

UNIVERSITY OF CALIFORNIA
Los Angeles

**Three Essays in Healthcare Operations
Management**

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Management

by

Sarang Deo

2007

UMI Number: 3299545

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 3299545

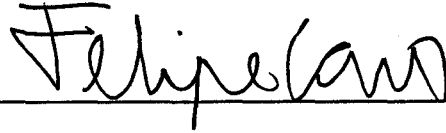
Copyright 2008 by ProQuest LLC.

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

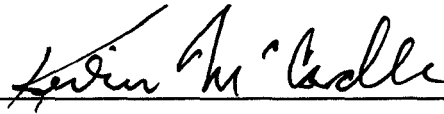
ProQuest LLC
789 E. Eisenhower Parkway
PO Box 1346
Ann Arbor, MI 48106-1346

© Copyright by
Sarang Deo
2007

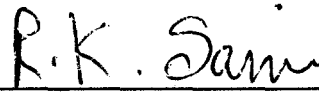
The dissertation of Sarang Deo is approved.



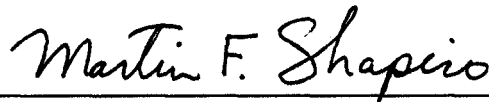
Felipe Caro



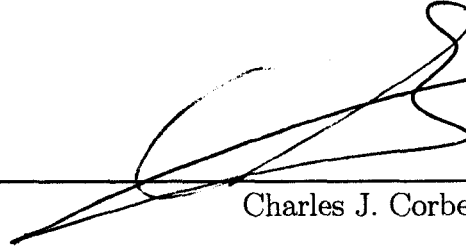
Kevin McCardle



Rakesh Sarin



Martin F. Shapiro



Charles J. Corbett, Committee Chair

University of California, Los Angeles

2007

*To my grandfather,
Late P. R. Sangitrao*

TABLE OF CONTENTS

1 Cournot competition under yield uncertainty: The case of the U.S. influenza vaccine market	1
1.1 Introduction	1
1.2 Background on influenza	4
1.3 Literature Review	6
1.4 Model Formulation	9
1.4.1 Modeling the supply	9
1.4.2 Modeling the demand	10
1.4.3 Modeling the market	10
1.5 Equilibrium of the two-stage game	11
1.5.1 Post entry competition under yield uncertainty	11
1.5.2 Entry game	12
1.5.3 Total vaccine supply	14
1.5.4 Social welfare	15
1.5.5 First best solution	16
1.5.6 Seond-best solution	18
1.6 Application to the U.S. influenza vaccine market	20
1.6.1 Model calibration	20
1.6.2 Analysis of market equilibrium	23
1.6.3 Evaluation of demand-side and supply-side policies	24
1.6.4 Sensitivity analysis	27

1.7	Concluding remarks	28
1.8	Proofs	29
2	Rationing of HIV treatment in resource-constrained settings un-	
	der supply uncertainty	35
2.1	Introduction	35
2.2	Background	38
2.3	Literature Review	40
2.4	Model Formulation	43
2.4.1	Drug supply	43
2.4.2	Patients	44
2.4.3	System dynamics	45
2.4.4	Objective function	46
2.5	Optimal policy	47
2.6	Model reformulation	49
2.6.1	Resource-constrained condition	49
2.6.2	Two-period model	50
2.6.3	Prioritization of current patients	53
2.6.4	Infinite horizon	56
2.7	Enrollment heuristics	57
2.7.1	Safety-stock policy	57
2.7.2	Myopic policy	58
2.8	Numerical illustrations	59

2.8.1	Setting parameter values	60
2.8.2	Results	62
2.9	Conclusion and future research	64
2.10	Proofs	67
2.11	Appendix: A conceptual model of HIV treatment scale-up in resource- constrained setting	76
2.11.1	Conceptual framework	77
2.11.2	Existing Literature	77
2.11.3	Agenda for future research	82
3	Organizational determinants of performance in quality improve- ment collaboratives	87
3.1	Introduction	87
3.2	Methods	90
3.2.1	Context	90
3.2.2	Data	91
3.2.3	Measures	92
3.3	Results	95
3.3.1	Descriptive Statistics	95
3.3.2	Number of interventions	96
3.3.3	Cross-departmental nature of interventions	96
3.3.4	Implementation success	97
3.4	Discussion and Limitations	97

References 103

LIST OF FIGURES

1.1	Equilibrium and socially optimal number of firms as a function of the coefficient of variation, δ , for different demand side interventions	25
1.2	Impact of supply-side and demand-side interventions for different values of the coefficient of variation, δ	26
2.1	Performance of heuristics as a function of the coefficient of variation in the supply distribution	63
2.2	Minimum gap across all heuristics as a function of the coefficient of variation in the supply distribution	66
2.3	Conceptual model of HAART scale-up	78
2.4	Treatment rationing at the clinic level	85

LIST OF TABLES

1.1	Parameter values for the U.S. influenza vaccine market	21
1.2	Quantity of influenza vaccine produced and distributed (million doses)	22
1.3	Equilibrium and second-best solutions for the influenza vaccine case	24
1.4	Equilibrium and socially optimal number of firms for different values of f and δ	27
2.1	Quality of Life estimates	61
2.2	Heuristic with the best performance for different values of C.V. and s_2 . ME denotes Maximal Enrollment Policy and SS (a) denotes Safety-stock Policy with a stock of “a” months	65
3.1	Comparison of characteristics of study clinics with all Title III clinics	95
3.2	Descriptive statistics	99
3.3	Negative binomial regression explaining the number of unique interventions (N = 42, Log Likelihood = 3960, Dispersion = 0.18) .	100
3.4	OLS Regression explaining the cross-departmental nature of interventions (N = 42, R2 = 0.55, Adj. R2 = 0.34, p-value = 0.02) . .	101
3.5	OLS regression explaining the implementation success (overall clinic rating) at clinics (N = 42, R2 = 0.57, Adj. R2 = 0.49, p-value ; 0.01)	102

ACKNOWLEDGMENTS

A number of persons have enriched my experience in the PhD program and made this dissertation possible. I owe invaluable debt to Charles J. Corbett, who agreed to be my dissertation advisor much earlier in the PhD program and has been a true mentor ever since. I am truly grateful to him for allowing me the freedom to choose a research agenda of my interest. His judgement in choosing the right questions and genuine curiosity and inquiry in pursuing them has been a constant guiding force in my research. During this journey, I hope to have imbibed these values in addition to the usual tricks and tools of the trade. I was equally fortunate to have a very helpful dissertation committee. Kevin McCardle and Felipe Caro provided numerous suggestions to tackle the technically difficult aspects on my analysis. Martin Shapiro helped me understand various healthcare related aspects of my dissertation and sensitized me to the broader context of my research.

I also benefitted greatly from my interactions with the other DOTM faculty members. On numerous occasions, they were kind enough to engage me in discussions though they did not result in any tangible outcome for them. Christopher Tang provided me with early advice on conducting scholarly research and publishing it. I am truly grateful to him for the opportunities. I am also grateful to Rakesh Sarin and Uday Karmarkar for their constant encouragement and kind concern for me throughout the PhD program. I would particularly like to thank Art Geoffrion for going out of his way on multiple occasions to encourage me. His thoughts have greatly influenced my research philosophy.

I want to thank a number of persons outside the Anderson School for providing me with substantial inputs for the dissertation. Paul Cleary at Yale School of Public Health and Bruce Landon, Ira Wilson and Keith McInnes at Harvard

School of Public Health shared data on quality improvement collaboratives and provided with many critical suggestions. Yasmin Chandani and her team at JSI shared their thoughts and insights on the supply chain management of antiretroviral drugs. Agnes Fiamma and Thomas Coates shared their knowledge about the complexities of HIV treatment in resource-constrained settings. Tony Jackson and his colleagues introduced me to the various challenges in the influenza vaccine market.

I have been fortunate to have worked with many worthy colleagues in the PhD program. I learnt a lot from my early conversations with past DOTM colleagues: Aydin, Ram, Murat, Thanos, Barbara and Deming. Soo-Haeng and Rui have been a constant source of encouragement and support; this would have been a very lonely voyage without their companionship. I also want to thank Steve, Suresh and Ravi for kindly listening to all the unwarranted advice that I have thrown at them and for telling me that it was actually useful.

In closing, I want to express my gratitude toward my parents and my wife's parents for their unfailing support during the PhD program; though it meant that they could not meet us as much as they would have liked. Spruha and Pranjala made this otherwise dry journey immensely enjoyable and fun-filled. Finally, my wife Prachi has been an exemplary figure of love, courage and sacrifice even as she put my PhD ahead of her own personal priorities. All this would have been a mere dream without her.

VITA

1978	Born, Nagpur, India
1999	B. Tech. in Chemical Engineering, IIT Bombay, Mumbai, India
2001	PGDM, IIM Ahmedabad, Ahmedabad, India
2001-2003	Business Analyst, Accenture, Mumbai, India
2003-2007	Graduate Student Researcher, UCLA Anderson School of Management, Los Angeles, California

PUBLICATIONS

C. S. Tang and S. Deo (2007). Rental price and rental duration under retail competition. *European Journal of Operational Research*, forthcoming.

S. Deo (2007). Cournot competition under yield uncertainty: The case of the U.S. influenza vaccine market (Extended Abstract). *Manufacturing & Service Operations Management*, 9, 114-117.

S. Deo and C. S. Tang (2005). Optimal procurement, disposal and pricing policies for managing rental goods. *International Transactions in Operations Research*, 12, 595-629.

ABSTRACT OF THE DISSERTATION

**Three Essays in Healthcare Operations
Management**

by

Sarang Deo

Doctor of Philosophy in Management

University of California, Los Angeles, 2007

Professor Charles J. Corbett, Chair

Operations Management can contribute greatly in understanding the various challenges currently faced by the healthcare sector worldwide including the U.S. In the three essays of this dissertation, I investigate three such challenges in diverse settings that require diverse methods of analysis.

In Chapter 1, I study the degree of concentration in the U.S. influenza vaccine market and its impact on the supply of vaccines. I show that interaction between yield uncertainty in the production process and firms' strategic behavior can contribute to a high degree of concentration in an industry and a reduction in the industry output and the expected consumer surplus in equilibrium. I analyze the social trade-off between risk pooling (by diversification of supply) and economies of scale (by avoiding duplication of fixed costs). Finally, I conduct numerical analysis with realistic parameters to assess the impact of yield uncertainty on the U.S. influenza vaccine market.

Chapter 2 presents a prescriptive model for rationing treatment for HIV+ patients in resource-constrained regions such as Asia and Africa. I consider an individual clinic facing an uncertain supply of drugs resulting from inadequate

supply management skills and a weak infrastructure. I model the clinic's trade-off between improving access to treatment for new patients and providing uninterrupted treatment for current patients and derive its optimal treatment rationing policy using stochastic dynamic programming. I show that under certain conditions the optimal policy coincides with the clinically preferred policy of prioritizing previously enrolled patients. Numerical illustrations suggest that the performance of enrollment policies used in practice can be substantially suboptimal.

In Chapter 3, I examine the relationship between organizational factors and quality of care in healthcare organizations. Using the data from a QIC conducted in Ryan White CARE Act funded clinics in the U.S. and an accompanying survey of clinicians, I find that organizations with more open culture, a higher focus on QI and multidisciplinary teams attempted higher number of interventions and attempted interventions that were more cross-departmental in nature. Controlling for number of interventions and mean importance rating of interventions, implementation success was significantly associated with cross-departmental nature of interventions, fraction of interventions repeated and evaluated and presence of multidisciplinary teams. These results provide one potential explanation for the heterogeneity of implementation performance across healthcare organizations.

CHAPTER 1

Cournot competition under yield uncertainty: The case of the U.S. influenza vaccine market

1.1 Introduction

The number of firms producing influenza vaccine for the U.S. has been declining steadily in the recent past. Two manufacturers, Sanofi Pasteur and Chiron, had been supplying all injectible vaccine since 2002, down from around five in the 1990s and more than a dozen in the 1970s (Brown, 2004), and a third manufacturer (Glaxo-SmithKline) entered the U.S. market after the supply crisis in the 2004-2005 season. The non-injectible vaccine “Flumist” still only accounts for 2% of the total market. Articles in the popular press have blamed low market price, insufficient incentives and uncertain demand for this high degree of concentration and for the frequent vaccine shortages observed in the recent years (Forbes, 2004; Newsweek, 2004; Time, 2004). However, the existing evidence does not conclusively support these claims. The price for influenza vaccine, unlike other vaccines, is not controlled by the government (Danzon et al., 2004) and has increased from \$2 to around \$8 per dose in the past five years (Forbes, 2004). Also, the demand for influenza vaccine has been increasing steadily over the past decade as can be seen from the immunization rates (O’Mara, 2003). Other possible reasons for exit of firms include mergers and acquisitions, plant closures resulting from

inability to meet stringent regulatory standards and the market for vaccines being less profitable and much smaller compared to that for other pharmaceutical products. Danzon et al. (2005) argue that high country-specific regulatory cost is one of the key factors that would drive the long term equilibrium in the U.S. flu vaccine market to be characterized by one or two suppliers.

While these hypotheses might provide some explanation for the reduction in equilibrium number of firms over time, they do not address the question of whether this equilibrium is socially optimal. The American Antitrust Institute has argued for more government involvement in order to build surge capacity, claiming that the free market process is not working satisfactorily (American Antitrust Institute, 2004). Economic theory, on the other hand, predicts that an oligopolistic market with unregulated but costly entry, such as the influenza vaccine market, will experience *excess* entry and oversupply compared to the social optimum.

One additional characteristic of the influenza vaccine market that further complicates the situation has received considerable attention recently in the trade literature but not yet in the academic literature: the yield uncertainty in the production process. The manufacturing process for influenza vaccine involves growing the virus in chicken eggs and later extracting, purifying, inactivating and packaging the vaccine (Gerdil, 2002). Due to the inherent uncertainty regarding the growth characteristics of the viral strains, the quantity of vaccine that can be obtained per chicken egg is uncertain (National Influenza Vaccine Summit, 2006; National Vaccine Advisory Committee, 2003; Powermed, 2005; Gerson Lehrman Group, 2005; GAO, 2001). The magnitude of the challenge posed by the yield uncertainty in influenza vaccine production is illustrated by quotes such as "...the yield of candidate strains sometimes is not as high as desired which results in fewer

doses, or strains may take additional time to obtain optimal yields, resulting in delays in the availability of vaccine” (National Vaccine Advisory Committee, 2003) and “The first [major factor contributing to the delay in vaccine availability in 2001] was that two manufacturers had unanticipated problems growing one of the two new influenza strains introduced into the vaccine for 2000-01” (GAO, 2001).

We model the effect of yield uncertainty on the influenza vaccine market using using a two-stage game of oligopolistic competition. In the first stage, firms simultaneously decide whether to enter the market by incurring a fixed cost of entry. In the second stage, each entering firm selects the target production quantity. Then each firm’s yield is realized, actual quantity produced is brought to the market and price emerges according to the traditional model of quantity (Cournot) competition. We employ this model to answer the following specific questions:

- (i) What is the impact of yield uncertainty on the quantity produced by each firm, total output of the industry and total number of firms in the market under competitive equilibrium i.e., without any intervention by the social planner?
- (ii) What is the impact of yield uncertainty on consumer welfare? How should society trade off the risk pooling value of supply diversity against the efficiency of having a single source?
- (iii) What conditions result in less entry and lower production as compared to the social optimum?
- (iv) Which regulatory interventions (supply-side or demand-side) are more effective and under what conditions?

Although originally inspired by the influenza vaccine market, our model applies to several other industries with yield uncertainty including bio-pharmaceuticals, semiconductor manufacturing and agriculture. The rest of the chapter is structured as follows. Section 2.2 contains background information on the influenza vaccine market. Section 2.3 reviews related literature from operations management (OM), economics and public health economics on yield uncertainty, competition and vaccinations respectively. In section 1.4, we present the basic model. Section 1.5 outlines the main results concerning the competitive equilibrium and socially optimal solutions while section 1.6 discusses numerical experiments with data pertinent to the U.S. influenza vaccine market. We provide some concluding remarks in section 2.9.

1.2 Background on influenza

The 20th century has seen several influenza pandemics with the most severe in 1918 causing close to 20 million deaths worldwide (WHO, 2002). Every year, 10-20% of the population gets influenza and nearly 36,000 people die of the resulting complications in the U.S. alone (Thomspon et al., 2003). Influenza and resulting complications are the sixth largest cause of death in the U.S. (Martone, 2002) with estimated annual costs of \$11-18 billion (WHO, 2002). The most important reason for the persistence of influenza epidemics is the uncanny ability of the virus to continuously adapt itself every season, a phenomenon called the antigenic drift. As a result, the composition of the vaccine has to be reviewed every year and can undergo frequent changes. The recommendations for vaccine composition are made every year by the WHO, in February for the northern hemisphere (for the flu season lasting from October to February of next year) and in August for the southern hemisphere (for the flu season lasting from May to August of

next year). In addition, periodically, antigenic shift, in which genetic material of different strains of virus are recombined results in pandemics. (The current public focus on avian flu stems from the anticipation of a similar pandemic due to antigenic shift in the H5N1 strain.)

The challenges faced by the influenza vaccination system in the U.S. can be categorized broadly into demand-side challenges and supply-side challenges. Supply-side challenges in an influenza supply chain arise primarily from the combination of long production lead-time, short immunization season and frequent changes in the vaccine composition. Production of influenza vaccine involves a long and complex biological process. The virus is grown in chicken eggs and later inactivated, purified and processed to manufacture the vaccine (Gerdil, 2002). The entire process takes six to eight months. Hence the manufacturers have to decide on the production quantity long before complete information about demand is available. In addition, due to the continuous change in the constituent strains, unused vaccine from the previous season cannot be utilized this season. Williams (2005) and Yadav (2005) provide a detailed discussion of these distinguishing features of the influenza vaccine supply chain and suggest various improvement opportunities.

These challenges are compounded by the high yield uncertainty in the production process, discussed earlier. Due to the long lead-time it is impossible to take any recourse later if faced with a particularly virulent strain of the virus or higher than expected demand or lower than expected yield (Danzon et al., 2005). The effects of uncertain yield on the supply of vaccine are clearly evident from the recent U.S. experience. In 2004-05, Chiron's vaccine manufacturing plant in the U.K. was shut down by regulators due to bacterial contamination resulting in a reduction of total supply to the US market by about 50%, causing unprecedented

shortages. The U.S. had also faced considerable shortages in the 2003-04 and 2001-02 flu seasons due to an early onset of the epidemic and unexpected delays in the production process respectively.

On the demand side, immunization rates are lower than is socially optimal. Immunization of elderly citizens has been shown to be cost beneficial (Nicol et al., 1998) and is recommended. However, in the U.S., as recently as 2002-03, the immunization rate was only around 60% among the elderly and even lower in other population groups (O'Mara et al., 2003). Most other countries have even lower immunization rates. Key factors cited by elderly people for low immunization rates include perceived good health, lack of advice from medical personnel and negative views on efficacy and safety of the vaccine (Evans and Watson, 2003). Other probable factors include lack of health insurance and high cost of vaccination. More generally, public health economists have long argued that individuals fail to internalize the positive externalities arising from vaccination, resulting in lower rates of immunization than is socially optimal. See Philipson (2003) for more details.

1.3 Literature Review

This work draws on and contributes to three distinct streams of literature. First, we extend the literature on the stochastically proportional yield model in operations management (OM) to a competitive setting. Second, we show how yield uncertainty affects the existing results in the oligopoly literature that discusses various models of competition with endogenous entry. Third, we build on the public health economics literature to which we also contribute by simultaneously studying supply and demand side factors traditionally analyzed in isolation.

The model of yield uncertainty employed here has been widely used in the OM literature and is referred to as the stochastically proportional yield model. However, most of this OM literature considers the impact of yield uncertainty on either the production planning decisions of a single firm or procurement decisions of a single firm buying from multiple non-competitive suppliers with uncertain yields (Yano and Lee, 1995). Henig and Gerchak (1990), in a single firm model, show using an approximation that a higher yield variance results in lower optimal target production quantity. The first part of our analysis shows that this result extends to a competitive setting. Anupindi and Akella (1993) and Gerchak and Parlar (1990) discuss the value of diversification in the case of a given number of unreliable suppliers. Recently Federgruen and Yang (2005) and Dada et al. (2007) consider the problem of procurement from multiple suppliers with differing reliability and cost. In the second part of our analysis, the number of suppliers is endogenously determined through an entry game.

Carr et al. (2005) consider a competitive model of demand and capacity uncertainty. They show that a reduction in yield uncertainty can reduce the firm's profit, if process improvement leads to an effective over-capacity in the industry resulting in stiffer price competition. One of our results is consistent with this, but in our model the increase in quantity produced is a rational decision rather than a direct outcome as in Carr et al. (2005). Moreover, Carr et al. (2005) do not consider entry decisions and their main focus is on studying the interaction between process improvement and competitive forces. In short, we contribute to the OM literature by studying the impact of yield uncertainty on strategic decisions of the firm such as entry and production quantity.

Other applications of OR/OM models to the influenza context have focused on control (Finkelstein et al., 1981) and management (Longini et al., 1977) of

epidemics and on strain selection (Wu et al., 2005). Williams (2005) and Yadav (2005) provide a detailed review of the structure of the influenza vaccine supply chain and propose that an “information hub” and “government buy-back” scheme would improve its performance.

Vives (1999) provides a detailed account of the vast literature related to the Cournot (1838) model of oligopolistic competition. Part of this essay focuses on the question of firm entry in the context of oligopoly. Mankiw and Whinston (1986) compare the number of firms in the “free-entry equilibrium” with the number of firms that a social planner would choose. They show that under a decreasing inverse demand function and convex cost structure entry of an additional firm reduces the output of incumbents. Ignoring the integer constraint on the number of firms, this effect is sufficient to ensure that there will always be excess entry relative to the social optimum. Von Weizsäcker (1980) reaches similar conclusions using a linear inverse demand function and numerical examples. We show that in contrast to Mankiw and Whinston (1986), adding yield uncertainty leads to less entry than optimal in a homogeneous goods market with business stealing effect. Thus, we contribute to the literature on oligopoly by including yield uncertainty, so far not considered in the context of Cournot competition. Traditionally, the uncertainty studied in Cournot models has been related to demand or cost or that resulting from players’ private information (Leland, 1972; Klemperer and Meyer, 1986). However, yield uncertainty is fundamentally different from demand or cost uncertainty as it relates to a decision variable (production quantity) rather than an exogenous parameter; this has some important consequences.

Our model of consumer demand for vaccines is closely related to Brito et al. (1991), where consumers differ in the cost of vaccination, and a consumer’s likelihood of contagion from the unvaccinated population depends on the number

of unvaccinated individuals. This interaction between individually rational decisions and epidemiological dynamics has recently received attention in the public health economics literature, reviewed by Philipson (2003).

1.4 Model Formulation

1.4.1 Modeling the supply

We assume that the industry consists of n manufacturing firms denoted by $i \in \{1, 2, \dots, n\}$ possessing identical manufacturing process. If \bar{q}_i is the production quantity targeted by firm i (as reflected by the total number of chicken eggs chosen), then the actual quantity produced is given by $q_i = \alpha_i \bar{q}_i$, where α_i is a random variable reflecting the random yield per egg for firm i . Since the yield uncertainty results in a random proportion of the target quantity being produced, this multiplicative model is also known as the stochastically proportional yield model. Yano and Lee (1995) mention that this model is appropriate when relatively large batch sizes are used, when the variation of the batch size from production run to production run tends to be small or when the yield losses might be relatively predictable for any particular set of conditions, but the conditions are not predictable. All these criteria are met in the case of influenza vaccine production. Since all firms have the same production technology, α_i is identically distributed for all firms. In addition, we assume that the α_i are independent (GAO, 2001). Let $\mu = E[\alpha_i]$ and $\sigma^2 = \text{Var}[\alpha_i] \quad \forall i$. We include two marginal costs: (i) c_1 per unit target quantity and (ii) c_2 per unit actually produced. In our context, the first cost is driven by the number of chicken eggs and the second cost corresponds to the cost of bottling and packaging the actual vaccine produced.

1.4.2 Modeling the demand

We assume a linear inverse demand function given by $p = \hat{a}(e) - bq$, where e is a random variable denoting the efficacy of the vaccine. The reservation price \hat{a} depends on how effective the vaccine is against the circulating virus strain in the coming season and allows for the fact that the customers would be willing to pay more for a more effective vaccine. While selecting the target production quantity, the firm does not know what the actual efficacy will be but does know the underlying distribution. We assume that e is independent of the yield uncertainty variable α_i , i.e., the efficacy of the viral strains selected for the vaccine is independent of the production characteristics for those strains. In some seasons due to the antigenic drift the strains in the vaccine will usually be less effective, but that does not appear to be related to the production yield that will be obtained with these strains.

1.4.3 Modeling the market

We model competition among firms as a two-stage game. First, the firms simultaneously decide whether to enter the industry. Each entering firm incurs a fixed cost f . The manufacturers for influenza vaccine decide on production quantities six to eight months before the onset of the flu season. Hence the Cournot (1838) model is reasonable for the competition among the entering firms in the second stage. Each firm sets its target production quantity, \bar{q}_i . After that, each firm's yield α_i is realized, total production $q = \sum_i \alpha_i \bar{q}_i$ occurs and price p is set according to the above inverse demand function.

We solve this two-stage game using backward induction. We first solve the second stage game for a given number of firms in the industry and derive the equilibrium target production quantities and profits as a function of this number.

Then we analyze the first stage game to find the equilibrium number of firms in the market.

1.5 Equilibrium of the two-stage game

1.5.1 Post entry competition under yield uncertainty

In the second stage, given that there are n firms in the industry, each firm decides a target quantity \bar{q}_i at a cost of $c_1\bar{q}_i$. The uncertainty is resolved during the production process and $q_i = \alpha_i\bar{q}_i$ is the actual quantity produced at a cost of c_2q_i . The market price is given by the inverse demand function $p = \hat{a}(e) - b\left(\sum_{j=1}^n q_j\right)$ and the expected profit of the i^{th} firm is given by $\Pi_i(q_i) = E_{e,\alpha_i} \left[\left(\hat{a}(e) - b\left(\sum_{j=1}^n q_j\right) \right) q_i - c_1\bar{q}_i - c_2q_i \right]$. Substituting for $q_i = \alpha_i\bar{q}_i$ in this expression, defining $c = \frac{c_1}{\mu} + c_2$ and noting that e and α_i are independent, we obtain

$$\Pi_i(\bar{q}_i) = E \left[\left(a - b \left(\sum_{j=1}^n \alpha_j \bar{q}_j \right) \right) \alpha_i \bar{q}_i - c \bar{q}_i \right] \quad (1.1)$$

where $a = E_e [\hat{a}(e)]$. Let Π_i^* denote the maximum profit of the i^{th} firm. Then, writing \bar{q}_{-i} to denote the decisions of all firms other than i , the decision problem of the i^{th} firm is

$$\Pi_i^* = \max_{\bar{q}_i \geq 0} \Pi_i(\bar{q}_i, \bar{q}_{-i}^*) \quad (1.2)$$

The equilibrium is found by solving the following set of equations:

$$\frac{\partial \Pi_i(\bar{q}_i)}{\partial \bar{q}_i} \Big|_{\bar{q}_i = \bar{q}_i^*} = (a - c)E[\alpha_i] - 2b\bar{q}_i^* E[\alpha_i^2] - bE \left[\alpha_i \sum_{i \neq j} \alpha_j \bar{q}_j^* \right] = 0 \quad \forall i \quad (1.3)$$

Since $E[\alpha_i] = \mu$ and $\text{Var}[\alpha_i] = \sigma^2 \quad \forall i$, the above system of equations has a unique solution that is symmetric. Define the coefficient of variation $\delta = \frac{\sigma}{\mu}$. Since we are primarily interested in analyzing the impact of yield uncertainty on market and socially optimal solutions, we keep μ constant and only analyze

the changes in σ . Hence we can express all our results in terms of δ rather than σ which simplifies the exposition considerably. The unique equilibrium of the Cournot game is straightforward:

Lemma 1. *The second stage Cournot game with yield uncertainty has a unique equilibrium in which:*

- (i) *The target quantity of each firm is given by $\bar{q}_i^* = \frac{(\alpha-c)\mu}{b[(n+1)\mu^2+2\sigma^2]} \quad \forall i$.*
- (ii) *The expected quantity produced by each firm is given by $E[q_i^*] = \mu\bar{q}_i^* = \frac{\alpha-c}{b(n+1+2\delta^2)} \quad \forall i$.*
- (iii) *For given n , each firm's target quantity and expected quantity is decreasing in the yield uncertainty as measured by δ .*
- (iv) *The expected profit of each firm is given by $\Pi_i^*(n) = \frac{(\alpha-c)^2(\delta^2+1)}{b(n+1+2\delta^2)^2} \quad \forall i$.*
- (v) *For given n , each firm's expected profit is first increasing and then decreasing in δ .*

All proofs are provided in Appendix. Note that in the absence of any uncertainty, i.e., $\delta = 0$, the expected quantity produced reduces to $q_i^* = \frac{\alpha-c}{b(n+1)}$, while expected profit reduces to $\Pi_i^* = \frac{(\alpha-c)^2}{b(n+1)^2}$: both familiar from Cournot competition without yield uncertainty. Moreover, for given n , higher yield uncertainty leads each firm to produce lower expected quantity.

1.5.2 Entry game

Next, we focus on the first stage of the game. We assume that there is a large population of identical potential entrants. Each of these potential entrants has a reservation profit level of zero. All firms simultaneously decide whether to

enter the market or not. We are not interested in which specific firms out of the potential population enter, but only in the equilibrium number of entrants. For $n^* \in \mathbb{N}$ to be the equilibrium number of firms in the industry, we must have $\Pi_i^*(n^*) \geq f$ and $\Pi_i^*(n^* + 1) \leq f$, as otherwise entering firms are losing money or earning sufficient profits to attract additional entrants. Temporarily relaxing the integer constraint, the equilibrium number of entrants, $x^* \in \mathbb{R}_+$ satisfies $\Pi_i^*(x^*) = f$. Let $n_u^* \in \mathbb{N}$ denote the equilibrium number of firms under yield uncertainty and $n_d^* \in \mathbb{N}$ be the corresponding equilibrium number of firms for the deterministic case. Similarly, let $x_u^*, x_d^* \in \mathbb{R}_+$ be the respective equilibrium numbers after relaxing the integer constraints. Let $\lfloor n \rfloor$ denote the largest integer less than or equal to n .

Lemma 2. *The number of firms in the industry at equilibrium with and without yield uncertainty is given by $n_u^* = \left\lfloor \left(\frac{a-c}{\sqrt{bf}} \sqrt{1 + \delta^2} \right) - (1 + 2\delta^2) \right\rfloor$ and $n_d^* = \left\lfloor \left(\frac{a-c}{\sqrt{bf}} \right) - 1 \right\rfloor$ respectively.*

We now use these results to determine the impact of yield uncertainty on the equilibrium number of firms n_u^* using the deterministic equilibrium number n_d^* as benchmark. One might expect that uncertainty always (weakly) reduces the number of entrants in equilibrium, but the following proposition shows that that is not necessarily true.

Proposition 1. *The equilibrium number of firms under uncertainty (n_u^*) and in the deterministic case (n_d^*) satisfy (i) $n_u^* \leq n_d^*$ if $\left\{ \frac{a-c}{\sqrt{bf}} > 4 \text{ and } \delta \geq \delta_1^* \right\}$ or $\frac{a-c}{\sqrt{bf}} \leq 4$ and (ii) $n_u^* \geq n_d^*$ if $\left\{ \frac{a-c}{\sqrt{bf}} > 4 \text{ and } \delta \leq \delta_1^* \right\}$, where $\delta_1^* \triangleq \sqrt{\left(\frac{a-c}{2\sqrt{bf}} - 1 \right)^2 - 1}$.*

Recall that $\left\lfloor \left(\frac{a-c}{\sqrt{bf}} \right) - 1 \right\rfloor = n_d^*$. Thus, if the industry can support at most three firms at equilibrium without uncertainty ($n_d^* \leq 3$), then any amount of yield uncertainty (weakly) reduces the number of firms at equilibrium. However, if the

industry can support three or more firms at equilibrium without uncertainty, then the equilibrium number of firms is (weakly) greater than n_d^* if the uncertainty is lower or equal to a certain threshold. In other words, large uncertainty always results in exit of firms from the industry relative to the deterministic case, while limited uncertainty can actually cause more firms to enter in certain cases.

To understand this, observe that yield uncertainty has two effects on firms' expected profits. First, yield uncertainty reduces the expected quantity produced by each firm. This has a negative effect on the expected profit. A secondary effect is that yield uncertainty increases the market price by reducing output; this affects all the units and not just the marginal units, which has a positive effect on the expected profit. For small levels of uncertainty, the positive effect can dominate the negative effect and hence cause a net increase in the expected profit, thus attracting new entrants. However, for large uncertainty the net effect is always negative and hence lowers the equilibrium number of firms. Moreover, from Proposition 1, it follows that the threshold level of uncertainty δ_1^* is non-decreasing in the intercept of the inverse demand function a and non-increasing in the fixed cost of entry f , marginal cost c , and price sensitivity b . In other words, if the industry has very high cost of entry, then relatively small uncertainty can reduce the number of entrants.

1.5.3 Total vaccine supply

While clearly related to the number of entrants, we are ultimately interested in the impact of yield uncertainty on total expected quantity of vaccine produced in equilibrium, since in our model that is directly linked to the number of vaccinations and hence to the health care outcome for society. Let q_d^* be the total quantity produced at equilibrium in the absence of uncertainty and $E[q_u^*]$ the to-

tal expected quantity produced at equilibrium under uncertainty. In Proposition 1, we already saw that limited levels of uncertainty can lead to increased entry. However, the next proposition shows that this is not true for total vaccine supply.

Proposition 2. $E[q_u^*] \leq q_d^* \quad \forall \delta \geq 0$, i.e., the expected quantity produced by the market under yield uncertainty is lower than or equal to that in the deterministic case.

The above result is true even for some levels of uncertainty where the number of firms at equilibrium is higher than in the base case, i.e., $\delta_1^* > \delta > 0$. One might expect that higher quantities imply better societal outcomes, as production quantity equals number of vaccinations in our model. However, as seen below, we can prove this only in certain range of the yield uncertainty.

1.5.4 Social welfare

To characterize the impact of these effects on consumers, we compare the consumer welfare with and without yield uncertainty. At equilibrium, let $E[CS_u(q_u^*)]$ denote the expected consumer welfare under uncertainty and let $CS_d(q_d^*)$ denote the consumer welfare in the absence of uncertainty. Formally, when total quantity produced is q , total expected consumer utility from vaccination is $E \left[\int_0^q (a - bu) du \right]$ and the expected amount paid by the consumers for vaccination is $E[(a - bq)q]$. Hence, the expected consumer welfare in the case of yield uncertainty is given by $E[CS_u(q_u^*)] = E \left[\int_0^{q_u^*} (a - bu) du - (a - bq_u^*)q_u^* \right] = \frac{b}{2} E[(q_u^*)^2]$. Similarly, in the deterministic case, the expected consumer welfare is given by $CS_d(q_d^*) = \frac{b}{2} (q_d^*)^2$. Then;

Proposition 3. Define $\delta_2^* \triangleq \min \left\{ \delta > 0 : \left[\sqrt{1 + \delta^2} \left(\frac{a-c}{\sqrt{bf}} \right) - (1 + 2\delta^2) \right] + \delta^2 \leq \left[\frac{a-c}{\sqrt{bf}} - 1 \right] \right\}$.

Then, the expected consumer welfare in equilibrium with and without uncertainty satisfy $E[CS_u(q_u^*)] \leq CS_d(q_d^*)$ if $\delta > \delta_2^*$. Also, $\delta_2^* \geq \delta_1^*$

This result states that large uncertainty reduces the expected consumer welfare when compared to the deterministic case. The contrast with Proposition 2 (which holds for any level of uncertainty) arises because consumer welfare depends on $E[q^2]$, not on $E[q]$. We are unable to compare $CS_d(q_d^*)$ and $E[CS_u(q_u^*)]$ when $\delta < \delta_2^*$.

1.5.5 First best solution

Having understood the impact of yield uncertainty on equilibrium entry and quantity supplied, we can now consider various interventions that could drive the market outcomes closer to the socially optimal ones. First, we formulate and solve the decision problem of a social planner who wants to maximize the total social welfare, then we compare the socially optimal solution to the equilibrium outcomes derived above.

“First-best” denotes the solution to the social planner’s problem of maximizing the total social welfare or the total surplus of society by choosing the number of firms (n) and the target production quantity of each firm (\bar{q}_i) (Vives, 1999). This presupposes the existence of an omnipotent and omniscient benevolent agency, possibly government, that can costlessly and perfectly control both the structure of the industry and the conduct of the firms in the industry. Then the social planner’s problem can be formulated as:

$$\max_{\bar{q}_i, n} E[W(q, n)] = \max_{\bar{q}_i, n} E \left[\int_0^q (a - bq) dq \right] - E[cq] - nf$$

where $q = \sum_{i=1}^n q_i = \sum_{i=1}^n \alpha_i \bar{q}_i$ denotes the total quantity produced by n firms. The first term is the total expected consumer utility, i.e., the area under the de-

mand curve for consumers who do purchase, and the second term is the expected variable cost of production if $q(n)$ is the total quantity produced. The third term is the total cost of entry incurred by society if n firms enter the industry. Simplifying we obtain the following social planner's problem:

$$\max_{\bar{q}_i, n} E [W(q(n), n)] = \max_{\bar{q}_i, n} (a - c)E(q) - \frac{b}{2}E(q^2) - nf \quad (1.4)$$

This problem can be solved optimally by first fixing n and optimizing over \bar{q}_i , which we call the quantity problem. In the second step, we substitute the optimal \bar{q}_i in the original problem and optimize over n . We call this the structural problem.

Lemma 3. *Let \bar{q}_i^{fb} denote the first-best planned production quantity of the i^{th} firm and let q_i^{fb} denote the corresponding actual quantity produced. Then (i) $\bar{q}_i^{fb} = \frac{2(a-c)\mu}{b[(n+1)\mu^2 + 2\sigma^2]} = 2\bar{q}_i^*$ and (ii) $E [q_i^{fb}] = \frac{2(a-c)}{b(n+1+2\delta^2)}$; where $\delta = \frac{\sigma}{\mu}$ as defined earlier.*

For a given number of firms, the socially optimal target quantity and expected production quantity for each firm is twice that under competition. The next step is to characterize the socially optimal number of firms. Substituting the expression for \bar{q}_i^{fb} in (1.4) and simplifying, the structural problem is:

$$\max_n E [W(n)] = \max_n \frac{2(a-c)^2 n(1+\delta^2)}{b(n+1+2\delta^2)^2} - nf \quad (1.5)$$

and the result is characterized in the following proposition.

Proposition 4. *Let n^{fb} denote the number of firms in the first best solution. Then $1 \leq n^{fb} < 1 + 2\delta^2$. Also $n^{fb} = 1$ if $\frac{2(1+\delta^2)}{\delta} \geq \frac{(a-c)}{\sqrt{bf}}$.*

In the deterministic case, where $\delta = 0$, it is easy to see that $n^{fb} = 1$. This is in accordance with the existing intuition that the first-best solution in a deterministic setting involves having a benevolent monopoly which produces the

socially optimal quantity since society then incurs the fixed cost of entry only once. $n_d^* = \left\lfloor \frac{(a-c)}{\sqrt{bf}} - 1 \right\rfloor \leq 3$ implies that $n^{fb} = 1 \quad \forall \delta > 0$. In other words, if the fixed costs are high enough that the market can support at most three entrants in the deterministic case, then the first-best solution is to have benevolent monopoly regardless of the level of uncertainty. However, when $n_d^* > 3$, there exist levels of uncertainty for which society prefers multiple suppliers since the value of diversity of supply is greater than the additional fixed cost of entry. This risk pooling effect occurs due to the concavity of consumer welfare. Each incremental unit of vaccine produced less leads to higher social costs.

1.5.6 Seond-best solution

Now, we turn to the case where the social planner can regulate the number of firms in the industry, but cannot regulate their conduct, so that the entering firms engage in Cournot competition in the post-entry game. This solution is referred to as the second-best structural regulation or simply “second-best” (Vives, 1999). We shall focus on this case in greater detail due to the restrictive assumptions required for the first-best. The social planner’s problem in this case is given by

$$\max_n E [W(q(n), n)] = \max_n (a - c)E(q) - \frac{b}{2}E(q^2) - nf \quad (1.6)$$

Substituting $E(q) = \frac{n(a-c)}{b(n+1+2\delta^2)}$ from Lemma (1) and $E(q^2) = \frac{(a-c)^2 n(n+\delta^2)}{b^2(n+1+2\delta^2)^2}$ in (1.6), we obtain:

$$\max_n E [W(q(n), n)] = \max_n \frac{(a-c)^2}{2b} \left[1 - \frac{(1+2\delta^2)^2 + n\delta^2}{(n+1+2\delta^2)^2} \right] - nf \quad (1.7)$$

Relaxing n to $x \in \mathbb{R}_+$, it can be verified that $E [W(q(x), x)]$ is strictly concave. Hence by restricting $n \in \mathbb{N}_+$, $E [W(q(n), n)]$ can have at most two maximizers. Let n_u^{sb} denote the element of this set of two maximizers and let n_d^{sb} denote the

deterministic optimum. We begin by analyzing the deterministic case and then extend the analysis to the case with uncertainty.

Proposition 5. *In the absence of yield uncertainty, the second-best number of firms is given by $n_d^{sb} \in \left\{ \left[\left(\frac{a-c}{\sqrt{bf}} \right)^{\frac{2}{3}} - 1 \right] - 1, \left[\left(\frac{a-c}{\sqrt{bf}} \right)^{\frac{2}{3}} - 1 \right] + 1 \right\}$. Also, $n_d^{sb} - 1 \leq n_d^*$.*

In other words, in the absence of yield uncertainty, the equilibrium number of firms can be less than the second-best number of firms but not by more than one. This is because in the second-best outcome, the firms are making positive profits causing more firms to enter. This result is identical to that of Mankiw and Whinston (1986). However, we show that including yield uncertainty can change the relationship between the second-best number of firms n_u^{sb} and the equilibrium number of firms under uncertainty n_u^* . The result is summarized in the following proposition:

Proposition 6. *The number of firms in equilibrium n_u^* and in the second-best solution n_u^{sb} satisfy (i) $n_u^{sb} > n_u^*$ if $\delta \geq \delta_3^*$ and (ii) $n_u^{sb} - 1 \leq n_u^*$ if $\delta < \delta_3^*$, where $\delta_3^* > 0$ solves $\frac{a-c}{\sqrt{bf}} = \frac{2(1+2\delta^2)\sqrt{1+\delta^2}}{2+\delta^2}$.*

This shows that if the yield uncertainty is larger than a certain threshold, then the number of firms at unregulated equilibrium will be less than in the second-best case. Since the total expected quantity produced $\frac{n(a-c)}{b(n+1+2\delta^2)}$ is increasing in n , the industry undersupplies at equilibrium whenever uncertainty is higher than that threshold. In contrast, for a low level of uncertainty, the outcome is the same as that in the deterministic case.

1.6 Application to the U.S. influenza vaccine market

In this section, we apply our model to the U.S. influenza vaccine market. Our objective is to determine whether and how much yield uncertainty might have contributed to the observed exit from the U.S. influenza vaccine market. We start with calibrating our model, then analyze the market equilibrium, evaluate the impact of demand-side and supply-side public policies on social welfare, and conduct sensitivity analyses.

1.6.1 Model calibration

In this section, we derive the demand function for the U.S. market by assuming uniformly distributed consumer valuations which leads to a linear inverse demand function. Consider M individuals, who each demand zero or one unit of vaccine in a single-period context. The individuals differ only in the expected cost incurred if they do not get vaccinated, denoted by v . This cost reflects the likelihood of getting infected and the resulting costs of health care, lost income etc. We assume perfect vaccination, i.e., after vaccination consumers stay perfectly healthy and do not incur any health care or other costs; relaxing this would not change our results.

Following Brito et al. (1991), we assume that v follows a uniform distribution $F(v)$ on the range $[\underline{v}, \bar{v}]$. Let p be the price of one dose of vaccine and let v^* be the valuation of the threshold consumer who is indifferent between getting vaccinated and not getting vaccinated, given price p . Then for a rational consumer, who does not account for the positive externality of vaccination mentioned earlier, $p = v^*$ and $q(p) = (1 - F(v^*)) M$ is the total demand at that price. Substituting $F(v^*) = \frac{v^* - \underline{v}}{\bar{v} - \underline{v}}$ and $p = v^*$, we get $q(p) = M \left(\frac{\bar{v} - p}{\bar{v} - \underline{v}} \right)$. Defining $a \triangleq \bar{v}$ and $b \triangleq \frac{(\bar{v} - \underline{v})}{M}$,

the following inverse demand function is obtained:

$$p = a - bq \tag{1.8}$$

In our subsequent analysis we assume that the individuals are rational, but including a positive externality does not change our results. While we chose parameter values (summarized in Table 1.1) that represent the U.S. situation to the degree possible, some of the values are inevitably, at best, very rough estimates.

Table 1.1: Parameter values for the U.S. influenza vaccine market

Parameter	$\underline{v}(\$)$	$\bar{v}(\$)$	$c (\$)$	$f (\$ \text{ million})$	$M (\text{million})$	δ
Value	0	8	3	40	300	0.64

The population (M) was chosen to be the U.S. population, which is approximately 300 million (U.S. Census Bureau, 2004). The lower limit of the customer valuation (\underline{v}) can be normalized to zero. The upper limit of the customer valuation (\bar{v}) of \$8 was chosen based on anecdotal evidence (Nichol, 2001) of the direct cost of vaccination and the fact that vaccination is a covered benefit under insurance for many customers. The true value of \bar{v} is likely to be much higher. But, the underlying distribution is also unlikely to be uniform. If \bar{v} is interpreted as the upper limit of the mass market's willingness to pay, then \$8 seems to be reasonable. The value of $u = 0$ was based on the fact that individuals behave in a self-interested manner with respect to vaccination.

The variable cost (c) of \$3 per dose was based on the costs of procurement from the manufacturers (O'Mara et al., 2003) and assuming around 50% gross margin. The value of f is also not directly available. Gottlieb (2004) reports that an investment of around \$300 million is required to build a new influenza vaccine plant. A 10-20% cost of capital on this investment translates into an annual fixed cost of \$30 - \$60 million dollars. During the 2000-01 season, Parkedale announced

its departure from the influenza vaccine market, writing off \$45 million (Danzon et al., 2004). Based on these data, we chose an annual fixed cost of \$40 million, but let it vary from \$20 million to \$100 million in our sensitivity analysis.

The yield uncertainty (δ) is even less observable. We first estimated industry-wide yield uncertainty using the data in Table 1.2 (Strikas, 2005). We used the time-variation of quantity produced and supplied as a proxy for the underlying yield uncertainty. Since vaccination begins in October, manufacturers aim to supply all the vaccine by that time. The year-to-year variability in the degree to which supply is late is an indicator of the yield uncertainty. However, in order to control for idiosyncratic variations such as that in 2004 due to the Chiron failure, we normalized the quantity supplied until October (A) by the total supply for that year (B).

Table 1.2: Quantity of influenza vaccine produced and distributed (million doses)

Year	Supplied by Oct. (A)	Total supplied, entire season (B)	Total produced (C)	$D = \frac{A}{B}$
1999	75.8	76.8	77.2	0.987
2000	26.6	70.4	77.9	0.378
2001	43.0	77.7	87.7	0.553
2002	82.7	83.0	95.0	0.996
2003	80.0	83.1	86.9	0.963
2004	51.0	57.1	61.0	0.893

We then calculated the standard deviation and mean of this normalized quantity (D) to estimate δ for the industry. We corroborated these values of δ using simply the standard deviations for total annual quantity produced (C) and total annual quantity supplied (B) during 1999-2004. All these values were in the

range 0.33 – 0.38. This would be an underestimate for the true coefficient of variation δ , since it does not take into account the variation between the targeted quantity and the final quantity produced in each year. We assumed the value of $\delta = 0.45$ based on this exercise. This value is the δ for the industry comprising two firms for the period under consideration. Assuming that the yields for the two existing firms are independent of one another, we calculated the δ for each firm as $\sqrt{2} * 0.45 = 0.64$, but let δ vary from 0 to 3 in the sensitivity analysis in the light of uncertainty about the true value of δ .

1.6.2 Analysis of market equilibrium

Here, we calculate the equilibrium for the U.S. influenza vaccine market as predicted by our model. Table 1.3 summarizes the difference between the deterministic ($\delta = 0$) and the stochastic yield case ($\delta = 0.64$) for the equilibrium and optimal (second-best) solutions. The results show that even a relatively low level of yield uncertainty ($\delta = 0.64$) can eliminate the excess number of entrants at equilibrium predicted by the deterministic model, in this case because the number of entrants in the optimal (second-best) solution increases. More importantly, yield uncertainty results in a substantial reduction (17%) of expected total quantity produced and a corresponding reduction in social welfare (27%) and consumer welfare (19%).

The equilibrium prediction from our model under uncertainty matches fairly closely with the observations from the U.S. market. While two firms had been in the market from 1999 to 2004, a third firm has entered in 2005. The real equilibrium industry output is around 100 million doses, in the same ballpark as the predicted 117 million. While this in itself certainly does not imply that our model and parameter values are correct, it seems to indicate some face validity.

Table 1.3: Equilibrium and second-best solutions for the influenza vaccine case

Solution	$\delta = 0$	$\delta = 0.64$	Difference
Socially optimal (second-best) number of firms	2	3	50%
Equilibrium number of firms	3	3	0%
Equilibrium industry output (million doses)	141	117	-17%
Equilibrium price (\$/dose)	4.25	4.88	15%
Equilibrium profit per firm (\$ million)	58.6	57.0	-3%
Equilibrium consumer surplus (\$ million)	180	146	-19%
Equilibrium social welfare (\$ million)	319	258	-24%

1.6.3 Evaluation of demand-side and supply-side policies

Having calculated the equilibrium outcomes, we now compare the performance of various policy interventions intended to improve the social welfare. These policies can be classified into demand-side and supply-side measures. An example of a demand-side policy is to improve awareness about the benefits of vaccination and thus implicitly shift the demand curve upward through an increase in a in (1.8). A supply-side intervention discussed in this essay is structural regulation, where the government regulates market entry through instruments such as entry taxes, subsidies or regulatory costs. Technological interventions such as a new production process to reduce the yield uncertainty are not considered here. We vary a (or correspondingly \bar{v}) and δ from Table 1.1, keeping other parameters fixed, to compare the demand- and supply-side interventions under different levels of yield uncertainty. We only compare the outcomes of these policies, not the costs, since that is not our focus and we are not aware of any reasonable data to estimate these costs. Clearly any actual policy decision would require analysis of the costs as well as benefits.

First, we study the impact of demand-side policies on the structure of the industry (n_u^* and n_u^{sb}) for different values of a and δ . In Figure 1.1, θ represents a \$ increase in the value of a (or correspondingly \bar{v}) above the base case considered in section 1.6.2, where $\bar{v} = \$8$. We consider $\theta = 0.4, 0.8, 1.2$ and 1.6 . This corresponds to a 10% to 40% increase in the valuation of the average consumer given our initial range of $[\underline{v}, \bar{v}] = [0, 8]$. Figure 1.1 verifies the result proved in Proposition 6: yield uncertainty beyond a particular threshold can cause less entry than is socially optimal. This threshold increases as the demand curve shifts up, i.e., as a increases.

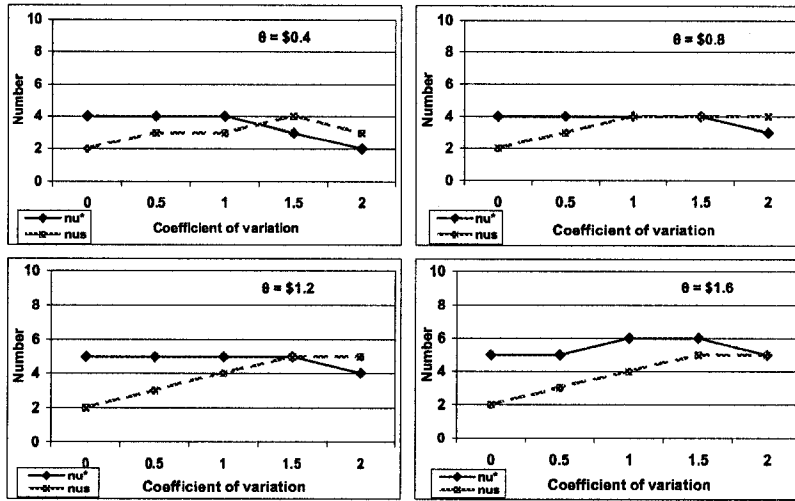


Figure 1.1: Equilibrium and socially optimal number of firms as a function of the coefficient of variation, δ , for different demand side interventions

Next, we study the impact of supply- and demand-side interventions on social welfare under various conditions depending on the value of δ . Let $E[W(n, \theta)]$ denote the social welfare when n firms enter the market and θ is as defined above. This can be interpreted as the effect of demand-side interventions such as improved awareness or better insurance coverage. The base case is given

by $E[W(n_u^*, 0)]$, i.e., no supply- or demand-side intervention. We measure the performance of demand-side intervention by calculating $r_d \triangleq \frac{E[W(n_u^*, \theta)]}{E[W(n_u^*, 0)]}$ and the performance of supply-side intervention by calculating $r_s \triangleq \frac{E[W(n_u^{s\theta}, 0)]}{E[W(n_u^*, 0)]}$. We plot r_d and r_s as a function of δ in Figure 1.2.

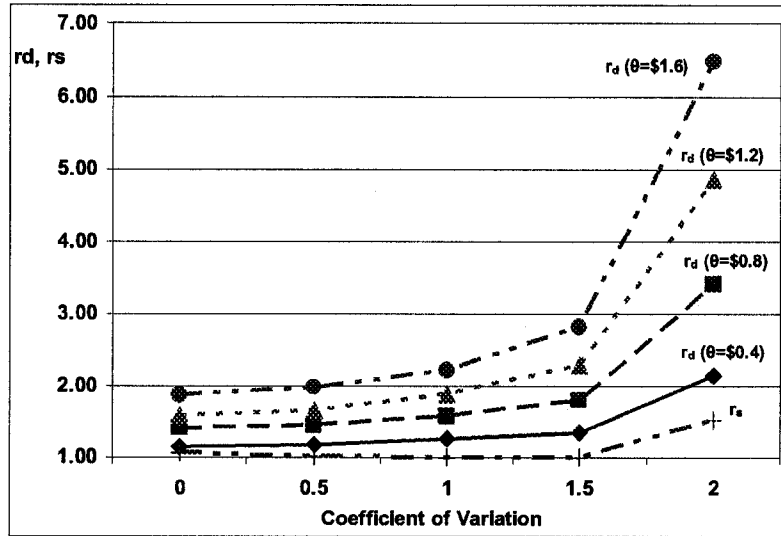


Figure 1.2: Impact of supply-side and demand-side interventions for different values of the coefficient of variation, δ

The results show that in the given parameter range, a demand-side policy causing a 10% increase in the valuation of an average consumer already results in a higher increase in social welfare than a supply-side policy of regulated entry; higher increases in valuation result in even higher social welfare. Recall, though, that the costs of implementing these policies are not included in our model and we do not know whether achieving a 10% increase in the valuation of an average consumer may be much more complex and costly than implementing a supply-side policy of regulated entry.

1.6.4 Sensitivity analysis

Since we are most uncertain about our estimates of f and δ , we conducted a sensitivity analysis to ascertain the impact of these parameters on our results. Recall that we have assumed μ to be constant and hence the change in δ is affected through change in σ . The results are summarized in Table 1.4. Each cell contains a pair (n_u^*, n^{sb}) , i.e., the equilibrium number followed by the second-best number of firms for each combination of f and δ .

Table 1.4: Equilibrium and socially optimal number of firms for different values of f and δ

f (\$ millions)	$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 2$	$\delta = 2.5$	$\delta = 3$
20	5, 5	6, 3	6, 5	6, 6	6, 6	4, 6	X
30	4, 3	4, 3	4, 4	4, 4	3, 4	1, 3	X
40	3, 2	3, 2	3, 3	3, 3	1, 3	1, 2	X
50	3, 2	3, 2	3, 3	2, 2	1, 2	X	X
60	2, 2	2, 2	2, 2	1, 2	1, 1	X	X
80	2, 1	2, 2	1, 2	1, 1	X	X	X
100	2, 1	1, 1	1, 1	1, 1	X	X	X

The cells with X indicate that the industry is not viable for those combinations of f and δ : even one firm producing the vaccine would result in negative social welfare. We pay particular attention to cases where $n_u^* < n^{sb}$, highlighted in bold, as those are the cases where yield uncertainty can help explain lower entry than is socially optimal. The sensitivity analysis demonstrates that at higher levels of the fixed cost of entry, f , even relatively low levels of yield uncertainty can result in less than optimal entry at equilibrium. Table 1.4 also shows that for a given level of yield uncertainty, an increase in fixed cost f reduces the

equilibrium number of firms. This provides an indirect support for the hypothesis that increased regulatory cost could explain the exit of firms from the U.S. market. Conversely, for relatively low levels of yield uncertainty, including those in the range we estimate and for most values of fixed cost, the presence of yield uncertainty does not change the equilibrium number of firms.

To summarize, the predictions of our model match reasonably closely with the observations from the U.S. market as seen from Table 1.3. Without accounting for the cost of implementation, we found that demand-side policies had higher impact on social welfare than the supply-side policies over the range of parameters valid for the U.S. market, although our analysis only addresses a part of this issue and suffers from significant limitations.

1.7 Concluding remarks

In this essay we analyze the effect of yield uncertainty in Cournot competition. The model is based on the context of the market for influenza vaccine, but applies to other settings with yield uncertainty, fixed cost of entry and Cournot competition. We show that if yield uncertainty is sufficiently large, less firms will enter in equilibrium than at the social optimum with regulated entry. This is in contrast to the traditional result (on Cournot competition without yield uncertainty) that excess entry will occur at equilibrium relative to the second-best social optimum. We also show that this uncertainty can reduce the expected total industry output and the expected consumer surplus in equilibrium. These results continue to hold even in the presence of positive externalities of vaccination.

We report numerical analyses with parameter values pertinent to the U.S. influenza vaccine market. The predictions for number of firms and quantity

supplied are broadly comparable to what we actually observe in practice. In the relevant range of parameters, we find that including yield uncertainty eliminates the excess entry compared to the socially optimal number of firms predicted by the traditional oligopoly model. However, in explaining the exit of firms from the market over years, increasing fixed cost appears to be a more significant factor than the yield uncertainty. We also compare the performance of demand-side and supply-side policies aimed at improving social welfare. We find that demand-side policies, though possibly much more difficult and costly to implement, are likely to be significantly more effective than supply-side policies for the range of parameters characterizing the U.S. market.

1.8 Proofs

Proof of Lemma 1: The proof is analogous to that for the deterministic case. The i^{th} firm solves the concave maximization problem in \bar{q}_i^* given by (1.2). Hence the first order condition in (1.3) is necessary and sufficient to obtain the equilibrium target quantities. Using $E[\alpha_i \alpha_j] = E[\alpha_i] E[\alpha_j]$ (random variables α_i and α_j are independent), $Var(\alpha_i) = E[\alpha_i^2] - (E[\alpha_i])^2 = \sigma^2$, $E[\alpha_i] = \mu \quad \forall i$ (each firm has the same yield distribution), and simplifying (1.3), we obtain a unique solution to the above set of equations given by:

$$\bar{q}_i^* = \frac{(a-c)\mu}{b[(n+1)\mu^2 + 2\sigma^2]} \quad \forall i \quad (1.9)$$

The expected production quantity of the i^{th} firm is given by

$$E[q_i^*] = E[\alpha_i] \bar{q}_i^* = \frac{(a-c)\mu^2}{b[(n+1)\mu^2 + 2\sigma^2]} = \frac{(a-c)}{b(n+1 + 2\delta^2)} \quad \forall i \quad (1.10)$$

which is decreasing in δ . This completes the proof for parts (i), (ii) and (iii). Substituting (1.9) in (1.1) and simplifying yields

$$\Pi_i^*(n) = \frac{(a-c)^2(\delta^2+1)}{b(n+1+2\delta^2)^2} \quad (1.11)$$

which proves part (iv). Differentiating (1.11) w.r.t. δ , we obtain $\frac{\partial \Pi_i(\bar{q}_i^*)}{\partial \delta} = \frac{(a-c)^2 2\delta(n-3-2\delta^2)}{b(n+1+2\delta^2)^3}$. Hence, $\frac{\partial \Pi_i(\bar{q}_i^*)}{\partial \delta} < 0$ for $\delta > \sqrt{\frac{n-3}{2}}$ if $n > 3$ and $\frac{\partial \Pi_i(\bar{q}_i^*)}{\partial \delta} < 0$ for $\forall \delta > 0$ if $n \leq 3$, which proves part (v).

Proof of Lemma 2: We first ignore the integer constraint on the number of firms and solve for x_u^* by using the condition $\Pi_i^*(x_u^*) = f$. Using (1.11) and rearranging the terms we get:

$$x_u^* = \frac{(a-c)}{\sqrt{bf}} \sqrt{1+\delta^2} - (1+2\delta^2) \quad (1.12)$$

Since $\Pi_i^*(x)$ is decreasing in x , $n_u^* = \lfloor x_u^* \rfloor = \frac{(a-c)}{\sqrt{bf}} \sqrt{1+\delta^2} - (1+2\delta^2)$, using (1.12).

Proof of Proposition 1: Since the derivatives w.r.t. $\delta \geq 0$ and δ^2 have the same signs, we focus on the latter due to ease of analysis. Differentiating (1.12) w.r.t. δ^2 , we get $\left. \frac{dx_u^*}{d\delta^2} \right|_{\delta^2=0} = \frac{(a-c)}{2\sqrt{bf}} \frac{1}{\sqrt{1+\delta^2}} - 2 \Big|_{\delta^2=0} = \frac{(a-c)}{2\sqrt{bf}} - 2$. For $\frac{(a-c)}{\sqrt{bf}} \leq 4$,

$$\left. \frac{dx_u^*}{d\delta^2} \right|_{\delta^2=0} \leq 0 \implies x_u^* \leq x_d^* \implies n_u^* \leq n_d^* \quad \forall \delta > 0$$

Next, consider the case $\frac{(a-c)}{\sqrt{bf}} > 4$. Substituting $\delta = 0$ in (1.12), we get $x_d^* = \frac{(a-c)}{\sqrt{bf}} - 1$. Comparing this with (1.12) and simplifying, we obtain $x_u^* \leq x_d^* \iff \delta \geq \delta_1^*$, where $\delta_1^* = \sqrt{\left(\frac{a-c}{2\sqrt{bf}} - 1\right)^2 - 1}$. Since $\lfloor x_u^* \rfloor = n_u^*$ and $\lfloor x_d^* \rfloor = n_d^*$, we obtain $\delta \geq \delta_1^* \implies n_u^* \leq n_d^*$ and $\delta \leq \delta_1^* \implies n_u^* \geq n_d^*$.

Proof of Proposition 2: With yield uncertainty, the total quantity produced at equilibrium is $E[q_u^*] = \sum_{i=1}^{n_u^*} E[q_i^*] = \frac{n_u^*(a-c)}{b(n_u^*+1+2\delta^2)}$, using (1.10). Without yield uncertainty ($\delta = 0$ and $n_u^* = n_d^*$), we get $q_d^* = \frac{n_d^*(a-c)}{b(n_d^*+1)}$. Comparing the two and

using $n_d^* = \left\lfloor \frac{a-c}{\sqrt{bf}} - 1 \right\rfloor$ and $n_u^* = \left\lfloor \frac{(a-c)}{\sqrt{bf}} \sqrt{1+\delta^2} - (1+2\delta^2) \right\rfloor$:

$$q_d^* \geq E[q_u^*] \iff (1+2\delta^2) \left\lfloor \frac{a-c}{\sqrt{bf}} - 1 \right\rfloor \geq \left\lfloor \frac{(a-c)}{\sqrt{bf}} \sqrt{1+\delta^2} - (1+2\delta^2) \right\rfloor \quad (1.13)$$

Consider the following inequality:

$$(1+2\delta^2) \left\lfloor \frac{a-c}{\sqrt{bf}} - 1 \right\rfloor \geq \frac{(a-c)}{\sqrt{bf}} \sqrt{1+\delta^2} - (1+2\delta^2) \quad (1.14)$$

Write LHS and RHS for left and right hand side of (1.14) respectively. $RHS = LHS$ at $\delta^2 = 0$. Again, since we are only interested in the sign of the derivative, we can differentiate w.r.t. δ^2 . Note that $\frac{\partial LHS}{\partial \delta^2} = 2 \left\lfloor \frac{a-c}{\sqrt{bf}} - 1 \right\rfloor$ and $\frac{\partial RHS}{\partial \delta^2} = \frac{a-c}{\sqrt{bf}} \frac{1}{2\sqrt{1+\delta^2}} - 2$. So, $\frac{\partial RHS}{\partial \delta^2} \Big|_{\delta^2 > 0} < \frac{\partial RHS}{\partial \delta^2} \Big|_{\delta^2 = 0} < \frac{\partial LHS}{\partial \delta^2} \Big|_{\delta^2 = 0} = \frac{\partial LHS}{\partial \delta^2} \Big|_{\delta^2 > 0}$. Hence (1.14) and consequently (1.13) holds $\forall \delta > 0$. Hence, we have $q_d^* \geq E[q_u^*] \forall \delta \geq 0$.

Proof of Proposition 3: Begin with calculating $E[(q_u^*)^2]$ and $E[CS_u(q_u^*)]$. Using (1.9) and simplifying we obtain:

$$E[CS_u(q_u^*)] = \frac{(a-c)^2 n_u^* (n_u^* + \delta^2)}{2b(n_u^* + 1 + 2\delta^2)^2} = \frac{n_u^* (n_u^* + \delta^2) f}{2} \quad (1.15)$$

Hence, for the deterministic case ($\delta = 0$ and $n_u^* = n_d^*$),

$$CS_d(q_d^*) = \frac{(n_d^*)^2 f}{2} \quad (1.16)$$

Next, define $h(\delta) \triangleq \left(\frac{a-c}{\sqrt{bf}} \right) \sqrt{1+\delta^2} - (1+2\delta^2) + \delta^2 - \left\lfloor \frac{a-c}{\sqrt{bf}} - 1 \right\rfloor$, which is unimodal in δ for $\delta > 0$ and $h(0) = 0$. So, $\delta_2^* = \min \{ \delta > 0 : h(\delta) \leq 0 \}$ exists. Thus, $\delta > \delta_2^* \implies \left\lfloor \frac{a-c}{\sqrt{bf}} - 1 \right\rfloor > \left(\frac{a-c}{\sqrt{bf}} \right) \sqrt{1+\delta^2} - (1+2\delta^2) + \delta^2 \geq \left\lfloor \sqrt{1+\delta^2} \left(\frac{a-c}{\sqrt{bf}} \right) - (1+2\delta^2) \right\rfloor + \delta^2$. This, along with the expressions for n_d^* and n_u^* and (1.15) and (1.16), proves the first part of the result. Also, $\delta > \delta_2^* \implies n_d^* \geq n_u^* + \delta^2 \implies n_d^* \geq n_u^* \iff \delta > \delta_1^*$. Hence $\delta_2^* > \delta_1^*$.

Proof of Lemma 3: Note that $E(q) = E[\sum_{i=1}^n q_i] = E[\sum_{i=1}^n \alpha_i \bar{q}_i] = \mu \sum_{i=1}^n \bar{q}_i$.

Similarly,

$$\begin{aligned} E(q^2) &= E \left[\left(\sum_{i=1}^n q_i \right)^2 \right] = E \left[\sum_{i=1}^n q_i^2 \right] + E \left[\sum_{i \neq j} q_i q_j \right] \\ &= E [\alpha_i^2] \sum_{i=1}^n \bar{q}_i^2 + E [\alpha_i] E [\alpha_j] \sum_{i \neq j} \bar{q}_i \bar{q}_j = (\sigma^2 + \mu^2) \sum_{i=1}^n \bar{q}_i^2 + \mu^2 \sum_{i \neq j} \bar{q}_i \bar{q}_j \end{aligned}$$

Substituting the expressions for $E(q)$ and $E(q^2)$ in (1.4) and simplifying, we obtain

$$\max_{\bar{q}_i, n} \left\{ E [W(q(n), n)] = (a - c)\mu \sum_{i=1}^n \bar{q}_i - \frac{b}{2} \left[(\sigma^2 + \mu^2) \sum_{i=1}^n \bar{q}_i^2 + \mu^2 \sum_{i \neq j} \bar{q}_i \bar{q}_j \right] - nf \right\}$$

We first maximize over \bar{q}_i , keeping n fixed. The resulting objective function is jointly concave in \bar{q}_i with first order condition

$$(a - c)\mu = \frac{b}{2} \left[(2\sigma^2 + \mu^2) \bar{q}_i + \mu^2 \sum_{k=1}^n \bar{q}_k \right] \quad \forall i$$

This condition and consequently the optimal solution \bar{q}_i^* is symmetric in i . Summing over i and utilizing symmetry we obtain $\bar{q}_i^{fb} = \frac{2(a-c)\mu}{b[(n+1)\mu^2 + 2\sigma^2]}$. Similarly, the expected quantity produced by each firm is given by $E[\bar{q}_i^*] = \frac{2(a-c)}{b(n+1+2\delta^2)}$.

Proof of Proposition 4: First, consider the continuous relaxation of (1.5) to $x \in \mathbb{R}_+$. The first order condition w.r.t. x gives $\frac{2(a-c)^2(1+\delta^2)(1+2\delta^2-x)}{bf} = (x+1+2\delta^2)^3$. Note that the left hand side is decreasing in x and positive only for $x < 1+2\delta^2$. The right hand side is increasing in x and always positive. Since we require that the expected quantity produced is non-negative for any n , it is required that $n \geq 1$. Thus, if a solution exists to this equation, it is unique and lies in the range $1 \leq n < 1+2\delta^2$. Also, a necessary and sufficient condition for a solution to exist is given by $\frac{2(a-c)^2(1+\delta^2)(1+2\delta^2-x)}{bf} \Big|_{x=1} > (x+1+2\delta^2)^3 \Big|_{x=1}$ or alternately $\frac{2(1+\delta^2)}{\delta} < \frac{(a-c)}{\sqrt{bf}}$. If this condition is not satisfied, then $n = x = 1$, since otherwise would imply that $\frac{dE[W(x)]}{dx} \Big|_{x=1} \leq 0$.

Proof of Proposition 5: Consider the continuous relaxation of (1.7) in $x \in \mathbb{R}_+$ instead of $n \in \mathbb{N}$. Clearly, $E[W(q(x), x)]$ is also a concave function of x and has a unique optimum given by the following first-order condition,

$$\frac{\partial E[W(q(x), x)]}{\partial x} = \frac{(a-c)^2}{2b} \left[\frac{2(1+2\delta^2)^2 + x\delta^2 - (1+2\delta^2)\delta^2}{(x+1+2\delta^2)^3} \right] - f = 0 \quad (1.17)$$

For $\delta = 0$, this yields $x_d^{sb} = \left(\frac{a-c}{\sqrt{bf}}\right)^{\frac{2}{3}} - 1$ and hence $n_d^{sb} \in \{[x_d^{sb}], [x_d^{sb}] + 1\}$. Recall that $n_d^* = \left\lfloor \left(\frac{a-c}{\sqrt{bf}}\right) - 1 \right\rfloor$ and $x_d^* = \left(\frac{a-c}{\sqrt{bf}}\right) - 1$. First check that $x_d^{sb} \leq x_d^*$. Now $n_d^{sb} - 1 \leq [x_d^{sb}] \leq x_d^{sb} \leq x_d^*$. Clearly, $n_d^{sb} - 1 \leq x_d^* \implies n_d^{sb} - 1 \leq n_d^*$.

Proof of Proposition 6: We define x_u^* and x_u^{sb} corresponding to n_u^* and n_u^{sb} respectively. Thus $x_u^{sb} = \left\{ x : \frac{\partial EW(x)}{\partial x} \Big|_{x=x_u^{sb}} = 0 \right\}$. First, in order to compare x_u^* and x_u^{sb} , we calculate $\frac{\partial EW(x)}{\partial x} \Big|_{x=x_u^*}$. Using (1.17) we get:

$$\frac{\partial E[W(q(x), x)]}{\partial x} \Big|_{x=x_u^*} = \frac{(a-c)^2}{2bf} \left[\frac{2(1+2\delta^2)^2 + x_u^*\delta^2 - (1+2\delta^2)\delta^2}{(x_u^*+1+2\delta^2)^3} \right] - f$$

Also, since $f = \Pi_i^*(x_u^*) = \frac{(a-c)^2(1+\delta^2)}{b(x_u^*+1+2\delta^2)^2}$ at equilibrium:

$$\begin{aligned} & \frac{\partial E[W(q(x), x)]}{\partial x} \Big|_{x=x_u^*} \\ &= \frac{(a-c)^2}{2bf} \left[\frac{2(1+2\delta^2)^2 + x_u^*\delta^2 - (1+2\delta^2)\delta^2}{(x_u^*+1+2\delta^2)^3} \right] - \frac{(a-c)^2(1+\delta^2)}{b(x_u^*+1+2\delta^2)^2} \\ &= \frac{(a-c)^2}{2bf(x_u^*+1+2\delta^2)^3} \left[2(1+2\delta^2)^2 - 3(1+2\delta^2)\delta^2 - x_u^*\delta^2 - 2(x_u^*+1+2\delta^2) \right] \end{aligned}$$

A sufficient condition for $x_u^* \leq x_u^{sb}$ is $\frac{\partial E[W(q(x), x)]}{\partial x} \Big|_{x=x_u^*} \geq 0$. Substituting x_u^* and simplifying gives $x_u^* \leq x_u^{sb}$ if $\frac{2(1+2\delta^2)\sqrt{1+\delta^2}}{2+\delta^2} \geq \frac{a-c}{\sqrt{bf}}$. In addition, $n_u^* = [x_u^*]$ and $n_u^{sb} \in \{[x_u^{sb}], [x_u^{sb}] + 1\}$. Thus, $x_u^* \leq x_u^{sb} \implies n_u^* \leq n_u^{sb}$. Combining these two conditions we obtain $\frac{2(1+2\delta^2)\sqrt{1+\delta^2}}{2+\delta^2} > \frac{a-c}{\sqrt{bf}} \implies n_u^* \leq n_u^{sb}$. It is easy to check that $\frac{2(1+2\delta^2)\sqrt{1+\delta^2}}{2+\delta^2}$ is an increasing function of δ . Defining

$\delta_3^* = \left\{ \delta > 0 : \frac{2(1+2\delta^2)\sqrt{1+\delta^2}}{2+\delta^2} = \frac{a-c}{\sqrt{bf}} \right\}$, the above condition is equivalent to $n_u^* \leq n_u^{sb}$

if $\delta \geq \delta_3^*$. This proves the first part of the proposition. For the second part of the proposition, combining the fact that $\frac{2(1+2\delta^2)\sqrt{1+\delta^2}}{2+\delta^2} \leq \frac{a-c}{\sqrt{bf}} \implies x_u^* \geq x_u^{sb}$ and $x_u^* \geq x_u^s \implies n_u^* \geq n_u^{sb}$, we obtain $\frac{2(1+2\delta^2)\sqrt{1+\delta^2}}{2+\delta^2} \leq \frac{a-c}{\sqrt{bf}} \implies n_u^* \geq n_u^{sb}$. Note that the left hand side of this expression is increasing in δ and using the same argument as above we conclude that $n_u^* \geq n_u^{sb}$ if $\delta \leq \delta_3^*$.

CHAPTER 2

Rationing of HIV treatment in resource-constrained settings under supply uncertainty

2.1 Introduction

Access to highly active antiretroviral therapy (HAART) for HIV+ patients in Sub-Saharan Africa and other resource-constrained regions has barely reached 1 million patients despite growing attention from the international donor community (PEPFAR¹, GFATM², CHAI³, etc.) and WHO's ambitious "3 by 5" program which aimed to get 3 million people on treatment by 2005. According to WHO (2005b), the gap between available supply of drugs and demand is unlikely to be eliminated in the near future. Hence, governments and clinics in these countries have to make difficult choices related to the selection of new patients (or treatment rationing) while scaling up HAART programs. Various qualitative guidelines developed by WHO and UNAIDS (Rosen et al., 2005; McGough et al., 2005; Macklin, 2004) exist for the selection of new patients from different segments of the population based on various social, economic and clinical criteria. However, they provide no guidance on how many new patients to enroll in a given

¹President's Emergency Plan for AIDS Relief

²Global Fund for AIDS, Tuberculosis and Malaria.

³Clinton HIV AIDS Initiative.

period.

On the other hand, extensive field work aimed at quantifying and planning for new enrollments has been done by organizations like John Snow Inc. (JSI) and Management Sciences for Health (MSH) as a part of their partnership with PEPFAR. The approach followed in most of this work, so far, is static and deterministic, hence ignoring two important characteristics of the operating systems under consideration. Firstly, in addition to the aggregate shortage, the supply received at individual clinics is highly variable and unpredictable, leading to treatment interruptions (ITPC, 2005; BBC News, 2004; IRINNews.org, 2005). These unanticipated treatment interruptions can lead to adverse clinical outcomes such as treatment failure and drug resistance (IOM, 2005; WHO, 1998). Secondly, enrollment decisions in the current period have a direct impact on the clinic's ability to guarantee treatment continuity in the future periods in the wake of supply uncertainty and shortage.

Thus a clinic administering HAART faces a trade-off between expanding treatment to new patients in the current period but facing an increased probability of stock-outs or treatment interruptions in future periods and restricting access to new patients in the current period but being more likely to offer continuous treatment for current patients. The resulting tension can be sensed in the following quote from a health care provider in Amajuba District, South Africa (Wu, 2004): "...when you run out of stock you begin to stress. You don't know when the stock is coming. We counsel patients so closely on adherence, and on what happens if you miss a dose. Then they come in all frantic and we have to deal with the problems." Our discussions with JSI revealed that decision rules (policies) used in practice to manage this trade-off include 'always enroll a new patient if drugs are available', 'stop enrollment of new patients if the available inventory

reaches a predefined safety stock level' and 'enroll new patients upto a predefined enrollment cap'. However, without formal analysis, it is not immediately apparent how these policies perform relative to each other and which of these, if any, is optimal.

To investigate this issue, we model the clinic's trade-off using a discrete time stochastic dynamic program. At the beginning of each period, the clinic receives a drug shipment of uncertain quantity over which it has little or no control. The clinic faces a deterministic demand from patients who have been treated in previous periods (current patients). In addition, the clinic can initiate treatment for patients from a large pool of previously untreated patients (new patients) reflecting the situation that total supply is not enough to meet total demand. Knowing the available inventory of drugs and the demand from current patients at the beginning of each period, the clinic needs to decide how many current and new patients to treat in each period to maximize the quality adjusted life years of its patients over the planning horizon. We employ this model to answer the following questions:

- (i) What treatment rationing policy maximizes the clinic's objective? What are its salient characteristics?
- (ii) How do above policies followed in practice compare to the optimal policy?
- (iii) What is the impact of supply uncertainty on the performance of various rationing policies?

We first show that the optimal policy for our problem is a 'modified base-stock' type policy where the base-stock levels correspond to the desired level of inventory before receiving supply and the size of current patient pool. Then, we derive conditions on our problem parameters under which it is optimal to

prioritize current patients over new patients, an accepted standard of care. Under these conditions, the optimal policy is characterized by an enrollment cap in each period. If the available inventory is greater than this threshold it is optimal to carry over the excess inventory to the next period as a safety-stock. For the finite horizon formulation, we show that the size of this safety-stock is state-dependent and dynamic.

The remainder of the chapter is organized as follows. In section 2.2, we describe the operational challenges of delivering HAART in resource-constrained settings in greater detail. Section 2.3 provides a brief review of the various streams of literature related to this chapter and outlines our contribution to them. The model formulation is described in section 2.4. The optimal policy and its properties are derived in section 2.5. In section 2.6, we introduce the resource-constrained condition which simplifies the model formulation and allows to study various properties of the optimal policy. Section 2.7 describes heuristics which are either used in practice or have practical appeal. We provide numerical illustrations to compare these heuristics with the optimal policy in section 2.8. Section 2.9 provides concluding remarks. Proofs for all the theoretical results are provided in the appendix.

2.2 Background

Acquired Immunodeficiency Syndrome (AIDS) has caused more than 25 million deaths over the past 25 years. As of 2005, close to 40 million people were living with Human Immunodeficiency Virus (HIV) and around five million were newly infected in 2005 (WHO, 2005a). Sub-Saharan Africa is the worst affected region with just over 10% of the world's population but more than 60% of all the people living with HIV and AIDS. Prevalence rates in many countries in this region

are in the range of 20% to 30%. WHO (2005a, 2005b) provide a more detailed discussion of the epidemiology of the disease and the unprecedented social and economic damage caused by it.

HAART, a complex regimen comprising multiple antiretroviral (ARV) drugs, has been available for treatment against HIV / AIDS in many developed countries since the late 1990s (Bartlett, 2006). While HAART can neither cure nor prevent HIV infection and AIDS, it has considerably reduced mortality and morbidity in HIV+ patients in the U.S. (Palella et al., 1998) and saved more than three million life years (Walensky et al., 2006). However, less than 20% of the eligible patients in Sub-Saharan Africa and other developing regions of the world, receive HAART despite the recent expansion of treatment because of (i) a reduction of drug prices by around 37% - 53% (WHO, 2006), (ii) a multifold increase in long-term funding by GFATM, World Bank and PEPFAR, and (iii) an increased awareness as a result of the WHO's "3 by 5 initiative".

The conceptual model presented in section 2.11 elaborates on various challenges that contribute to this slow progress. In this chapter, we abstract from these complexities to focus our attention on the operational bottlenecks such as limited capacity for activities such as storage and inventory control, quantification and reporting, and security of commodities (GAO, 2006). A major consequence of these bottlenecks is the uncertainty in the supply of drugs received by the clinics. This supply uncertainty has important implications for the treatment rationing decisions made by the clinics through periodic stock-outs of drugs. To maintain our focus, we do not model the impact of treatment on prevention through modified behavior of patients, reduced viral load and increased number of patients willing to test. We also do not consider the impact of current program outcomes on future resource availability.

Various incidents of drug stock-outs have been reported in various parts of the world including India, Russia, Dominican Republic (ITPC, 2005), Nigeria (Ekong et al., 2004), South Africa (BBC News, 2004), Kenya (IRINNews.org, 2002) and Swaziland (IRINNews.org, 2005). In addition to this anecdotal evidence, logistics assessment surveys commissioned by the USAID and conducted by JSI provide systematic evidence of stock-outs and supply uncertainty in Zimbabwe (Nyenwa et al., 2005) and Tanzania (Amenyah et al., 2005). These stock-outs cause interruption of treatment for patients which could lead to drug resistance and / or treatment failure (Bartlett, 2006). Oyugi et al. (2007) and van Oosterhout et al. (2005) provide systematic evidence of this phenomenon in Uganda and Malawi. In extreme cases, drug shortages due to supply interruptions have also resulted in the death of patients in South Africa (Health Systems Trust, 2005).

However, this underlying supply uncertainty has not received enough attention in the quantification and forecasting tools used by clinics or in the academic literature. Current approaches include using informal guidelines for deciding a safety stock to manage this uncertainty. There is an urgent need for formal models to quantify the safety stock required to optimally manage the underlying supply uncertainty while scaling up HAART (Daniel, 2006). This is important because overdesigning the safety stock would mean blocking scarce funds in nonproductive assets and slowing treatment expansion, while underestimating the safety stock could result in extremely undesirable stock-outs and treatment interruptions.

2.3 Literature Review

The mathematical model presented in this chapter extends the existing inventory rationing models by explicitly modeling the conversion of customers from one segment to the other. This chapter contributes to the broader literature on health

care rationing by incorporating quality and access considerations in the objective function. In particular, it contributes to the literature on resource allocation for HIV / AIDS interventions, which has predominantly focused on prevention. In the context of HAART in resource-constrained settings, it complements the existing qualitative discussion by providing a quantitative framework for rationing treatment between new and current patients at the clinics. Our model can also be used to analyze the resource allocation decisions for non-profit organizations where uninterrupted service provision is crucial to meeting the organization's social objective and thus contributes to the yet sparse operations research literature on non-profits.

Our model is related to the models of inventory rationing among customer classes of differing priorities (Topkis, 1968; Evans, 1969, Nahmias and Demmy, 1981; Ha, 1997a, 1997b; de Véricourt et al., 2002). The optimal allocation policy in these models consists of a threshold or reservation level for each segment such that it is optimal to stop serving a segment if the on-hand inventory drops below the threshold associated with that segment. Frank et al. (2003) and Zhang and Sobel (2001) study inventory rationing schemes where demand from one segment has to be met while demand from the other segment can be either backlogged or lost at a penalty. The customer segments in these models are unrelated, i.e., customers do not move from one segment to the other as a result of receiving service. In contrast, in our model, the two customer segments are inherently related as customers from one pool (previously untreated) are moved permanently to another pool (previously treated) as a result of the treatment decisions. Olsen and Parker (2006) model flows of customers from one segment to the other but do not consider rationing.

Most literature on health care rationing focuses on developed countries with

the key tradeoff being between efficiency and equity. Wagstaff (1991) provides a detailed discussion of the underlying concepts. In the operations literature, Zenios et al. (2000) present a normative model for allocating cadaveric kidneys among various patient segments. In their context, the key trade-off between equity and efficiency is exacerbated because of the constrained supply of kidneys. In contrast, the key trade-off in our model is between access (enrolling more patients) and quality (providing uninterrupted treatment to enrolled patients) which is exacerbated by the uncertainty in the future supply of drugs.

Considerable work has been done in combining epidemiological models and optimal control theory to study dynamic allocation of resources in the case of HIV epidemics (Richter et al., 1999; Kaplan and Pollack, 1998). However, the focus of these models is on prevention interventions and there is no uncertainty regarding the availability of resources in these models. In contrast, we focus on treatment programs for HIV. We considerably simplify the epidemiological component of our model and choose to focus instead on the uncertainty in resource availability (drug supply) as a key dimension of our resource allocation problem.

There has been recent qualitative discussion on rationing strategies for HAART in developing countries that focus on the issue of “which” new patients to enroll (Rosen et al., 2005; Bennet and Chanfreau, 2005). However, it pays inadequate attention to two important characteristics of HAART scale-up - (i) patients once enrolled have to be treated continuously through their life and (ii) there is a variability in supply of drugs in addition to the aggregate shortage. We complement this literature by incorporating these characteristics in a quantitative model that to help clinics decide “how many” new patients to enroll when accurate information about the future supply of drugs is not available.

The model in this essay could also be applied to resource allocation problems

in non-profits. To our knowledge, the only other paper that uses operations research methods to analyze resource allocation decision in non-profit organizations is de Véricourt and Lobo (2006). They study the allocation of the organization's assets among mission and revenue customers so as to maximize the total discounted social benefit. However, not all non-profit organizations can engage in for-profit activities due either to lack of requisite skills or the domain of their activities (Dees, 1998; Foster and Bradach, 2005). Such non-profits have to depend entirely on external funding sources which are known to be highly unreliable and variable (Gronbjerg, 1992). Also, in homeless shelters and drug rehabilitation programs, it is critical to maintain continuity of service to current beneficiaries while expanding service to new beneficiaries (Scott, 2003). Our model could be adapted to incorporate these concerns and complement the model in de Véricourt and Lobo (2006).

2.4 Model Formulation

In this section, we present the formal model for the decision problem of an individual clinic in a resource-constrained setting that wants to maximize the expected total discounted quality adjusted life years (QALYs) of its patients. Let T denote the length of the problem horizon consisting of discrete decision making epochs $t = 1, 2, 3 \dots T$ where $t = 0$ denotes the end of the horizon.

2.4.1 Drug supply

Current distribution systems for ARTs in resource-constrained settings consist of central medical depots that are typically situated at the provincial or district headquarters. The drugs are “pushed” from these depots to the sites of health

care delivery (WHO, 2005a; WHO, 2003). The ultimate goal is to move to a more formal system where clinics order drugs based on their requirements. However, inadequate inventory management skills at the clinics make this transition from “push” to “pull” both difficult and slow (JSI, 2006; WHO, 2003). Also, due to a weak transport infrastructure, the drug supply actually received at the clinics is uncertain.

To reflect this situation, we model the supply of drugs as exogenous but stochastic; order quantity is not a decision variable for the clinic. Extending the model to include ordering decision by clinics would be interesting but appears to be analytically intractable. Let \tilde{Z}_t be independently (not necessarily identically) distributed random variables that denote the supply of drugs received by the clinic in period t with cumulative distribution $\Phi_t(\cdot)$ and support on $[z_t^L, z_t^U]$. Thus at the time of deciding the number of new and current patients to treat in period t , the clinic does not know the actual quantity of drugs it will receive in the future periods ($1 \leq u < t$) but only knows the cumulative distribution $\Phi_u(\cdot)$. Let I_t and W_t denote the inventory of drugs before and after receiving the supply in period t so that $W_t = I_t + \tilde{Z}_t$.

2.4.2 Patients

The demand for drugs consists of patients who have been diagnosed as HIV+ and are eligible for treatment based on the national guidelines. We model the demand at the clinic to be composed of two pools of patients: $y_{t,c}$ denotes the number of current (previously treated) patients and $y_{t,n}$ denotes the number of new (previously untreated) patients at the time of deciding treatment allocations. The decision on which segments of patients to prioritize, based on socio-economic and clinical characteristics (CD4+ count) is made at a national level (Bennet and

Chanfreau, 2005) and the clinic faces demand from patients which are roughly similar on these attributes. Moreover, the health status of patients as reflected by both CD4+ count and QOL score becomes reasonably homogenous after around six months of HAART (Cleary et al., 2006). Hence we model the pool of new and current patients for an individual clinic to be homogenous along these attributes and capture the average health status of each pool for our analysis. This substantially improves the analytical tractability of our model.

2.4.3 System dynamics

In each period t , knowing the available inventory W_t and the demand from current and new patients $y_{t,c}$ and $y_{t,n}$ respectively, the clinic decides on the number of current and new patients to treat denoted by $x_{t,c}$ and $x_{t,n}$ respectively. After the treatment decisions, the inventory of drugs drops to $I_{t-1} = W_t - x_{t,c} - x_{t,n}$ and the pool of new patients reduces to $y_{t,n} - x_{t,n}$. At the end of each period, a deterministic fraction β_1 of all current patients and β_2 of all new patients survive through to period $t - 1$ and the remaining patients die. Thus $\beta_2 x_{t,n}$ denotes the number of patients who were initiated on treatment in period t and survived, thus adding to the pool of current patients in period $t - 1$. The number of new diagnoses occurring at the beginning of the next period $t - 1$ is denoted by $\alpha y_{t,n}$. Thus the system dynamics are given by the following set of equations:

$$y_{t-1,c} = \beta_1 y_{t,c} + \beta_2 x_{t,n} \quad (2.1)$$

$$y_{t-1,n} = (\beta_2 + \alpha) (y_{t,n} - x_{t,n}) \quad (2.2)$$

$$W_{t-1} = W_t - x_{t,c} - x_{t,n} + \tilde{Z}_{t-1} = I_{t-1} + \tilde{Z}_{t-1} \quad (2.3)$$

There are several assumptions associated with the dynamics of our model. First, the average survival rates β_1 and β_2 and the rate of new diagnoses α

are assumed to be known with certainty. Including uncertainty in the survival rates is non-trivial but our model can be adapted to include survival rates that are independent of the uncertainty in the drug supply. Second, β_1 and β_2 are not impacted by the treatment status of current and new patients in period t . Inclusion of survival rates that depend on the treatment status in the current period would not change our results but would make the analysis algebraically more cumbersome. Third, we assume that $\beta_2 + \alpha > 1$ to reflect the situation that the rate of new diagnoses is higher than the mortality rate (WHO, 2005). With respect to the drug supply, we assume that there is no limit on the available storage for drugs and that the drugs are not perishable. This is true for all the drugs used in the first line of treatment, which is our focus here.

2.4.4 Objective function

As mentioned earlier, the objective of the clinic is to maximize the total quality adjusted life years (QALYs) for the patient population over the planning horizon T . While QALYs have been traditionally used for clinical decision making at an individual level, there has been a recent trend to use QALYs at a population level to evaluate alternate policy measures (Zenios et al., 2000; Richter et al., 1999). See Loomes and Mckenzie (1985) for a detailed discussion of the related issues.

As discussed before, our patient population consists of two patient pools - current and new - based on their treatment history. We further divide each of these pools into two subcategories based on the treatment status in the current period and assign a quality of life (QOL) score to each of these four categories. Thus s_1 denotes the QOL score for current patients who receive treatment in the current period, s_2 denotes the QOL score for current patients whose treatment is interrupted in the current period. Similarly s_3 denotes the QOL score for

new patients who receive treatment for the first time in the current period and s_4 denotes the QOL score for new patients who have never received treatment. Since we are not modeling the difference in health status of patients within each subcategory, s_1, s_2, s_3 and s_4 could be considered as average QOL scores for each of the four subcategories. Furthermore, since we assume that the underlying composition of these categories does not change over the problem horizon these parameters are time invariant. Thus the objective of the clinic for a finite horizon T is given by

$$\max_{x_{t,n} \geq 0, x_{t,c} \geq 0} E \left[\sum_{t=1}^T \delta^{T-t} h_t(x_{t,c}, x_{t,n}, y_{t,c}, y_{t,n}) \right] \quad (2.4)$$

where δ is a single period discount factor and h_t is the single period reward function given by:

$$\begin{aligned} h_t(x_{t,c}, x_{t,n}, y_{t,c}, y_{t,n}) &= s_1 x_{t,c} + s_2 (y_{t,c} - x_{t,c}) + s_3 x_{t,n} + s_4 (y_{t,n} - x_{t,n}) \\ &= (s_1 - s_2) x_{t,c} + (s_3 - s_4) x_{t,n} + s_2 y_{t,c} + s_4 y_{t,n} \end{aligned} \quad (2.5)$$

Similar objective function has been used previously to analyze resource allocation decisions in the context of epidemics (Brandeau et al., Richter et al.). To denote that treatment in the current period has positive benefit for both current and new patients, it is reasonable to assume $(s_1 - s_2) > 0$ and $(s_3 - s_4) > 0$.

2.5 Optimal policy

Using the above building blocks, we now state the decision problem for the clinic under consideration. Let $V_t(W_t, y_{t,c})$ denote the maximum net benefit from the clinic's treatment decisions for the remaining t periods till the end of the horizon. Then the clinic's decision problem in period t is given by:

$$\begin{aligned}
V_t(W_t, y_{t,c}, y_{t,n}) &= \max_{x_{t,n} \geq 0, x_{t,c} \geq 0} \{h_t(x_{t,c}, x_{t,n}) + \delta E[V_{t-1}(W_{t-1}, y_{t-1,c}, y_{t-1,n})]\} \\
&\text{s.t. (2.1), (2.2) and (2.3)} \\
&\quad x_{t,c} \leq y_{t,c} \\
&\quad x_{t,n} \leq y_{t,n} \\
&\quad x_{t,n} + x_{t,c} \leq W_t \quad (2.6)
\end{aligned}$$

where $h_t(x_{t,c}, x_{t,n}) = (s_1 - s_2)x_{t,c} + (s_3 - s_4)x_{t,n} + s_2y_{t,c} + s_4y_{t,n}$.

Equations (2.1), (2.2) and (2.3) are system dynamics described earlier. The next two constraints state that the number of current and new patients treated cannot be more than the total number of current and new patients in that period respectively. The last constraint states that the total number of treatments delivered in period t is limited by the available inventory. We also define $V_0(\cdot) \equiv 0$.

The model (2.6) is similar to the two-product inventory control models studied by Deuermeyer and Pierskalla (1978), Evans (1967) and Simpson (1978) and the multi-location multi-period inventory model studied by Karmarkar (1981). We use this similarity in the structure to derive the optimal policy for our problem in the next section. Let us define the maximand in (2.6) as

$$f_t(x_{t,c}, x_{t,n}) \triangleq h_t(x_{t,c}, x_{t,n}) + \delta E[V_{t-1}(W_{t-1}, y_{t-1,c}, y_{t-1,n})]$$

Proposition 7. (i) $V_t(W_t, y_{t,c}, y_{t,n})$ is jointly concave in its arguments. (ii) $f_t(x_{t,c}, x_{t,n})$ is jointly concave in its arguments. (iii) The optimal policy for (2.7) is characterized by the existence of a vector $S^* = (I_t^*, y_{t,c}^*, y_{t,n}^*)$ such that it is optimal to move from the state vector $(W_{t+1}, y_{t+1,c})$ to S^* at the beginning of period t before supply is received, provided such a move is feasible. If this move is not feasible, then it is optimal to reach to the vertex of the feasible region that is closest to vector S^* .

The proof for the first two parts of Proposition 7 follows from the usual inheritance properties of dynamic programs. Thus in each period t , (2.7) is a concave optimization problem on a set of linear constraints. Hence the third part follows directly from the multi-location multi-period inventory model of Karmarkar (1981). Thus the optimal policy for our problem is similar to the modified base-stock policy for inventory problems with side-constraints.

A more explicit characterization of the optimal policy is possible using geometric interpretation if the state-space is two-dimensional (Deuermeyer and Pier-skalla, 1978; Evans, 1967 and Simpson, 1978). Their method essentially involves considering different regions of the state-space depending on whether one or more constraints are tight and then using KKT conditions to derive the optimal solution in each of the regions, since in each period, (2.8) is a constrained concave maximization problem. Then induction would be used to show that this structure of the optimal policy is preserved through the state transitions.

2.6 Model reformulation

The method for characterizing the optimal policy discussed above is not directly applicable in our case because our state-space is three dimensional and hence geometric interpretation becomes very complex. Next, we use the characteristics of the resource-constrained setting to reduce the state-space of our problem and thus make the problem more tractable.

2.6.1 Resource-constrained condition

At present, drug supply in many developing countries is enough to reach only a small fraction of all the eligible patients. WHO does not expect to reach its

target of universal access until atleast 2010 (WHO, 2006) and PEPFAR targets to put only 2 million people on treatment by 2009 (PEPFAR, 2006). Moreover, according to the epidemiological update by WHO (2005b): “*Indications are that some of the treatment gaps will narrow further in the immediate years ahead, but not at the pace required to effectively contain the epidemic*”. To reflect this situation, we assume that $y_{t,n} > W_t \forall t$, i.e., the demand for drugs will outstrip supply over the planning horizon. In the appendix we provide an upper bound on the support of distributions $\Phi_t(\cdot)$ so that this condition is satisfied. Since $y_{t,n} > W_t \forall t$ the feasible set $x_{t,n}$ does not depend on $y_{t,n}$. Substituting $y_{t,n} = (\beta_2 + \alpha)^{T-t} y_{T,n} - \sum_{i=t+1}^T (\beta_2 + \alpha)^{i-t} x_{i,n}$ and leaving out the constant term involving $y_{T,n}$ in (2.4), the objective function becomes

$$\max_{x_{t,n} \geq 0, x_{t,c} \geq 0} E \left[\sum_{t=1}^T \delta^{T-t} \left((s_1 - s_2) x_{t,c} + \left(s_3 - s_4 \sum_{i=0}^{t-1} (\delta (\beta_2 + \alpha))^i \right) x_{t,n} + s_2 y_{t,c} \right) \right]$$

Defining $s_4 \sum_{i=0}^{t-1} (\delta (\beta_2 + \alpha))^i \triangleq s_4^t$ and $\hat{h}_t(x_{t,c}, x_{t,n}) = (s_1 - s_2) x_{t,c} + (s_3 - s_4^t) x_{t,n} + s_2 y_{t,c}$, the formulation in (2.6) can be modified as

$$V_t(W_t, y_{t,c}) = \max_{x_{t,n} \geq 0, x_{t,c} \geq 0} \left\{ \hat{h}_t(x_{t,c}, x_{t,n}) + \delta E[V_{t-1}(W_{t-1}, y_{t-1,c})] \right\} \quad (2.7)$$

s.t. (2.1), and (2.3)

$$x_{t,c} \leq y_{t,c}$$

$$x_{t,n} + x_{t,c} \leq W_t$$

Thus, we have used the resource-constrained condition to reduce the state-space, but this introduces non-stationarity in one of the problem parameters, s_4^t . For the remainder of the chapter, we shall focus on this formulation.

2.6.2 Two-period model

In this section, we shall solve the most simple, yet non-trivial instance of 2.7 for $T = 2$ to highlight some of the difficulties associated with the formulation.

Specifically, we shall show that the optimal solution can be clinically unacceptable. When then derive additional conditions on the range of parameters that are sufficient to ensure that the optimal policy possess structural properties that are desirable from the perspective of clinical standpoint. We use the same notation as described in section 2.4. The only source of uncertainty in this problem is the quantity of drugs received at $t = 1$ whose cumulative distribution is denoted by $\Phi_1(\cdot)$. At $t = 2$, the available inventory W_2 and size of the current patient pool $y_{2,c}$ are known. The clinic has to decide the number of new and current patients to be treated in period 2 and 1 denoted by $x_{2,c}$, $x_{2,n}$, $x_{1,c}$, $x_{1,n}$. We need the following definitions in order to characterize the optimal policy:

$$k_1 = \frac{(s_2 - s_4(1 - \delta\alpha)) - (1 - \delta\beta_2)(s_1 - s_3)}{\delta\beta_2((s_1 - s_2) - (s_3 - s_4))}$$

$$k_2 = \frac{s_3 - s_4(1 + \delta(\beta_2 + \alpha)) + \delta(s_1 - (s_3 - s_4)(1 + \beta_2))}{((s_1 - s_2) - (s_3 - s_4))(1 + \delta\beta_2)}$$

Proposition 8. *The optimal policy for the two period problem is given by:*

Case I $s_1 - s_2 > s_3 - s_4$:

$$x_{2,c}^* = \min \{y_{2,c}, [W_2 - \eta]^+\}, \quad x_{2,n}^* = \min \{W_2 - \min \{y_{2,c}, [W_2 - \eta]^+\}, \theta\},$$

$$x_{1,c}^* = \min \{y_{1,c}, W_1\}, \quad x_{1,n}^* = [W_1 - y_{1,c}]^+ \text{ where}$$

$$\eta = \min \left\{ x_{t,n} \geq 0 : \frac{\partial f_2}{\partial x_{2,n}} \leq \frac{\partial f_2}{\partial x_{2,c}} \Big|_{x_{2,n} + x_{2,c} = W_2} \right\} = \left[\frac{\Phi_1^{-1}(k_1) - \beta_1 y_{2,c}}{\beta_2} \right]^+$$

$$\theta = \min \left\{ x_{2,n} \geq 0 : \frac{\partial f_2}{\partial x_{2,n}} \Big|_{x_{2,c} = y_{2,c}} \leq 0 \right\} = \left[\frac{\Phi_1^{-1}(k_2) + W_2 - (1 + \beta_1) y_{2,c}}{1 + \beta_2} \right]^+$$

Case IIA $s_3 - s_4 > s_1 - s_2$ and $s_4 - s_2 < \frac{(s_3 - s_4) - (s_1 - s_2)}{\delta\beta_2}$:

$$x_{2,c}^* = 0, \quad x_{2,n}^* = W_2, \quad x_{1,c}^* = [W_1 - y_{1,n}]^+, \quad x_{1,n}^* = \min \{W_1, y_{1,n}\}$$

Case IIB $s_3 - s_4 > s_1 - s_2$ and $s_4 - s_2 > \frac{(s_3 - s_4) - (s_1 - s_2)}{\delta\beta_2}$:

$$x_{2,c}^* = \min \{W_2, y_{2,c}\}, x_{2,n}^* = [W_2 - y_{2,c}]^+, x_{1,c}^* = 0, x_{1,n}^* = W_1$$

Proposition 8 shows that even for the simple two period problem, the structure of the optimal policy is quite complicated. The optimal policy depends not only on the relative values of the QOL parameters but also on the current system state.

In case I, $s_1 - s_2 > s_3 - s_4$ implies that the value of treatment is higher for previously treated patients. Then, as expected it is optimal to prioritize current patients in $t = 1$. However, the prioritization is not unambiguous for $t = 2$ and it depends on the values of the thresholds η and θ . Consider the case when $W_2 - y_{2,c} < \eta < W_2$ and $\eta < \theta$. Then $x_{2,c}^* = W_2 - \eta$ and $x_{2,n}^* = \eta$. Thus in this case, new patients are enrolled before all the current patients have been treated. This is because the marginal value from treating a current patient is equal to the marginal value from treating a new patient when the supply constraint is tight. Thus shifting the treatments from new to current patients would reduce the objective function due to its concavity.

In cases IIA and IIB, the condition $s_3 - s_4 > s_1 - s_2$ implies that the value of treatment is higher for previously untreated patients. In both these cases, it is optimal to prioritize treatment for new patients for $t = 1$. However, this condition is not sufficient to maintain the prioritization for both periods. Prioritization for new patients is maintained for $t = 2$ only if s_4 is not sufficiently greater than s_2 (Case IIA). However, if s_4 is sufficiently greater than s_2 , the prioritization is reversed and it is optimal to prioritize current patients for $t = 2$ (Case IIB).

2.6.3 Prioritization of current patients

The optimal policy in Proposition 8 is undesirable, partly because of its complexity, but also because it allows allocation policies that are clinically unacceptable. Recent clinical studies have clearly shown that even structured treatment interruptions can drastically increase the mortality and morbidity in HIV+ patients (El-Sadr et al., 2006). Hence, continuous treatment for life is the recommended practice once a patient is enrolled for treatment (IOM, 2005). In terms of the model, this is equivalent to saying $x_{2,n}^* > 0$ only if $x_{2,c}^* = y_{2,c}$ irrespective of the state variables W_2 and $y_{2,c}$. Here we investigate additional conditions on parameter values to guarantee that the optimal policy has this structure. It is clear that $s_1 - s_2 > s_3 - s_4$ is required for prioritization of current patients in the last period. However, Proposition 8 noted that this is not sufficient to ensure prioritization of current patients for $t = 2$. Note that if $\eta = 0$ then the optimal policy in Case I would be equivalent to prioritizing current patients for $t = 2$. In the next proposition, we build on this idea to derive sufficient conditions on the parameter values so that prioritization of current patients is optimal in all periods.

Proposition 9. *It is optimal to prioritize current patients over new patients in every period if the following conditions are satisfied:*

$$(C1) \quad (s_1 - s_2) > (s_3 - s_4)$$

$$(C2) \quad s_2(1 - \delta(\beta_1 - \beta_2)) < (s_1 - s_3)(1 - \delta\beta_1) + s_4(1 - \delta(\beta_2 + \alpha - \beta_1))$$

Moreover, the optimal solution under these conditions is given by

$$x_{t,c}^* = \min\{y_{t,c}, W_t\} \text{ and } x_{t,n}^* = \min\{\theta_t, [W_t - y_{t,c}]^+\} \text{ where}$$

$$\theta_t = \min\left\{x_{t,n} \geq 0 : \left.\frac{\partial f_t}{\partial x_{t,n}}\right|_{x_{t,c}=y_{t,c}} \leq 0\right\}.$$

Condition (C1) states that, on average, the health benefit from treating a previously treated patient is higher than the health benefit from treating a new patient. This is reasonable since not treating a previously treated patient can lead to development of drug resistance and eventually lead to transmission of these resistant strains to the susceptible population, a clear negative externality. Condition (C2) is less easy to interpret, but it helps to consider a few special cases. For $\delta = 0$, (C2) reduces to (C1) confirming that for a single period problem only (C1) is sufficient to guarantee prioritization of current patients. For $\beta_1 = \beta_2$ and $s_1 = s_3$, (C2) reduces to $s_2 < s_4(1 + \delta\alpha)$ which implies that the average QOL score of patients with interrupted treatment should not be too much higher than the average QOL score of unenrolled patients. Intuitively, if (C2) is not satisfied then the penalty from treatment interruption is not high enough to warrant prioritization of current patients over new patients.

Now let us analyze the implications of the optimal solution. Treating an additional new patient in the current period has three effects. First, there is an immediate social benefit of the treatment, s_3 . Second, it reduces the available inventory to be carried over to the next period. Third, it increases the pool of current patients by β_2 (adjusting for mortality). The uncertainty regarding the supply of drugs in the future periods implies that there is an increased chance that this newly converted patient might go untreated in the future. Thus the optimal policy balances the expected penalty of interrupting the treatment of previously treated patients in the future periods with the immediate benefit of treating an additional new patient. The quantity $[W_t - \theta_t - y_{t,c}]^+$ could be interpreted as the safety stock to be carried over to protect current patients against the future supply uncertainty.

It is instructive to contrast Proposition 9 with the results from traditional

models of ordering and rationing inventory across multiple customer classes. The optimal policy in these models usually involves thresholds, one corresponding to each customer segment, such that it is optimal to not serve a particular segment if the on-hand inventory falls below the corresponding threshold (Topkis, 1969; Ha, 1997a, 1997b; de Véricourt et al., 2001). This enables the decision maker to carry enough safety stock to protect against the uncertain demand from high value customers in future periods. However, in our model, since the supply is uncertain and beyond the control of the clinic, a safety stock is built and maintained by restricting the enrollment of new patients. This serves to protect the “higher value” patients (previously treated patients) from any unanticipated supply interruptions in future periods. Next proposition provides more insight into the structure of the threshold θ_t for the special case of $\beta_1 = \beta_2$.

Proposition 10. *Let $\beta_1 = \beta_2$. If $\theta_t > 0$, then (i) $\theta_t = \psi_t(W_t) - y_{t,c}$ and (ii) $F(z_{t-1}) \underset{fsd}{>} G(z_{t-1})$ implies that $\theta_t(F) > \theta_t(G)$.*

When $\beta_1 = \beta_2$ and $\theta_t > 0$, the system states in period $t - 1$ depend only on $y_{t,c} + \theta_t$. Hence, what matters is how much total inventory was dispensed rather than how this was divided between new and current patients. Part (i) of Proposition 10 shows that in this case, the optimal policy is equivalent to carrying over a fraction of the available inventory to the next period as a safety stock. This fraction is given by $\frac{W_t - (\theta_t + y_{t,c})}{W_t} = \frac{W_t - \psi_t(W_t)}{W_t}$. Since the supply is stochastic and dynamic, this fraction is not a constant but depends on the available inventory in that period. Part (ii) shows that if the next period’s drug supply is stochastically greater then everything else being equal, the safety stock would be reduced, or equivalently, the enrollment cap θ_t would increase.

2.6.4 Infinite horizon

The model discussed so far is for a horizon of finite length denoted by T . However, analysis of an infinite horizon model could be appropriate if T is not known with certainty or if T is long enough so that the infinite horizon problem can be considered as an approximation to the finite horizon problem. The infinite horizon problem corresponding to (2.7) is stated below:

$$\begin{aligned}
 V^* &= \max_{x_{t,n} \geq 0, x_{t,c} \geq 0} \sum_{t=1}^{\infty} \delta^{t-1} \hat{h}_t(x_{t,c}, x_{t,n}) & (2.8) \\
 & \text{s.t. } x_{t,c} \leq y_{t,c} \quad \forall t \\
 & \quad x_{t,n} + x_{t,c} \leq W_t \quad \forall t
 \end{aligned}$$

The corresponding recursive equation in the infinite horizon case is given by:

$$\begin{aligned}
 V(W, y_c) &= \max_{x_n \geq 0, x_c \geq 0} \left\{ \hat{h}(x_c, x_n) + \delta E[V(W - x_c - x_n + Z, \beta(y_c + x_n))] \right\} \\
 & \text{s.t. } x_c \leq y_c \\
 & \quad x_n + x_c \leq W \quad (2.9)
 \end{aligned}$$

However, for the infinite horizon formulation to be meaningful, the resource-constrained condition needs to be satisfied for all periods. A sufficient condition for this to happen is provided in the appendix. Other technical challenges in our formulation, which make the infinite horizon problem difficult are (i) the single period reward function $h(\cdot)$ and hence the value function $V(\cdot)$ is unbounded since W is not uniformly bounded from above, and (ii) the underlying state-space is continuous.

Following the approach by Lippman (1974) and Van Nunen and Wessels (1978) among others, we define a modified sup norm that bounds V . Also, to resolve the issue of continuous state-space, we allow only Borel measurable policies to ensure that the underlying transition functions have the Feller property

(Stokey et al., 1989). Using these modifications, we can redefine a Banach space over which the contraction mapping approach (Denardo, 1967) can be applied to show that the equation (2.9) has a unique fixed point. This result is summarized in the following Proposition and the details of our approach are given in the appendix. A similar approach has been used for consumption investment problems by Abrams and Karmarkar (1979) and Miller (1974).

Proposition 11. *The recursive equation (2.9) has a unique solution \hat{V} , which satisfies $\hat{V} = V^* = \lim_{t \rightarrow \infty} V_t$ and there exists a unique optimal policy such that V^* is attained.*

2.7 Enrollment heuristics

Proposition 9 describes that under conditions (C1) and (C2) it is optimal to prioritize the treatment for current patients and the enrollment of new patients is characterized by a threshold θ_t . While prioritization of current patients is followed in practice, the enrollment policies that are actually implemented have much simpler structure compared to the threshold policy, which involves solving the recursive dynamic program (2.7). In this section, we describe two such heuristics that have practical appeal and contrast them with the optimal prioritization policy from Proposition (9); in the next section we report numerical illustrations to evaluate the heuristics.

2.7.1 Safety-stock policy

A common approach recommended in real life scale up situations is to maintain a safety-stock equivalent to a few months of demand to buffer against supply uncertainty and probable treatment interruptions in the future periods (Chandani

and Muwonge, 2003; WHO, 2003; WHO, 2004). We assume that even under this policy, current patients are always prioritized over new patients. Thus, using our previous notation, a safety-stock policy can be denoted by

$$x_{t,c}^H = \min \{y_{t,c}, W_t\} \text{ and } x_{t,n}^H = [W_t - (a + 1) y_{t,c}]^+ ; a > 0 \quad (2.10)$$

where the superscript H denotes heuristic and $ay_{t,c}$ is the safety-stock, equivalent to a periods of demand from current patients. The popularity of this approach is largely due to its simplicity and intuitive appeal and widespread use in traditional inventory systems. However, even among organizations that carry out logistics and supply chain implementations, there is a recognition that this simple approach might not be optimal and a more scientific approach is needed (Daniel, 2006). We shall compare the performance of this policy with that of the optimal policy in Section 2.8.

2.7.2 Myopic policy

As seen from Proposition 9, the optimal policy involves the possibility of holding on to scarce drugs even though there is an inexhaustible pool of new patients that could be enrolled for treatment. This aspect of the optimal policy could be unappealing to health care practitioners for ethical reasons. Moreover, many health care programs including WHO's 3-by-5 campaign and PEPFAR programs have explicitly focused on number of enrolled patients as a measure of program success. Our interaction with supply chain consultants working in this area revealed that there is a lot of political pressure to put as many people on treatment as possible without fully considering the potential future impact of these enrollment decisions.

An extreme form of such a policy that focuses only on the current period and completely ignores the impact of new enrollments on the ability to continue

treatment in the future is obtained by solving the single period problem. This myopic policy is given by $x_{t,c}^m = \min\{y_{t,c}, W_t\}$ and $x_{t,n}^* = [W_t - y_{t,c}]^+$. The next proposition provides a sufficient condition for such a myopic policy to be optimal.

Proposition 12. *A myopic policy is optimal if (C1) and (C2) and the following condition is satisfied:*

$$(C3) \quad (s_3 - s_4) \geq \delta (s_1 - s_2) + [s_4 - s_2]^+ \sum_{u=1}^{T-1} (\delta (\beta_2 + \alpha))^u$$

First note that for $\delta = 0$, (C3) reduces to $s_3 \geq s_4$ which we have assumed to be true. Thus if the future is completely discounted, myopic policy is optimal, as expected. Now for $\delta > 0$, if $s_2 > s_4$, (C3) reduces to $(s_3 - s_4) \geq \delta (s_1 - s_2)$. This is because enrolling a new patient transfers the patient from a pool of low QOL score and survival rate into a pool of high QOL score and high survival rate thus increasing the total QALY score of the clinic. Hence the only relevant comparison is between improving the QOL score of a new patient today and improving the QOL score of a current patient tomorrow. On the other hand, if $s_2 < s_4$, it implies that the average QOL score of patients with treatment interruptions is worse than that of the new patient pool. Hence a myopic policy would be optimal only if the benefit from treating a new patient today outweighs the cost from interrupting treatment for a current patient in all the future periods.

2.8 Numerical illustrations

In this section, we provide numerical illustrations to evaluate the performance of the two enrollment heuristics (myopic policy and safety-stock policy) described in Section 2.6.3 relative to the optimal enrollment policy. Our primary objective in this exercise is to examine the impact of the supply uncertainty on the performance of these heuristics.

2.8.1 Setting parameter values

In our model, the patient segments are defined on the basis of current and past treatment status and not directly on health status. As a result, the QOL parameters required in our model were not directly available in the existing literature. Table 2.1 shows the parameter values chosen for the numerical illustrations and the source for each of them. The first two studies (Tengs and Lin, 2002; Holtgrave and Pinkerton, 1997) are meta-analyses of various studies conducted in the U.S. Jelsma et al. (2005) examined the health status of HIV+ patients in South Africa using a visual analog scale (VAS) which were then converted into utilities using time-tradeoff method. These scores and the methodology are reported in Cleary et al. (2006). We assumed that patients starting on HAART would have CD4+ counts less than 200 or show clinical symptoms of AIDS. Hence based on the three sources, s_4 was chosen to 0.65. For estimating s_3 , we assumed that after one month of treatment, CD4+ count of patients would increase and on an average be between 200 and 399 but would be symptomatic. For s_1 , we assumed that previously enrolled patients who receive uninterrupted treatment in this period would be asymptomatic. However, since Cleary et al. (2006) reported lower QOL for patients on HAART for a year, we adjusted our estimate downwards from 0.94 to 0.90. It was relatively difficult to estimate s_2 using the available data as it would depend on the fraction of patients developing drug resistance or other adverse outcomes as a result of treatment interruption. Hence, we decided to use values between 0.65 and 0.80 and examine the sensitivity of our results to this variation.

We choose $T = 24$ to reflect a time horizon of 2 years which is quite natural for resource-limited setting. Thus each period can be thought of as equivalent to a month. Use of discounting in health economics is not free of controversy

Table 2.1: Quality of Life estimates

Parameter	Tengs and Lin (2002)	Holtgrave and Pinkerton (1997)	Jelsma et al. (2005); Cleary et al. (2006)	Values chosen
s_1	0.93 (Asymptomatic HIV infection)	0.94 (Asymptomatic HIV infection)	0.85 (ART >12 months)	0.90
s_2	0.81 (Symptomatic HIV infection)	No appropriate estimate	No appropriate estimate	0.65-0.80
s_3	0.81 (Symptomatic HIV infection)	0.70-0.80 (200 <CD4 < 499)	0.71 (ART 0-3 months)	0.75
s_4	0.60-0.70 (CD4 < 200 or Clinical AIDS)	0.60-0.65 (CD4 < 200 or Clinical AIDS)	0.71 (HIV+; no ART)	0.65

(Krahn and Gafni, 1993). We follow the conventional approach (Shepard and Thompson, 1979; Drummond et al., 1980; Drummond, 1980) and set discount rate $\delta = 0.99$. However, our results are not sensitive to the actual choice of the discount rate. We model supply as a three-point distribution with support over the set $\{0, 6, 12\}$. We consider symmetric probability distributions of the form $\Pr(z = 0) = \Pr(z = 12) = p$ and $\Pr(z = 6) = 1 - 2p$. The coefficient of variation for this supply distribution is given by $C.V. = \sqrt{2p}$. Using this form of the supply distribution allow us to change the variance of the distribution without changing the mean. Also since the coefficient of variation is independent of the mean, our results do not depend on the absolute value of the mean. We consider three versions of the safety-stock policy depending on the level of safety-stock a in (2.10): $a = 1$, $a = 2$ and $a = 3$.

2.8.2 Results

The performance of each heuristic is evaluated using the formulation in (2.7). This captures the increase in QALY score over the baseline of no treatment. Then the performance each heuristic relative to the optimal enrollment policy is calculated as: $\% \text{ gap} = \frac{V(\text{optimal}) - V(\text{heuristic})}{V(\text{optimal})}$. Figure 2.1 shows the $\%$ gap plotted as a function of the coefficient of variation of the supply distribution.

First, the behavior of the performance gap is different for values of s_2 greater than s_4 and less than s_4 . For higher values of s_2 , the gap decreases with uncertainty. The maximal enrollment policy gives the best performance, but the performance gap is overall higher. For lower values of s_2 , performance for all the heuristics increases as supply uncertainty increases. In other words, the value of using the optimal policy increases with supply uncertainty. Second, comparison among the heuristics reveals that maximal enrollment policy performs worse

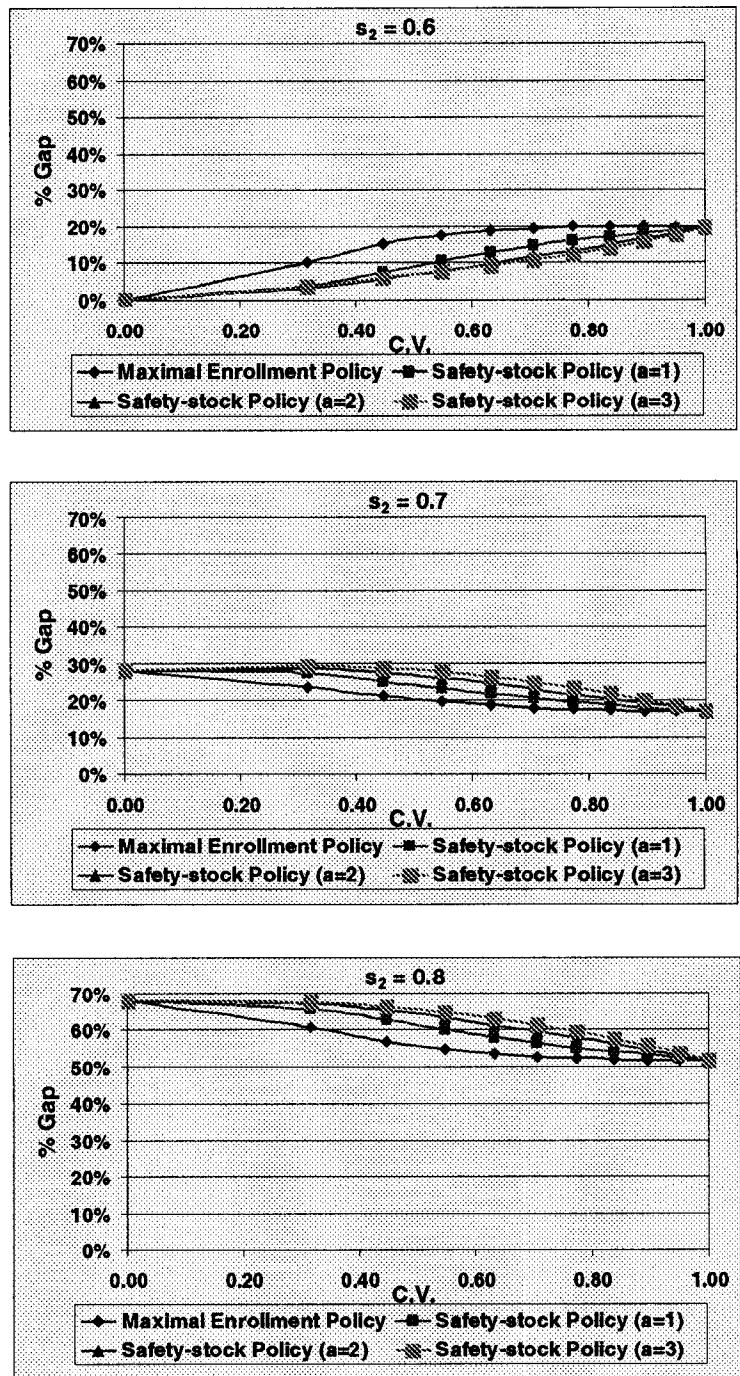


Figure 2.1: Performance of heuristics as a function of the coefficient of variation in the supply distribution

than the safety-stock policy for a large range of the supply uncertainty. Third, the choice of the actual safety-stock factor between $a = 1, 2, 3$ does not substantially impact the performance of the safety-stock policy.

Next we examine the best performance across all the heuristics as a function of the supply uncertainty. We find that heuristics perform best when s_2 is lower than s_4 and the supply uncertainty is low. In all other cases, even the best performance heuristics performs much worse compared to the optimal enrollment policy.

Next we examine the sensitivity of our results with respect to the parameter s_2 , which is the average QOL score of patients with interrupted treatment. A lower value of s_2 indicates that the problem of treatment interruption is more severe. Figure 2.2 plots the minimum % gap across all heuristics as a function of the coefficient of variation for different values of s_2 . We find that heuristics perform reasonably well as s_2 increases. However for lower values of s_2 even the choice of best among all the heuristics considered here yields large performance gap with respect to the optimal enrollment policy.

Table 2.2 shows the heuristic that gives the minimum % gap for the above values of s_2 and the coefficient of variation. We find that which heuristic performs the best depends on the value of s_2 as well as on the extent of supply uncertainty.

2.9 Conclusion and future research

In this essay, we study a clinic's problem of optimally allocating scarce and unreliable supply of antiretroviral drugs between new and current patients when continuity of treatment for previously treated patients is essential. We use dynamic programming to derive the optimal policy of the clinic with the objective of maximizing the total discounted quality adjusted life years of its patients. Our analysis

Table 2.2: Heuristic with the best performance for different values of C.V. and s_2 . ME denotes Maximal Enrollment Policy and SS (a) denotes Safety-stock Policy with a stock of “a” months

C.V.	$s_2 = 0.60$	$s_2 = 0.70$	$s_2 = 0.80$
0.00	All	All	All
0.32	SS (2)	ME	ME
0.45	SS (2)	ME	ME
0.55	SS (3)	ME	ME
0.63	SS (3)	ME	ME
0.71	SS (3)	ME	ME
0.77	SS (3)	ME	ME
0.84	SS (3)	ME	ME
0.89	SS (3)	ME	ME
0.95	SS (3)	All	ME
1.00	All	All	All

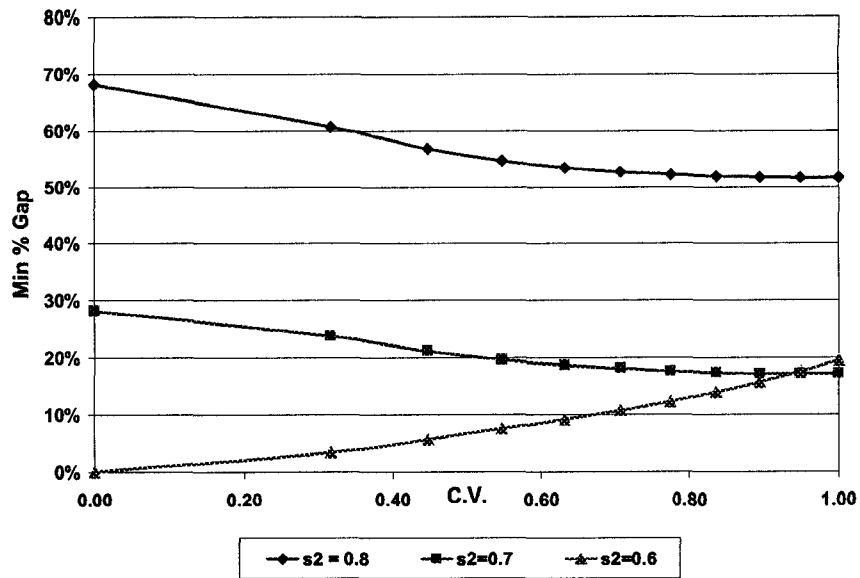


Figure 2.2: Minimum gap across all heuristics as a function of the coefficient of variation in the supply distribution

shows that the optimal policy under certain conditions results in prioritization of current patients, an accepted standard of care. But it also creates the possibility of restricting access to treatment for new patients. In our numerical illustrations the optimal enrollment policy (with enforced prioritization of current patients) performs substantially better than enrollment heuristics followed in practice for a wide range of parameter values. We find that supply uncertainty can greatly exacerbate the suboptimality gap of these heuristics. However, as mentioned earlier, our model is a simplified representation of the resource-constrained context which abstracts from various links between diagnosis, prevention and treatment. An explicit inclusion of these links would be required before the findings from this model could be used in practical settings.

Our work can be extended in several different directions. In the context of rationing of HAART, the demand model could be made more realistic, at

the expense of tractability, by considering attributes of patients other than the treatment status, such as their social status. Another extension would be the empirical determination of the actual rationing policies followed by clinics and the impact of supply uncertainty on these policies. We are currently in the process of preparing this, including attempting a limited empirical validation of the model proposed here.

2.10 Proofs

Resource-constrained condition: In our model described in (2.7), we assumed that $y_{t,n} > W_t \forall t$. Since $y_{t,n}$ and W_t are both random variables, this is true if certain restrictions are placed on the supply distributions $\Phi_t(\cdot)$. Here, we derive one such restriction in the form of an upper bound on the support of $\Phi_t(\cdot)$. Consider the finite horizon problem with initial conditions $y_{T,n}$ and I_T before the shipment in period T is received. Then $y_{T,n} > W_T$ if $z_T < y_{T,n} - I_T$. Suppose this is true. Then for period $T - 1$ under any feasible solution $x_{T,n}$ and $x_{T,c}$; $y_{T-1,n} = (y_{T,n} - x_{T,n})(\beta_2 + \alpha)$ and $W_{T-1} = W_T - x_{T,n} - x_{T,c} + z_{T-1}$. Now

$$\begin{aligned}
 y_{T-1,n} > W_T &\iff \\
 (y_{T,n} - x_{T,n})(\beta_2 + \alpha) > W_T - x_{T,n} - x_{T,c} + z_{T-1} &\iff \\
 z_{T-1} < y_{T,n}(\beta_2 + \alpha) - (\beta_2 + \alpha - 1)x_{T,n} + x_{T,c} - W_T &\quad (2.11)
 \end{aligned}$$

Since (2.11) has to be true for all feasible $x_{T,n}$, $x_{T,c}$ and $\beta_2 + \alpha - 1 > 0$ we substitute $x_{T,c} = 0$ and $x_{T,n} = W_T$ to obtain a lower bound on RHS. Thus (2.11) is satisfied for all feasible $x_{T,n}$, $x_{T,c}$ if $(\beta_2 + \alpha)z_T + z_{T-1} < (\beta_2 + \alpha)(y_{T,n} - I_T)$. Continuing this inductively, we find that a sufficient condition to ensure $y_{t,n} > W_t$

$\forall t$ is given by

$$\sum_{i=t}^T (\beta_2 + \alpha)^i z_i^U < (\beta_2 + \alpha)^T (y_{T,n} - I_T) \quad \forall t \quad (2.12)$$

where z_i^U is the upper bound on the support of z_i . A less tight bound is obtained by replacing each z_i^U by $\max_{t \leq i \leq T} z_i^U$ in (2.12) to obtain

$$\max_{t \leq i \leq T} z_i^U < \frac{(\beta_2 + \alpha)^T (y_{T,n} - I_T)}{\sum_{i=t}^T (\beta_2 + \alpha)^i} = \frac{(y_{T,n} - I_T) \left(1 - \frac{1}{(\beta_2 + \alpha)}\right)}{\left(1 - \frac{1}{(\beta_2 + \alpha)^{T-t+1}}\right)} \quad \forall t \quad (2.13)$$

Now since (2.13) has to be satisfied $\forall t$, we substitute $t = 1$ to obtain

$$\max_{1 \leq i \leq T} z_i^U < \frac{(y_{T,n} - I_T) \left(1 - \frac{1}{(\beta_2 + \alpha)}\right)}{\left(1 - \frac{1}{(\beta_2 + \alpha)^T}\right)} \quad (2.14)$$

Note that for $T \rightarrow \infty$, RHS of (2.14) $\rightarrow (y_{T,n} - I_T) \left(1 - \frac{1}{(\beta_2 + \alpha)}\right)$ and max operator in the LHS has to be replaced by sup. Thus the equivalent condition for infinite horizon problem is

$$\sup_{1 \leq i \leq T} z_i^U < (y_{T,n} - I_T) \left(1 - \frac{1}{(\beta_2 + \alpha)}\right) \quad (2.15)$$

While analyzing the infinite horizon problem in Section 11, we assume that (2.15) is satisfied.

Proof of Proposition 7:

(i) We use induction to prove this. Let

$$\mathcal{S}_t = \{(x_{t,n}, x_{t,c}) : x_{t,n} + x_{t,c} \leq W_t, 0 \leq x_{t,c} \leq y_{t,c}, 0 \leq x_{t,n}\}$$

Note that \mathcal{S}_t is a convex set. Using this notation, the recursive equation for $t = 1$ is given by:

$$\begin{aligned} V_1(W_1, y_{1,c}) &= \max_{(x_{1,n}, x_{1,c}) \in \mathcal{S}_1} s_c x_{1,c} + s_n x_{1,n} - k_c y_{1,c} \\ &= s_c \min\{y_{1,c}, W_1\} + s_n [W_1 - y_{1,c}]^+ - k_c y_{1,c} \end{aligned}$$

Thus $V_1(W_1, y_{1,c})$ is jointly concave in its arguments. Now assume that the result holds for $t - 1$. Let

$$\begin{aligned} V_t(y_{t,c}^i, W_t^i) &= f_t(x_{t,c}^i, x_{t,n}^i) \quad i = 1, 2 \\ \text{and } V_t(y_{t,c}^\lambda, W_t^\lambda) &= f_t(x_{t,c}^*, x_{t,n}^*) \end{aligned}$$

where $(y_{t,c}^\lambda, W_t^\lambda) \triangleq \lambda(y_{t,c}^1, W_t^1) + (1 - \lambda)(y_{t,c}^2, W_t^2)$. Also define $(x_{t,c}^\lambda, x_{t,n}^\lambda) \triangleq \lambda(x_{t,c}^1, x_{t,n}^1) + (1 - \lambda)(x_{t,c}^2, x_{t,n}^2)$. Since \mathcal{S}_t is a convex set $(x_{t,c}^\lambda, x_{t,n}^\lambda) \in \mathcal{S}_t$. Using this notation

$$\begin{aligned} &\lambda V_t(y_{t,c}^1, W_t^1) + (1 - \lambda) V_t(y_{t,c}^2, W_t^2) \\ &= \lambda h_t(x_{t,c}^1, x_{t,n}^1) + (1 - \lambda) h_t(x_{t,c}^2, x_{t,n}^2) \\ &\quad + \delta E [\lambda V_{t-1}(\beta(y_{t,c}^1 + x_{t,n}^1), W_t^1 - x_{t,c}^1 - x_{t,n}^1)] \\ &\quad + \delta E [(1 - \lambda) V_{t-1}(\beta(y_{t,c}^2 + x_{t,n}^2), W_t^2 - x_{t,c}^2 - x_{t,n}^2)] \\ &\leq h_t(x_{t,c}^\lambda, x_{t,n}^\lambda) + \delta E [V_{t-1}(\beta(y_{t,c}^\lambda + x_{t,n}^\lambda), W_t^\lambda - x_{t,c}^\lambda - x_{t,n}^\lambda)] \\ &= f_t(x_{t,c}^\lambda, x_{t,n}^\lambda) \leq f_t(x_{t,c}^*, x_{t,n}^*) = V_t(y_{t,c}^\lambda, W_t^\lambda) \end{aligned}$$

Thus $V_t(\cdot)$ is jointly concave in its arguments.

(ii) From (i), V_{t-1} is jointly concave in its arguments for all realizations of Z_{t-1} . Thus, V_{t-1} is also jointly concave in $(x_{t,c}, x_{t,n})$ since W_{t-1} and $y_{t-1,c}$ are linear transformations of $(x_{t,c}, x_{t,n})$ as seen from (2.3) for all realizations of Z_{t-1} . Since the expectation operator preserves concavity, $E[V_{t-1}]$ is also jointly concave in $(x_{t,c}, x_{t,n})$. $h_t(x_{t,c}, x_{t,n})$ is linear and hence jointly concave in its arguments. Since $f_t(x_{t,c}, x_{t,n})$ is a sum of two concave functions, it is also jointly concave in $(x_{t,c}, x_{t,n})$.

(iii) We show that our problem can be reformulated in the form described in Karmarkar (1981). Introduce slack variables $\rho_i^1, \rho_i^2, \rho_i^3$ in the constraints in (2.6),

and $u_{t,c} = y_{t,c} + \frac{\beta_2}{\beta_1} x_{t,n}$. Then define

$$\mathbf{y}_t = \begin{bmatrix} W_t \\ y_{t,c} \\ y_{t,n} \\ y_{t,c} \\ y_{t,n} \\ W_t \end{bmatrix}; \mathbf{x}_t = \begin{bmatrix} x_{t,c} \\ x_{t,n} \\ \rho_t^1 \\ \rho_t^2 \\ \rho_t^3 \end{bmatrix}; \mathbf{u}_t = \begin{bmatrix} I_t \\ u_{t,c}^1 \\ u_{t,c}^2 \\ 0 \\ 0 \\ 0 \end{bmatrix}; \mathbf{A} = \begin{bmatrix} -1 & -1 & 0 & 0 & 0 \\ 0 & \frac{\beta_2}{\beta_1} & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 \\ -1 & -1 & 0 & 0 & -1 \end{bmatrix}$$

and a transition function ω such that

$$\omega(\mathbf{u}_t, \tilde{z}_{t-1}) = \begin{bmatrix} I_t - \tilde{z}_{t-1} \\ \beta_1 u_{t,c}^1 \\ (\beta_2 + \alpha) u_{t,c}^2 \\ \beta_1 u_{t,c}^1 \\ (\beta_2 + \alpha) u_{t,c}^2 \\ I_t - \tilde{z}_{t-1} \end{bmatrix}$$

With these definitions, the problem formulation (2.7) becomes

$$V_t(W_t, y_{t,c}) = \max_{\mathbf{x}_t \geq 0} \left\{ \hat{h}_t(x_{t,c}, x_{t,n}) + \delta E[V_{t-1}(W_{t-1}, y_{t-1,c})] \right\}$$

s.t. $\mathbf{u}_t = \mathbf{A}\mathbf{x}_t + \mathbf{y}_t$
 $\mathbf{y}_{t-1} = \omega(\mathbf{u}_t, \tilde{z}_{t-1})$

This formulation is equivalent to the formulation (MP) in Karmarkar (1981) and hence Proposition 8 applies completing the result.

Proof of Proposition 8: First consider the case when $s_3 - s_4 > s_1 - s_2$. For $t = 1$, the optimal solution is trivial. Thus

$$E_{z_1}[V_1(W_1, y_1)] = s_2 y_{1,c} + (s_3 - s_4)(I_1 + E[z_1]) \quad (2.16)$$

Then the objective function for $t = 2$ becomes:

$$f_2(x_{2,c}, x_{2,n}) = ((s_1 - s_2) - \delta(s_3 - s_4))x_{2,c} + (s_3 - s_4^2) + \delta(s_2\beta_2 - (s_3 - s_4)) + (s_3 - s_4)E[z_1] \quad (2.17)$$

Since (2.17) is linear in its arguments, comparing the coefficients gives the desired result.

Now consider the case of $s_1 - s_2 > s_3 - s_4$. Again the optimal policy for $t = 1$ is straightforward. Thus evaluating $E[V_1(I_1, y_{1,c})]$ and adding the single period reward function, the objective function for $t = 2$ becomes

$$f_2(x_{2,c}, x_{2,n}) = (s_1 - s_2)x_{2,c} + (s_3 - s_4^2)x_{2,n} + \delta \left((s_1 - (s_3 - s_4))y_{1,c} + (s_3 - s_4)I_1 - ((s_1 - s_2) - (s_3 - s_4)) \int_{z_1^L}^{y_1 - I_1} \Phi_1(\tilde{z}_1) d\tilde{z}_1 \right)$$

Since this is a constrained optimization problem with concave objective function and linear constraints, KKT conditions are both necessary and sufficient for optimality. Let λ_1 and λ_2 be the lagrangean multipliers associated with the constraints in (2.7). Then forming the lagrangean in the usual manner and taking the first order derivatives

$$\frac{\partial L_2}{\partial x_{2,c}} = (s_1 - s_2) - \delta(s_3 - s_4) - \lambda_1 - \lambda_2 = 0 \quad (2.18)$$

$$\begin{aligned} \frac{\partial L_2}{\partial x_{1,n}} &= (s_3 - s_4^1) + \delta(-(s_3 - s_4) - ((s_1 - s_2) - (s_3 - s_4))\Phi_1(y_1 - I_1)) \\ &+ \beta_2((s_1 - (s_3 - s_4)) - ((s_1 - s_2) - (s_3 - s_4)))\Phi_1(y_1 - I_1) - \lambda_2 = 0 \end{aligned} \quad (2.19)$$

Then depending on which of the constraints are binding or slack we get following cases:

Case I: $\lambda_2 > 0; \lambda_1 = 0$

$x_{2,c}^* + x_{2,n}^* = W_2$ and $I_1 = 0$. Substituting $\lambda_2 > 0; \lambda_1 = 0$ and equating (2.18) and (2.19) we obtain $\Phi(y_1^*) = \frac{(s_2 - s_4(1 - \delta\alpha)) - (1 - \delta\beta_2)(s_1 - s_3)}{\delta\beta_2((s_1 - s_2) - (s_3 - s_4))} = k_1$. Thus $x_{2,n}^* = \frac{\Phi_1^{-1}(k_1) - \beta_1 y_{2,c}}{\beta_2}$ and $x_{2,c}^* = W_2 - \frac{\Phi_1^{-1}(k_1) - \beta_1 y_{2,c}}{\beta_2}$. For this to be feasible, we need the following conditions to be satisfied: (i) $\Phi_1^{-1}(k_1) - \beta_1 y_{2,c} \geq 0$ (ii) $\beta_2 W_2 + \beta_1 y_{2,c} \geq \Phi_1^{-1}(k_1)$ and (iii) $\beta_2 W_2 + (\beta_1 - \beta_2) y_{2,c} \leq \Phi_1^{-1}(k_1)$.

Case II: $\lambda_1 > 0; \lambda_2 = 0$

$x_{2,c}^* = y_{2,c}$. Substituting in (2.19) we obtain

$$\Phi(y_1^* - I_1^*) = \frac{s_3 - s_4(1 + \delta(\beta_2 + \alpha)) + \delta(s_1 - (s_3 - s_4)(1 + \beta_2))}{((s_1 - s_2) - (s_3 - s_4))(1 + \delta\beta_2)} = k_2$$

Simplifying and substituting the value of $x_{2,c}^*$, we get $x_{2,n}^* = \frac{\Phi_1^{-1}(k_2) + W_2 - (1 + \beta_1)y_{2,c}}{(1 + \beta_2)}$.

Again for this to be feasible, we need the following conditions to be satisfied: (i) $(1 + \beta_2)y_{2,c} - W_2 \leq \Phi_1^{-1}(k_2)$ (ii) $\beta_2 W_2 + (\beta_1 - \beta_2)y_{2,c} \geq \Phi_1^{-1}(k_2)$.

Case III: $\lambda_1 > 0; \lambda_2 > 0$

$x_{2,c}^* = y_{2,c}$ and $x_{2,n}^* = W_2 - y_{2,c}$ and. Substituting $\lambda_1 > 0; \lambda_2 > 0$ in (2.18) and (2.19) we get $\beta_2 W_2 + (\beta_1 - \beta_2)y_{2,c} \geq \Phi_1^{-1}(k_1)$. This is feasible if (i) $W_2 \geq y_{2,c}$.

Case IV: $\lambda_1 > 0; \lambda_2 > 0$

This is not possible since $(s_1 - s_2) - \delta(s_3 - s_4) > 0$.

Thus combining all the three cases, we get the desired form of the optimal policy, where η and θ are defined appropriately.

Proof of Proposition 9: We shall use induction to prove this result. First consider $t = 1$. Clearly since (C1) states that $s_1 - s_2 > s_3 - s_4$, the optimal policy has the desired form since $\frac{\partial f_1}{\partial x_{1,c}} > \frac{\partial f_1}{\partial x_{1,n}}$ and $\frac{\partial f_1}{\partial x_{1,c}} > 0$. Now consider period t . The partial derivatives of the maximand are given by:

$$\frac{\partial f_t}{\partial x_{t,n}} = (s_3 - s_4^t) + \delta E \left[\beta_2 \frac{\partial V_{t-1}}{\partial y_{t-1,c}} - \frac{\partial V_{t-1}}{\partial W_{t-1}} \right] \quad (2.20)$$

$$\frac{\partial f_t}{\partial x_{t,c}} = (s_1 - s_2) + \delta E \left[-\frac{\partial V_{t-1}}{\partial W_{t-1}} \right] \quad (2.21)$$

Now suppose that the following two results hold for $t - 1$:

$$E \left[\frac{\partial V_{t-1}}{\partial W_{t-1}} \right] < (s_1 - s_2) \quad (2.22)$$

$$E \left[\frac{\partial V_{t-1}}{\partial y_{t-1,c}} \right] < \frac{(s_1 - s_2) - (s_3 - s_4^t)}{\delta \beta_2} \quad (2.23)$$

These together imply that $\frac{\partial f_t}{\partial x_{t,c}} > 0$ and $\frac{\partial f_t}{\partial x_{t,c}} > \frac{\partial f_t}{\partial x_{t,n}}$ and hence the optimal policy has desired form in period t . In order to carry forward the induction we need to show that (2.22) and (2.23) hold for t . Using the structure of the optimal policy in period t , we can write the following:

$$V_t(W_t, y_{t,c}) = \begin{cases} (s_1 - s_2) W_t + s_2 y_{t,c} + \delta E [V_{t-1}(z_{t-1}, \beta_1 y_{t,c})]; \\ \quad \text{if } z_{t-1} < y_{t,c} - I_t \\ \\ s_1 y_{t,c} + (s_3 - s_4^t) (W_t - y_{t,c}) \\ + \delta E [V_{t-1}(z_{t-1}, \beta_2 W_t + (\beta_1 - \beta_2) y_{t,c})]; \\ \quad \text{if } y_{t,c} - I_t < z_{t-1} < y_{t,c} - I_t + \theta_t \\ \\ s_1 y_{t,c} + (s_3 - s_4^t) \theta_t \\ + \delta E [V_{t-1}(W_t - y_{t,c} - \theta_t + z_{t-1}, \beta_1 y_{t,c} + \beta_2 \theta_t)]; \\ \quad \text{if } z_{t-1} > y_{t,c} - I_t + \theta_t \end{cases} \quad (2.24)$$

Using the fact that $V_t(\cdot)$ is jointly concave in its arguments we can see that $E \left[\frac{\partial V_t}{\partial W_t} \right] < (s_1 - s_2)$ and $E \left[\frac{\partial V_t}{\partial y_{t,c}} \right] < s_1 - (s_3 - s_4^t) + \delta (\beta_1 - \beta_2) E \left[\frac{\partial V_{t-1}}{\partial y_{t-1,c}} \right]$. Clearly (2.22) holds for t . Now to prove the remaining, consider

$$\begin{aligned} E \left[\frac{\partial V_t}{\partial y_{t,c}} \right] &< s_1 - (s_3 - s_4^t) + \delta (\beta_1 - \beta_2) E \left[\frac{\partial V_{t-1}}{\partial y_{t-1,c}} \right] \\ &< s_1 - (s_3 - s_4^t) + \delta (\beta_1 - \beta_2) \frac{(s_1 - s_2) - (s_3 - s_4^t)}{\delta \beta_2} \\ &= \frac{(s_1 - (s_3 - s_4^t)) \beta_1 - (\beta_1 - \beta_2) s_2}{\beta_2} \end{aligned} \quad (2.25)$$

where we have used (2.23). Now rewriting (C2) we obtain

$$\begin{aligned}
s_2 (1 - \delta (\beta_1 - \beta_2)) &< (s_1 - s_3) (1 - \delta \beta_1) + s_4 (1 + \delta (\beta_2 + \alpha - \beta_1)) \\
&= (s_1 - s_3) (1 - \delta \beta_1) + s_4^2 - s_4^1 (\delta \beta_1) \\
&< (s_1 - s_3) (1 - \delta \beta_1) + s_4^{t+1} - s_4^t (\delta \beta_1)
\end{aligned} \tag{2.26}$$

where we have used the definition of s_4^t and that it is increasing in t if

$(\beta_2 + \alpha - \beta_1) > 0$. Now substituting (2.26) in (2.25) we obtain the desired result.

Proof of Proposition 10: θ_t is a solution to the equation $\frac{\partial f_t}{\partial x_{t,n}} = (s_3 - s_4^t) + \delta E \left[\beta \frac{\partial V_{t-1}}{\partial y_{t-1,c}} - \frac{\partial V_{t-1}}{\partial W_{t-1}} \right] = 0$, where the expectation is with respect to the distribution of Z_{t-1} i.e., F_{t-1} or G_{t-1} . First, using the implicit function theorem it is clear that $\frac{\partial \theta_t}{\partial y_{t,c}} = -\frac{\frac{\partial}{\partial y_{t,c}} E \left[\beta \frac{\partial V_{t-1}}{\partial y_{t-1,c}} - \frac{\partial V_{t-1}}{\partial W_{t-1}} \right]}{\frac{\partial}{\partial \theta_t} E \left[\beta \frac{\partial V_{t-1}}{\partial y_{t-1,c}} - \frac{\partial V_{t-1}}{\partial W_{t-1}} \right]} = -1$ since $\frac{\partial W_{t-1}}{\partial y_{t,c}} = \frac{\partial W_{t-1}}{\partial \theta_t}$ and $\frac{\partial y_{t-1,c}}{\partial y_{t,c}} = \frac{\partial y_{t-1,c}}{\partial \theta_t}$.

This proves the part (i). Now for part (ii) $\frac{\partial V_{t-1}}{\partial W_{t-1}}$ is a non-increasing function of W_{t-1} and hence of Z_{t-1} . Hence, $-\frac{\partial V_{t-1}}{\partial W_{t-1}}$ is a non-decreasing function of Z_{t-1} .

Also using $\beta_1 = \beta_2$ from (2.24), we have

$$\frac{\partial V_{t-1}}{\partial y_{t-1,c}} = \begin{cases} s_2 + \delta \beta_1 E \left[\frac{\partial V_{t-2}}{\partial y_{t-2,c}} \right]; & z_{t-1} < y_{t-1,c} - I_{t-1} \\ s_1 - (s_3 - s_4^t); & z_{t-1} \geq y_{t-1,c} - I_{t-1} \end{cases}$$

Now $\frac{\partial V_{t-1}}{\partial y_{t-1,c}}$ is decreasing in $y_{t-1,c}$ and hence increasing in z_{t-1} for a given $y_{t-1,c}$. Hence $E_F \left[-\frac{\partial V_{t-1}}{\partial W_{t-1}} \right] \geq E_G \left[-\frac{\partial V_{t-1}}{\partial W_{t-1}} \right]$ by first order stochastic dominance. This implies that $\left. \frac{\partial f_t}{\partial x_{t,n}} \right|_F > \left. \frac{\partial f_t}{\partial x_{t,n}} \right|_G$ and hence $\theta_t(F) > \theta_t(G)$. This proves the result.

Proof of Proposition 11: Define a borel measurable function $r : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ that chooses $(x_{t,c}, x_{t,n}) \in \mathcal{S}_t$ given $(W_t, y_{t,c})$. Let \mathcal{V} denote the space of continuous real-valued functions satisfying the following boundedness property:

$$\sup_{W,y} \left\{ \frac{\|V(W,y)\|}{\max\{W,D\}} \right\} < \infty \tag{2.27}$$

where D is a positive constant and define the distance function

$$q(V, V') = \sup_{W,y} \left\{ \frac{\|V(W,y) - V'(W,y)\|}{\max\{W,D\}} \right\}; \quad V, V' \in \mathcal{V}$$

Then (\mathcal{V}, q) is a Banach space (Miller, 1974). Define $g(\cdot)$ as a vector valued function defining the transitions in (2.1) and (2.3). Since, r is a measurable function, $g(\cdot)$ is a transition function (Theorem 8.9; Stokey et al., 1989). Moreover, since $g(\cdot)$ is continuous, it possesses the Feller Property (Stokey et al., 1989; p.237). Hence the operator

$$(H_r V)(W, y) = h(r(W, y)) + \delta E[V(g(W, y))]$$

maps into a continuous function. To show that $H_r V$ satisfies the boundedness property in (2.27) note that

$$\begin{aligned} (H_r V)(W, y) &= h(r(W, y)) + \delta E[V(g(W, y))] \\ &\leq s_c W + \delta E[V(W - x_c - x_n + z, \beta(y + x_n))] \\ &\leq s_c W + \delta V(W + E[z], \beta(y + x_n)) \\ &\leq s_c W + \delta M \max\{W + E[z], D\} \end{aligned} \quad (2.28)$$

where the last inequality is due to the boundedness condition (2.27) and the inequality before that is because $V(W, y)$ is concave and increasing in W . Clearly, (2.28) implies that $(H_r V)(W, y)$ is bounded since $E[z]$ is bounded. Thus $H_r V : \mathcal{V} \rightarrow \mathcal{V}$. Hence Theorem 9.6 and 9.2 from Stokey et al. (1989) prove the result.

Proof of Proposition 12: Using results (i) and (ii) of Proposition 9 in (2.20), we obtain:

$$\begin{aligned} \frac{\partial f_t}{\partial x_{t,n}} &\geq (s_3 - s_4^t) + \delta \left(s_2 \beta_2 \sum_{u=0}^{t-2} (\delta \beta_1)^u - (s_1 - s_2) \right) \\ &\geq (s_3 - s_4^t) + s_2 \sum_{u=1}^{t-1} (\delta \beta_2)^u - \delta (s_1 - s_2) \\ &= (s_2 - s_4) \sum_{u=0}^{t-1} (\delta \beta_2)^u - (s_2 - s_3) - \delta (s_1 - s_2) \end{aligned}$$

Now two cases are possible depending on the relative values of s_2 and s_4 . If $s_2 > s_4$, then $(s_2 - s_4) \sum_{u=0}^{t-1} (\delta \beta_2)^u \geq s_2 - s_4$ and hence condition (C3) in the

hypothesis implies $\frac{\partial f_t}{\partial x_{t,n}} \geq 0 \quad \forall t \leq T$. If $s_2 < s_4$, then $(s_2 - s_4) \sum_{u=0}^{t-1} (\delta\beta_2)^u \geq (s_2 - s_4) \sum_{u=0}^{T-1} (\delta\beta_2)^u$ and hence condition (C4) in the hypothesis implies $\frac{\partial f_t}{\partial x_{t,n}} \geq 0 \quad \forall t \leq T$.

2.11 Appendix: A conceptual model of HIV treatment scale-up in resource-constrained setting

Scaling up highly active antiretroviral therapy (HAART) for HIV+ patients in resource-constrained settings such as Sub-Saharan Africa has received enormous attention in the recent years. The number of patients receiving HAART in these regions has barely reached 1 million patients despite growing attention from the international donor community (PEPFAR , GFATM , CHAI , etc.) and WHO's ambitious 3 by 5" program which aimed to get 3 million people on treatment by 2005 (WHO, 2005). Not discouraged by this failure, the WHO is now aiming at universal coverage for all eligible patients by 2010 (WHO, 2007).

IOM (2005) provides a very detailed discussion of the ethical, clinical and social principles underlying scale-up efforts and elements of an integrated management framework necessary for long-term success of treatment programs. The report emphasizes the importance of (i) sustainability of response, (ii) integration of treatment and prevention, (iii) comprehensive supporting infrastructure, and (iv) continuous monitoring and evaluation.

While there is a growing body of research on HAART scale-up in resource-constrained settings, the level of integration between different streams is low. The objective of this essay is to provide a unifying framework for HIV / AIDS management in resource-constrained settings. Using a patient flow model as

the underlying framework we describe various potential links between treatment, prevention and diagnosis. It also provides a platform to discuss issues related to allocation of resources between these three activities. We also highlight the missing links in the current literature and suggest opportunities for future research. Finally, we briefly elaborate on one such potential research idea.

2.11.1 Conceptual framework

Figure 2.3 describes our conceptual model. It is a patient flow model, where rectangular boxes represent different patient segments, stars represent different resources, simple arrows represent patient flows and block arrows represent usage of resources in the program. The fundamental premise of the model is that HAART can not be analyzed in isolation; it is essential to highlight its interactions with other portions of the healthcare system (testing and prevention) through patient flows and shared resources. However, the model is not intended to display the minutest details. For example each patient segment is further divided into sub-categories based on their health status and socio-economic status. Further use of this model to answer specific research questions mentioned later in this chapter would require taking such sub-categories into account.

2.11.2 Existing Literature

In this section, we use our conceptual model to review the literature relevant to HAART scale up. We focus on links in the model that are particularly important to resource-constrained settings. We discuss evidence (or lack of it) for these links and wherever appropriate compare the situation in resource-constrained settings with that in developed countries.

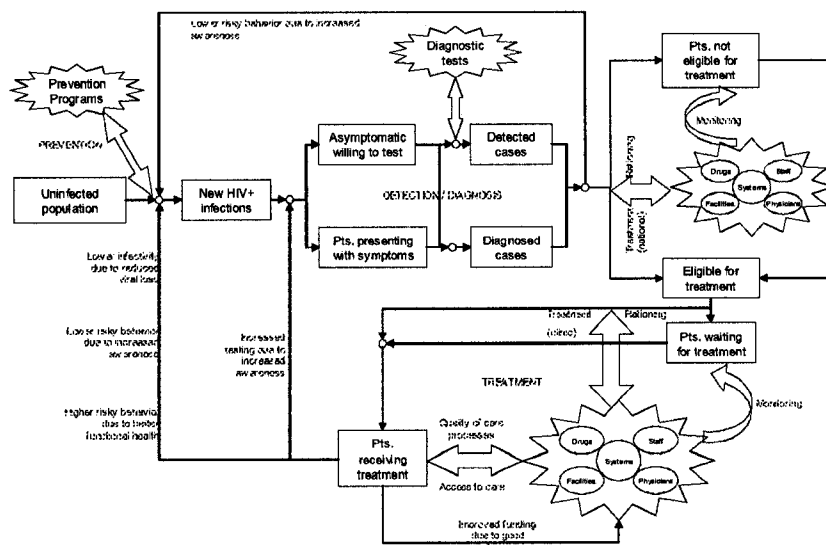


Figure 2.3: Conceptual model of HAART scale-up

Treatment Rationing

A particularly important issue related to initiation of therapy in resource-constrained setting is that of treatment rationing: given the anticipated gap between available supply of drugs and demand (WHO, 2005), the key question is which patients should be prioritized for treatment. Many countries have adopted national level treatment rationing policies based on non-clinical criteria (Bennet and Chanfreau, 2005). Typical factors used for rationing include adherence to treatment, social and economic benefits and financial factors in addition to clinical characteristics such as CD4 counts and disease staging. A detailed discussion of ethical principles involved in creating rationing guidelines is available in Macklin (2004). Based on these ethical principles, Rosen et al. (2005) and McGough et al. (2005) provide various qualitative mechanisms for selecting new patients from different patient segments. However, this discussion fails to recognize that HAART is a life-long treatment and selecting patients in the present has resource implications

for the future, especially if resource availability in the future is unknown. We elaborate on this further in a later section.

Impact of treatment on VCT uptake

Impact of improved treatment access on increased testing is especially significant in resource-constrained settings. Success stories of drastic improvements in health status of treated patients encourage more people to come forward for testing. While the importance of this link is widely recognized, the evidence is still scant. Experience with community roll-out of treatment in Haiti showed that uptake of voluntary counseling and testing (VCT) increased by 300% (WHO, 2003a). Similarly, in Khayelitsha, South Africa, the uptake of VCT increased 12 fold after treatment introduction (WHO, 2003b). Conceivably, expansion of VCT programs provides healthcare workers with new gateway for spreading prevention messages promoting less risky behaviors and reducing incidence of new HIV cases. However, there are very few studies that demonstrate this link between VCT uptake and improved prevention (Roth et al., 2001).

Impact of treatment on risky behavior

Similarly impact of HAART on the behavior of patients undergoing treatment is not unambiguous. Many studies in developed countries have found that risky behavior in treated patients increased as a result of improved functional health and the belief that HIV is not a life threatening infection (Katz et al., 2002; Dilley et al., 1998). However, these studies were focused on gay communities. Similar studies of heterosexual communities have shown mixed findings (Wilson and Minkoff, 2001; Flaks et al, 2003; van der Straten, 2000). In resource-constrained settings, recent evidence seems to point that HAART programs result in more

preventive behavior through counseling and raising awareness among discordant couples (Bunnell et al., 2006). However, other studies have only found no increase in risky behavior as a result of treatment (Moatti et al., 2003). Clearly, more studies are required before generalizing these findings and basing policy recommendations on them.

Impact of treatment on spread of HIV

A direct benefit of successful treatment regimen is the reduced viral load in the patients (Pereira et al., 1999) which can reduce infectivity and hence the spread of HIV, even if their risky behavior was unchanged (Hart et al., 1999; Musicco et al., 1994). However studies based in sub-Saharan Africa have been less optimistic. Using a stochastic simulation model and HIV transmission data from Uganda, Gray et al. (2003) found that HAART alone would not substantially reduce the prevalence of HIV infection in the population over the next 20 years. Auvert et al. (2004) reach to a similar conclusion regarding the impact of HAART on incidence of new HIV infections in South Africa.

Interdependence of treatment and prevention

The results from these studies have been incorporated to understand the epidemiological impact of HAART programs. Blower et al. (2005) and Blower et al. (2003) use mathematical models to predict the impact of HAART rollout on the evolution of drug-resistant HIV epidemics in resource-constrained settings and on number of new infections averted. They model three main effects of HAART on the epidemiology of the disease: (i) reduced infectivity due to reduced viral load of treated patients (ii) changed behavior as a result of improved health status (increase / decrease in risky behavior) (iii) acquired drug resistance due to long-

term exposure to treatment and transmitted drug resistance due to risky behavior. Salomon et al. (2005) use an epidemiological model to show that integration of treatment and prevention responses is highly beneficial in resource-constrained settings as treatment makes prevention efforts more effective and prevention efforts make treatment more affordable. However, as discussed later these models do not adequately incorporate the characteristics of resource constraints in these settings.

Characterizing resource-constrained settings

A growing stream of operational research literature focuses on detailed characterization of the organizational challenges associated with rapid HAART scale-up. Landman et al. (2006) surveyed and evaluated the capacity of 19 health care facilities in Tanzania to deliver HAART and found need for substantial improvement in several areas including staff training in HAART and laboratory facilities. Wester et al. (2005) found the lack of space and well-trained staff as the key bottlenecks during the expansion of HAART at a single clinic in Botswana. During a similar study of community based HAART program in South Africa, Bekker et al. (2003) found significantly increased staffing needs during the recruitment of new patients on HAART.

Another stream of this literature studies the impact of HAART programs on patient outcomes. Coetzee et al. (2004) found that standard approaches to patient preparation and adherence counseling resulted in high patient retention, improved viral load control and patient survival. In contrast, in a Malawian study (Oosterhout et al., 2005) several patients were lost to follow-up and unreliable drug supply and financial constraints were found to be the main causes of poor adherence. Oyugi et al. (2007) found similar causes for treatment interruptions

in a study in Uganda and also found that such interruptions were significantly related to development of drug resistance.

Thus the main contribution of this literature has been to establish the feasibility of HAART in resource-constrained settings and highlighting operational issues that can impact patient adherence and treatment outcomes. However, two important aspects have not received adequate attention. First, an important unanswered question in this context is which model of care is the most appropriate. The only study that compared different models of HAART provision at five leading centers in the Western Cape province of South Africa did not find significant difference in patient outcomes (Pienaar et al., 2006). Second, due to funding by external agencies, there is a strong link between future availability of resources and current performance of HAART programs. Consequently, while making resource allocation decisions, the healthcare facilities need to consider the current direct impact of their HAART programs as well as program sustainability through future availability of resources. However, currently there are no models that can support the implementing agencies in making such decisions.

2.11.3 Agenda for future research

Previous section points to important gaps in current literature on HAART scale up in resource-constrained settings. In this section, we present three important building blocks of HAART scale up in resource-constrained settings and argue that the main shortcoming of the current literature is that it does not combine these building blocks; an attribute that is essential to designing more relevant research questions.

1. Dynamics of HAART: HAART is life-long treatment and enrolled patients need to be treated without interruption for the rest of their life. This implies

that current enrollment decisions have implications on the sustainability of treatment in the future.

2. Interdependence of prevention, testing and treatment: Discussion in the previous section has highlighted the importance of these interdependencies. Any program that focuses on one aspect and fails to recognize the connection with other aspects will inevitably only achieve local optimum.
3. Characterization of resource availability: Operational research presents a complex picture of the resource availability in this context. In addition to aggregate shortage, availability of resources is also highly variable owing to multilateral coordination involved in garnering resources and a weak infrastructure involved in delivering resources. Moreover, future availability of resources might often depend on the current performance of programs. Next, we present an illustrative example of potential research projects which explicitly incorporate these building blocks.

Dynamic rationing of treatment at the clinic level

A majority of the current discussion on treatment rationing reviewed above is qualitative and focuses on which patient segments to select for treatment based on a combination of socio-economic and clinical criteria. The two main deficiencies of this approach are: (i) it is implicitly assumed that patients, once enrolled, are guaranteed to receive treatment in the future (ii) it primarily focuses on national level policies with the hope that these policies will eventually percolate down to the clinic level.

Studies of HAART programs in resource-constrained settings provide evidence that the first assumption is not true. Oyugi et al. (2007) and Oosterhout et al.

(2005) provide systematic evidence of treatment interruptions caused by drug stock outs resulting in drug resistance and treatment failure in Uganda and Malawi. The primary reason for these stock outs is a combination of supply and demand side issues. On the supply side, there is variability in the quantity of drugs delivered to clinics due to variability in the funding at the national level and deficiencies in the infrastructure available to deliver these drugs. On the demand side, there is inadequate understanding of how the future demand for drugs is shaped by the current enrollment decisions due to the chronic nature of HIV. As a result, the national level policies fail to provide concrete guidelines to clinics on how to plan for new enrollments in order to provide them and previous enrollees a sustainable treatment in the future.

A quantitative approach that incorporates all the above characteristics of the resource-constrained settings is required to provide this decision support to the clinics. Figure 2.4 provides a schematic representation of the proposed approach. Available resources need to be allocated between new (treatment naïve) patients and current (previously enrolled) patients. Note that both the patient pools are not homogenous but consist of patients with varying health status as characterized by CD4 count or disease stage. Initiating new patients on treatment has two consequences: (i) immediate improvement in the health status of these patients and (ii) increased likelihood of future deterioration in health status of patients due to shortage of drugs owing to supply variability. Thus deciding the number of new enrollments in the current period requires quantitatively trading off this immediate improvement with a possible future deterioration.

Calculating the second component – future deterioration in health status due to treatment interruption – is tricky and requires further exposition. The deterioration in health status could be due to direct causes such as development of drug

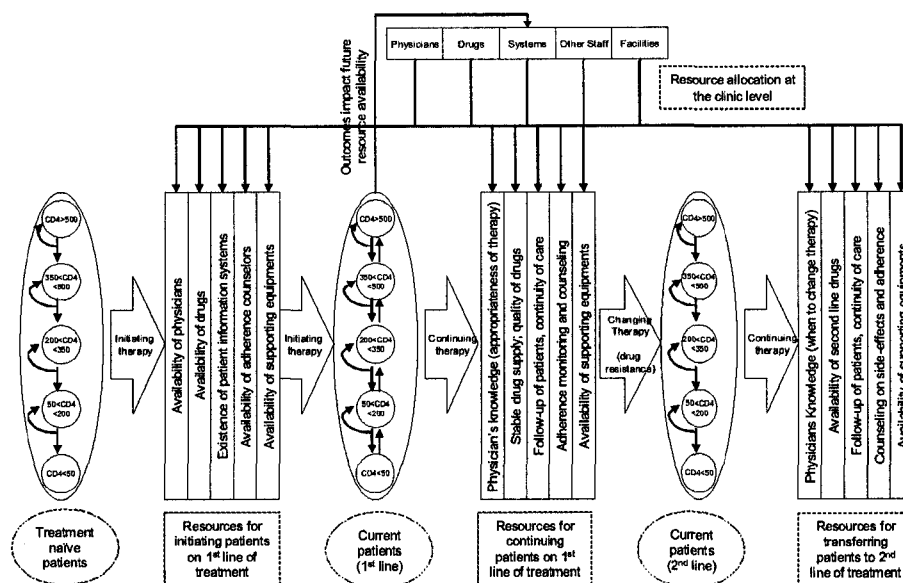


Figure 2.4: Treatment rationing at the clinic level

resistance and viral rebound resulting in therapy failure. A detailed model needs to capture that the likelihood of these consequences could vary depending on the current health status of patients, history of adverse events and the duration of the current treatment interruption. Moreover, the model also needs to capture the indirect causes for health deterioration such as reduced adherence due to a loss of trust in treatment and health care system in general.

In addition to these consequences on the health of patients under treatment, interruptions can have negative impact on other aspects of the system described in Figure 1. Interruptions could make untreated patients less optimistic about treatment programs resulting in lower testing rates, which in turn could negatively impact the effectiveness of prevention programs.

Another important aspect evident from the schematic representation is the fact that current enrollment decisions also influence future availability of resources through program outcomes. In some cases, aggressively increasing the number of

new enrollments might have the negative impact of increasing the likelihood of future treatment interruptions. However, at the same time, it might also result in more resources being made available to the clinic for future periods.

Apart from this normative model, there is an urgent need to understand how clinics actually ration treatment among different patient segments and between current and new patients. Empirical research in other contexts has shown that clinicians are guided more by equity than by efficiency in their rationing decisions (Ubel et al., 2000; Perneger et al., 2002; Hurst et al., 2005). It would be interesting to test the hypotheses coming out of this literature in the context of resource-constrained settings and the chronic nature of HAART.

Using this illustrative example, we find that there are ample opportunities for cross-functional research which can considerably improve our understanding of the situation and help us to design better informed policy recommendations.

CHAPTER 3

Organizational determinants of performance in quality improvement collaboratives

3.1 Introduction

Over the past decade, there has been a growing consensus that serious quality problems exist in the delivery of health care in the U.S. due to an increase in the prevalence of chronic conditions among patients and a poorly organized delivery system that is designed to deliver episodic care. (Chassin, 1996; Chassin and Galvin, 1998; IOM, 2001). This necessitates coordination among different areas of the health care delivery system in order to ensure high quality of care over a long period. Many have advocated a systems approach to quality improvement (Blumenthal, 1996; Chassin and Galvin, 1998; Shortell et al., 1998; Ferlie and Shortell, 2001; IOM, 2001; Berwick, 2002; Leape and Berwick, 2005).

Quality Improvement Collaboratives (QIC), an approach based on systems thinking, is arguably one of the most powerful tools available to the health care industry (Mittman, 2004). It has been adopted on a large scale by Health Resources and Systems Administration (HRSA) in the U.S. and National Health Service (NHS) in the U.K. The QIC method involves bringing together a number of health care delivery organizations who are committed to improving a certain aspect of health care. Representative teams from these organizations learn from

experts about the practical approaches to quality improvement and share their own findings and experiences with each other to create a collaborative learning experience. This approach in conjunction with the Chronic Care Model (CCM) has been widely employed to improve the quality of care for chronic conditions (Wagner et al., 2001).

However, there is limited rigorous evidence on the improvements achieved as a result of this approach (Mittman, 2004). One of the most rigorous evaluations of QIC in HIV care (Landon et al., 2004) did not find statistically significant improvement in the quality of processes and outcomes in the clinics that participated in the collaborative when compared with a set of matching clinics. These researchers have suggested that organizational factors might play an important role in predicting successful quality improvement. Other researchers have also commented on the variation in the performance of the organizations that take part in collaboratives (Ovretveit et al., 2002) which may be a result of the organizational culture and readiness to inculcate quality improvement philosophy (Plsek, 1997). However, there has been relatively little attention in the published literature to structural determinants of health care quality (Flood, 1994; Dudley et al., 2000; West, 2001).

In this chapter, we assess several organizational factors that we thought would be related to the performance of the health care delivery organizations participating in QICs. Specifically, we hypothesized that a supportive and open environment, organization's focus on quality improvement, presence of multidisciplinary teams and measurement of progress towards quantifiable goals would be associated with successful quality improvement efforts.

Importance of multidisciplinary teams as effective means of improving current practices in an organization (Argote et al., 2001) and their usage in health care

organizations (Horbar et al., 2001; Wagner et al., 1996) is well known. Thus we expect that the clinics that have multidisciplinary teams will perform better than those that do not have such teams. Similarly employees' willingness to engage in trial and error experimentation and collaborative problem-solving depends on a supportive organizational and interpersonal climate (Edmondson et al., 2001; Sarin and McDermott, 2003; Edmondson, 1999). Various aspects of this climate include organization's focus on quality improvement and psychological safety or a culture of openness in the organization. Hence we hypothesize that an open organizational culture, strong QI focus and presence of multidisciplinary teams will be associated with higher number of interventions. In the context of quality improvement collaboratives, reporting progress on targets is shown to keep teams focused on the collaborative objective and on the need for measurement, and helps them to learn the importance of objective assessment (Ovretveit et al., 2001). Hence we hypothesize that organizations that routinely measure their progress towards quantifiable goals will implement a higher number of interventions. (Hypothesis 1)

Multifaceted, interconnected changes to the organization and functioning of practice are essential to ensure routine performance of critical tasks in a chronic care setting. A study of 41 collaboratives in diabetes care found that interventions that tended to be complex and involved multiple areas of care (Rothman and Wagner, 2003). Since in chronic care, patients have repeated contacts with the health care system, possibly with different aspects of the system each time, it is very important to have excellent internal coordination between different areas of care giving within the organization. Clearly, presence of multidisciplinary teams and an open organizational culture can facilitate dialogue between different departments of the organization and pave the way for interventions which are more complex and span multiple areas of care. Hence we hypothesize that

organizations with more open culture and multidisciplinary teams will undertake interventions of more cross-departmental nature. (Hypothesis 2)

Choosing interventions that can have large potential impact on the quality of care and choosing a large number of interventions is clearly not sufficient to guarantee overall implementation success. The chosen interventions need to be implemented in the right way across the organization to actually see the impact. Moreover, critical evaluation of interventions is necessary in order to see if the potential impact of the interventions is being realized or not. Also, interventions focusing on chronic care need to be multifaceted and span multiple departments of the organization in order for them to have significant impact. Hence we hypothesize that controlling for the number of unique and repeated interventions and the mean importance of interventions, the overall clinic rating will be higher for those clinics that evaluated a higher percentage of interventions, had a multidisciplinary team and implemented interventions spanning multiple departments. (Hypothesis 3)

3.2 Methods

3.2.1 Context

The data for this study was collected as a part of the EQHIV project (Landon et al., 2004) from HIV clinics funded by the Ryan White CARE Act. The EQHIV project evaluated a quality improvement collaborative (QIC) that was conducted by Institute of Healthcare Improvement (IHI) and in which 62 clinics participated from July 2000 to November 2001. Details of the IHI's breakthrough collaborative process and the underlying conceptual framework of Chronic Care Model (CCM) are described elsewhere (IHI, 2003).

3.2.2 Data

In this chapter, we used data collected in the EQHIV study (Landon et al., 2004; Marsden et al., 2006). These included information on the quality improvement interventions or initiatives attempted by the clinics participating in the collaborative. Of the 62 clinics that participated in the collaborative, teams from 50 clinics submitted monthly Senior Leader Reports (SLRs) that contained information on the change initiatives attempted at the sites. Information about each initiative attempted was recorded using a standard data entry form. The coder classified each intervention in one or more of the six CCM categories namely Clinical Information Systems, Delivery System Design, Decision Support, Self Management / Adherence, Community Linkages and Organizational Leadership. Moreover, each intervention was also classified in to one or more target areas. These target areas could be areas of care such as screenings, immunization, women's health, antiretroviral therapy or could be aspects of organization such as team building, chart initiatives, case management, staffing etc.

Additional characteristics that were coded included whether the initiative was evaluated by the clinic, what phase of implementation was the intervention in, how did it impact the physician and non-physician staff and what was the overall importance of the intervention for quality improvement effort. In addition to the coding at the individual intervention level, the coder was asked to rate the overall activity level at the clinic in terms of the overall likely impact of the whole package of changes being carried out.

The second dataset contains a survey of clinicians from two sets of clinics: one that participated in the collaborative and second set of identified control sites who did not participate in the collaborative. Before beginning the collaborative, 337 clinicians were randomly sampled (up to 5 from each clinic) of whom

300 completed the questionnaire. Analysis here focuses only on 104 clinicians from the intervention clinics. Survey questionnaires covered the structure, culture and organization of the clinics and their services with focus on the decision making environment in the clinic (leadership and staff attitudes towards quality improvement) and patient education efforts at the clinic. See Marsden et al. (2006) for more details on the survey methodology. We found 42 sites that had matching data from both datasets.

Other demographic characteristics pertaining to the clinics such as organization type (for example whether the clinic is a community based organization or a part of a larger multispecialty hospital), size (number of HIV patients), clinician staff mix (physicians vs. non-physicians) and region were also obtained from the EQHIV database.

3.2.3 Measures

3.2.3.1 Intervention choices

The monthly reports contained multiple aspects of the intervention choices made by the clinics during the course of the collaborative. The overall activity level in the clinic was captured by the number of unique interventions that were attempted. Since each intervention could span multiple areas of care, we constructed a measure similar to standardized Herfindahl's index to capture whether the clinic focused on a narrow aspect of care or took a more holistic perspective. This measure has a theoretical range of 0 to 1 where 1 implies that the clinic focused only one area of care, while 0 implies that the clinic focused on all areas of care equally. Thus, a lower score on this measure implies interventions that are more cross-departmental. The other variables that could be important in explaining the implementation success included percentage of interventions that

were evaluated by the clinics and average importance of the attempted interventions for quality improvement.

3.2.3.2 Clinic characteristics

We used two different types of variables to describe the organizational characteristics of the clinics. First, we used the demographic variables of organization type (community based organization, county health center, health department, hospital and university medical center), region (north, south, west, east) as control variables and whether the site was a specialty site or a general site. Number of HIV patients was used as a proxy for the size of the organization. A 0-1 dummy variable was used to describe whether the organization used multidisciplinary teams.

In order to capture the organizational culture of the clinics, we considered clinician's responses to questions on the decision-making environment in the organization. These questions asked respondents about the leadership's vision and support for improving quality of care and staff's receptiveness to new ideas and initiative for change. We aggregated the clinician's responses within each site and constructed an average response for each clinic. The intraclass correlation for each of these measures was high; however the overall reliability was low to moderate (Marsden et al., 2006).

3.2.3.3 Implementation success

We used the overall rating of the clinic as a measure of the implementation success. The rater was asked to summarize each clinic's monthly Senior Leadership Report and rate the clinic on a scale of 1 to 5 (1=nothing, or almost nothing significant happening and 5=potential for "breakthrough" change). This is different

from most of the quality improvement studies where implementation success is judged either by objective data on outcomes and processes or by self reporting of the employees of the concerned organization. Since such a rating procedure could be highly subjective, it was later verified by ratings from members of the EQHIV team to improve reliability.

3.2.3.4 Statistical analysis

We conducted an exploratory factor analysis on responses to nine questions pertaining to the decision making environment in the organization. We performed the principal factor analysis and used varimax method of orthogonal rotation. Using the Kaiser criterion for minimum eigen value of 1.0 yielded two factors with eigen value of 5.5 and 1.1. The proportion of the variance explained by the two factors was around 67%. We constructed the factors using the factor score method. We labeled the first factor as organization's QI focus and it included items such as: leadership's clarity in stating its QI vision, leadership's ability to implement new QI programs, staff initiative in developing new ideas, staff cooperation to improve HIV care, staff training in QI, and patient involvement in QI activities. We named the second factor as openness in the organization's culture and it included items such as: responsiveness and support of leadership to new ideas, respondent's willingness to participate in policy decisions and receptiveness of staff to new ideas. Cronbach's alpha for the two factors was found to be 0.87 and 0.70 indicating good reliability (Nunnally, 1967).

We used count regression to test Hypothesis 1 since the dependent variable is a count variable. Since the descriptive statistics in Table 2 clearly indicate that overdispersion for our dependent variable, number of unique interventions, we used a negative binomial model (Cameron and Trivedi, 1998). For Hypothesis 2

and 3 we used regular OLS models.

3.3 Results

3.3.1 Descriptive Statistics

Table 3.1 contains a brief description of various clinic characteristics chosen for this study as compared to the overall sample.

Table 3.1: Comparison of characteristics of study clinics with all Title III clinics

Variable	All Title III sites	Study sites
Region, %		
Northeast	39.8	40.4
South	27.7	35.7
Midwest	15.0	16.7
West	17.5	7.1
Organization Type, %		
Community Health Center	38.9	30.
Hospital	11.1	11.9
Other	50.0	57.14
HIV infected patients \pm SD, n	623 \pm 733	682 \pm 758
Large Clinic (>400 patients), %	51.0	50.0
HIV speciality clinic, %	74.3	64.3

Table 3.2 contains descriptive statistics for the continuous variables used in our analysis. Clinics implemented around 35 unique interventions and repeated around 9 interventions on average. Only 16.66% or one in every six interventions was evaluated and only 0.25 or one in every four interventions were repeated by

the clinics on average. The variability on each of these dimensions was considerable as seen from the standard deviation and the range in Table 2.

3.3.2 Number of interventions

Table 3.3 shows that openness in the organizational culture and QI focus in an organization was associated with higher number of interventions attempted in the clinic as hypothesized. However, we could not find evidence for significant association of the other independent variables: presence of multidisciplinary teams and regular measurement of quantifiable goals. Also, size of clinics as measured by the number of HIV patients was not associated with the number of interventions. We also find that there is statistically significant difference in the number of interventions across different organization types and also across different regions. This indicates presence of other underlying predictor variables that might be correlated with organization types and regions.

3.3.3 Cross-departmental nature of interventions

The dependent variable for Hypothesis 3 is the standardized Herfindahl's index over target areas of interventions. Since this is a continuous variable, we used regular OLS regression model. Recall that a lower Herfindahl's index implies that the interventions are focused on a broader set of target areas in the context of chronic care model. Table 3.4 contains the coefficient values with the respective p-values. Thus, the negative sign of the coefficients implies that we find support for our hypothesis: open organizational culture, QI focus and presence of multidisciplinary team is associated with interventions that are more cross-departmental in nature.

3.3.4 Implementation success

Table 5 contains the coefficient values and associated p-values for testing Hypothesis 3. As expected, number of interventions and mean importance rating of interventions were significantly associated with the overall implementation success. Moreover, after controlling for these two aspects of intervention choices, we found that clinics that repeated and evaluated a higher fraction of the interventions and implemented more cross-departmental interventions were rated as more successful in their quality improvement efforts. We also estimated another model that included the organizational culture variables in addition to the above variables. We found that all the variables pertaining to the organizational culture turned out to be non-significant. The overall F-value and adjusted R^2 values dropped suggesting that these variables do not explain additional variation. The introduction of these variables also reduced the statistical significance of a number of the original variables (Data not shown). This could potentially be due to the effect of the organizational culture being already captured in the intervention choice variables such as the number of unique interventions attempted and the cross-departmental nature of the interventions.

3.4 Discussion and Limitations

In this chapter, we were able to identify the organizational characteristics that can have an impact on the success of a quality improvement collaborative in a chronic care setting. We also identified the mechanism through which these characteristics impact the implementation outcome - the intervention and implementation choices during the collaborative. This provides an explanation of the fact that organizational performance in a quality improvement collaborative depends on

various contextual factors that are specific to the organization's culture.

However, there are several limitations to our study. First, the implementation success in our study was using the ratings given by the coders to the overall quality improvement efforts at the clinics. We have some assurance from the fact that inter-rater reliability was tested during codification in the earlier studies. Moreover, we recognize that this measure need not be an indicator of how the quality of care actually improved at the patient level in the participating clinics as a result of the collaborative.

Second, due to the structure of the existing survey instrument, we could extract only limited information regarding the organization's culture. It would be interesting to also elicit the attitudes of the team members and their influence in the broader organization since these characteristics could also significantly impact the implementation outcome.

Third, previous study based on this survey data found low to modest reliability of responses for many items. However, given the relatively high intraclass correlation, we feel reasonably confident about the validity of aggregation of survey responses within the clinics.

Finally, Structured Equations Modeling presents a more rigorous approach to estimate multiple relationships among variables of our interest. However, this approach usually requires more observations in order to obtain sufficient statistical power. An alternative approach involving instrumental variables should be employed to check the robustness of our analysis.

Table 3.2: Descriptive statistics

Variable	Mean	Std. Dev.	Min.	Max.
Number of HIV patients	681.85	758.20	59.00	3500.00
Organization's QI focus	0.00	1.00	-2.02	1.56
Openness in organization's culture	0.00	1.00	-2.17	2.03
Measurement of progress towards quantifiable goals	1.14	0.25	1.00	2.00
Number of unique interventions	34.71	18.66	1.00	77.00
Fraction of repeated interventions	0.25	0.13	0.00	0.54
Fraction of evaluated interventions	0.17	0.12	0.00	0.58
Mean rating of interventions	2.45	0.44	1.00	3.59
Overall clinic rating	2.55	0.51	1.36	3.24

Table 3.3: Negative binomial regression explaining the number of unique interventions (N = 42, Log Likelihood = 3960, Dispersion = 0.18)

Variable	Coefficient	p-value
Intercept	4.15	< 0.01
<i>Predictor Variables</i>		
QI focus	0.29	0.02
Openness in organizational culture	0.37	< 0.01
Measuring quantifiable goals	0.26	0.47
Multidisciplinary team	0.16	0.47
<i>Control Variables</i>		
Large clinics (> 400 patients)	-0.03	0.88
Specialty site	-0.02	0.92
Organization Type:		
Community based organization	-0.51	0.03
Community health center	-0.69	0.01
Public health clinic	0.04	0.91
Hospital	-0.61	0.05
Region:		
Northwest	-0.68	0.07
South	-0.67	0.06
Midwest	-0.56	0.26

Table 3.4: OLS Regression explaining the cross-departmental nature of interventions (N = 42, R2 = 0.55, Adj. R2 = 0.34, p-value = 0.02)

Variable	Coefficient	p-value
Intercept	0.17	0.02
<i>Predictor Variables</i>		
QI focus	-0.03	0.09
Openness in organizational culture	-0.04	< 0.01
Measuring quantifiable goals	-0.07	0.12
Multidisciplinary team	-0.09	<0.01
<i>Control Variables</i>		
Large clinics (> 400 patients)	-0.02	0.39
Specialty site	0.01	0.92
Organization Type:		
Community based organization	-0.00	0.93
Community health center	0.04	0.31
Public health clinic	-0.04	0.36
Hospital	0.04	0.42
Region:		
Northwest	0.07	0.14
South	0.04	0.37
Midwest	0.02	0.70

Table 3.5: OLS regression explaining the implementation success (overall clinic rating) at clinics ($N = 42$, $R^2 = 0.57$, Adj. $R^2 = 0.49$, p -value ≤ 0.01)

Variable	Coefficient	p-value
Intercept	1.19	0.05
<i>Predictor Variables</i>		
Percentage of interventions repeated	0.89	0.08
Percentages of interventions evaluated	1.13	0.06
Multidisciplinary team	0.26	0.08
Cross-departmental nature of interventions (Herfindahl's index)	-7.91	0.03
Number of unique interventions	0.01	0.03
Mean importance rating of interventions	0.29	0.08

REFERENCES

- Abrams, R.A. and U. S. Karmarkar. 1979. Infinite horizon investment-consumption policies. *Management Sci.* **25** 1005-1013.
- Amenyah, J., B. Chovitz, E. Hasselberg, A. Karim, D. Mmari, S. Nyinondi, T. Rosche. 2005. Tanzania: Integrated Logistics System Pilot-Test Evaluation: Using the Logistics Indicator Assessment Tool. Arlington, Va.: DELIVER, for the U.S. Agency for International Development.
- American Antitrust Institute. 2004. Could the federal government have prevented the flu vaccine shortage? An industrial organization perspective. October 29.
- Anupindi, R., R. Akella. 1993. Diversification under supply uncertainty. *Management Sci.* **39** 944-963.
- Argote, L., D. Gruenfeld, C. Naquin. 2001. Group learning in organizations. In Turner, M. E. (ed.), *Groups at Work: Advances in Theory and Research*. 369-411. Erlbaum, New York, 369-411.
- Auvert B., S. Males, A. Puren, D. Taljaard, M. Carael, B. Williams. 2004. Can highly active antiretroviral therapy reduce the spread of HIV? A study in a township of South Africa. *JAIDS* **36** 613-621.
- Bartlett, J.G. 2006. Ten years of HAART: Foundation for the future.
- BBC News. 2004. Outrage at SA AIDS drug shortage. June 1.
- Bekker L.G., C. Orrell, L. Reader, K. Matoti, K. Cohen, R. Martell, F. Abdullah, R. Wood. 2003. Antiretroviral therapy in a community clinic - early lessons from a pilot project. *South African Medical Journal* **93** 458-462.
- Bennet, S. and C. Chanfreau. 2005. Approaches to rationing antiretroviral treat-

ment: ethical and equity implications. *Bulletin of the World Health Organization* **83** 541-547.

Berwick, D.M. 2002. A user's manual for the IOM's 'quality chasm' report. *Health Affairs* **21** 80-90.

Billeter, M. 2004. Meeting the quality standards for influenza, pneumococcal vaccination. *Infectious Disease News*.

Blumenthal, D. 1996. Quality of care – What is it? *The New Eng. J. Med.* **335** 891-894.

Blower, S, L. Ma, P. Farmer, S. Koenig. 2003. Predicting the Impact of Antiretrovirals in Resource-Poor Settings: Preventing HIV Infections whilst Controlling Drug Resistance. *Current Drug Targets – Infectious Disorders* **3** 345-353.

Blower, S., E. Bodine, J. Kahn and W. MacFarland. 2005. The antiretroviral rollout and drug-resistant HIV in Africa: insights from empirical data and theoretical models. *AIDS* **19** 1-14.

Brito, D.L., E. Sheshinski, M.D. Intriligator. 1991. Externalities and compulsory vaccinations. *Journal of Public Economics* **45** 69-90.

Brown, D., 2004. How U.S. got down to two makers of flu vaccine. *Washington Post* October 17, 2004. A01-04.

Bunnell R., J.P. Ekwaru, P. Solberg, N. Wamai, W. Bikaako-Kajura, W. Were, A. Coutinho, C. Liechty, E. Madraa, G. Rutherford, J. Mermin. 2006. Changes in sexual behavior and risk of HIV transmission after antiretroviral therapy and prevention interventions in rural Uganda. *AIDS* **20** 85-92.

Cameron, A.C. and P.K. Trivedi. 1998. Regression analysis of count data. Cambridge University Press, Cambridge, UK.

Chandani, Y. and M. Muwonge. 2003. Logistics and procurement decisions and

issues for consideration for introducing and expanding access to ARVs in Uganda.
JSI

Chick, S., H. Mamani, D. Simchi-Levi. 2006. Supply chain coordination and influenza vaccination. *Working paper*

Chassin, M. R. 1996. Improving the quality of care. *The New Eng. J. Med.* 335. 1060-1063.

Chassin, M. R., R. W. Galvin, The National Roundtable on Quality. 1998. The urgent need to improve health care quality. *JAMA* 280. 1000-1005.

Cleary, S.M., D. McIntyre. A. Boulle. 2006. The cost-effectiveness of antiretroviral treatment in Khayelitsha, South Africa - a primary data analysis. *Cost effectiveness and Resource Allocation* 4.

Coetzee D., A. Boulle, K. Hildebrand, V. Asselman, G. van Cutsem G, F. Goemaere. 2004. Promoting adherence to antiretroviral therapy: the experience from a primary care setting in Khayelitsha, South Africa. *AIDS* 18 S27-S31.

Colebunders R., T. Bukenya, N. Pakker, O. Smith, V. Boeynaems, J. Waldron, A. Muganzi Muganga, C. Twijukye, K. Mcadam, E. Katabira. 2007. Assessment of the patient flow at the infectious diseases institute out-patient clinic, Kampala, Uganda. *AIDS Care* 19 149-151.

Cournot A., 1838. *Recherches sur les principes mathematiques de la theorie des richesses* English edition (ed. N. Bacon): Researches into the Mathematical Principles of the Theory of Wealth. Macmillan, New York. 1987.

Dada, M., N. C. Petruzzi, L. B. Schwarz. 2007. A newsvendor's procurement problem when suppliers are unreliable. *Manufacturing & Service Operations Management* 9 9-32.

Daniel, G. 2006. Improving ARV Medicines and Information Management in

Ethiopia, January 8–March 11, 2006: Technical Assistance Update and Trip Report. Submitted to the U.S. Agency for International Development by the Rational Pharmaceutical Management Plus Program. Arlington, VA: Management Sciences for Health.

Danzon, P.M., N.S. Periera, S.S. Tejjwani. 2005. Vaccine supply: A cross national perspective. *Health Affairs* **24** 706-717.

de Véricourt, F., F. Karaesman, Y. Dallery. 2002. Optimal stock allocation for a capacitated supply system. *Management Sci.* **48** 1486-1501.

de Véricourt, F. and M. S. Lobo. 2006. Resource and revenue management in nonprofit operations. *Working paper* Fuqua School of Business.

Dees, J.G. 1998. Enterprising nonprofits. *Harvard Business Review* Jan-Feb 1998. 58-67.

Deliver report submitted to the logistics subcommittee of the ARV task force, Ministry of Health, Uganda.

Denardo, E. 1965. Contraction mappings in the theory underlying dynamic programming. *SIAM Review* **9** 165-177.

Deuermeyer, B.L., W.P. Pierskalla. 1978. A by-product production system with an alternative. *Management Sci.* **24** 1373-1383.

Dilley J, McFarland W, Sullivan P, Discepola M. 1998. Psychosocial correlates of unprotected anal sex in a cohort of gay men attending an HIV-negative support group. *AIDS Education and Prevention* **10** 317–326.

Drummond, M.F. 1980. *Studies in Economic Appraisal in Health Care*. New York: Oxford University Press.

Drummond, M.F., G. Stoddart, G.W. Torrance. 1980. *Principles of Economic Appraisal in Health Care*. Oxford: Oxford University Press.

- Dudley, R. A. B. E. Landon, H. R. Rubin, N. L. Keating, C. A. Medlin, H. S. Luft. 2000. Assessing the relationship between quality of care and the characteristics of health care organizations. *Med. Care Res. Rev.* **57** 116-135.
- Edmondson, A. C. 1999. Psychological safety and learning behavior in work teams. *Admin. Sci. Quart.* **44** 350-383.
- Edmondson, A. C., R. Bohmer, G. P. Pisano. 2001. Disrupted routines: Team learning and new technology adaptation. *Admin. Sci. Quart.* **46** 685-716.
- Ekong, E., V. Idemyor, O. Akinlade, A. Uwah. 2004. Challenges to antiretroviral (ARV) drug therapy in resource-limited settings - Progress and challenges in the Nigerian initiative. *11th Conference on Retroviruses and Opportunistic Infection*. San Francisco, USA
- El-Sadr W.M., J.D. Lundgren, J.D. Neaton et al. 2006. CD4+count-guided interruption of antiretroviral treatment *The New Eng. J. Med.* **22**. 2283-2296
- Evans, R.V. 1967. Inventory control of a multi-product system with a limited production resource. *Naval. Res. Logist. Quart.* **14** 173-184.
- Evans, R.V. 1969. Sales and restocking policies in a single item inventory system. *Management Sci.* **14** 463-472.
- Evans, M.R., P.A. Watson. 2003. Why do older people not get immunized against influenza? A community survey. *Vaccine* **21** 2421-2427.
- Federgruen, A., N. Yang. 2005. Optimal supply diversification under general supply risks. *Working paper*
- Ferlie, E. B. and S. M. Shortell. 2001. Improving the quality of health care in the United Kingdom and the United States: A framework for change. *The Milbank Quarterly* **79** 281-315.
- Finkelstein S., C.N. Smart, A.M. Gralla, C.R. d'Oliviera. 1981. A two-stage

- model for the control of epidemic influenza. *Management Sci.* **27**. 834-846.
- Flaks R.C., W. J. Burman, P. J. Gourley, C. A. Rietmeijer, D. L. Cohn. 2003. HIV transmission risk behavior and its relation to antiretroviral treatment adherence. *Sexually Transmitted Diseases* **30** 399-404.
- Flood, A.B. 1994. The impact of organisational and managerial factors on the quality of care in health care organisations. *Med. Care Rev.* **51**. 381-429.
- Forbes. 2004. Broken Eggs. November 1.
- Foster, W. and J. Bradach. 2005. Should nonprofits seek profits? *Harvard Business Review*. February 2005. 92-100.
- Frank, K.C., R.Q. Zhang, I. Duenyas. 2003. Optimal policies for inventory systems with priority demand classes. *Oper. Res.* **51** 993-1002.
- GAO. 2001. *Flu vaccine: Supply problems heighten need to ensure access for high-risk people*. Report to congressional requesters.
- Gerchak, Y., M. Parlar. 1990. Yield variability, cost trade-offs and diversification in the EOQ model. *Naval Res. Logist.* **37** 341-354.
- Gerdil, C. 2002. The annual production cycle for influenza vaccine. *Vaccine* **21** 1776-1779.
- Gottlieb, S. 2004. Vaccine makers get a shot in the arm. *Forbes.com* October 11.
- Gray R.H., Li X., Wawer M.J., Gange S.J., Serwadda D., Sewankambo N.K., Moore R., Wabwire-Mangen F., Lutalo T., Quinn T.C. 2003. Stochastic simulation of the impact of antiretroviral therapy and HIV vaccines on HIV transmission; Rakai, Uganda. *AIDS* **17** 1941-1951.
- Gronbjerg, K.A. 1992. *Nonprofit human service organizations: Funding strategies and patterns of adaptation*. In Y. Hasenfeld (Ed.), *Human Services as Complex Organizations*. 73-97. Newbury Park, CA: Sage.

- Ha, A. 1997a. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Sci.* **43** 1093–1103.
- Ha, A. 1997b. Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Res. Logist.* **44** 458–472.
- Hart C.E., J. L. Lennox, M. Pratt-Palmore, T. C. Wright, R. F. Schinazi, T. Evans-Strickfaden, T. J. Bush, C. Schnell, L. J. Conley, K. A. Clancy, T. V. Ellerbrock. 1999. Correlation of human immunodeficiency virus type 1 RNA levels in blood and the female genital tract. *Journal of Infectious Diseases* **179** 871-882.
- Health Systems Trust. 2005. Implementing the Comprehensive Care and Treatment Programme for HIV and AIDS patients in the Free State: Sharing experiences. Conference report.
- Henig M., Y. Gerchak. 1990. The structure of periodic review policies in the presence of random yield. *Oper. Res.* **38** 634-643.
- Holtgrave, D.R. and S.D. Pinkerton. 1997. Updates of cost illness and quality of life estimates for use in economic evaluations of HIV prevention programs. *JAIDS* **16** 54-62.
- Horbar, J. D., J. Rogowski, P. Plesk, P. Delmor, W. H. Edwards, J. Hocker, A. D. Kantak, P. Lewallen, W. Lewis, E. Lewit, C. J. McCarroll, D. Majsce, N. R. Payne, P. Shiono, R. F. Soll, K. Leahy, J. H. Carpenter. 2001. Collaborative quality improvement for neonatal intensive care. *Pediatrics* **107** 14-22.
- Hurst S.A., S. C. Hull, G. DuVal, M. Danis. 2005. Physicians' response to resource constraints. *Archives of Internal Medicine* **165** 639-644.
- Institute of Medicine (IOM). 2001. Crossing the Quality Chasm: A New Health System for the 21st Century. National Academy Press, Washington, D.C.

- Institute Of Medicine. 2005. *Scaling Up Treatment for the Global AIDS Pandemic: Challenges and Opportunities*. The National Academies Press.
- IRINNews.org. 2005. Swaziland: HIV positive Swazis take govt to task over ARV supply. December 6.
- IRINNews.org. 2002. Coalities decries “Regular” shortage of HIV drugs. April 16.
- ITPC. 2005. Missing the target: A report on HIV/AIDS treatment access from the frontlines.
- Institute for Healthcare Improvement. 2003. Breakthrough series collaboratives. Available at www.ihl.org/collaboratives/
- Jelsma, J., E. Maclean, J. Hughes, X. Tinise, M. Darder. 2005. An investigation into the health-related quality of life of individuals living with HIV who are receiving HAART. *AIDS Care* **17** 579-588.
- JSI. 2006. Personal communication.
- Kaplan, E. H. and H. Pollack. 1998. Allocating HIV prevention resources. *Socio-Econ. Plann. Sci.* **32** 257-263.
- Karmarkar, U.S. 1981. The multiperiod multilocation inventory problem. *Oper. Res.* **29** 215-228.
- Katz M.H., S. K. Schwarcz, T. A. Kellogg, J. Klausner, J. W. Dilley, S. Gibson, W. McFarland. 2002. Impact of highly active antiretroviral treatment on HIV seroincidence among men who have sex with men in San Francisco. *American Journal of Public Health* **92** 388-394.
- Klemperer P., M. Meyer. 1986. Price-competition vs. quantity competition - The role of uncertainty. *The RAND Journal of Economics* **17** 618-638.
- Kurian, S., D.S. Blog, K.M. Sherin. 2004. Optimizing vaccine availability and

utilization: Position statement of the American College of Preventive Medicine. *American Journal of Preventive Medicine* **26** 372-374.

Landman K.Z., G. D. Kinabo, M. Schimana, W. M. Dolmans, M. E. Swai, J. F. Shao and J. A. Crump. 2006. Capacity of health-care facilities to deliver HIV treatment and care services, Northern Tanzania, 2004. *International Journal of STD & AIDS* **17** 459-462

Landon, B.E., I.B. Wilson, K. McInnes, M.B Landrum, L. Hirschhorn, P.V. Marsden, D. Gustafson, P.D. Cleary. 2004. Effects of a quality improvement collaborative on the outcome of care of patients with HIV infection: the EQHIV study. *Ann. Intern. Med.* **140** 887-896.

Leape L.L., D.M. Berwick. 2005. Five years after to err is human: What have we learned? *JAMA* **293** 2384-2390.

Leland, H.E. 1972. Theory of the firm facing uncertain demand. *Amer. Econ. Rev.* **62** 278-291.

Lippman, S.A. 1974. Semi-markov decision processes with unbounded rewards. *Management Sci.* **19** 717-731.

Macklin, R. 2004. Ethics and equity in access to HIV treatment - 3 by 5 initiative. Background paper for the consultation on equitable access to treatment and care for HIV/AIDS.

Mankiw, N.G., M.D. Whinston. 1986. Free entry and social inefficiency. *The RAND Journal of Economics* **17** 48-58.

Marsden, P. V., B. E. Landon, I. B. Wilson, K. McInnes, L. R. Hirschorn, L. Ding, P. Cleary. 2006. The reliability of survey assessments of characteristics of medical clinics. *Health Services Research* **41** 265-283.

Martone. 2000. Influenza: The virus, the disease and how to protect yourself.

National Foundation of Infectious diseases.

McGough, L.J., S. J. Reynolds, T.C. Quinn, J.M. Zenilman. 2005. Which patients first? Setting priorities for antiretroviral therapy where resources are limited. *American Journal of Public Health* **95** 1173-1180.

Miller, B. 1974. Optimal consumption with stochastic income stream. *Econometrica* **42** 253-266.

Longini, I.M., E. Ackerman, L.R. Elveback. 1977. An optimization model for influenza epidemics. *Math. Biosci.* **38** 141-157.

Mittman, B.S. 2004. Creating the evidence base for quality improvement collaboratives. *Ann Intern Med* **140** 897-901.

Moatti J.P., J. Prudhomme, D. C. Traore, A. Juillet-Amari, H. A. D. Akribi, P. Msellati. 2003. Access to antiretroviral treatment and sexual behaviours of HIV-infected patients aware of their serostatus in Cote d'Ivoire. *AIDS* **17** S69-S77.

Musicco M., A. Lazzarin, A. Nicolosi, M. Gasparini, P. Costigliola, T. C. Arici, A. Saracco. 1994. Antiretroviral Treatment Of Men Infected With Human-Immunodeficiency-Virus Type-1 Reduces The Incidence Of Heterosexual Transmission. *Arch. Int. Med.* **154** 1971-1976.

Nahmias, S. and W. S. Demmy. 1981. Operating characteristics of an inventory system with rationing. *Management Sci.* **27** 1236-1245.

National Influenza Vaccine Summit. 2006. Comments on the status of prebooked influenza vaccine for 2006-2007.

<http://www.ama-assn.org/ama/pub/category/13732.html>

National Vaccine Advisory Committee. 2003. Strengthening the Supply of Routinely Recommended Vaccines in the United States.

<http://www.hhs.gov/nvpo/bulletins/nvac-vs-r.htm#intro>

- Newsweek. 2004. The flu shot fiasco. November 1.
- Nichol, K.L. 2001. Cost-benefit analysis of a strategy to vaccinate healthy working adults against influenza. *Arch. Int. Med.* **161** 749-759.
- Nichol, K.L., J. Wuorenma, T. von Sternberg. 1998. Benefits of influenza vaccination for low-, medium- and high-risk senior citizens. *Arch. Int. Med.* **158** 1769-1776.
- Nunnally, J. C. 1967. Psychometric Theory. McGraw-Hill, New York.
- Nyenwa, J., D. Alt, A. Karim, T. Kufa, J. Mboyane, Y. Ouedraogo, T. Simoyi. 2005. Zimbabwe HIV & AIDS Logistics System Assessment. Arlington, Va.: John Snow, Inc./DELIVER, for the U.S. Agency for International Development.
- Olsen, T.L. and R.P. Parker. 2006. Customer behavior in inventory management. *Working paper* Yale School of Management.
- O'Mara, D., K. Fukuda and J.A. Singleton. 2003. Influenza vaccine: ensuring timely and adequate supply. *Infect. Med.* **20** 548-554.
- Ovretveit, J., P. Bate, P. Cleary, S. Cretin, D. Gustafson, K. McInnes, H. McLeod, T. Molfenter, P. Plsek, G. Robert, S. Shortell, T. Wilson. 2002. Quality collaboratives: lessons from research. *Qual. Saf. Health Care* **11** 345-351.
- Oyugi, J.H., J. Byakika-Tusiime, K. Ragland, O. Laeyendecker, R. Mugerwa, C. Kityo, P. Mugenyi, T.C. Quinn, D.R. Bangsberg. 2007. Treatment interruptions predict resistance in HIV-positive individuals purchasing fixed-dose combination antiretroviral therapy in Kampala, Uganda. *AIDS* **21** 965-971.
- Palella et al. 1998. Declining morbidity and mortality among patients with advanced Human Immunodeficiency Virus infection. *The N. Eng. J. Med* **338** 853-860.
- Pereira A.S., A. D. M. Kashuba, S. A. Fiscus, J. E. Hall, R. R. Tidwell, L. Troiani,

- J. A. Dunn, J. J. Eron, M. S. Cohen. 1999. Nucleoside analogues achieve high concentrations in seminal plasma: Relationship between drug concentration and virus burden. *Journal of Infectious Diseases* **180** 2039-2043.
- Perneger T.V., D. P. Martin, P. A. Bovier. 2002. Physicians' attitude toward health care rationing. *Medical Decision Making* **22** 65-72.
- Pienaar D., L. Myer, S. Cleary, D. Coetzee, D. Michaels, K. Cloete, H. Schneider, A. Boulle. 2006. Models of Care for Antiretroviral Service Delivery. Cape Town: University of Cape Town.
- Philipson, T. 2003. Economic epidemiology and infectious diseases. *Handbook of Health Economics. Vol 1B.* (eds. Culyer A.J., J.P. Newhouse)
- Plsek, P. 1997. Collaborating across organizational boundaries to improve quality of care. *Am. J. Infect. Control* **25** 85-95.
- Powermed. 2005. Pandemic influenza and bioterror preparedness: Role of PMEDTM DNA vaccines. www.powermed.com
- Richter, A., M.L. Brandeau, and D.K. Owens. 1999. An analysis of optimal resource allocation for HIV prevention among injection drug users and nonusers. *Med. Decis. Making* **19** 167-179.
- Rosen, S. I. Sanne, A. Collier, J.L. Simon. 2005. Rationing antiretroviral therapy for HIV / AIDS in Africa: Choices and consequences. *PLoS Medicine* **2** e303.
- Roth D.L., K. E. Stewart, O. J. Clay, A. van der Straten, E. Karita, S. Allen. 2001. Sexual practices of HIV discordant and concordant couples in Rwanda: effects of a testing and counselling programme for men. *International Journal of STD & AIDS* **12** 181-188.
- Rothman, A. A. and E. H. Wagner. 2003. Chronic illness management: What is the role of primary care? *Annals of Internal Medicine* **138** 256-262.

- Salomon J.A., D. R. Hogan, J. Stover, K. A. Stanecki, N. Walker, P. D. Ghys, B. Schwartlander. 2005. Integrating HIV prevention and treatment: From slogans to impact. *PLOS Medicine*. **2** 50-56.
- Sarin, S., C. McDermott. 2003. Effect of team leader characteristics on learning, knowledge application, and performance of cross-functional new product development teams. *Decision Sci.* **34** 707-739.
- Scott, K. 2003. Funding matters: The impact of Canada's new funding regime on nonprofit and voluntary organizations.
- Shepard D.S. and M.S. Thompson. 1979. First Principles of Cost-Effectiveness in Health. *Public Health Reports* **6** 535-543.
- Shortell, S.M., C.L. Bennett, G.R. Byck. 1998. Assessing the impact of continuous quality improvement on clinical practice: What it will take to accelerate the progress. *The Milbank Quarterly* **76** 593-624.
- Simpson, V.P. 1978. Optimum Solution Structure for a Repairable Inventory Problem. *Oper. Res.* **26** 270-281.
- Stokey, N.L., R.E. Lucas, Jr., E.C. Prescott. 1989. *Recursive Methods in Economic Dynamics*. Harvard University Press.
- Strikas, R.A. 2005. Driving increased influenza vaccine uptake. The 2005 *National Influenza Summit, Chicago, IL*.
- Suzumura K., K. Kiyono. 1987. Entry barriers and economic welfare. *Rev. Econ. Studies* **54** 157-167.
- Thompson W.W., D.K. Shay, E. Weintraub, L. Brammer, N. Cox, L.J. Anderson, K. Fukuda. 2003. Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA* 179-186.
- Time. 2004. The flu Snafu. November 1.

- Topkis, D. M. 1968. Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Management Sci.* **15** 160–176.
- Ubel P.A, DeKay M.L., Baron J., D. A. Asch. 1996. Cost-effectiveness analysis in a setting of budget constraints: Is it equitable? *The N. Eng. J. Med.* **334** 1174-1177.
- U.S. Census Bureau. 2004. Nation adds 3 million people in last year; Nevada again fastest-growing state.
www.census.gov/Press-Release/www/releases/archives/population/003153.html.
- Vives, X. 1999. *Oligopoly Pricing: Old Ideas and New Tools.* The MIT Press. Cambridge, Massachusetts.
- van der Straten A., C. A. Gomez, S. J. Saul, J. Quan, N. Padian. 2000. Sexual risk behaviors among heterosexual HIV serodiscordant couples in the era of post-exposure prevention and viral suppressive therapy. *AIDS* **14** F47-F54.
- van Nunen, J.A.E.E. and J. Wessels. 1975. A note on dynamic programming with unbounded rewards. *Management Sci.* **24** 576-580.
- van Oosterhout, J.J., N. Bodasing, J.J. Kumwenda, C. Nyirenda, J. Mallewa, P.R. Cleary, M.P. de Baar, R. Schuurman, D.M. Burger, E.E. Zijlstra. 2005. Evaluation of antiretroviral therapy results in a resource-poor setting in Blantyre, Malawi. *Tropical Medicine and International Health* **10** 464-470.
- von Weizsäcker, C.C. 1980. A welfare analysis of barriers to entry. *Bell Journal of Economics* **11** 399-420.
- Wagner, E. H., B. T. Austin, C. Davis, M. Hindmarsh, J. Schaefer, A. Bonomi. 2001. Improving chronic illness care: Translating evidence into action. *Health Affairs* **20** 64-78.
- Wagner, E. H., B. T. Austin, M. von Korff. 1996. Organizing care for patients

- with chronic illness. *The Milbank Quarterly* **74** 511-544.
- Wagstaff, A. 1991. QALYs and equity-efficiency trade-off. *Journal of Health Economics* **10** 21-41.
- Walensky R.P., A.D. Paltiel, E. Losina. 2006. Three million years of life saved: The survival benefits of AIDS therapy in the United States. *J Infect Dis.* forthcoming.
- West, E. 2001. Management matters: the link between hospital organisation and quality of patient care. *Qual. Health Care* **10** 40-48.
- Wester C.W., H. Bussmann, A. Avalos, N. Ndwapi, T. Gaolathe, P. Cardiello, C. Bussmann, H. Moffat, Mazonde P and Marlink RG. 2005. Establishment of a Public Antiretroviral Treatment Clinic for Adults in Urban Botswana: Lessons Learned. *Clinical Infectious Diseases* **40** 1041-1044.
- WHO. 1998. Guidance modules on antiretroviral treatments. Module 8: ARVs - Regulation, distribution and control.
- WHO. 2002. Influenza vaccines: WHO position paper. *Weekly Epidemiological Record* **28** 230-239.
- WHO. 2003. Emergency scale up of antiretroviral therapy in resource-limited settings: technical and operational recommendations to achieve 3 by 5.
- WHO. 2003a. Antiretroviral Therapy In Primary Health Care: Experience Of The Khayelitsha Programme In South Africa.
- WHO. 2003b. Access to Antiretroviral Treatment and Care: The Experience of the HIV Equity Initiative, Cange, Haiti.
- WHO. 2004. Patient monitoring guidelines for HIV care and antiretroviral therapy (ART).
- WHO. 2005a. Progress on global access to HIV antiretroviral therapy: An update

on “3 by 5”.

WHO. 2005b. AIDS epidemic update: December 2005.

WHO 2006. Progress on global access to HIV antiretroviral therapy: A report on “3 by 5” and beyond.

WHO. 2007. Towards Universal Access: Scaling up priority HIV/AIDS interventions in the health sector. Progress Report.

Williams, D. G. 2005. The influenza vaccine supply chain: structure, risk and coordination. Unpublished thesis. Massachusetts Institute of Technology - Zaragoza International Logistics Program.

Wilson T.E. and H. Minkoff. 2001. Brief report: condom use consistency associated with beliefs regarding HIV disease transmission among women receiving HIV antiretroviral therapy. *JAIDS* **27** 289–291.

Wu, J.T., L.M. Wein, A.S. Perelson. 2005. Optimization of influenza vaccine selection. *Oper. Res.* **53** 456-476.

Yadav, P. 2005. Value of Creating a Redistribution Network for Influenza Vaccine in the United States. Fifth Workshop on Business Aspects of Closed Loop Supply Chains. Nashville TN.

Yano C.A and H.L. Lee. 1995. Lot sizing with random yields: A review. *Oper. Res.* **43** 311-334.

Zenios, S.A., L.M. Wein, G.M Chertow. 2000. Dynamic allocation of kidneys to candidates on the transplant waiting list. *Oper. Res.* **49** 549-569.

Zhang, R.Q. and M. Sobel. 2001. Inventory policies for systems with stochastic and deterministic demand. *Oper. Res.* **49** 157-162.