI of Canada [84] for 1998). We averaged the instantaneous rates of developing cancer for these age ranges to get $s_{12} := 0.0030122$. The 5 year survival probabilities for different cancer stages are taken from ACS data [3] in Table 10, so $b_{72} := 0.0081$ and $b_3 = b_{73} = 0.0915$ are weighted averages from the regional and distant categories.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

A wide range of estimates for p_2 has been proposed, and precise estimates of r_1, p_1, r_2 are not yet available. We therefore set the default values of these parameters to match flows. Death rate estimations are done using 5-year mortality figures, so we set $r_1 = 1/5$ and $p_1 = 0.05$ to get a recurrence rate of $r_1 p_1 = 0.01$ from early stage treatment. We assumed the recurrence rate tripled after treatment for late stage cancer, with $r_1 = 1/5$, $p_2 = 0.15$, so that $r_2 p_2 = 0.03$. That same product $r_2 p_2$ is obtained if $r_2 = 0.05$, $p_2 = 0.6$. Runs with the latter values would place a slightly higher screening load on the mammogram facility and a slight increase in waiting times, but wait times were not a significant deleterious factor in Section 3.3.1, so the results would differ litte with those parameter values. The rate r_0 of treatment completion after a false positive diagnostic was set very high to model the continuing potential for the onset of preclinical cancer.

We assumed that the probability of joining the program later, and of asking for a diagnosis out of the screening schedule at an early stage of cancer, are small, so we set $\gamma = m_1 = 0.01$. Diagnostic follow up tests may include one or more of the diagnostic mammogram, fine needle aspirations, sonography, or biopsy [97]. Sensitivity estimates are 0.858 for diagnostic mammogram [12], 0.95 for fine needle aspiration [91], and 0.97 for biopsy [110]. As an average, sensitivity and diagnostic follow-up tests were set to 0.90 for preclinical cancer. In some rare cases, diagnostic tests may miss cancers, so we set the sensitivity of diagnostic test to 0.99 for clinical cancer. Specificity of diagnostic test is set to 0.95. The probability of having cancer at the entry to the target population is estimated using the data from Canadian Organized Breast Cancer Screening Program: cancer detection rate at first screen is 4.4/1000. With a sensitivity of 0.80 we get $p = 0.0055$.

Discrete changes for health status and position in the service system are essentially

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 9: Comparison of Country Statistics with Simulation Results (per 100,000)

Source of Estimate	Incidence	Breast Cancer Deaths
[2] (for 1998)	302.18	70.39
Model Estimate	324	67.8
% Error	7.2%	3.6%

Table 10: American Cancer Society [3] Survival Data

Stage	Pct. at Diagnosis	5-Year Survival Rate	Death Rate
local	65%	96%	0.0081
regional	30%	76%	0.0548
distant	5%	21%	0.312

Markovian in continuous time, conditional on the reading quality, which may also vary through time. The quality-volume relationship is modeled as in Section 3.2.2. Quality, as measured by sensitivity and specificity, influence the type of transition when a screening mammogram is performed. Overall, there are 10 reasons for state changes and each occurs with the instantaneous transition rates given in Table 11, where X_{ij} represents the size of the compartment (i, j). The rates are sums, each summand representing flows out of individual compartments.

Table 11: Event Rates

Event			Rate
Ask for screening mammogram	ϕ_1	=	$f\sum_{j=1}^{3} X_{2j} + \gamma\sum_{j=1}^{3} X_{1j}$
Ask for diagnostic mammogram	ϕ_2	=	$m_1\sum_{i=1}^{2} X_{i2} + m_2\sum_{i=1}^{3} X_{i3}$
Screening mammogram completion	ϕ_3	=	$\mu_S\sum_{j=1}^{3} X_{4j}$
Diagnostic mammogram completion	ϕ_4	=	$\mu_D\sum_{j=1}^{3} X_{6j}$
Develop preclinical breast cancer	ϕ_5	=	$s_{12}\sum_{i=1}^{6} X_{i1}$
Progress from preclinical to clinical stage	ϕ_6	=	$s_{23}\sum_{i=1}^{6} X_{i2}$
Treatment completion with preclinical stage	ϕ_7	=	$r_1 X_{72}$
Treatment completion with clinical stage	ϕ_8	=	$r_2 X_{73}$
Cancer death	ϕ_9	=	$b_3\sum_{i=1}^{6} X_{i3} + b_{72}X_{72} + b_{73}X_{73}$
Other death	ϕ_{10}	=	$g\sum_{i=1}^{7}\sum_{j=1}^{3} X_{ij}$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

4 A Model for Disease Screening: Quality Decision Under Competition

4.1 Introduction

This chapter looks at disease screening and treatment services in different industrial organization settings. The main question of interest is a comparison of the decision about quality levels in a setting where a social planner provides these services versus a setting where private health care service providers compete given reimbursements from a social planner. The motivation comes from breast cancer screening services, which was discussed in Chapter 3, and recent questions raised about the quality of screenings and the complaints of private providers about reimbursement levels [82]. We consider a health care system in a public-contract model [35] for the competitive setting, i.e. the social planner (public policy maker) pays private providers for the health service, as in the case of Medicare and Medicaid programs in the US. In this chapter we do not consider a separate payer like an insurance company.

We focus on screening test quality in this model. More specifically, the quality of the test is represented by a single variable, α, sensitivity of the screening test. There is evidence that many factors affect sensitivity (machine quality, radiologist training and experience, number of readings per year etc.) [14], and a provider can increase sensitivity by improving these factors.

We account for the following elements of the problem in this model, which we will describe in more detail later in section 4.3:

- The decision variable is quality level α, a measure of sensitivity.
- People have a preference to get the test from the provider with the higher quality level

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

and quality levels can be observed.

- People may have a disutility for getting the test, including social distrust for program, difficulty with access, other personal factors or distance to facility. We aggregate all these factors under the general term travel distance as also done by [38].

- Detecting the disease early results in lower treatment costs compared to late detection.

We present three models, starting from an overly-simplified model and changing one assumption at a time in order to capture more of the above mentioned points. Section 4.2 briefly discusses the related work in the health economics literature. Model assumptions and research questions are stated in Sections 4.3 and 4.4. Sections 4.5, 4.6 and 4.7 present and analyze the three models. Finally in Section 4.8 we conclude.

4.2 Literature Review

There are several articles in the health economics literature that model the health services in a competitive environment and investigate reimbursement mechanisms to improve social welfare and control decisions taken by profit maximizing providers. [38] analyzes three provider strategies, with respect to how they treat patients with different severities of illness: Creaming (over-provision of services to low severity patients), skimming (under-provision of services to high severity patients), and dumping (the explicit avoidance of high severity patients). The social optimum is compared with the private Cournot-Nash solution under different reimbursement schemes. They show that cost based reimbursement results in over-provision of services. In the operations literature, [34] models how system congestion and customer valuations may influence over-provision of services, where health care services is an example. [20] models

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the provision of two services by two providers, with customers that value the screening service of each provider differently. They investigate market structures where the two services are bundled or separated. We model two services as well, screening and treatment, and examine cases where customers value the screening service differently. In our model, the second service is not demanded by all the consumers, in contrast with the model in [20]. [15] studies a model similar in spirit to ours, a hospital's choice of quality and competition being the central issue. However, that paper focuses more on the effects of asymmetric information about costs and models the hospital's choice of quality. We do not have asymmetric information in our model, and verifiability of quality is assumed not to be an issue for the reimbursement scheme decision. In our model, reimbursements are not based on quality itself, but the outcomes of it, observed in terms of the early detection of disease. [77] models effects of cost reimbursement and prospective payment systems in the health care industry, on the cost and the incentives for quality , and shows that the latter can be used to give efficient quality and cost reduction efforts.

One common assumption in the above studies is that the type of service in question is a treatment. This implies the service is desirable in itself and a high quality of that service has direct influence on the patient welfare, hence also on the social welfare. The difference in our model comes from the fact that there are two services explicitly modeled; screening and treatment, and the first service is good only if it leads to a more appropriate delivery of the second. Provision of screening-type public health services by private providers has not been studied before to our knowledge.

There is a vast literature on disease screening in medical, operations and public health literatures. A review focusing on research on breast cancer screening is given in Chapter 3. Disease screening has been studied mostly from a public health or policy control and

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

optimization points of view, while the effects of leaving this public health service to profit maximizing providers was not yet questioned.

4.3 Assumptions

We examine three models, presented in the order of increasing complexity. The basic assumptions are common for all the three models, except when mentioned otherwise.

- A target population of size N is uniformly distributed in a linear city of length 1. There are two providers at the edges of the city. This is a very common assumption in health economics literature [38, 20, 15]. The transportation cost is t per unit distance. With the transportation cost, we model the cost of access in general, which represents patient disutility from getting the screening service. But throughout this chapter we will use the distance interpretation. In the first model, we assume $t = 0$, in the second and third, we assume $t > 0$.

- Patient willingness to get screened is modeled by a utility for screening determined by their valuation of the screening test, which is increasing in the quality of the test. Utility for a screening test of quality α ($0.5 \leq \alpha \leq 1$), at distance x from the facility is modeled as:

$$u(\alpha, x) = v\alpha - tx. \tag{34}$$

  Patients get screened if $u(\alpha, x) \geq 0$ and choose the facility with highest $u(.)$ if more than one facility is viable.

- Patients can be in one of three stages: healthy with probability p_0, have the disease at early stage with probability p_1, and have the disease at late stage with probability p_2.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

- The cost of a screening test is c_0 per test.
- Two services are offered by each provider: screening and treatment. The provider which detects the disease does the treatment as well, i.e. patients do not switch providers for treatment.
- All people with late stage disease will be treated even if they were not screened. This assumption is consistent with the assumption of a fully insured population, and it helps to write the benefits of early detection as a difference of the costs early and late stage treatment.
- The only decision variable is screening test quality, as represented by sensitivity, α. The cost of quality is a convex increasing function of α, $C(\alpha)$. Quality can be improved with more training or hiring more experienced radiologists, using better technology, or having more than one radiologist check the test results. We do not explicitly model how the quality is improved here. We model the cost of quality as a cost scaled with the population size. Since it is fixed with respect to the demand observed, we refer to it as "fixed cost". Fixed quality cost is a common assumption in economics literature [106]. This is motivated from the assumption of facilities having ample capacity. We assume that the capacity of both providers is enough to serve to the whole population, N. The investment in quality will be a fixed cost for that supply, no matter how many people are served. If there were no effort on quality, a sensitivity of $\alpha = 0.5$ would be achieved at zero cost (we assume specificity of 1 (no false positives, due to presence of highly specific follow-up tests). We set

$$C_i(\alpha) = Nc_i(\alpha^2 - \frac{1}{4})$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

for facility $i = 1, 2$, and restrict α to be at least 0.5. The feasible set for quality is $\alpha \in [0.5, 1]$.

- We assume a prospective payment system, i.e. the reimbursement is not based on the actual costs. Instead, a reimbursement rate is identified for each service provided.

- In the analysis the decision of a social planner and the decision of profit maximizing providers are investigated. We make the following assumptions for each case:

  - **Social planner**: The social planner owns both facilities. The objective is to maximize social welfare, i.e. benefits from treating the disease at early stage (this is from reduced treatment cost as well as the gains in life years), minus the costs of screening and quality. The benefits and costs of screening and treatment services for the social planner are as follows:

    * b_1 : Benefits of early treatment for the social planner (per treatment). This includes treatment cost reduction and the increase in life expectancy by early detection. In other words,

    $$b_1 = \gamma_2 - \gamma_1 + \Delta l \tag{35}$$

    where γ_1 and γ_2 are the cost of early and late stage treatment and Δl is the value of life years gained by early detection.

    * c_0 : Cost of screening test (per test).

  - **Private providers case**: We assume duopoly competition on quality; two providers want to maximize profits. A social planner reimburses the providers for each screening and treatment, which determines the cost benefit structure for the private

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

providers. Given the reimbursement levels, the net profit per each service provider is determined by:

* r_0 : Net revenue from screening, per screening = Reimbursement + Co-payment per screening test - cost per screen. I.e.

$$r_0 = r_s - c_0 \tag{36}$$

where r_s is the reimbursement per screen (co-payment is normalized to zero).

* r_1 : Net revenue of early detection = Revenue increase for every correctly diagnosed and treated early stage disease + treatment cost reduction by early detection. I.e.

$$r_1 = r_{t1} - r_{t2} + \gamma_2 - \gamma_1 \tag{37}$$

where r_{t1} and r_{t2} are the reimbursements per early and late stage treatment respectively (co-payment is normalized to zero).

- In practice, net reimbursement levels r_0 and r_1 need not be the same as b_1 and c_0. b_1, the benefit of early detection for the social planner, includes the long term benefits from increased life expectancy, Δl, which is typically not directly reflected in the reimbursement levels.

- All parameters are common knowledge.

- Quality levels are observed. This assumption may seem strong at first sight since it would be difficult to observe the exact quality levels in reality. On the other hand, there is a regular certification process for mammogram facilities that signals the quality level. In general health care setting, a good example of how quality levels can be reported

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

is given in the Shouldice Hospital Case [58]. For screening services, the experience of radiologists and machine technology are easily observable factors that effect quality. Hence, we can argue that it is a reasonable first order approximation.

- Social welfare is defined as the sum of total profits for the two providers and the total utility for the patients (from increased life expectancy) minus the costs for the social planner. The utility of patients for screening defined in (34) is not included in the social welfare calculation. For social welfare calculation purposes, patient utility is only the increase in life time expectancy, and is valued as Δl per early detection. Thus the social welfare for both the social planner case and private provider are equal to the following, as shown in the Appendix, where D_1 and D_2 are the demands for facilities 1 and 2 :

$$\Pi^{SP} = D_1(b_1 p_1 \alpha_1 - c_0) + D_2(b_1 p_1 \alpha_2 - c_0) - C_1(\alpha_1)) - C_2(\alpha_2) \qquad (38)$$

When there are two private providers, each maximizes its own profits. The profit for provider i is:

$$\Pi_i = D_i(\alpha_1, \alpha_2)(p_1 r_1 \alpha_i + r_0) - C_i(\alpha_i) \qquad (39)$$

where $D_i(\alpha_1, \alpha_2)$ is the demand for provider i, given quality levels of both providers.

4.4 Research Questions

The basic questions we are interested in are: What is the effect of competition on the social welfare and the quality levels set for a screening service? Can we get profit-maximizing health care providers to offer a quality level that is socially optimal, i.e. that maximizes social welfare, or is it better to have the social planner provide this public health service? How do disease and

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

population characteristics (in terms of the disease prevalence and willingness to get screening test) effect the optimal quality levels in different health systems?

Section 4.5 analyzes the basic model where there is no access cost and everyone in the population has the same valuation for screening test quality. Section 4.6 adds a transportation cost, which creates differentiation between the two providers. Section 4.7 presents a preliminary model for the case of heterogeneous valuations for quality in the target population.

4.5 Case I: No transportation cost

In this section, we analyze the optimal quality selection of a social planner (maximizing social welfare), and the decision of two competing private providers (maximizing profits), for given reimbursement rates, for the model without transportation costs ($t = 0$).

Utility of patients: The utility of a patient at a distance x_i from provider i for getting screening from provider i is $u(\alpha_i) = v\alpha_i$ and α_i is the quality level of provider i. Since there is no disutility of screening when $t = 0$, all customers are willing to get screened.

Demand: The provider with higher quality would attract all the patients since there is no access cost to differentiate the providers. We assume that when the quality levels are the same, patients choose one of the two providers randomly. Demand for provider 1 given the quality levels for both providers, α_1 and α_2 is:

$$D_1(\alpha_1, \alpha_2) = \begin{cases} N & if \quad \alpha_1 > \alpha_2 \\ \frac{N}{2} & if \quad \alpha_1 = \alpha_2 \\ 0 & if \quad \alpha_1 < \alpha_2 \end{cases} \tag{40}$$

The demand for provider 2 is $D_2(\alpha_1, \alpha_2) = N - D_1(\alpha_1, \alpha_2)$. This is a typical winner-takes all demand allocation.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

4.5.1 Case I: Social Planner's Decision

For the social planner, the objective is to maximize social welfare, i.e. the benefits from early detection minus costs of screening and quality investment for both facilities. Note that the benefits of the social planner per early detection, b_1 also include the utility of patients in terms of life expectancy.

The Socially Optimal quality levels α_1, α_2 maximize Π^{SP} in equation 38. If the quality for both providers were the same, then the demand would be shared equally. Otherwise, the facility with the highest quality would attract all the demand. We analyze these two cases separately to find the optimum:

i) If $\alpha_1 = \alpha_2 = \alpha$ then:

$$\max_{\alpha} \left(\Pi^{SP} = N(p_1 b_1 \alpha - c_0) - Nc_1(\alpha^2 - \frac{1}{4}) - Nc_2(\alpha^2 - \frac{1}{4}) \right)$$

$$\text{First order condition } (f.o.c.): \quad \frac{d\Pi^{SP}}{d\alpha} = b_1 p_1 - 2c_1\alpha + 2c_2\alpha = 0$$

$$\Rightarrow \alpha^{SP} = \max(0.5, \min(\frac{b_1 p_1}{2c_1 + 2c_2}, 1))$$

ii) $\alpha_1 > \alpha_2 \Rightarrow D_1(\alpha_1, \alpha_2) = N$, $D_2(\alpha_1, \alpha_2) = 0 \Rightarrow \alpha_2 = 0.5$; then

$$\max_{\alpha_1} \left(\Pi^{SP} = N(p_1 \alpha_1 b_1 - c_0) - Nc_1(\alpha_1^2 - \frac{1}{4}) \right)$$

$$f.o.c. \quad : \quad \frac{d\Pi^{SP}}{d\alpha} = p_1 b_1 - 2c_1\alpha_1 = 0$$

$$\Rightarrow \quad \alpha^{SP} = \left(\max(0.5, \min(\frac{b_1 p_1}{2c_1}, 1)), 0.5 \right)$$

Proposition 9 *When there is no transportation cost, the social optimum is to make all of the investment in quality in one facility, i.e.* $\alpha_1^{SP} = \max(0.5, \min(\frac{b_1 p_1}{2c_1}, 1))$ *at the optimum. Operating two facilities is inefficient.*

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Proof. If the cost of a given quality level is fixed regardless of demand, it is suboptimal to invest in two facilities. This is because the net benefits from screening would be the same, since the whole population is served in either case, whereas the total quality cost would be higher if there is investment in both facilities. ■

Remark 10 *The optimal quality level increases as the prevalence of the disease* (p_1) *increases.*

4.5.2 Case I: Profit Maximizing Providers' Decision

The demand function is the same as in a typical Bertrand type competition, i.e. a provider can get the whole market by setting its quality level slightly above its competitor, so each provider will want to increase the quality above the competitor's quality level [106]. However, there is a limit to the point up to which they can increase their quality level without making losses. The optimization problem for provider i, given the competitor's quality is:

$$\max_{\alpha_i} \Pi_i(\alpha_1, \alpha_2) = D_i(\alpha_1, \alpha_2)(r_0 + p_1 r_1 \alpha_i) - C_i(\alpha_i) \tag{41}$$

s.t.

$$\alpha_i \in [0.5, 1]$$

We define two reference quality levels to facilitate the discussion of equilibrium. Let α_i^N be the profit maximizing quality level that provider i would set when its demand is N. $\alpha_1^{N,0}$ is the maximum quality level that provider i would set without making losses, when its demand is N. These quality levels are found by examining first order optimality conditions and taking into account the constraint $\alpha_i \in [0.5, 1]$:

$$\alpha_1^N : \frac{d\Pi_1}{d\alpha_1} = 0 \qquad \text{for } D_1 = N,\ \alpha_1^N = \max(\frac{1}{2}, \min(\frac{p_1 r_1}{2c_1}, 1)) \tag{42}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\alpha_1^{N,0} \quad : \quad N(r_0 + p_1\alpha_1 r_1) - Nc_1(\alpha_1^2 - \frac{1}{4}) = 0$$

$$\alpha_1^{N,0} = \begin{cases} \max\left(0.5, \min\left(\frac{1}{4c_1}\left(2p_1r_1 + 2\sqrt{(p_1^2r_1^2 + 4cr_0 + c_1^2)}\right), 1\right)\right) \\ \max\left(0.5, \min\left(\frac{1}{4c_1}\left(2p_1r_1 - 2\sqrt{(p_1^2r_1^2 + 4cr_0 + c_1^2)}\right), 1\right)\right) \end{cases} \tag{43}$$

The same arguments apply for provider 2:

$$\alpha_2^N = \max\left(0.5, \min(\frac{p_1r_1}{2c_2}, 1)\right)$$

$$\alpha_2^{N,0} = \begin{cases} \alpha_{21}^{N,0} = \max\left(0.5, \min\left(\frac{1}{4c_2}\left(2p_1r_1 + 2\sqrt{(p_1^2r_1^2 + 4cr_0 + c_2^2)}\right), 1\right)\right) \\ \alpha_{22}^{N,0} = \max\left(0.5, \min\left(\frac{1}{4c_2}\left(2p_1r_1 - 2\sqrt{(p_1^2r_1^2 + 4cr_0 + c_2^2)}\right), 1\right)\right) \end{cases},$$

There are two solutions for $\alpha_2^{N,0}$, where $\alpha_{21}^{N,0} \geq \alpha_{22}^{N,0}$. For the equilibrium discussion, only the larger point, $\alpha_{21}^{N,0}$ will be valid. We analyze the equilibrium in two cases:

i) Identical Providers: $c_1 = c_2 = c$

If $c_1 = c_2 = c$, then $\alpha_1^N = \alpha_2^N = \alpha^N$, and $\alpha_1^{N,0} = \alpha_2^{N,0} = \alpha^{N,0}$. Consider provider 1's actions given provider 2's decision α_2: If $\alpha_2 < \alpha^N$, then provider 1 could set $\alpha_1 = \alpha^N$, to serve the whole population and maximize profits. If provider 2 increases its quality such that $\alpha_2 \geq \alpha^N$, provider 1 increases its quality too in order to beat provider 2 and attract patients. As soon as α_2 reaches the maximum quality level $\alpha^{N,0}$, it is not profitable for provider 1 to continue competition and increase its quality, since if provider 1 sets $\alpha_1 = \alpha^{N,0}$, they would share the market, and both would make negative profits. Setting α_1 greater than α_2, i.e. setting $\alpha_1 > \alpha^{N,0}$ is not profitable for provider 1 either, given that $\alpha^{N,0}$ is the maximum affordable quality level. So, if provider 2 sets $\alpha_2 = \alpha^N$, provider 1 will prefer to set $\alpha_1 = 0.5$ and make zero profit. However, this is not an equilibrium, since if provider 1 sets $\alpha_1 = 0.5$, provider 2 would prefer to decrease its quality and set $\alpha_2 = \alpha^N$ (note that α^N is the profit maximizing quality and $\alpha^{N,0} > \alpha^N > 0.5$). These reactions would start again, which leads us

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

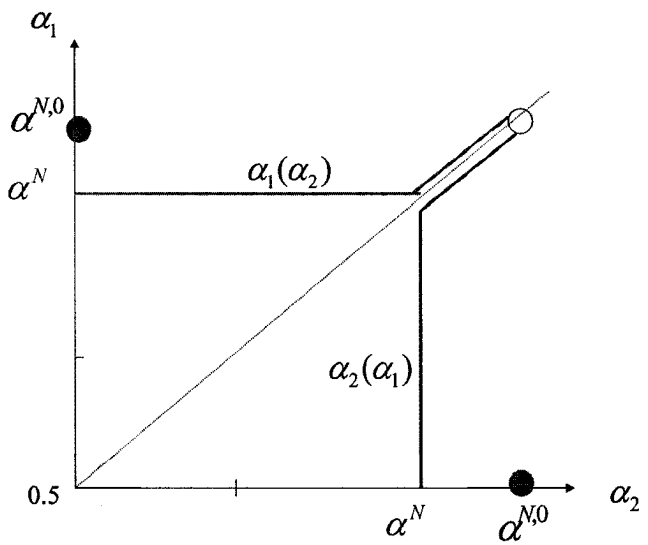

Figure 12: Reaction Functions for Case 1. Two providers are assumed to be identical ($c_1 = c_2$).

to the conclusion that an equilibrium cannot be reached when two providers are symmetric and quality costs are fixed regardless of the demand. Figure 12 illustrates the lack of equilibrium in this system, since the reaction functions never intersect.

The reaction functions explored in the above discussion are summarized below and illustrated in Figure 12, for two identical providers, i.e. $c_1 = c_2 = c$.

$$\alpha_i = \begin{cases} \max\left(0.5, \min(\frac{p_1 r_1}{2c}, 1)\right) & \text{for} \quad \alpha_j < \frac{p_1 r_1}{2c} = \alpha^N \\ \max\left(0.5, \min(\alpha_j + \varepsilon\ , 1)\right) & \text{for } \alpha_i^{N,0} - \varepsilon \geq \alpha_j \geq \frac{p_1 r_1}{2c} \\ 0.5 & \text{for } \alpha_j > \alpha_1^{N,0} - \varepsilon \end{cases}$$

Remark 11 *When the quality investment should be done for the whole population regardless of the demand (i.e. cost of quality is fixed), then there is no pure strategy equilibrium for the two firms competing on quality.*

ii) Non-identical Providers: $c_2 > c_1$

The reaction functions for non-identical providers are shown in Figure 13. The result is the same as the identical provider case, i.e. there is no pure strategy equilibrium. The analysis

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

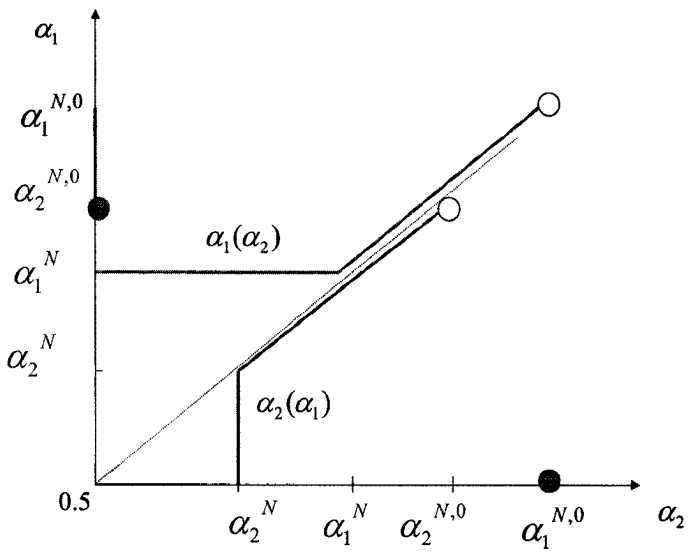

Figure 13: Reaction functions for non-identical providers ($c_2 > c_1$).

of this case shows that having two providers is not a viable policy if both have enough capacity to meet the demand from the whole market, and there is no difference between them except for their quality from the patients' perspective (i.e. no transportation cost in our model). Indeed, from the social planner's perspective, having two facilities at the same location while one has enough capacity is suboptimal (as in section 4.5).

The result suggests that two providers competing on quality cannot co-exist in a market, when the cost of quality is scaled with the capacity and not with the demand. The assumption of not having any differentiation except for quality is a strong one which drives this result. In practice, factors other than quality would create differentiation between providers, such as agreements of insurance companies or individual preferences for doctors in different facilities.

If we change the assumption that both facilities invest for the whole market and instead assume that the quality investment is a variable cost, then we can achieve an equilibrium as shown in the Appendix. In that case, we find that there is a unique equilibrium quality level, where both providers make zero profits. This is also analogous to the classical Bertrand

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

competition result. For disease screening problem, we interpret the quality investment as a fixed cost for the existing capacity (for e.g. hiring a more experienced doctor or buying a more advanced equipment).

4.6 Case II: Transportation cost $t > 0$

In this section we change the basic model discussed in Section 4.5 by assuming a positive transportation cost, $t > 0$. This reflects customers preferences between the two providers for factors other than quality. We then follow a similar analysis to compare the optimal quality levels.

The patient utility from getting screening from provider i is $u_i(\alpha_i) = v\alpha_i - tx_i$, where v and t are positive constants and x_i is the distance to provider i, $x_i \in [0, 1]$. We look at two cases: (1) Both facilities are owned by the social planner, and (2) Facilities are owned by competing private providers.

Demand. A patient at a distance x_1 to provider 1 accepts to get a screening test if and only if his/her utility from getting screening from provider 1 or provider 2 is non-negative. We assume initially that all the population is willing to get the screening test from at least one of the providers, i.e.:

$$\max(v\alpha_1 - tx_1, v\alpha_2 - t(1 - x_1)) \geq 0 \qquad \text{(Assumption 1)}$$

Remark 12 *If Assumption 1 is not true, then there will be some patients in the middle of the linear city, who will not get screening test from either of the firms, and there will be no real competition (each will behave like a monopolist for patients in their catchment area). Since we are interested in the effect of competition, we analyze only the case with assumption 1 here.*

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Since the minimum quality allowed is 0.5, Assumption 1 implies $v \geq t$.

Provider 1 is preferred to provider 2 if and only if:

$$v\alpha_1 - tx_1 \geq v\alpha_2 - t(1 - x_1) \tag{44}$$

Then the value of x_1 that leads to equality in inequality (44) determines the demands for both providers: $D_1(\alpha_1) = Nx_1^*$ and $D_2(\alpha_2) = Nx_2^* = N(1 - x_1^*)$ with:

$$x_1^* = \frac{1}{2} + \frac{v(\alpha_1 - \alpha_2)}{2t} \tag{45}$$

$$x_2^* = \frac{1}{2} + \frac{v(\alpha_2 - \alpha_1)}{2t} \tag{46}$$

Equations (45) and (46) assume that transportation cost t is sufficiently large so that $x_1^* \in [0, 1]$ is in the interior of the unit city. (Otherwise, for sufficiently small t, the provider with greater quality would serve the whole population. The limiting case of $t \to 0$ was analyzed as Case I in Section 4.5, so we do not discuss it here).

4.6.1 Case II: Social Planner's Decision

The social planner's objective is to maximize total welfare, Π^{SP}, i.e. the sum of total net benefits from services of the two facilities:

$$\begin{aligned}\Pi^{SP} = {} & N\left(\frac{1}{2} + \frac{1}{2}v\frac{\alpha_1 - \alpha_2}{t}\right)(b_1 p_1 \alpha_1 - c_0) - C_1(\alpha_1) \\ & + N(\frac{1}{2} + \frac{1}{2}v\frac{\alpha_2 - \alpha_1}{t})(b_1 p_1 \alpha_2 - c_0) - C_2(\alpha_2)\end{aligned}$$

From first order optimality conditions we find the optimal values as:

$$\alpha_1^{SP} = \max(0.5, \min(\frac{1}{2}b_1 p\frac{c_2 t - b_1 pv}{2c_1 tc_2 - b_1 pvc_2 - c_1 b_1 pv}, 1))$$

$$\alpha_2^{SP} = \max(0.5, \min(\frac{1}{2}b_1 p\frac{c_1 t - b_1 pv}{2c_1 tc_2 - b_1 pvc_2 - c_1 b_1 pv}, 1))$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

If the two facilities are identical ($c_1 = c_2 = c$), then

$$\alpha_1^{SP} = \alpha_2^{SP} = \max(0.5, \min(1, \frac{b_1 p_1}{4c}))$$

The objective function is concave if:

$$c > \frac{b_1 p_1 v}{2t} \tag{47}$$

If (47) does not hold, the solution would be at the upper-bound, 1. If we assume (47) holds, then we know that $\min(1, \frac{b_1 p_1}{4c}) = \frac{b_1 p_1}{4c}$, and there are 2 possible optima for a social planner operating two identical facilities, with symmetric costs:

$$1 \ : \ (\alpha_1^{SP}, \alpha_2^{SP}) = (0.5, 0.5)$$

$$2 \ : \ (\alpha_1^{SP}, \alpha_2^{SP}) = (\frac{b_1 p_1}{4c}, \frac{b_1 p_1}{4c})$$

The first optimum ($\alpha_1 = \alpha_1 = 0.5$) refers to a case in which benefits from early detection and the prevalence of the disease are so low compared to the cost of quality that it is not worth making any effort for screening test quality. The second optimum is the interior solution ($\alpha_1 = \alpha_2 \in (0.5, 1)$), where the optimal level of quality is determined by the ratio of marginal benefits over marginal costs of screening quality. The cost of screening does not affect the results, because everyone gets screened by assumption and the screening cost is a sunk cost for the social planner.

If the two facilities were owned by a monopolist who wants to maximize profits given reimbursements (as opposed to a social planner who wants to maximize social welfare), then the quality level chosen by the monopolist would be $\alpha^M = \alpha^M = \frac{r_1 p_1}{4c}$. This is found by using r_1 and r_0 in place of b_1 and $-c_0$ for the profit function found in (47). Note that difference between a profit maximizer and a social planner comes from the difference in their utility functions, a monopolist only considers the revenues and not life gains.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

4.6.2 Case II: Profit Maximizing Providers' Decision

Private providers each maximize their own profits which consist of net revenues from screening and treatment and the cost of providing a certain quality level for the population N.

$$\Pi_1^C = N\left(\frac{1}{2}+\frac{1}{2}v\frac{\alpha_1-\alpha_2}{t}\right)(r_0+r_1p_1\alpha_1) - Nc_1(\alpha_1^2-\frac{1}{4}) \quad (48)$$

$$\Pi_2^C = N\left(\frac{1}{2}+\frac{1}{2}v\frac{\alpha_2-\alpha_1}{t}\right)(r_0+r_1p_1\alpha_2) - Nc_2(\alpha_2^2-\frac{1}{4}) \quad (49)$$

Second order optimality conditions are satisfied if

$$c_i > \frac{r_1p_1v}{2t} \qquad \text{for} \quad i=1,2 \quad (50)$$

We assume (50) is true. Then the reaction functions of the two firms, given the decision of their competitor, are found by solving for the first order optimality conditions for maximizing profits given above for each firm:

$$\alpha_1(\alpha_2) = \max\left(0.5, \min\left(1, \frac{r_0v+r_1p_1t}{4c_1t-2r_1p_1v} - \frac{r_1p_1v}{4c_1t-2r_1p_1v}\alpha_2\right)\right) \quad (51)$$

$$\triangleq \max\left(0.5, \min\left(1, a_1-\beta_1\alpha_2\right)\right) \quad (52)$$

$$\alpha_2(\alpha_1) = \max\left(0.5, \min\left(1, \frac{r_0v+r_1p_1t}{4c_2t-2r_1p_1v} - \frac{r_1p_1v}{4c_2t-2r_1p_1v}\alpha_1\right)\right) \quad (53)$$

$$\triangleq \max\left(0.5, \min\left(1, a_2-\beta_2\alpha_1\right)\right) \quad (54)$$

and the a_i, β_i are defined by Equations (52) and (54). Figures 14 and 15 show the reaction functions for two cases (one with stable equilibrium), if the coefficients in the linear term of the reaction functions, β_1 and β_2 are different. Figures 14 and 15 assume parameter values which ensure that the equilibrium quality levels stay in the feasible range [0.5, 1],

where

$$1 \geq a_1, a_2 \geq 0.5 \quad \text{and} \quad 1 \geq \frac{a_1}{\beta_1} = \frac{a_2}{\beta_2} = \frac{r_0v+r_1p_1t}{r_1p_1v} \geq 0.5$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

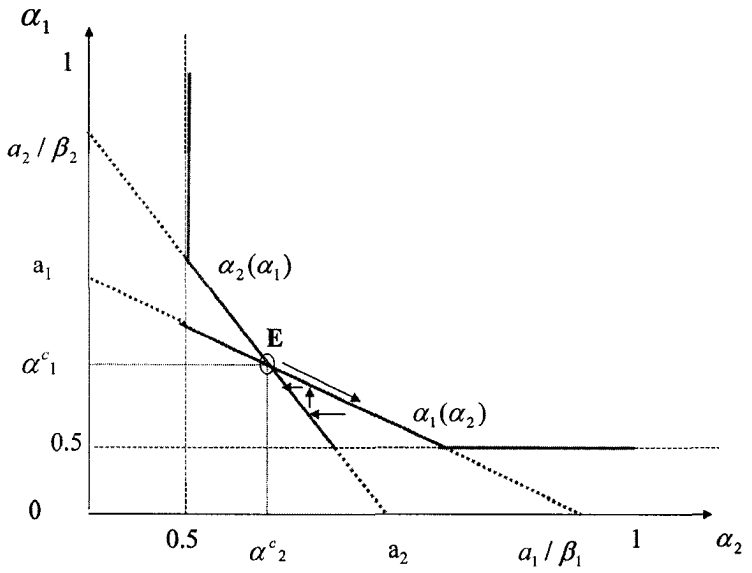

Figure 14: Reaction Functions of two firms when $\beta_1, \beta_2 < 1$, given by the solid lines. Equilibrium E is stable.

4.6.3 Stability of Equilibria

In this section, the stability of equilibria is investigated. To better interpret the conditions, we first define quality elasticity of demand:

Definition 13 *Quality elasticity of demand shows how sensitive the demand is to a change in quality and is equal to:*

$$\frac{\textit{proportionate change in demand}}{\textit{proportionate change in quality}} = \frac{\frac{1}{2} + \frac{v(\alpha_1-\alpha_2)}{2t} - (\frac{1}{2} + \frac{v(\alpha'_1-\alpha_2)}{2t})}{\alpha_1 - \alpha'_1} = \frac{v}{2t}$$

If the reaction functions intersect at an interior point, then stability is ensured by the condition $\beta_1, \beta_2 < 1$, i.e. $2c_1 > 3r_1 p_1 \frac{v}{2t}$ and $2c_2 > 3r_1 p_1 \frac{v}{2t}$, as shown by Proposition 14.

Proposition 14 *If (α^c_1, α^c_2) is an interior equilibrium point then it is asymptotically stable if and only if $2c_1 > 3r_1 p_1 \frac{v}{2t}$ and $2c_2 > 3r_1 p_1 \frac{v}{2t}$.*

Proof. The evolution of decisions $\alpha = (\alpha_1, \alpha_2)$ given by equations (52) and (54) can be represented by the following dynamic system: $\alpha(k+1) = B(k)\alpha(k) + A$ where

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

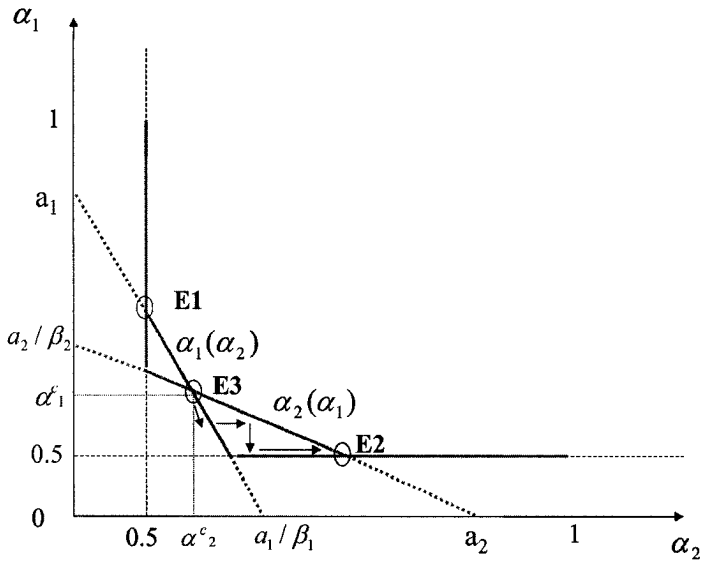

Figure 15: Reaction Functions of two firms when β_1 or $\beta_2 \geq 1.$, given by the solid lines. Equilibrium E3 is not stable.

$B(k) = \begin{bmatrix} -\beta_1 & 0 \\ 0 & -\beta_2 \end{bmatrix}$ and $A = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$, assuming the decisions evolve in the interior of $(\alpha_1, \alpha_2) \in [0.5, 1]^2$. The eigenvalues of B are equal to $-\beta_1$ and $-\beta_2$. The system is asymptotically stable if and only if both eigenvalues have magnitude less than 1 [76], i.e. if and only if $2c_i > 3r_1 p_1 \frac{v}{2t}$ for $i = 1, 2$.

$$\beta_1 = \frac{r_1 p_1 v}{4c_1 t - 2r_1 p_1 v} < 1 \Rightarrow 2c_1 > 3r_1 p_1 \frac{v}{2t} \quad (55)$$

$$\beta_2 = \frac{r_1 p_1 v}{4c_2 t - 2r_1 p_1 v} < 1 \Rightarrow 2c_1 > 3r_1 p_1 \frac{v}{2t} \quad (56)$$

■

The stability conditions given in equations (55) and (56) compare the marginal cost of quality with the marginal revenues of quality, scaled with the elasticity factor. The conditions are satisfied for a high cost, low elasticity, and low marginal revenue. A high quality cost c_i or a low elasticity $\frac{v}{2t}$ reduces incentives of the two providers to increase quality and try to serve to the whole market. In that case two providers would co-exist and share the market. On the other

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

hand, if the quality can be improved at low cost (low c_i) or if the demand is highly sensitive to changes in quality (high elasticity), then having two competing providers is not a sustainable solution.

Figure 14 illustrates the stable equilibrium case, where the only candidate for equilibrium is the point (α_1^c, α_2^c), denoted by "E", where the two lines intersect. Stability of this equilibrium is illustrated by arrows in the figure. These arrows show evolution of reactions triggered by a deviation of firm 1 from the equilibrium point, and the two firms' decisions eventually converge to E again. However this point is an equilibrium only if it satisfies the feasibility conditions. Then the equilibrium quality levels would be:

$$\alpha_1^C = \max(0.5, \min(\frac{(4c_2t - 3r_1p_1v)(r_0v + r_1p_1t)}{16c_2t^2c_1 - 8c_2tr_1p_1v - 8r_1p_1vc_1t + 3r_1p_1^2v^2}, 1))$$
$$\alpha_2^C = \max(0.5, \min(\frac{(4c_1t - 3r_1p_1v)(r_0v + r_1p_1t)}{16c_2t^2c_1 - 8c_2tr_1p_1v - 8r_1p_1vc_1t + 3r_1p_1^2v^2}, 1))$$

If the cost of quality is the same ($c_1 = c_2 = c$) for both firms, then the equilibrium is symmetric ($\alpha_1^C = \alpha_2^C$) with

$$\alpha_1^C = \alpha_2^C = \max(0.5, \min(\frac{r_0v + r_1p_1t}{4ct - r_1p_1v}, 1)) = \max(0.5, \min(\frac{r_0\frac{v}{2t} + \frac{r_1p_1}{2}}{2c - r_1p_1\frac{v}{2t}}, 1)) \quad (57)$$

In equation (57), for the interior solution, we can see each factor affecting the result explicitly: $\frac{v}{2t}$ represents the quality elasticity of demand, r_1p_1 is the marginal revenues of quality increase and $2c$ is the marginal cost of quality increase. This shows, the competitive quality level would increase with an increase in reimbursement for screening test, or early stage treatment, as well as the prevalence of the disease. On the other hand increasing cost of quality would decrease the equilibrium quality level, as expected.

If the condition given in Proposition 14 is not satisfied (β_1 or $\beta_2 \geq 1$), then the interior equilibrium is not stable and the reaction functions can look like Figure 15. In this case, the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

reaction functions can intersect at three points, E1, E2 and E3. The arrows in Figure 15 indicate the reactions triggered by a deviation from point E3 illustrating that E3 is not a stable equilibrium. The system stabilizes either at point E1 or E2. The possible equilibria are:

$$(\alpha_1^C, \alpha_2^C) = (0.5, \max(0.5, \min(\frac{r_0 v + r_1 p_1 t}{4ct - 2r_1 p_1 v}, 1))) \quad (58)$$

$$\text{or}(\alpha_1^C, \alpha_2^C) = (\max(0.5, \min(\frac{r_0 v + r_1 p_1 t}{4ct - 2r_1 p_1 v}, 1)), 0.5) \quad (59)$$

This implies that at most one firm would invest in quality if $2c_i \leq 3r_1 p_1 \frac{v}{2t}$, i.e. if the cost of quality and transportation were small compared to the potential benefits of it.

4.6.4 Discussion

Assume that the socially optimal quality level is $\frac{b_1 p_1}{4c} > 0.5$ (that social optimum was analyzed in section 4.6.1). The social welfare for private providers is less than the optimal social welfare, if the quality levels set by the private providers at equilibrium is different than the socially optimal quality level. If the following condition holds, then it is best to offer the screening service as a public service rather than letting private providers do it, since the socially optimal quality level would not be reached in competition:

$$\alpha^{SP} = \frac{b_1 p_1}{4c} > \alpha^C = \frac{r_0 v + r_1 p_1 t}{4ct - r_1 p_1 v} \quad (60)$$

This condition holds if there is a big gap between the social benefits b_1 and the net revenue for treatment r_1, and when the valuation of patients for screening quality, v is small. Since b_1 includes the long run benefits of an increased life time, a social planner would not pay this amount as a reimbursement increase for each early detection, which suggests that $b_1 > r_1$. Depending on how much value a social planner assigns for the life years, the gap can be big, in which case public provision of screening should be preferred.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Screening costs and net reimbursements for screening (r_0) also play a role in private providers' decision for a quality level. The social planner should make sure that the reimbursement for the screening test is high enough to make $\frac{r_0 v + r_1 p_1 t}{4ct - r_1 p_1 v}$ equal to the socially optimal level $\frac{b_1 p_1}{4c}$. This shows the interaction between the two services offered: While the first one (screening) contributes to the costs, the benefits from the second service (treatment) compensates for these costs, and the reimbursements should ensure the right balance of the costs and benefits in order to achieve the desired quality levels. If the screening test cost is reimbursed fully, i.e. $r_0 = 0$ (defined in (36)), then the condition in (60) is satisfied if and only if $\frac{1}{r_1} - \frac{1}{b_1} \geq \frac{vp_1}{4ct^2}$. I.e.

$$\text{If} \left(\frac{1}{r_1} - \frac{1}{b_1} \right) \geq \frac{vp_1}{4ct^2} \quad \text{then} \quad \alpha^{SP} > \alpha^C \tag{61}$$

Other factors that may potentially create a difference between the social planner solution and private provider solution are the valuation of patients for quality (v) and the access cost (t): With competition, providers have more incentive to increase their quality if v is high and t is low, i.e. if demand is more sensitive to quality differences, then competing providers would achieve equilibrium at a higher quality level. On the other hand for the social planner we do not observe this elasticity effect since the whole population is served anyway. This result is stated in proposition 15:

Proposition 15 *If $t > 0$, $c_1 = c_2 = c$, $r_0 \geq \frac{-r_1^2 p_1^2}{4c}$, and $c_i > \frac{r_1 p_1 v}{2t}$, then the equilibrium quality level under competition, $\alpha^C = \frac{r_0 v + r_1 p_1 t}{4ct - r_1 p_1 v}$ increases with v and decreases with t. That is, quality under competition increases in "quality elasticity of the demand", $\frac{v}{2t}$.*

Proof. Given the assumptions stated in the proposition,

$$\frac{d}{d(\frac{1}{2}\frac{v}{t})} \left(\frac{r_0 v + r_1 p_1 t}{4ct - r_1 p_1 v} \right) = \frac{1}{2} \frac{4r_0 c + r_1^2 p_1^2}{\left(2c - r_1 p_1 \frac{1}{2}\frac{v}{t}\right)^2} \geq 0 \text{ if } r_0 \geq \frac{-r_1^2 p_1^2}{4c}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

■

Proposition 15 suggests that a policy maker can ensure higher quality levels by increasing the sensitivity of population to screening test quality when they choose between the two providers. This can be done by increasing awareness, or decreasing the transportation costs. However, there is a limit to this intervention without disturbing stability. If the demand becomes too sensitive to quality (i.e. if $\frac{v}{2t} \geq \frac{2c_1}{3r_1p_1}$), then in equilibrium, only one provider would invest in quality. Moreover, the condition $r_0 \geq \frac{-r_1^2p_1^2}{4c}$ should be satisfied to guarantee that reimbursement for screening is high enough for the providers to make nonnegative profits by offering the screening service.

The discussion so far was based on given parameters of the utility for the private providers and the social planner (r_0, r_1, b_1 and c_0). A social planner should solve the following program to find the optimal reimbursement levels:

$$\max_{r_0,r_1} : \ \Pi^{SP} \tag{62}$$

$$\text{s.t.} \tag{63}$$

$$\Pi^C \geq 0 \qquad \text{(Individual Rationality)}$$

$$\alpha^C = \alpha^{SP} \qquad \text{(Incentive Compatibility)}$$

The first constraint in the above program is the participation constraint, i.e the providers should make non-negative profits. The second constraint shows incentive compatibility, i.e. optimal competitive quality level should be equal to the socially optimal quality. The optimal r_0 and r_1 should satisfy:

$$\frac{r_0}{2} + \frac{1}{4} \geq \frac{p_1b_1(b_1 - 2r_1)}{16c} \tag{64}$$

$$\frac{b_1p_1}{4c} = \frac{r_0v + r_1p_1t}{4ct - r_1p_1v} \tag{65}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The solution of equations 64 and 65 is found as:

$$b_1 = \rho \tag{66}$$

$$r_0 = -\frac{1}{4}p_1\frac{-4\rho ct + \rho r_1 p_1 v + 4tr_1 c}{cv} \tag{67}$$

where there are two possible ρ values:

$$Z = \frac{1}{2vp_1}\left(2r_1p_1v + 8p_1ct - 2p_1^2r_1v + 2\sqrt{(r_1^2p_1^2v^2 - 2r_1^2p_1^3v^2 + 16p_1^2c^2t^2 - 8p_1^3ctr_1v + p_1^4r_1^2v^2 + 4v^2p_1c)}\right)$$

$$Z = \frac{1}{2vp_1}\left(2r_1p_1v + 8p_1ct - 2p_1^2r_1v - 2\sqrt{(r_1^2p_1^2v^2 - 2r_1^2p_1^3v^2 + 16p_1^2c^2t^2 - 8p_1^3ctr_1v + p_1^4r_1^2v^2 + 4v^2p_1c)}\right)$$

We leave a more involved discussion of reimbursement schemes for future work.

4.7 Case III: Patients are not Homogeneous in Their Utility from Screening

In this section we add one additional feature to the model in case II: A customer's valuation of screening, v is assumed uniform in [0, 1]. Then the linear city of length 1 in the previous section becomes a unit square, with N people uniformly distributed on $(v, x) \in [0, 1]^2$ (as also used in [38]).

Unlike case II, there will now be some customers who do not want to get screening from any provider since their valuation is not large enough to compensate for the burden of travel. The region marked "0: No screening" represents those people in Figure 16. We denote the demand for provider i with D_i, as a shorthand notation of $D_i(\alpha_1, \alpha_2)$, where $D_1 + D_2 < N$ (in Cases I and II, the assumption was $D_1 + D_2 = N$). We define $D_0 = N - D_1 - D_2$, the number of people that do not want to get screening from either provider.

Let the quality for the two providers be α_1 and α_2. At any point $x \in [0, 1]$ between provider 1 and 2, the demands are determined by the customer valuations. A customer at

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

distance x_1 to the provider 1 is willing to get screening if the utility of getting screening from at least one of the providers is greater than zero, i.e.:

$$\max(v\alpha_1 - tx_1, v\alpha_2 - t(1 - x_1)) \geq 0 \tag{68}$$

If the utility for both providers is positive, then a customer at point x_1 prefers provider 1 for values of v that satisfies:

$$v\alpha_1 - tx_1 \geq v\alpha_2 - t(1 - x_1)$$

If $\alpha_1 = \alpha_2$ then customers at points $x_1 \leq 0.5$ prefer provider 1. If $\alpha_2 > \alpha_1$ then customers who prefer provider 1 have valuations

$$v \geq \frac{2t}{\alpha_2 - \alpha_1}x_1 - \frac{t}{\alpha_2 - \alpha_1} \tag{69}$$

The preferences of customers on the unit square, where the x axis is the distance and y axis is the valuation of screening are shown in Figure 16. The demands are determined by three lines, d_{12}, d_{10} and d_{20}. Figure 16 is drawn assuming $\alpha_2 > \alpha_1$. When α_1 and α_2 are equal, the line d_{12} becomes a straight line passing from $x = 0.5$, and $D_1 = D_2$.

For given quality levels α_1 and α_2, the demands D_1, D_2, D_0 for a uniformly distributed population equal the area of the respective regions in Figure 16:

if $\alpha_1 < \alpha_2$:

$$D_0 = \frac{t}{2(\alpha_1 + \alpha_2)}N \tag{70}$$

$$D_1 = \frac{t\alpha_1}{2(\alpha_2^2 - \alpha_1^2)}N \tag{71}$$

$$D_2 = \frac{2(\alpha_2^2 - \alpha_1^2) - \alpha_2 t}{2(\alpha_2^2 - \alpha_1^2)}N \tag{72}$$

If $\alpha_1 = \alpha_2 = \alpha$ then

$$D_0 = \frac{t}{4\alpha} \quad \text{and} \quad D_1 = D_2 = \frac{4\alpha - t}{8\alpha} \tag{73}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

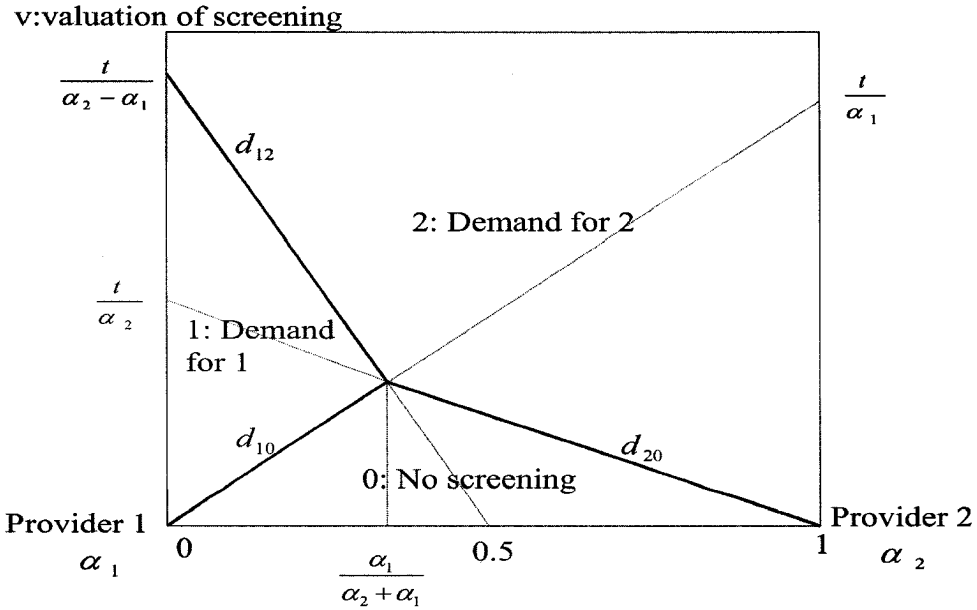

Figure 16: Demands for providers for given quality levels, $\alpha_2 > \alpha_1$

Demands for each provider increase in their own quality and decreases in the competitor's quality. Number of people who do not get screened (D_0) decreases as the quality of screening tests increases, and as the transportation cost decreases.

If a continuum of values of α are used, the analysis of optimal decisions given the demand structure is not analytically tractable, since there are many special cases for different α values. We make an assumption of having two levels of quality, high, α^H and low, α^L for the rest of the analysis. The assumption of having discrete (high and low) quality levels is a reasonable one. Although the screening test sensitivity would have continuous values, the quality measure in this model should be seen as an expected value. In practice, sensitivity values for screening tests would be around a high value if a certain investment is made, while it would be around a low value otherwise. For example, for breast cancer screening, hiring an experienced radiologist would provide an expected sensitivity value of 0.80 while with an experienced radiologist the sensitivity could take values around 0.65 [42].

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Set

$$\alpha^H - \alpha^L = \Delta \tag{74}$$

$$u(\alpha) = v\alpha - tx \tag{75}$$

We assume $t \leq \Delta$ and $\Delta > 0$ so that the demand functions given in equations (70-72) are nonnegative. There can be four solutions, defined as the quality levels chosen by the two providers (α_1, α_2), which are: (L,L), (H,H), (L,H) and (H,L).

4.7.1 Case III: Social Planner's Decision

The Social planner's objective is to maximize social welfare as in the previous sections. Assume $c_1 = c_2 = c$, hence $C_i(\alpha) = C(\alpha)$, and so(L,H) and(H,L) are equivalent. We only consider the cases (L,L), (L,H) and (H,H). We denote the contribution of provider i to social welfare for $(\alpha_1, \alpha_2) = (\alpha^L, \alpha^L)$ by Π_i^{LL}.

1.(L, L) : If $(\alpha_1, \alpha_2) = (\alpha^L, \alpha^L)$, then the social welfare is $2\Pi_1^{LL} = 2\Pi_2^{LL}$ where

$$\Pi_1^{LL} = \Pi_2^{LL} = N\frac{4\alpha^L - t}{8\alpha^L}(b_1 p_1 \alpha^L - c_0) - C(\alpha^L)$$

2.(H, H) :If $(\alpha_1, \alpha_2) = (\alpha^H, \alpha^H)$, then the social welfare is $2\Pi_1^{HH} = 2\Pi_2^{HH}$ where

$$\Pi_1^{HH} = \Pi_2^{HH} = N\frac{4\alpha^H - t}{8\alpha^H}(b_1 p_1 \alpha^H - c_0) - C(\alpha^H)$$

3.(L, H): If $(\alpha_1^{SP}, \alpha_2^{SP}) = (\alpha^L, \alpha^H)$, then the social welfare is $\Pi_1^{LH} + \Pi_2^{LH}$ where

$$\begin{aligned} \Pi_1^{LH} &= Nt\frac{\alpha^L}{2(\alpha^H)^2 - 2(\alpha^L)^2}\left(b_1 p_1 \alpha^L - c_0\right) - C(\alpha^L) \\ \Pi_2^{LH} &= N\frac{2(\alpha^H)^2 - 2(\alpha^L)^2 - t\alpha^H}{2(\alpha^H)^2 - 2(\alpha^L)^2}\left(b_1 p_1 \alpha^H - c_0\right) - C\left(\alpha^H\right) \end{aligned} \tag{76}$$

The conditions for each solution to be the optimum are found by comparing the social welfare in each of the three cases.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

4.7.2 Case III: Profit Maximizing Providers' Decision

We denote the profit of provider i for the solution $(\alpha_1^C, \alpha_2^C) = (\alpha^L, \alpha^L)$ by Π_i^{LL} and likewise for the other solutions. Assume $c_1 = c_2 = c$, hence $C_i(\alpha) = C(\alpha)$. The profits of the two providers for each of the four cases are:

1.(L, L) : If $(\alpha_1, \alpha_2) = (\alpha^L, \alpha^L)$, then profits are:

$$\Pi_1^{LL} = \Pi_2^{LL} = N\frac{4\alpha^L - t}{8\alpha^L}(r_1 p_1 \alpha^L + r_0) - C(\alpha^L)$$

2.(H, H) : $(\alpha_1, \alpha_2) = (\alpha^H, \alpha^H)$.

$$\Pi_1^{HH} = \Pi_2^{HH} = N\frac{4\alpha^H - t}{8\alpha^H}(r_1 p_1 \alpha^H + r_0) - C(\alpha^H)$$

3.(L, H) : $(\alpha_1, \alpha_2) = (\alpha^L, \alpha^H)$, there are two cases, $t \leq \Delta$ and $t > \Delta$. If $t \leq \Delta$ then

$$\begin{aligned} \Pi_1^{LH} &= Nt\frac{\alpha^L}{2(\alpha^H)^2 - 2(\alpha^L)^2}\left(r_1 p_1 \alpha^L + r_0\right) - C(\alpha^L) \\ \Pi_2^{LH} &= N\frac{2(\alpha^H)^2 - 2(\alpha^L)^2 - t\alpha^H}{2(\alpha^H)^2 - 2(\alpha^L)^2}\left(r_1 p_1 \alpha^H + r_0\right) - C\left(\alpha^H\right) \end{aligned} \tag{77}$$

Remark 16 *It is always true that:* $\Pi_2^{LH} \geq \Pi_2^{HH}$ *and* $\Pi_1^{LL} \geq \Pi_1^{LH}$. *A provider always prefers its competitor to have a low quality level.*

The analysis of this model will involve finding conditions for each of the four decisions for both social planner's and private providers case and finding cases when they coincide and when they do not coincide.

4.8 Summary of Results and Concluding Remarks

In this chapter, we introduced a model to investigate the impact of different industrial organizations for public health services provision, in the context of disease screening. Motivated from

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the results of Chapter 3, we focused on the sensitivity of screening tests as the main decision variable.

We analyzed two models and introduced a more complex third model. The analysis of the third model and extensions to optimal reimbursement are areas for further research. The results presented in this Chapter are summarized in Table 12 and 13 (for the case of identical providers for simplicity in exposition):

Case	Basic Assumption	Social Optimum
I	$t = 0, v > 0$	$\alpha^{SP} = (\alpha_1^{SP}, \alpha_2^{SP}) = \max(0.5, \min(\frac{p_1 b_1}{2c}, 1))$
II	$t > 0, v > 0$	$\alpha_1^{SP} = \alpha_2^{SP} = \max(0.5, \frac{b_1 p_1}{4c})$

Table 12: Summary of Analysis and Conclusions for the Social Planner Case

Case	Basic Assumption	Competitive Solution	Condition
I	$t = 0, v > 0$	no equilibrium	no equilibrium
II	$t > 0, v > 0$	$\alpha_1^C = \alpha_2^C = \max(0.5, \min(\frac{r_0 v + r_1 p_1 t}{4ct - r_1 p_1 v}, 1))$	if $c > \frac{3 r_1 p_1 v}{4t}$
II	$t > 0, v > 0$	$(\alpha_1^C, \alpha_2^C) = (0.5, \max(0.5, \min(\frac{r_0 v + r_1 p_1 t}{4ct - r_1 p_1 v}, 1)))$	if $c \leq \frac{3 r_1 p_1 v}{4t}$

Table 13: Summary of Analysis and Conclusions for the Competitive Case

For the analysis completed to date, here is a summary:

- If the cost of screening quality is a fixed cost (that is a cost scaled with the total population size as opposed to the demand) and there is no patient disutility for participation, then a social planner operating two facilities is inefficient. The optimal decision is to invest in only one facility.

- If the cost of screening quality is a fixed cost and if there is no cost of access (or

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

other differentiation between the firms), then duopoly competition does not have an equilibrium. Competition ensues and collaboration is required to obtain profits.

- A high disease prevalence increases the optimal quality level for both the social planner and competitive models.

- When the two facilities are located in separate locations, the decision for the quality level, given competition, is affected by the net reimbursements for screening and valuations of patients, which are factors that do not affect the social planner's decision. Therefore, optimal reimbursements should provide a balance between screening costs and treatment revenues in order to provide the right incentives for the private providers.

- A high elasticity of demand for quality increases the quality under competition. That is, if customers have a high valuation for screening quality and low transportation costs, there will be harsher competition for attracting demand between the two providers. This result is not valid for the social provider, who cares only for the total demand for both facilities.

- If the marginal cost of quality is very low, then the equilibrium with two providers in competition is not stable. One of the providers would be driven out of competition and only one provider will make investment for quality in equilibrium.

In this chapter, we focused on points that were ignored in Chapter 3, including competition between two providers. The comparison of optimal quality levels depends on the valuation of the social planner for life gains from early detection and the reflection of this on reimbursements. In order for competition to be realized, certain conditions regarding the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

relative costs of quality and transportation are necessary.

Chapter 3 showed that increasing quality or increasing outreach of screening results in improved health outcomes, while increasing quality is more beneficial in terms of reducing system costs (screening and treatment costs). In Chapter 4, an interaction between these two dimensions, quality and outreach, is shown: a higher quality would increase outreach too, since people would be more willing to get screened. This means that the benefit of increasing quality is even stronger than that estimated in Chapter 3.

4.9 Appendix to Chapter 3

Social Welfare calculation

For a population of size N, the total benefit of screening and treating when screening quality is α is found as; the cost of screening minus cost of treatment in case of early and late treatment plus the value of increase in life expectancy minus the cost of quality:

$$\begin{aligned}
\Pi^{SP} &= N(-c_0 - p_1\alpha\gamma_1 - p_1(1-\alpha)\gamma_2 + p_1\alpha\delta l - p_2\gamma_2) - C(\alpha) \\
&= N(p_1\alpha(\gamma_2 - \gamma_1 + \delta l) - \gamma_0) - C(\alpha) - N\gamma_2(p_2 + p_1) \\
&= N(b_1 p_1 \alpha_1 - c_0) - Nc(\alpha^2 - \frac{1}{4}) - Nc_2(p_2 + p_1)
\end{aligned}$$

By assumption that everyone is treated when they have the disease at late stage, the benefits of screening can be written as the marginal gains by early detection. The last terms is a constant which does not affect quality decision, so we drop it from the analysis, normalizing to zero.

Analysis of Case II with Variable Cost for Quality Let the total cost of quality equal the realized demand times the cost of quality, i.e. let $C_i(\alpha) = D_i c_2(\alpha_1^2 - \frac{1}{4})$. We look at only the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

case with identical providers in this section.

Socially Optimal α maximizes Π^{SP}, for the whole population N, which is.

$$\alpha^{SP} = \max(0.5, \min(\frac{b_1 p_1}{2c}, 1)) \tag{78}$$

This is the same solution as when only one facility serves the whole demand, when the quality cost was scaled for the whole population

Profit Maximizing Providers' Decision (Variable Quality Cost)

The demand is the same as in Section 4.5.2. We have the Bertrand-type of competition again, where the quality level is escalated until the point where provider have zero profits and cannot increase the quality any more since they make losses. In the Bertrand model, the equilibrium is achieved if price equals marginal cost. For our model the result is similar. Equilibrium is achieved at quality level that results in zero profit, α^C.

Proposition 17 *Given the assumptions of this section, the only competitive equilibrium is achieved at:*

$$\alpha^C = \max(0.5, \min(1, \frac{1}{4c}\left(2p_1 r_1 + 2\sqrt{(p_1^2 r_1^2 + 4cr_0 + c^2)}\right)))$$

Proof. The quality level that results in zero profits for each provider ($\alpha^{N,0}$) is the solution to $r_0 + p_1 \alpha r_1 - c(\alpha^2 - \frac{1}{4}) = 0$, provided that it is in the feasible range [0.5, 1]. There are two potential solutions: equation:

$$\alpha^C = \begin{cases} \alpha_1^C = \frac{1}{4c}\left(2p_1 r_1 + 2\sqrt{(p_1^2 r_1^2 + 4cr_0 + c^2)}\right) \\ \alpha_2^C = \frac{1}{4c}\left(2p_1 r_1 - 2\sqrt{(p_1^2 r_1^2 + 4cr_0 + c^2)}\right) \end{cases}$$

In the absence of competition, it is optimal to set $\alpha = \frac{p_1 r_1}{2c}$, which is greater than α_2^C, so α_2^C cannot be an equilibrium. If one firm sets $\alpha = \alpha_2^C$, the other one can set $\alpha = \frac{p_1 r_1}{2c} \geq \alpha_2^C$,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

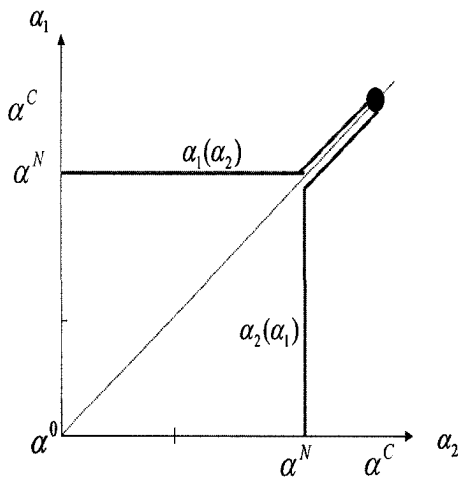

Figure 17: Reaction functions when quality cost is variable for population served and two providers are identical.

get all the market, and gain the maximum profit. Hence the only competitive equilibrium is α_1^C,provided it is feasible. If it is not feasible, one of the boundaries will be the equilibrium. Note that α_2^C is not necessarily negative since we do not restrict r_0 to have a positive value. ■

By setting $\alpha_1 = \alpha_2 = \alpha^C$, both firms make zero profit, but they do not have an incentive to deviate from this equilibrium. The reaction functions are shown in Figure 17.

Remark 18 *Quality levels under quality competition (given in proposition 17) are higher than the socially optimal level (given in equation 78), if all the costs and benefits of the social planner are transferred to the private providers, i.e. if* $r_1 = b_1$ *and* $r_0 = -c_0$.

Remark 19 *If social planner transfers all benefits and costs to the private providers as reimbursements* ($r_1 = b_1$ *and* $r_0 = -c_0$.), *then the total social welfare under competition will be*

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

less than the maximum social welfare under social planning because there will be an over-investment on quality.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

5 Conclusion

The thesis presented three essays, in three chapters, to analyze problems in service operations, specifically in the call center and the health care service contexts. The common theme in all the problems was to decide on a good match between different service levels and the needs of a heterogeneous customer base, which brought up the importance of the role of the front-line employee. We discussed two necessary conditions for the service employee to be able to accomplish this role (given the directions of the manager): the ability to detect the customer characteristics and decide on an appropriate service level, and the incentive to provide that service level.

In each of the chapters only one of these conditions were incorporated in the broader context of the specific problem studied: In Chapter 2, the server incentives were analyzed while customer characteristics were assumed to be observed perfectly by the server, in Chapter 3, the hypocratic oath in principle eliminates the incentive issue, but the ability to detect the customer characteristics (existence of the disease) depended on the system design and was not perfect. In Chapter 4, we considered the incentives of a service provider to invest in the ability of the employees to recognize customer characteristics. In the disease screening context, this essay investigated the effects of competition on public health services.

We used different methodologies throughout the thesis. In Chapter 2, a stylistic model was built using queuing and principal agent theory. In Chapter 3, a fairly complicated stochastic compartmental model was built and simulation experiments were used for its analysis. Parameters were estimated using country statistics and published data wherever possible. Chapter 3 and 4 focused on the same service context, public health care services. While Chapter 3

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

analyzed a specific problem, breast cancer screening, Chapter 4 modeled the disease screening problem for a general disease. Although stylistic in approach, the model captured elements (competition, effect of quality on willingness to get screened) that were ignored in Chapter 3 and in that sense complemented it.

The analysis for the call center context illustrated the importance of accounting for the system costs and server incentives in starting a value creation initiative. Moreover, it included the market segmentation as an endogenous decision in the value creation strategy determination, which was an improvement over existing practice. For the disease screening problem, main result was the importance of screening test quality on health outcomes and system costs. We showed that contrary to the anxiety about the waiting times for mammogram service in many countries, waiting is a less significant concern for health outcomes than participation and screening quality, unless there is a severe capacity problem. The analysis of disease screening delivery with competition showed that when there were no transportation costs, having two facilities do the screening was suboptimal for a social planner and was unsustainable for the private providers. When there was a transportation cost, elasticity of screening demand to quality was an important factor for the incentives of profit maximizing providers to invest on quality.

This elasticity effect on the quality levels show that public authorities have a lever to influence the quality provided by private health providers, other than reimbursement mechanisms. Increasing public awareness for importance of quality, and reducing transportation costs would increase incentives of private providers to invest in quality.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

5.1 Limitations and Further Research Directions

The models in this dissertation present and define several service operations problems formally for the first time. They provide a model of interacting dimensions of service delivery systems; they are simplistic in certain respects, and these limitations provide avenues for further research.

The model analyzed in Chapter 2 is stylistic and is a first step in bridging market segmentation, incentives and operational performance. It does not model repetitive interactions with customers, which is an important consideration for customer relationship management (CRM) initiatives. Increasing customer profitability can happen in two ways: by increasing the number of transactions, or by increasing the value of each transaction. We looked at only a single service interaction. Moreover, the policy that determines the service level for each customer type is static in Chapter 2. A dynamic policy that considers the congestion level in the system could improve the performance: when there are few customers in the system it might be profitable to try to cross sell to low potential customers. When it is highly congested it may not be profitable to try it for even high potential customers. Modeling repetitive interactions and dynamic policies with dynamic incentive schemes would be the next step to develop this research.

Another interesting issue that we have not discussed in the first essay is the information structure and its effects on the server behavior and incentives. It was assumed that servers have perfect information on customer characteristics, and the information on queue length was irrelevant. It would be interesting to see if there is a specific information structure that a manager should prefer, in terms of what to let the servers know about the customers and the system state. The result should depend on how server behavior is affected by the information

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

set. The psychology of servers as well as the incentive structure would play a role in their behavior, so the analysis of this problem requires an investigation of behavioral implications of different information sets for the employees.

The model in Chapter 3 does not include a cost-effectiveness analysis, mainly because of the difficulty in assessing the costs of achieving specific changes in quality and outreach dimensions. Although the model incorporates many aspects of breast cancer screening service delivery and health status dynamics, it does not consider the fact that different age groups would have different characteristics in terms of prevalence and mortality. The model assumes homogeneous compartments and Markovian dynamics for transitions between compartments. Another limitation comes from the complexity of the model and the methodology chosen to analyze it. The model is descriptive, and simulation analysis is done to compare the impact of different interventions, but there is no optimization analysis. We took the screening test frequency as a parameter of the model, while in reality it is a decision variable for the policy makers.

Three dimension for extending the analysis in Chapter 3 are: (1) Making a more complex dynamics model to more accurately represent health, (2) Getting a better handle on the costs and effects of health program interventions, and (3) Simplifying the model in parts to enable optimization. To this end a deterministic differential equation model, rather than a stochastic model could be used, since the variability leading to waits seems to have a second-order effect on health outcomes.

Finally in Chapter 4, a simplified model and a basic analysis is presented. The next step should be to fully analyze the case with heterogeneous valuations. The analysis can be taken further by investigating different reimbursement schemes. The model assumed full insurance

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

for all the target population and normalized the co-payments from the patients to zero. Letting the private providers decide on the co-payment amounts would introduce a second dimension for competition (in addition to quality) to capture another important dimension of the problem. Also, the current model looks at sensitivity of the screening test as the only factor, while specificity is assumed to be perfect (i.e. no false positives). Looking at the false positives as well as false negatives would make the model more realistic. The model currently considers a single period and does not consider the disease progression and the effects of policies on the future health states of the population. A multi period model can help completing the picture by incorporating the disease progression dimension.

The subject of how to organize public health services in the broader context of health services poses interesting questions for further research. Disease screening services should not be considered in isolation. Health care providers usually have to allocate budget between several health services. From a social planner's perspective, allocating resources between public health care and acute health care can be explored in the disease screening context. Similarly, for profit maximizing providers, the quality of treatment can be another investment opportunity, and if patients can switch providers for different services the problem becomes even more challenging and interesting. The model introduced in Chapter 4 can be a starting point to this end.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

References

[1] Organized breast cancer screening programs in Canada: 1997 and 1998 report, 1998. Health Canada.

[2] Cancer surveillance on-line, 2003. http://dsol-smed.hc-sc.gc.ca/dsol-smed/cancer/index_e.html, retrieved on 17/05/2004.

[3] ACS. Cancer facts and figures. Atlanta, GA, 2003. American Cancer Society.

[4] O.Z. Akşin and P.T. Harker. To sell or not to sell: Determining the trade-offs between service and sales in retail banking phone centers. *Journal of Service Research*, 2(1):19–33, 1999.

[5] M.R. Andersen, M. Hager, C. Su, and N. Urban. Analysis of the cost-effectiveness of mammography promotion by volunteers in rural communities. *Health Education and Behavior: The Official Publication of the Society for Public Health Education*, 29(6):755–770, 2002.

[6] Anonymous. Mammography quality standards act regulations. FDA Website, 1992. http://www.fda.gov/cdrh/mammography/frmamcom2.html#s90012, retrieved on 8/5/2004.

[7] Anonymous. Call center statistics report highlights. Tower Group Website, 1998. http://www.towergroup.com/public/ras/default_ras.asp?main=search.asp, retrieved on 8/7/2004.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[8] Anonymous. World developmet indicators 2000. World Bank Website, 2000. http://www.worldbank.org/data/wdi2000/pdfs/tab2_14.pdf, retrieved on 12/7/2004.

[9] Anonymous. Call center statistics report highlights. Data Monitor website, 2002. http://www.datamonitor.com/~1b1047c222d14e69b21c31222c928f1e~/industries/research/?pid=DMTC0852&type=Report, retrieved on 8/7/2004.

[10] Anonymous. World developmet indicators 2004. World Bank Website, 2004. http://www.worldbank.org/depweb/beyond/beyondbw/begbw_09.pdf, retrieved on 12/7/2004.

[11] R. Baker. Use of a mathematical model to evaluate breast cancer screening policy. *Health Care Management Science*, 1:103–113, 1998.

[12] W.E. Barlow, C.D. Lehman, Y. Zheng, R. Balard-Barbash, B.C. Yankaskas, G.R. Cutter, P.A. Carney, B.M. Geller, R. Rosenberg, K. Kerlikowske, D.L. Weaver, and S. Taplin. Performance of diagnostic mammography for women with signs or symptoms of breast cancer. *Journal of the National Cancer Institute*, 94(15):1151–1159, 2002.

[13] A.K. Basu, R. Lal, V. Srinivasan, and R. Staelin. Salesforce compensation plans: An agency theoretic perspective. *Marketing Science*, 4(4):267–291, 1985.

[14] C. Beam, E. Conant, and E. Sickles. Factors affecting radiologists inconsistency in screening mammography. *Academic Radiology*, 9(5):531–540, 2002.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[15] A. Beitia. Hospital quality choice and market structure in a regulated duopoly. *Journal of Health Economics*, 22:1011–1036, 2003.

[16] K. Benjamin. Customer relationship management: Building a strategy. special report: Understanding crm. *Financial Times*, November 26 2001.

[17] W.A. Berg, C.J D'Orsi, V.P. Jackson, L.W. Bassett, C.A. Beam, R.S. Lewis, and P. Crewson. Does training in the breast imaging reporting and data system (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers or mammography? *Radiology*, 224:871–880, 2002.

[18] P.D. Berger and N.I. Nasr. Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing*, 12(1):17–26, 1998.

[19] L.A. Bettencourt and K. Gwinner. Customization of the service experience: The role of the frontline employee. *International Journal of Service Industry Management*, 7(2):3–20, 1996.

[20] G. Biglaiser and C. Ma. Price and quality competition under adverse selection: Market organization and efficiency. *RAND Journal of Economics*, 34(2):266–286, 2003.

[21] W.C. Black, D.A. Hagstrom, and H.G Welch. All-cause mortality in randomized trials of cancer screening. *Journal of the National Cancer Institute*, 94(3):167–173, 2002.

[22] F. Boer, H. de Koning, P. Warmerdam, A. Street, E. Friedman, and C. Woodman. Cost effectiveness of shortening screening interval or extending age range of NHS breast screening programme: Computer simulation study. *British Medical Journal*, 317:376–379, 1998.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[23] R.M. Bradford. Pricing, routing, and incentive compatibility in multiserver queues. *European Journal of Operational Research*, 89(2):226–236, 1996.

[24] J. A. Buzacott. Commonalities in reengineered business processes: Models and issues. *Management Science*, 42(5):768–782, 1996.

[25] J. A. Buzacott. The impact of worker differences on production system output. *International Journal of Production Economics*, 78:37–44, 2002.

[26] L.S. Caplan, K.J. Helzlsouer, S. Shapiro, L.S. Freedman, R.J. Coates, and B.K. Edwards. System delay in breast cancer in whites and blacks. *American Journal of Epidemiology*, 142(8):804–812, 1995.

[27] L.S. Caplan, K.J. Helzlsouer, S. Shapiro, M.N. Wesley, and B.K. Edwards. Reasons for delay in breast cancer diagnosis. *Preventive Medicine*, 25(2):218–224, 1996.

[28] P. Carroll and M. Tadikonda. Customer profitability: Irrelevant for decisions? *Banking Strategies*, 75(6):77–82, 1997.

[29] J.P. Caulkins and G. Tragler. Dynamic drug policy: an introduction and overview. *Socio-Economic Planning Sciences*, 38:1–6, 2004.

[30] S.E. Chick, S. Soorapanth, and J.S. Koopman. Microbial risk assessment for drinking water. In M. Brandeau, F. Sainfort, and W. Pierskalla, editors, *Operations Research and Health Care: A Handbook of Methods and Applications*, pages 467–494. Kluwer Academic Publishers, 2004.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[31] P.M. Clarke. Cost benefit analysis and mammographic screening: A travel cost approach. *Journal of Health Economics*, 17(6):7367–787, 1998.

[32] A. Coughlan. Salesforce compensation: A review of ms/or advances. In *Handbooks in OR & MS*, volume 5, pages 611–652. J. Eliashberg and G.L. Lilien Eds., 1993.

[33] N.E. Day and S.D. Walter. Simplified models of screening for chronic disease: Estimation procedures from mass screening programs. *Biometrics*, 40:1–14, 1984.

[34] L. Debo and B. Toktay and L. Van Wassenhove. Queueing for Expert Services. INSEAD Working paper 2004/46/TM, 2004.

[35] E. Docteur and H. Oxley. Health Care Systems: Lessons From the Reform Experience, OECD Health Working Papers, 2003.

[36] S.W. Duffy, H.H. Che, L. Tabar, and N.E. Day. Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. *Statistics in Medicine*, 14(4):1531–1543, 1995.

[37] The Economist. Keeping the customers satisfied, July 14 2001.

[38] R. Ellis. Creaming, skimping and dumping: Provider competition on the intensive and extensive margins. *Journal of Health Economics*, 17:537–555, 1998.

[39] J.G. Elmore, D.L. Miglioretti, and A.P. Carney. Does practice make perfect when interpreting mammography?, part II. *Journal of the National Cancer Institute*, 95(2):250–252, 2003.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[40] J.G. Elmore, D.L. Miglioretti, L.M. Reisch, M.B. Barton, W. Kreuter, C.L. Christiansen, and Fletcher S. Screening mammograms by community radiologists: variability in false-positive rates. *Journal of the National Cancer Institute*, 94(18):1373–1380, 2002.

[41] K. Engelman, D.B. Hawley, R. Gazaway, M.C. Mosier, J.S. Ahluwalia, and E.F. Ellerbeck. Impact of geographic barriers on the utilization of mammograms by older rural women. *American Geriatrics Society*, 50(1):62–68, 2002.

[42] L. Esserman, H. Cowley, C. Eberle, A. Kirkpatrick, S. Chang, K.Berbaum, and A. Gale. Improving the accuracy of mammography: Volume outcome relationship. *Journal of National Cancer Institute*, 94(5):369–375, 2002.

[43] J.U. Farley. An optimal plan for salesmen's compensation. *Journal of Marketing Research*, 1(2):39–43, 1964.

[44] A.M. Faux, G.M. Lawrence, M.G. Wheaton, C.L. Jeffery C.L., and R.K. Grifitihs. Slippage in the NHS breast screening programme: An assessment of whether a three year screening round is being achieved. *Journal of Medical Screening*, 5:88–91, 1998.

[45] B.H. Fireman, C. Quesenberry, C. Somkin, A. Jacobson, D. Baer, D. West, A. Potosky, and M. Brown. Cost of care for cancer in a health maintenance organization. *Health Care Financing Review*, 18(4):51–76, 1997.

[46] J. Fischman. New-style mammograms detect cancer. So do the old. Either way you wait, 2001. `http://nl.newsbank.com/`, retrived 08/05/2004.

[47] A. Fisher. Customer relationship management: The personal touch. Special report: Understanding crm. *Financial Times*, November 26 2001.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[48] J.A. Fitzsimmons and M.J. Fitzsimmons. *Service Management:Operations, Strategy, and Information Technology*. McGraw-Hill, New York, 2000.

[49] G. Foster, M. Gupta, and L. Sjoblom. Customer profitability analysis: Challenges and new directions. *Journal of Cost Management*, 10(Spring):5–17, 1996.

[50] F.X. Frei, R. Kalakota, A.J. Leone, and L.M. Marx. Process variation as a determinant of bank performance: Evidence from the retail banking study. *Magement Science*, 45(9):1210–1220, 1999.

[51] S. Chiu Fridgeirsdottir, K. Pricing and marketing decisions in delay sensitive markets. Working Paper, Department of Management Science and Engineering, Stanford University, 2001.

[52] S.M. Gilbert and Z.K. Weng. Incentive effects favor nonconsolidating queues in a service system: The principal-agent perspective. *Management Science*, 44(12):1662–1668, 1998.

[53] R. Giltner and R. Ciolli. Rx for segmentation. *Banking Strategies*, 75(6):43–50, 1999.

[54] S. Grossman and O. Ciolli. An analysis of the principal-agent problem. *Econometrica*, 51:7–45, 1983.

[55] D. Gyrd-Hansen and J. Sogaard. Analyzing public preferences for cancer screening programs. *Health Economics*, 10:617–634, 2001.

[56] P. Heidelberger and P. Welch. Simulation run length control in the presence of an initial transient. *Operations Research*, 31:1109–1144, 1983.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[57] J.L. Heskett, T.O. Jones, G.W. Loveman, W.E. Sasser, and L.A. Schlesinger. Putting the service profit chain to work. *Harvard Business Review*, 72(2):164–175, 1994.

[58] J. Heskett and A. O'Dell, Shouldice Hospital Limited. Harvard Business School Case 683-068, 1983.

[59] B. Holmstrom and P. Milgrom. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics and Organizations*, 7:24–53, 1991.

[60] Imaginis. Breast cancer: Statistics on incidence, survival and screening, 2002. `http://imaginis.com/breasthealth/statistics.asp#1`, retrieved on 17/05/2004.

[61] National Cancer Institute. NCI statement on mammography screening: 1/31/2002 update, 2002. `http://www.nci.nih.gov/newscenter/content_nav.aspx?viewid=dcda4c3d-b4d2-4625-9d30-35721d6af68a`.

[62] C. Haigneré J-F. Mattei. Cancer: Une mobilisation nationale, tous ensemble, March 2003. `http://www.sante.gouv.fr/htm/dossiers/cancer/index2.htm`, retrieved on 17/05/2004.

[63] K. Josep and A. Thevaranjan. Monitoring and incentives in sales organizations: An agency-theoretic perspective. *Marketing Science*, 17(2):107–123, 1998.

[64] E. Kalai, M.T. Kamien, and M. Rubinovitch. Optimal service speeds in a competitive environment. *Management Science*, 38(8):1154–1163, 1992.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[65] A. Kan, I. Olivotto, L.W. Burhenne, E. Sickles, and A. Coldman. Standardized abnormal interpretation and cancer detection ratios to assess reading volume and reader performance in a breast screening programme. *Radiology*, 215:563–567, 2000.

[66] E.H. Kaplan, D.L. Craft, and L.M. Wein. Emergency response to a smallpox attack: The case for mass vaccination. *PNAS: Proc. National Academy of Sciences*, 99:10935–10940, 2002.

[67] U. Karmarkar. Will you survive the services revolution? *Harvard Business Review*, June:100–107, 2004.

[68] R.L.A. Kirch and M. Klein. Surveillance schedules for medical examinations. *Management Science*, 20(10):1403–1409, 1974.

[69] C. Klabunde, F. Bouchard, S. Taplin, A. Scharpantgen, and R. Ballard-Barbash. Quality assurance for screening mammography: an international comparison. *Journal of Epidemiology and Community Health*, 55:204–212, 2001.

[70] G. Kolata. 50 and ready for a colonoscopy? Doctors say wait is often long. *NY Times*, December 2003.

[71] R. Lal and V. Srinivasan. Compensation plans for single and multi-product salesforces: An application of the holmstrom-milgrom model. *Management Science*, 39(7):777–793, 1993.

[72] R. Lal and V. Srinivasan. Operations management and reengineering. *European Management Journal*, 16(3):306–316, 1998.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[73] S. Lapierre, D. Ratliff, and D. Goldsman. The delivery of preventive health services: A general model. Technical report, Georgia Institute of Technology, 1997.

[74] A.M. Law and W.D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, New York, 2000.

[75] H. Lee and W. Pierskalla. Mass screening models for contagious diseases with no latent period. *Operations Research*, 36(6):917–928, 1998.

[76] D.G. Luenberger. *Introduction to Dynamic Systems: Theory, Models, and Applications*. John Wiley and Sons, New-York, 1979.

[77] C. Ma. Health care payment systems: Cost and quality incentives. *Journal of Economics and MAnagement Strategy*, 3(1):93–112, 1994.

[78] H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations Research*, 38(5):870–883, 1990.

[79] J.A. Van Mieghem. Price and service discrimination in queueing systems: Incentive compatibility of gcμ scheduling. *Management Science*, 46(9):1249–1267, 2000.

[80] M. Montella, A. Crispo, G. Botti, M. De Marco, G. Bellis, G. Fabbrocini, M. Pizzorusso, M. Tamburini, and G. Daituo. An assessment of delays in obtaining definitve breast cancer treatment in southern italy. *Breast Cancer Research and Treatment*, 66:209–215, 2001.

[81] B. Morris and R. Johnston. Dealing with inherent variability:the difference between manufaturing and service? *Magement Science*, 7(4):13–22, 1986.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[82] M. Moss. Spotting breast cancer: Doctors are weak link. *NY Times*, June 27 2002.

[83] F.J. Mulhern. Customer profitability analysis: Measurement, concentration, and research directions. *Journal of Interactive Marketing*, 13(1):25–40, 1999.

[84] NCI. Canadian cancer statistics. Toronto, Canada, 2003. National Cancer Institute of Canada.

[85] W. Nie and D. L. Kellog. How professors of operations management view service operations? *Production and Operations Management*, 8(3):339–355, 1999.

[86] R. Niraj, M. Gupta, and N. Chakravarthi. Customer profitability in a supply chain. *Journal of Marketing*, 65:1–16, 2001.

[87] C.F. Nodine, H.L. Kundel, C. Mello-Thoms, S.P. Weinstein, S.G. Orel, D.C. Sllivan, and E.F. Conant. How experience and training influence mammography expertise. *Academic Radiology*, 6(10):575–585, 1999.

[88] P. Nutting, D. Iverson N. Calonge, and L. Green. The danger of applying uniform clinical policies across populations: The case of breast cancer in american indians. *American Journal of Public Health*, 84(10):1634–1636, 1994.

[89] World Health Organization. Screening for various cancers, 2003. `http://www.who.int/cancer/detection/breastcancer/en/`, retrieved on 17/05/2004.

[90] S. Ozekici and S. Pliska. Optimal scheduling of inspections: A delayed Markov model with false positives and negatives. *Operations Research*, 39(2):261–273, 1991.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[91] R. Pijnappel, M. van den Donk, R. Holland, W.P. Mali, J.L. Peterse, J.H.C.L. Hendriks, and P.H.M. Peeters. Diagnostic accuracy of different strategies of image guided breast intervention in cases of nonpalpable breast lesions. *British Journal of Cancer*, 90:595–600, 2004.

[92] E. Pinker and R. Shumsky. The efficiency-quality trade-off of cross-trained workers. *Manufacturing and Service Operations Management*, 2(1):32–48, 2000.

[93] E. Plambeck and S. Zenios. Performance based incentives in a dynamic principal-agent model. *Manufacturing and Service Operations Management*, 2(3):240–263, 2000.

[94] Canadiaon Population and Public Health Branch. Organized breast cancer screenning programs in canada, 1997-1998 report, 1998. `http://www.hc-sc.gc.ca/pphb-dgspsp/publicat/obcsp-podcs98/index.html`, retrieved on 17/05/2004.

[95] S.G. Powell. Specialization, teamwork, and production efficiency. *International Journal of Production Economics*, 67:205–218, 2000.

[96] M.A. Richards, A.M. Westcombe, S.B. Love, P. Littlejohns, and A.J. Ramirez. Influence of delay on survival in patients with breast cancer: A systematic review. *The Lancet*, 353:1119–1126, 1999.

[97] P. Salzman, K. Kerlikowske, and K. Phillips K. Cost-effectiveness of extending screening mammography guidelines to include women 40 to 49 years of age. *Annals of Internal Medicine*, 127:955–965, 1997.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[98] R. Saywell, V. Champion, T. Zollinger, M. Maraj, C. Skinner, K. Zoppi, and C. Muegge. The cost effectiveness of 5 interventions to increase mammography adherence in a managed care population. *The American Journal of Managed Care*, 9(1):33–44, 2003.

[99] M. Schwartz. A mathematical model used to analyze breast cancer screening strategies. *Operations Research*, 26(6):937–955, 1978.

[100] R. Shumsky and E. Pinker. Gatekeepers and referrals in services. *Management Science*, 49(7):839–856, 2001.

[101] E.A. Sickles, D.E. Wolverton, and K.E. Dee. Performance parameters for screening and diagnostic mammography: Specialist and general radiologists. *Radiology*, 224(3):861–869, 202.

[102] K. Storbacka. Segmentation based on customer profitability- A retrospective analysis of retail bank customer bases. *Journal of Marketing Management*, 13:479–492, 1997.

[103] X. Su and S.A. Zenios. Allocation of kidneys to autonomous transplant candidates: A sequential stochastic assignment model, 2004. submitted to Operations Research.

[104] P. Thongsuksai, V. Chongsuvivatwong, and H. Sriplung. Delay in breast cancer care: A study in Thai women. *Medical Care*, 38(1):108–114, 2000.

[105] H. Thornton, A. Edwards, and M Baum. Women need better information about routine mammography. *British Medical Journal*, 327:101–103, 2003.

[106] J. Tirole. *The Theory of Industrial Organization*. The MIT Press, Cambridge, Massachusetts, 1988.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[107] National Health Service (UK). Cancer screening programmes, 2002. http://www.cancerscreening.nhs.uk/breastscreen/index.html#how-org, retrieved on 17/05/2004.

[108] U.S. General Accounting Office. Report to the chairman: Mammography capacity generally exists to deliver services, February 19 2002.

[109] U.S. Health Program Office of Technology Assessment. Screening mammography in primary care settings: Implications for cost, access and quality, 1991. J. Wagner, (study director).

[110] H.M. Verkooijen, P.H. Peeters, E. Buskens, V.C. Koot, B. Rinkes, and T.J. van Vroonhiven. Diagnostic accuracy of large-core needle biopsy for nonpalpa breast disease: A meta analysis. *British journal of Cancer*, 82(5):1017–1022, 2000.

[111] V. Verter and S. Lapierre. Location of preventive health care facilities. *Annals of Operations Research*, 110:121–130, 2002.

[112] J. Voelker and W. Pierskalla. Test selection for a mass screening program. *Naval Research Logistics Quarterly*, 27:43–56, 1980.

[113] S.D. Walter and N.E. Day. Estimating the duration of a pre-clinical disease state using screening data. *American Journal of Epidemiology*, 118:865–886, 1983.

[114] J. Xu, M.R. Fagerstrom, and P. Prorok. Estimation of post-lead-time survival under dependence between lead-time and post-lead-time survival. *Statistics in Medicine*, 18(100):155–162, 1999.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[115] M. Zelen. Optimal scheduling of examinations for the early detection of disease. *Biometrika*, 80(2):279–293, 1993.

[116] S.A. Zenios and P.C. Fuloria. Managing the delivery of dialysis therapy: A multiclass fluid model. *Management Science*, 46:1317–1336, 2000.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.