TESIS DE LA UNIVERSIDAD DE ZARAGOZA

2014

65

Gerardo Pelayo Rubio

Analizando los efectos del diseño contractual y estructural en la cadena de suministro del sector salud

Departamento
Zaragoza Logistics Center

Director/es
Gürbüz, Mustafa Çagri

ISSN 2254-7606

Prensas de la Universidad Universidad Zaragoza

Universidad Zaragoza
1542

Tesis Doctoral

ANALIZANDO LOS EFECTOS DEL DISEÑO CONTRACTUAL Y ESTRUCTURAL EN LA CADENA DE SUMINISTRO DEL SECTOR SALUD

Autor

Gerardo Pelayo Rubio

Director/es

Gürbüz, Mustafa Çagri

UNIVERSIDAD DE ZARAGOZA

Zaragoza Logistics Center

UNIVERSIDAD DE ZARAGOZA

TESIS DOCTORAL

ANALIZANDO LOS EFECTOS DEL DISEÑO CONTRACTUAL Y ESTRUCTURAL EN LA CADENA DE SUMINISTRO DEL SECTOR SALUD

Gerardo Pelayo Rubio

Máster de Ingeniería en Logística y Gestión de la Cadena de Suministro,

Programa Internacional de Logística MIT-Zaragoza

Zaragoza Logistics Center (ZLC), Universidad de Zaragoza (España)

Licenciado en Ingeniería Industrial y de Sistemas,

Instituto Tecnológico y de Estudios Superiores de Monterrey (México)

24 de Marzo de 2014

Autor: D. Gerardo Pelayo Rubio, Doctorando

Director de tesis: Dr. Mustafa Çagri Gürbüz
Profesor de Gestión de la Cadena de Suministro,
Programa Internacional de Logística MIT-Zaragoza
Zaragoza Logistics Center (ZLC)

Director del Zaragoza Logistics Center (ZLC): Dr. David Gonsalvez

Analizando los efectos del diseño contractual y estructural en la cadena de suministro del sector salud

por

Gerardo Pelayo Rubio

en relación con el cumplimiento parcial de los requisitos para la obtención del título de Doctor en Logística y Gestión de las Cadenas de Suministro

Resumen

Introducción

Balancear el acceso a medicamentos necesarios contra el aumento en los costes es uno de los retos fundamentales en el diseño y la reforma de los sistemas de salud. Del año 2000 al 2008, el crecimiento promedio en el gasto per capita en productos farmacéuticos para los países de la Organización para la Cooperación y Desarrollo Económico (OCDE) fue de casi 60 %. La problemática se encuentra especialmente presente en el proceso de introducción de nuevos medicamentos orientados al tratamiento de condiciones crónicas. Aquí los precios de lista propuestos por los productores farmacéuticos tienden a ser altos para recuperar la inversión, en algunas ocasiones contrastando con la falta de evidencia robusta con respecto a la coste-efectividad del tratamiento al momento de la negociación del precio de transferencia. Más aún, tal coste-efectividad puede variar a través de las diferentes indicaciones terapéuticas de un mismo medicamento, *i.e.,* para distintos grupos de pacientes. Como resultado, en los acuerdos tradicionales el pagador de salud (*e.g.,* Sistemas Nacionales de Salud, Organizaciones de Mantenimiento de la Salud, grandes empresas aseguradoras) puede verse atrapado entre restringir el acceso al medicamento o arriesgar el pago de altos precios que pueden no justificarse *ex-post* debido a la incertidumbre sobre el valor real de la innovación terapéutica del medicamento, la falta de solidez en los resultados presentados por el productor, o la replicabilidad de esos resultados en la práctica clínica. En respuesta a la creciente presión para controlar el gasto en el sector salud, los pagadores de salud han empujado a los productores farmacéuticos a reducir los precios, potencialmente reduciendo los incentivos para invertir en tratamientos innovadores, y continuamente resultando en la (temporal o definitiva) ausencia de un acuerdo entre ambos agentes involucrados con la consecuente pérdida de bienestar para los pacientes potenciales y de beneficios financieros para el productor. Lo anterior ha motivado a los productores - particularmente aquellos en los sectores cardiovasculares y de oncología - a explorar acuerdos más sofisticados donde los riesgos puedan ser compartidos de una manera más eficiente.

Motivados por la tendencia mencionada, reconocemos que un pagador de salud debe decidir no únicamente si aprobar o no un nuevo medicamento para su (parcial o total) reembolso por consumo para la población de pacientes que sirve, sino también determinar el nivel de servicio (cuál será el volumen adquirido para satisfacer la demanda de los pacientes), el nivel de acceso (cuáles grupos de pacientes estarán cubiertos por el pagador de salud), y las condicio-

nes de reembolso a los productores (los parámetros del contrato). Además reconocemos que un pagador de salud puede tener diferentes prioridades según el ambiente social e industrial donde opere (*e.g.,* maximizar la eficiencia de los recursos versus maximizar el bienestar de los pacientes), así como restricciones (*e.g.,* límite máximo de gastos por periodo de demanda para algún medicamento o innovación terapeútica, y un límite mínimo de coste-efectividad). Con respecto a los productores farmacéuticos, consideramos que: la determinación del precio de transferencia puede ser exógena (a través de precios de referencia externos) o endógena (a través de acuerdos directos con los pagadores de salud); que pueden internalizar (parcial o totalmente) el riesgo de mantener el inventario; y que en algunos casos son capaces de segmentar el mercado a través de la creación de productos o canales distinguibles enfocados a cada grupo de pacientes.

Preguntas de Investigación

En el contexto descrito donde un medicamento innovador con múltiples aplicaciones terapéuticas busca su introducción al mercado, la presente investigación pretende responder de manera analítica las preguntas mostradas a continuación.

- En un sistema verticalmente integrado, ¿cómo interactúan los niveles de acceso y de servicio en función de las prioridades y restricciones del sistema?
- En una cadena de tipo productor - pagador de salud, ¿qué cambia cuando el precio es determinado de manera exógena (*vs.* endógena) y el productor está (*vs.* no está) dispuesto a compartir los riesgos asociados a la incertidumbre en la magnitud de la demanda y en los resultados observados en los pacientes?
- Para un medicamento con múltiples aplicaciones terapéuticas, ¿cómo se refleja la decisión de segmentar *vs.* consolidar el diseño/canal de distribución, en el nivel de servicio y los incentivos para ejercer esfuerzo orientado a la innovación?
- ¿Cuál es el efecto de todo lo anterior en los beneficios del productor farmacéutico, los gastos del pagador de salud, y el bienstar de los pacientes?

De este modo, la investigación espera contribuir a una comprensión más amplia del comportamiento del sistema, y así eventualmente orientar el diseño de la estructura y los contratos en las cadenas de suministro del sector salud, de modo que exista una mejor alineación con los objetivos de los agentes involucrados.

Metodología y Suposiciones Fundamentales

El procedimiento general para responder a las preguntas anteriores se basa en una modelación matemática de las situaciones previamente descritas utilizando la estructura del modelo del vendedor de periódicos (o *newsvendor*, como se le conoce normalmente en inglés). Esta elección se debe a: i)los tiempos de espera extensos (aproximadamente 4 meses) para la construcción de capacidad productiva, aprovisionamiento de materias primas, producción,

y envío de los medicamentos; ii)la práctica común en la industria de ofrecer precios preferenciales para órdenes de gran tamaño, respaldando la suposición sobre la división de la demanda en periodos largos de tiempo; iii)los altos niveles de utilización que son típicos en la industria, limitando la suposición de una amplia capacidad productiva; y iv)la baja probabilidad de, y las consecuencias negativas en tema de salud asociadas con, retrasar el tratamiento médico de un paciente. La cadena de suministro considerada se compone de un productor farmacéutico que ofrece la venta de un medicamento a un pagador de salud quien está a cargo de la disponibilidad de dicho medicamento para la población de pacientes. Se asume que existe heterogeneidad de pacientes de modo que al menos dos grupos de pacientes pueden verse beneficiados al recibir el medicamento, donde se espera que cada grupo obtenga beneficios clínicos diferentes entre sí al consumir el mismo medicamento. Analizamos el problema de optimización con restricciones para el productor, el pagador de salud, o el sistema integrado (según sea el caso en cuestión), utilizando conceptos de teoría de juegos para caractrizar la solución de equilibrio en la toma de decisiones tanto simultáneas como secuenciales.

Contribución Teórica

En su artículo seminal, Arrow (1963)[1] sostiene que la incertidumbre tanto en la incidencia de la enfermedad (*i.e.,* el tamaño de la demanda) como en la eficacia del tratamiento (*i.e.,* el ingreso/beneficio clínico por unidad de tratamiento) genera adaptaciones que limitan el poder descriptivo del modelo tradicional de competencia y sus implicaciones para la eficiencia económica. Tomando esto en cuenta, la disertación contribuye primordialmente a tres vertientes de investigación.

Primeramente, la literatura en economía de la salud se concentra sea en la determinación del nivel de acceso dada la heterogeneidad en las características de los pacientes y la incertidumbre en la eficacia del tratamiento (*e.g.,* Barros, 2011[2]; Zaric, 2008[3]), o en la decisión binaria de incluir un medicamento en la lista de tratamientos reembolsables por un pagador de salud dada la incertidumbre en la demanda (*e.g.,* Zhang et al., 2011[4]). En constraste, la tesis analiza de manera simultánea el problema del nivel de acceso e incertidumbre en la demanda, bajo las características específicas del sector.

Tal situación es similar al problema planteado en administración de operaciones donde el precio de venta y la cantidad de inventario disponible son determinadas de manera simultánea en la presencia de demanda aleatoria y dependiente del precio. La disertación contribuye a

[1]Arrow, K. 1963. Uncertainty and the welfare economics of medical care. *The American Economic Review* 53(5): 941-973.

[2]Barros, P. P. 2011. The simple economics of risk-sharing agreements between the NHS and the pharmaceutical industry. *Health Economics* 20: 461-470.

[3]Zaric, G. S. and B.J. O'Brien. 2005. Analysis of a pharmaceutical risk sharing agreement based on the purchaser's total budget. *Health Economics* 14: 793-803.

[4]Zhang, H., G.S. Zaric, and T. Huang. 2011. Optimal design of a pharmaceutical price-volume agreement under asymmetric information about expected market size. *Production and Operations Management* 20(3): 334-346.

tal línea de investigación (*e.g.,*., Petruzzi et al., 1999[5]; Salinger et al., 2011[6]) al analizar dicha interacción de decisiones según diferentes diseños de contratos entre el productor y el pagador de salud, bajo una combinación de objetivos y restricciones. Adicionalmente, contribuye a los trabajos en coordinación de la cadena de suministro (*e.g.,*., Bernstein et al., 2005[7]; Cachon et al., 2005[8]) al permitir que el "ingreso" por unidad "vendida" , *i.e.,* los beneficios clínicos, sea un valor no determinístico, limitando además el espacio de los posibles "precios de venta" a un subconjunto de valores discretos, siendo estos una función del nivel de acceso seleccionado.

Finalmente, se contribuye a la literatura de agregación de inventarios (*e.g.,* Eppen, 1979)[9] al incorporar la heterogeneidad de pacientes en un sistema de primeras-llegadas primeros-servicios sin posibilidad de reserva, demostrando resultados contrastantes con respecto a las preconcepciones sobre los beneficios generales de la agregación.

Estructura de la Disertación

El Capítulo 1 ofrece una introducción extendida de la problemática. Los Capítulos 2 y 3 se concentran en la decisión simultánea de los niveles de servicio y de acceso, mientras que el Capítulo 4 considera la decisión de acceso como un parámetro predefinido y analiza el diseño estructural y el nivel de esfuerzo óptimos. Comentarios finales aparecen en el Capítulo 5. La Figuras 1 ofrece una imagen instantánea del análisis, junto con las suposiciones correspondientes.

Capítulo 2

El análisis comienza con la modelación del proceso de introducción de un nuevo medicamento que puede ser utilizado por múltiples categorías de pacientes que se benefician del medicamento en diferente grado. Un productor farmacéutico, quien busca maximizar sus beneficios financieros, ofrece vender el nuevo medicamento a un pagador de salud, quien decide los niveles de acceso y de servicio para la población de pacientes de la que está a cargo. Se realiza una comparación analítica suponiendo que el pagador de salud maximiza ya sea el bienestar de los pacientes, o la función total de utilidad (*i.e.,* incorporando los costes de adquisición del medicamento). Bajo ambos criterios de decisión, se incluyen dos restricciones: una restricción de presupuesto absoluto que establece un límite superior en el gasto del sector salud, y una restricción de coste-efectividad orientada a mantener un balance mínimo entre los costes y los beneficios de una intervención. Inicialmente se presenta el análisis para la cadena verticalmente integrada, sirviendo tanto como una referencia de máxima

[5]Petruzzi, N., M. Dada. 1999. Pricing and the newsvendor problem: a review with extensions. *Operations Research.* 47(2): 183-194.

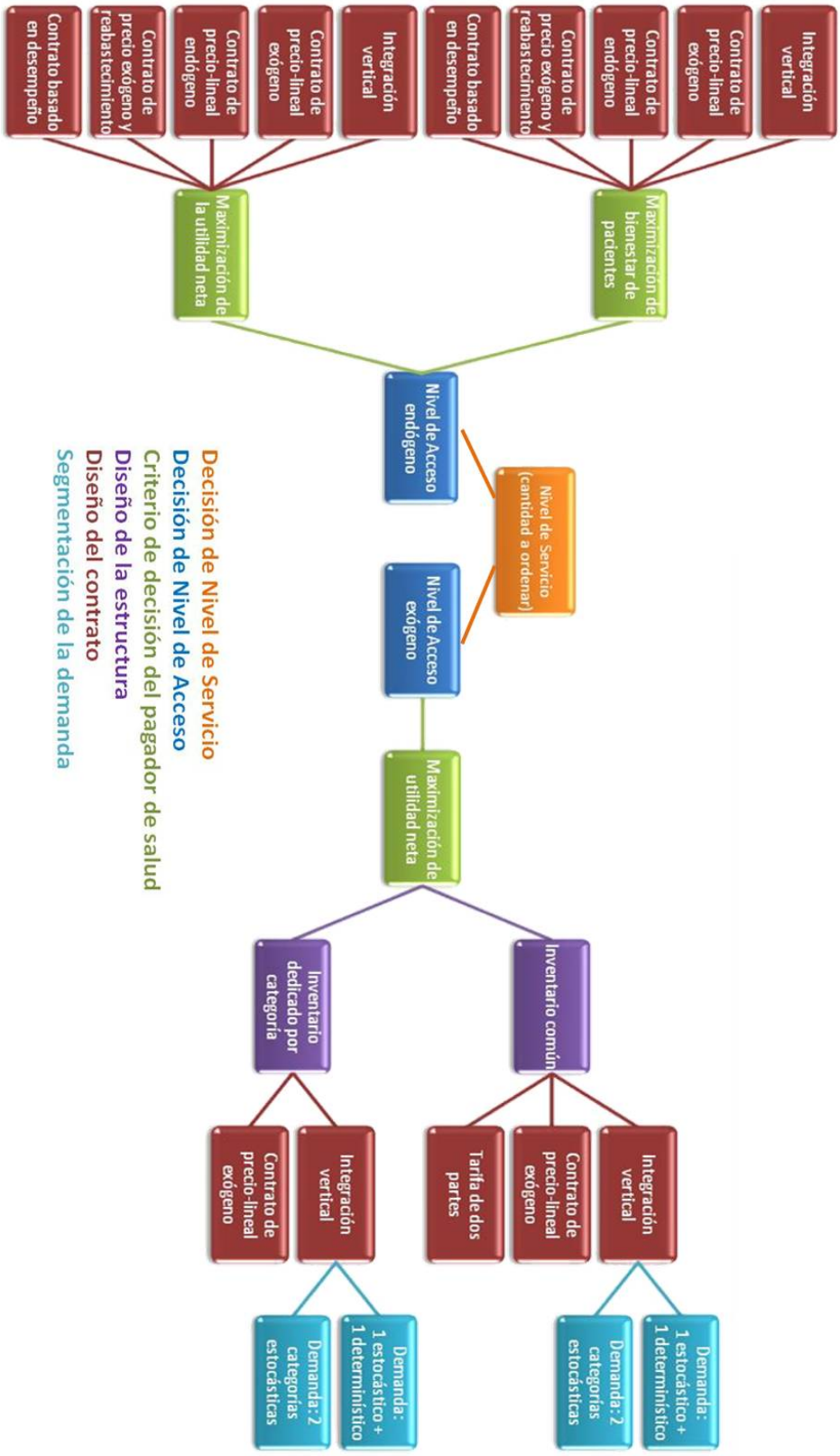
[6]Salinger, M. and M. Ampudia. 2011. Simple economics of the price-setting newsvendor problem. *Management Science.* 57(11): 1996-1998.

[7]Bernstein, F., A. Federgruen. 2005. Decentralized Supply Chains with Competing Retailers Under Demand Uncertainty. 2005. 51(1): 18-29.

[8]Cachon, G. and M. Lariviere. 2005. Supply Chain Coordination with Revenue-Sharing Contracts: Strengths and Limitations. *Management Science* 51(1): 30-44.

[9]Eppen, G. D. 1979. Effects of centralization on expected costs in a multi-location newsboy problem. *Management Science.* 25(5), 498-501.

Figura 1: Instantánea de la Estructura de la Tesis



eficiencia, así como un escenario simplificado para comprender la dinámica entre los niveles de acceso y de servicio bajo las características partículares del problema. Se caracteriza el comportamiento del sistema en función de la interacción de los parametros y se obtienen dos puntos de referencia (uno para cada prioridad del pagador de salud) que permiten una rápida detección de los niveles óptimos de acceso y de servicio considerando dos grupos de pacientes; posteriormente se propone un algoritmo que permite aplicar la misma lógica cuando el número de categorías de pacientes es mayor a dos. Finalmente, se formula y resuelve el contrato de precio-lineal exógeno, estableciendo la base para el análisis de contratos en los cuales el productor tenga la posibilidad de determinar de manera endógena al menos uno de los parámetros.

Capítulo 3

En este capítulo, conservamos la estructura del modelo previamente introducido, y llevamos el análisis a las decisiones hechas por el productor farmacéutico considerando que puede anticipar las decisiones en los niveles de acceso y de servicio del pagador de salud. Específicamente, se analizan tres mecanismos de contratos: contratos de precio-lineal endógeno; contratos de precio exógeno con capacidad de reabastecimiento; y contratos basados en el desempeño. Los contratos de precio-lineal endógeno han sido ampliamente estudiados en la literatura, pero no un en contexto de cadena de suministro comparable con el escenario de interés a la tesis; concretamente, no donde el agente que adquiere el producto tenga el espacio de decisión mencionado en el Capítulo 2. Uno de los resultados más relevantes es que cuando el nivel óptimo de acceso es restringido, el precio de transferencia varía poco en función de la prioridad del pagador de salud, de modo que el productor es capaz de extraer el excedente en la transacción. Sin embargo, si el productor desea inducir niveles de acceso mayores, el precio óptimo de transferencia es menor cuando el pagador de salud maximiza el beneficio neto que cuando maximiza el bienestar de los pacientes. Esta restricción de compatibilidad de incentivos provoca que, irónicamente, el bienestar esperado de los pacientes pueda ser mayor cuando el pagador de salud busca maximizar la utilidad neta que cuando busca maximizar el mismo bienestar de los pacientes. Los otros dos contratos son propuestas nuevas basadas en contratos previamente estudiados, pero adaptadas a las necesidades y limitantes del sistema. El contrato de precio exógeno con capacidad de reabastecimiento es especialmente útil para incrementar el nivel de acceso y la cantidad de medicamentos disponibles para los pacientes. Por su parte, el contrato basado en desempeño es especialmente útil cuando por una parte existe una amplia diferencia entre los beneficios esperados declarados por el productor y el equivalente de certidumbre para el pagador de salud, y por otra parte el productor posee información privada que le hace tener un alto grado de confianza en el desempeño del medicamento. A lo largo del análisis se detectan las virtudes e inconvenientes de cada mecanismo, enfatizando la búsqueda de mecanismos de mejora en el sentido Pareto, de modo que los contratos propuestos no sean simples mecanismos de control de gastos, sino que sean una genuina estructura de riesgos y beneficios compartidos.

Capítulo 4

La última parte de la disertación se distancia del análisis de la decisión del nivel de acceso, y captura una consecuencia diferente de la heterogeneidad de pacientes al comparar analíticamente el desempeño de dos diseños de cadena de suministro. Bajo el primer diseño,

(hasta) dos categorías de pacientes son atendidas por un inventario único bajo un esquema de primeras-llegadas primeros-servicios, mientras en el segundo diseño existe un inventario dedicado a cada categoría de pacientes, sin posibilidad de compartir el inventario entre categorías. Se asume que la existencia de la segunda categoría de pacientes depende de manera estocástica del nivel de esfuerzo de innovación realizado por el productor farmacéutico, que tal productor selecciona el diseño de la cadena de suministro al tener la opción de comercializar dos productos diferentes (*e.g.,* a través de diferentes presentaciones, diferentes formatos de administración, canales de distribución exclusivos), y que el pagador de salud es responsable de decidir el nivel de inventario. Primeramente se analiza la estrategia óptima cuando la cadena está verticalmente integrada. Los resultados fundamentales son la demostración analítica de que la selección del diseño estructural óptimo depende únicamente de la relación entre los beneficios ofrecidos por ambos medicamentos, y que el diseño estructural óptimo requiere un menor nivel de inventario total e induce un mayor nivel de esfuerzo, que el diseño ineficiente. En una segunda parte se detectan los conflictos en los incentivos generados cuando el productor y el pagador de salud actúan de manera independiente. Finalmente se presentan extensiones teóricas y aplicaciones prácticas del modelo en otros contextos.

Director de tesis: Dr. Mustafa Çagri Gürbüz
Cargo: Profesor de Gestión de la Cadena de Suministro, Zaragoza Logistics Center

Agradecimientos

Durante estos años, he tenido la fortuna de cruzar caminos con una amplia variedad de personas, muchas de las cuales han dejado huella en mi desarrollo personal y profesional (una parte de lo cual está reflejado en estas páginas). A todos ustedes, mi gratitud y la esperanza de haber ofrecido algo valioso en reciprocidad. Más allá de esto, deseo dedicar un breve mensaje a un subconjunto de estas personas sin cuyo soporte probablemente no habría iniciado este camino, y seguramente no lo habría disfrutado como lo he hecho.

A mi director de tesis, Çagri, por su paciencia, su atención al detalle, y su preocupación por mi bienestar.

A Santiago Kraiselburd y Prashant Yadav, con quienes las raíces de la tesis fueron inicialmente desarrolladas, por ser mis primeros mentores y por continuar siendo una inspiración.

A Gastón Cedillo, por su consejo, su apoyo incondicional, y por introducirme al mundo académico.

A Mozart Menezes, quien siempre tuvo fe.

A Mario Monsreal, por todas sus lecciones y por creer en mi.

A Florian Schick, por su amistad desinteresada.

A mi familia, por estar siempre presente.

Y a usted, lector, por tomarse el tiempo.

Índice general

Índice de figuras

Analyzing the effects of contract and structural design in health care supply chains

A thesis presented by

Gerardo Pelayo Rubio

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Logistics and Supply Chain Management

in the

MIT-Zaragoza International Logistics Program

at the

Zaragoza Logistics Center, a research institute associated with the

University of Zaragoza

March 2014

[page left intentionally blank]

To my parents: Gerardo and Jessica,

and to my sisters: Jackeline and Fernanda.

Thank you, always.

[page left intentionally blank]

Executive Summary

Introduction

Balancing access to needed medicines against escalating costs is one of the most challenging tasks in health care system design and reform. From 2000 to 2008, the average growth in the per capita spending on pharmaceuticals for Organisation for Economic Co-operation and Development (OECD) countries was almost 60%. The trade-off is particularly present in the introduction of new drugs aimed at treating chronic conditions where list prices proposed by the pharmaceutical manufacturers tend to be high in order to recoup their investment, sometimes contrasting with a lack of robust evidence regarding the cost-effectiveness of the treatment at the time when price is negotiated; moreover, such cost-effectiveness may vary across a drug's different indications, i.e., for different patient groups. As a result, in traditional agreements a health-payer - e.g., National Health Systems, Health Maintenance Organizations, large insurance companies - may be forced either to restrict access or to risk paying high prices that are not ex-post justified due to the uncertainty about the real value of a drug's therapeutic innovation, the lack of solidity of the results presented by the manufacturer, or the replicability of those results in clinical practice. But as pressures to control health care spending keep increasing, health-payers have pushed pharmaceutical manufacturers to decrease prices, potentially decreasing the incentives to invest in innovative treatments, and often resulting in the (temporary or definitive) absence of an agreement between both players at the loss of patient welfare and manufacturer's profits. This has motivated manufacturers - particularly those in the cardiovascular or oncology sectors - to

explore more sophisticated agreements where risks can be more efficiently shared.

Motivated by the above trend, we understand that a health-payer must decide not only whether to accept a new drug under (partial or full) reimbursement for the patient population it serves, but also determine the service level (what will be the volume purchased to satisfy patient demand), access level (which patient groups will be serviced by the health-payer), and reimbursement conditions to the manufacturers (contract parameters). Furthermore we acknowledge that a health-payer may have different priorities affected by the social and industry environment where it operates (e.g., maximizing resource efficiency versus maximizing social welfare), and constraints (e.g., expenditure cap per demand period for some drug/therapeutic indication, and minimum cost-effectiveness threshold). As for pharmaceutical manufacturers, we consider: that price-setting may occur exogenously (through external reference pricing) or endogenously (through direct negotiations with the health-payers); that they may internalize (partially or fully) the risk of holding inventory; and that they are able to segment the market through the creation of distinguishable products targeted at each patient group.

Research Questions

Within the above context where an innovative drug with multiple therapeutic indications looks to enter the market, the research aims at analytically responding to the questions below. Thus, it expects to contribute to a wider understanding of the system's behavior, eventually leading to structural and contract designs in health care supply chains which are better aligned with the players' objectives.

- In a vertically integrated system, how do access and service levels interact as a function of the system's priorities and constraints?

- In a manufacturer-health payer system, what changes as the selling price is exogenously (*vs.* endogenously) set, and the manufacturer is (*vs.* isn't) willing to share some of the risks associated with demand and health outcomes?

- How does the decision of segmenting vs. consolidating the design/distribution channel for a drug with multiple therapeutic indications reflect on the service level and the incentives for innovation effort?

- What is the effect of all the above on: pharmaceutical manufacturer's profits, health payer's expenditures, and patient welfare?

Methodology and key assumptions

The approach followed in this thesis is to mathematically model the described situations based on the newsvendor model framework. This choice is driven by: i)the long lead times (approximately 4 months) for capacity building, sourcing, manufacturing, and delivery of drugs; ii)the industry's common practice to offer preferential pricing for large orders, thus supporting the partition of demand into long periods; iii)the industry's high utilization levels, limiting the ample supply assumption; and iv)the low probability of, and negative health implications associated with, delaying a patient's treatment. The supply chain considered is that of a single pharmaceutical manufacturer that offers to sell a drug to a health-payer who is in charge of making that drug available to the patient population. Patient heterogeneity is assumed so that at least two patient groups could potentially benefit by receiving the drug, where each group is expected to receive different health-benefits by consuming the same drug. We analyze the constrained optimization problem for the manufacturer, health-payer, or the integrated system (depending on the case), making use of game theory concepts to characterize the equilibrium solution under simultaneous and sequential decisions.

Theoretical Contribution

In his seminal paper, Arrow (1963) sustains that the uncertainty both in the incidence of disease (i.e., the size of the demand) and in the efficacy of treatment (i.e., the revenue/health benefit per unit of treatment) generates adaptations that limit the descriptive power of the traditional competitive model and the implications for economic efficiency. Taking this into account, the dissertation contributes mainly to three research streams. First, the health economics literature focuses on determining access level given the heterogeneity in patients' characteristics and the uncertainty in the treatment's efficacy (e.g., Barros, 2011; Zaric, 2008), or on the binary decision to include a drug in a health payer's list of reimbursable treatments given demand uncertainty (e.g., Zhang et al., 2011). In contrast, the thesis simultaneously analyzes the problem of access level and demand uncertainty under the sector's particularities. Such situation is similar to the problem studied in operations management where selling price and stocking quantity are simultaneously determined in the presence of random, price-dependent demand. The thesis further expands the latter line of research (e.g., Petruzzi et al., 1999; Salinger et al., 2011) by analyzing the interaction with different contract designs under a combination of objectives and constraints, and additionally contributes to the supply chain coordination works (e.g., Bernstein et al., 2005; Cachon et al., 2005) by forcing the feasible "selling prices" *i.e., the access level, as will be explained in detail*, to be discrete and allowing the "revenue" per unit "sold" to be a random variable, *i.e., the health benefits*. Finally, we contribute to the inventory pooling literature (e.g., Eppen, 1979) by incorporating patient heterogeneity in a first-come first-serve system with no possibility of reservations, providing contrasting results with popular belief regarding the benefits of aggregation.

Acknowledgements

Over the past five years, I've been fortunate to cross paths with a large number of individuals, many of whom have left a mark in my personal and professional development (and a part of which is reflected on these pages). To all of you, my gratitude and the hope of having had offered something valuable in return. Still, I wish to dedicate a few words to a subset of those individuals without whom I probably wouldn't have started this path, and certainly wouldn't have enjoyed it as much.

To my advisor, Çagri, for his unending patience, his attention to detail, and his sincere concerns for my personal and professional well-being.

To Santiago and Prashant, with whom the roots of the dissertation was originally developed, for being my first mentors and for continuing to be an inspiration to this day.

To Gaston, for his advice, his unconditional support, and for opening my eyes to the world of academia.

To Mozart, who always had faith.

To Mario, for all his lessons and for believing in me.

To Florian, for his selfless friendship.

To my family, who was always present.

And to you, the reader, for taking the time.

[page left intentionally blank]

Author . . . . . .

Gerardo Pelayo Rubio, PhD Candidate

MIT-Zaragoza International Logistics Program

Thesis Supervisor . . . . . .

Dr. M. Çagri Gürbüz

Professor of Supply Chain Management, Zaragoza Logistics Center

Director, Ph.D. Program . . . . . .

Dr. Maria Jesus Saenz

Professor of Supply Chain Management, Zaragoza Logistics Center

[page left intentionally blank]

Contents

[page left intentionally blank]

List of Figures

Chapter 1

Introduction to the Problem

1.1 Motivation

Ever since Arrow (1963), the health care sector has been recognized as having particularities that often require tailored economic models in order to better approximate the dynamics in the system. He sustains that the uncertainty both in the incidence of disease (i.e., the size of the demand) and in the efficacy of treatment (i.e., the health benefit per unit of treatment) generates adaptations that limit the descriptive power of the traditional competitive model and the implications for economic efficiency. From an operations management perspective, Arrow's argument can be interpreted as demand and price uncertainty. While the former is one of the most common assumptions in supply chain analysis, the latter is much less studied. Price is typically assumed to be an exogenously determined parameter, and even when it is included as a decision variable (e.g., the works in supply chain contracts and coordination), such price is deterministically known to the seller at the time of the transaction. That is not necessarily the case in health care since a patient's reaction to a given treatment may be the output of a probabilistic function, *i.e.,* there may not be a guarantee over the effectiveness of a drug (when the outcome is binary), or the extent of the effectiveness (when the outcome is measured over a continuous range). The causes for such randomness include the particular characteristics of the patient, the drug, the disease, and the interactions between them.

However when analyzing health care from a supply chain perspective, demand and outcome uncertainty are far from being the only sources of complexity. While there are a number of dimensions under which the challenges in health care supply chains can be categorized, three such dimensions are proposed next as these are considered particularly relevant from a strategic perspective for the problem that will be introduced below. The first dimension is the *source of uncertainty*, which includes:

- *Supply uncertainty*: the discovery process of new drugs is not an exact science. Neither the timing of new discoveries nor the future number of suppliers available (*i.e.,* the level of competition) can be fully predicted. This poses difficulties for pharmaceutical manufacturers who must incur costly investments on the hope that a large enough fraction of their research and development efforts are successful enough to financially justify their actions. As a result, health-payers need to provide sufficient incentives for the manufacturers, where the patent system has been the main strategy used to protect and reward successful innovations. Patients are dependent on the efficacy of these processes to gain access to new treatments that are more effective, more practical, safer, or more accessible than the existing outside option. The implication for our problem is that manufacturers may have an incentive to segment the market and/or price the drugs sufficiently high in order to recoup their investments and justify the R&D risk.

- *Demand uncertainty*: the number of patients that are candidates for receiving a given treatment, or the extent of the treatment for each patient (*e.g.,* the dosage) may also be unknown to both pharmaceutical manufacturers and health-payers. Such uncertainty will influence the production capacity and inventory stock that will be maintained at the manufacturer level, the inventory ordered by the health-payers, and consequently, the availability of drugs to the end patients. The probability of a small number of patients demanding the drug (possibly via their physicians) may affect a manufacturer's

revenue and a health-payer's costs (depending on the conditions of the agreement between these players), while the probability of a large realized demand can create stock-outs, expediting or higher purchasing costs for the health-payers, and possibly lost sales and negative goodwill for the manufacturers.

- *Outcome uncertainty*: the complexity of determining the value of a therapeutic innovation, the solidity (or the lack of it) in the results provided by the manufacturer, or the doubts about the replicability of results in clinical practice, are all factors leading to an uncertain benefit derived from a drug intervention. In a setting where the product being sold is an individual's health (and the treatment is a means to deliver such health), the source of revenue for the player paying for the medical treatment can be considered a mapping of the patient's realized health outcomes. The argument holds true both if a health-payer charges the patient as a function of the efficacy of the service provided, or if the health-payer internalizes the patient's health outcomes as the revenue in his utility function. Health economists have developed a variety of techniques for translating health outcomes into economic terms, where the two most commonly used are QALY's (Quality Adjusted Life Years) and DALY's (Disability Adjusted Life Years). Each technique weighs the impact of an intervention differently and can result in a different measurement of the health benefits. However, their importance lies on their ability to map health outcomes into financial terms that allow for an economic assessment of a drug's intervention.

A second important dimension that is present in health care supply chains is the *level of heterogeneity* which can be present at different points in the chain:

- *Health-payer heterogeneity*: health-payer's vary widely in their nature, and as a result, in their decision-making process and priorities. National Health Systems, HMO's (Health Maintenance Organizations), and large insurance companies are the typical

health-payers that can play an active role in the negotiations with pharmaceutical manufacturers to determine access levels, service levels, and contract conditions for a given drug. Depending on whether a health-payer is a for-profit or a not-for-profit organization, the driving force of the decisions can be closer to the financial returns or the social benefits, respectively.

- *Patient heterogeneity*: patients can be categorized into mutually exclusive categories depending on their present condition. Such categories can determine the treatment recommended, the dosage of the treatment, and the probability of a successful intervention. For example, heterogeneity may exist for patients who share a common disease but are at different stages of disease progression, or for patients with different diseases who share a common drug treatment but benefit differently from the drug given the economic measure used (*e.g.* the QALY's achieved are different).

- *Product heterogeneity*: a single pharmaceutical compound can often be used for different treatment indications, and the pharmaceutical manufacturer may in some cases be able to create differentiated products for each of these indications (or for subsets of them). Different delivery methods, commercial brands, packaging, or distribution channels are some ways in which the pharmaceutical manufacturer may create differentiated products. The decision to have a single common product versus creating a number of dedicated products is a critical design decision with possible consequences on inventory and access levels.

Finally, the *level of vertical integration* is presented as a last key dimension in the analysis of health care supply chains. Namely, depending on the level of integration, the following factors become crucial:

- *Information asymmetry*: a variety of parameters and probabilistic beliefs about the sources of uncertainty can be either openly shared, privately held, or learned as a

function of a player's actions; even the same knowledge can be interpreted differently by each supply chain member depending on his degree of risk aversion. As motivating examples, the manufacturer may not openly announce its production capacity or may hold an informational advantage about the drug's probabilistic distribution of health outcomes. This can affect the level of trust between the players, and therefore the input that they will use to determine their optimal actions.

- *Contract parameters*: every key decision regarding the interaction between the health-payers and the pharmaceutical manufacturers will depend on the conditions negotiated in their agreement. The transaction price, minimum commitments, and risk ownership, are known in the operations management literature to have an important impact on the system's behavior. Taking advantage of the existing knowledge and adjusting to the particularities of the health care sector, it is expected that supply chain performance and profit allocation may benefit greatly from a contract design that meets the system's goals and needs.

1.2 Structure of the Dissertation

Considering the above, the dissertation focuses on a particular situation within the health care sector: the introduction of new drugs aimed at treating patients with chronic conditions. The initial motivation for focusing on this problem is the continuous rejection of new drug treatments for chronic conditions on the basis of lack of cost-effectiveness evidence along with the growing pressures to cut health-care spending. By definition, following the complete evolution of a chronic disease from the beginning of the treatment requires time, whereas pharmaceutical manufacturers have an incentive to introduce the drug to the market as early as possible, sometimes at the expense of statistically significant data regarding the true value of a drug's intervention. As such, the manufacturer attempts to commercialize

the drug in order to enjoy the benefits of patent protection for a period sufficiently long and at a price sufficiently high to justify the research, development, manufacturing and marketing costs associated with introducing a new drug to the market. Additionally, on top of the direct expenses associated with launching a new drug, the pharmaceutical manufacturer must also consider the cost of failed projects, *i.e.,* those compounds that at some point during the clinical discovery process were considered unsuccessful and had to be divested. On the part of the health-payers, the growing costs of providing health care has increased their awareness of the risks of paying a high price for a drug. The latter could be due to the value being lower than expected. Also, demand could be either larger than expected thus affecting the budget, or lower than expected which carries an associated opportunity cost for other products or services that may have been cut due to budget constraints, and can become an issue when there exists a minimum sales commitment. In short, there is a growing tension between the pharmaceutical manufacturers who want an early introduction of the drug at a sufficiently high price and the health-payers who want to make sure they are not paying an amount beyond the drug's realized value nor beyond their financial capabilities.

To make matters more complicated, the multiple dimensions of health care supply chain complexity previously discussed can all exist in the above setting. As exploring all of them simultaneously is an extremely challenging task, we focus on subsets of those dimensions in each of the following three chapters. Table 1 shows how the different aspects are included in each chapter. In general, they are aimed at understanding some of the structural causes leading to delayed, limited, or inefficient introduction of new drug treatments given uncertainty in the incidence of disease and/or the efficacy of treatment. Using the newsvendor framework, the focus is on the impact of: i)patient heterogeneity, ii)different decision-making priorities and constraints that pharmaceutical manufacturers and health-payers may have, and iii)design of the supply chain structure and contract agreements. Key strategic decisions including capacity and inventory investments, access level, contract parameters, and

innovation efforts are modeled. The results, which hope to orient public-policy making, are provided in terms of firm profitability, health care spending, drug access, and patient welfare. Methodologically, the research contributes firstly to the operations literature by translating familiar concepts such as the price-dependent newsvendor and the pooling effect in a new setting with particular characteristics for which results are not fully explained in previous works; and secondly, to the health economics literature by simultaneously modeling demand and outcome uncertainty. The detailed discussion of the theoretical contributions and conclusions derived from the analysis is provided within each chapter.

Figure 1.1: Structure of the dissertation

		CHAPTER 2	CHAPTER 3	CHAPTER 4
Source of uncertainty	*Demand*	X	X	X
	Supply			X
	Health outcome	X	X	
Level of heterogeneity	*Health-payer's priority*	X	X	
	Patient characteristics	X		X
	Product			X
Level of vertical integration	*Information asymmetry*	X	X	
	Contract design	X	X	X

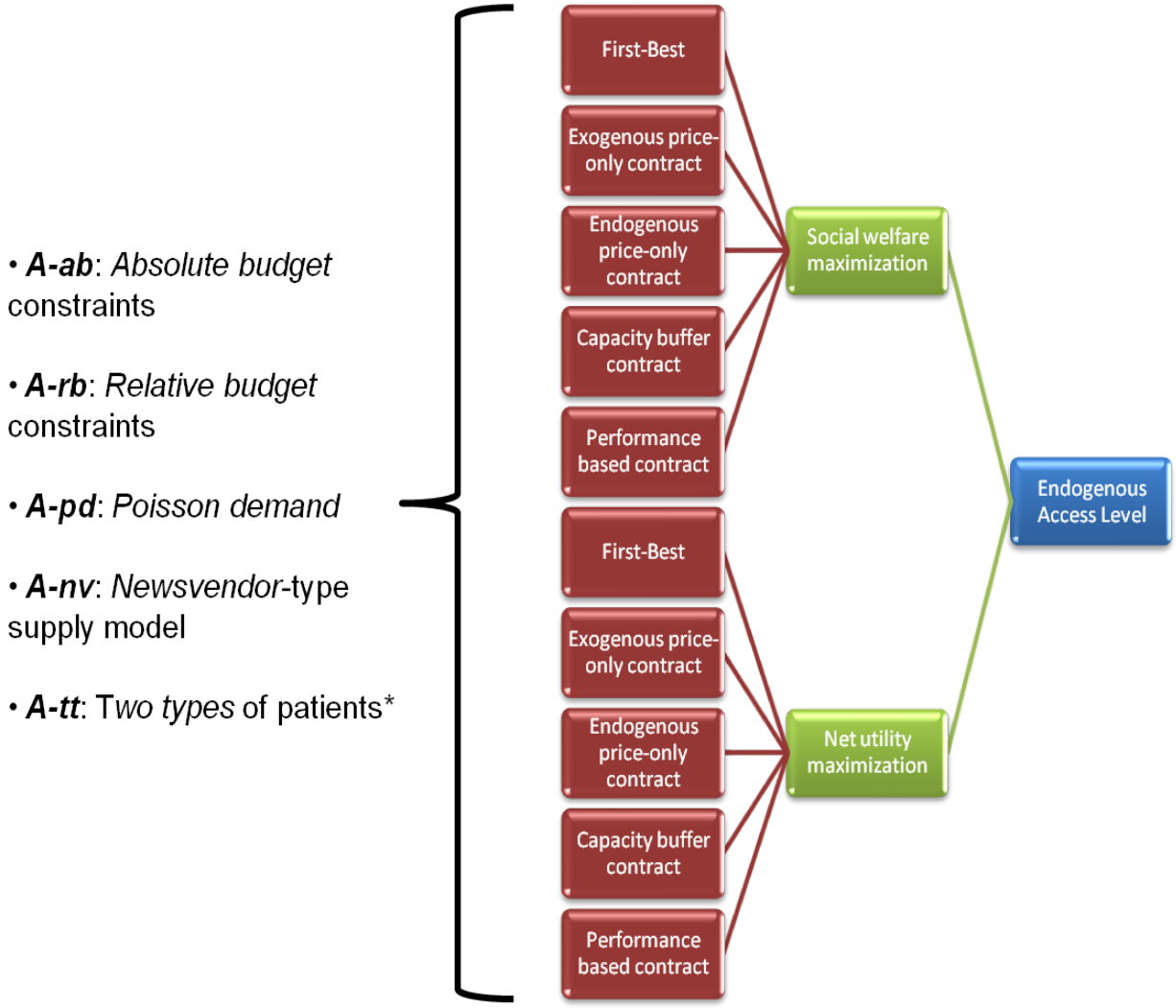
1.3 The Analysis

Chapter 1 provides an extended introduction to the problem. Chapters 2 and 3 focus on the simultaneous access and service level decisions, while Chapter 4 takes the access level as exogenously given and analyzes the optimal structural design and effort decisions. Table 2 and Table 3 provide a snapshot of the analysis, along with the key associated assumptions.

1.3.1 Chapter 2

We begin the analysis by modeling the introduction process of a new drug treatment that can be used by multiple patient categories who benefit differently from it. A profit-maximizing pharmaceutical manufacturer offers to sell the new drug to a health-payer, who decides the access and service levels for the patient population he serves. An analytical comparison is done assuming that the health-payer either maximizes patient welfare, or maximizes the entire utility function (i.e., incorporating purchasing costs). Under both decision-making criteria two constraints are included: an absolute budget constraint to set a limit on health care spending, and a cost-effectiveness constraint to maintain a balance between costs and bene-

Figure 1.2: Snapshot of Chapters 2 and 3



fits. First, the analysis for the vertically integrated chain is presented both as an efficiency benchmark and as a simplified setting for understanding the dynamics between access and service level under the problem's particular characteristics. Second, the exogenous price contract is formulated, setting the grounds for the analysis of contracts where the manufacturer can endogenously determine at least some contract parameters.

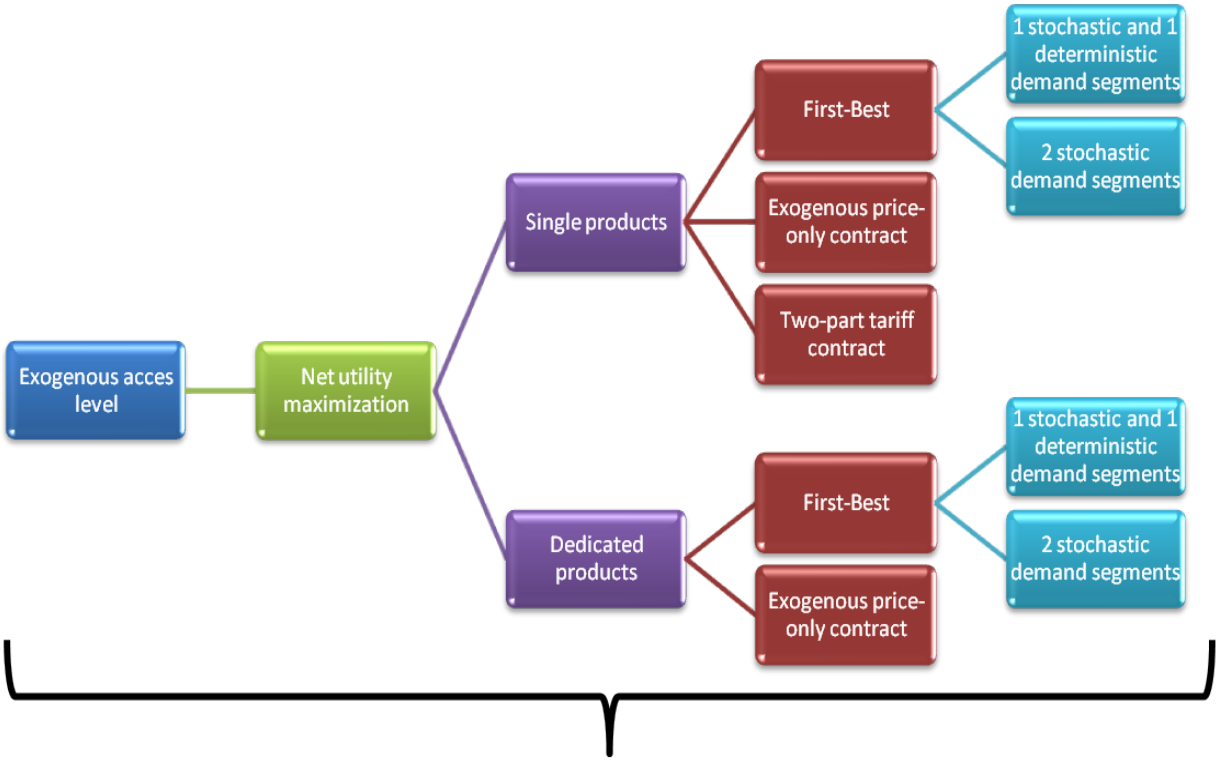
1.3.2 Chapter 3

In this chapter we keep the structure of the model presented earlier but focus on the decisions made by the pharmaceutical manufacturer, given that she can anticipate the health-payer's access and service level decisions. Specifically, we analyze three contracting mechanisms: endogenous price-only contracts; exogenous price contracts with capacity buffer; and performance-based contracts. Endogenous price-only contracts have been thoroughly studied, but not in a supply chain setting as the one we consider where the downstream player has such decision space (as mentioned above for Chapter 2). The last two contracts are novel proposals based on existing models, but adapted to the needs of the system. The virtues and drawbacks of each mechanism are detected, with an emphasis on the search for Pareto improvements.

1.3.3 Chapter 4

The last part of the dissertation departs from the analysis of the access level decision and captures a different consequence of patient heterogeneity by analytically comparing the performance of two supply chain designs. Under the first design, (up to) two patient categories are served by a single inventory stock on a first-come first-serve basis, while on the second design a dedicated inventory stock is used to serve each patient category. It is assumed that the realization of the second category is stochastically contingent on innovation efforts made by the pharmaceutical manufacturer, that such manufacturer chooses the supply chain

Figure 1.3: Snapshot of Chapter 4



design by having the option to commercialize two differentiated products (*e.g.,* through different presentations, different delivery formats, exclusive distribution channels), and that the health-payer is responsible for making the inventory decision. First, the optimal decision path for a vertically integrated chain is analyzed, and then the incentive misalignments derived from vertical separation are explained along with some theoretical and managerial extensions to the model.

Chapter 2

Analyzing the joint access and coverage problem in health care

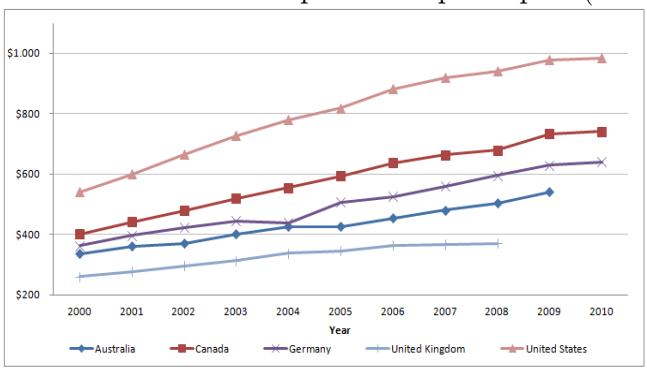
2.1 Introduction

A report in the United Kingdom (UK) by the Rarer Cancers Foundation shows that since 2009, 18 new treatments were rejected by the National Institute of Clinical Excellence (NICE) - many because they were not deemed "cost-effective" - out of 34 put forward.[1] In Australia, the federal government announced on February 2011 that to return the budget to a surplus, no new drugs would be added to the Pharmaceutical Benefits Scheme (PBS) until 2013, irrespective of the recommendations of the Pharmaceutical Benefits Advisory Committee (PBAC).[2] Balancing access to needed medicines against escalating costs is one of the most challenging tasks in health care reform (Chalkidou, Lopert and Gerber, 2012). From 2000 to 2008, the average growth in the per capita spending on pharmaceuticals for Organisation for Economic Co-operation and Development (OECD) countries was almost 60% (and exceeding 70% by 2010, considering the subset of countries for which information is available); Figures

[1] http://www.dailymail.co.uk/health/article-2194998/Number-cancer-drugs-rejected-health-watchdog-rises-50-years.html

[2] Following criticism by pharmaceutical industry, providers, and patients, the measure was removed on September the same year.

Figure 2.1: Pharmaceutical expenditure per capita (2000 - 2010)



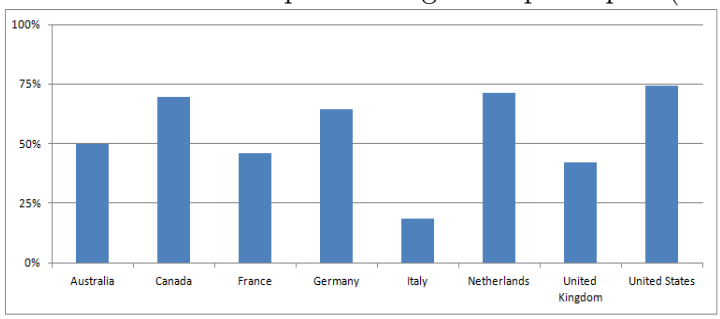
Note: Expenditures are expressed in U.S. dollar purchasing power parity.

Source: 2012 OECD Health Data, http://www.oecd.org/health/healthpoliciesanddata/oecdhealthdata2012-frequentlyrequesteddata.htm

2.1 and 2.2 provide more specific information about some of the largest/most influential markets. The trade-off is particularly present in the introduction of new drugs aimed at treating chronic conditions where list prices proposed by the pharmaceutical manufacturers tend to be high in order to recoup their investment, sometimes contrasting with a lack of robust evidence regarding the cost-effectiveness of the treatment at the time when price is negotiated; moreover, such cost-effectiveness may vary across a drug's different indications, *i.e.,* for different patient groups. As a result, in traditional agreements a health-payer - *e.g.,* National Health Systems, Health Maintenance Organizations, large insurance companies - may be forced either to restrict access or to risk paying high prices that are not ex-post justified due to the uncertainty about the real value of a drug's therapeutic innovation, the lack of solidity of the results presented by the manufacturer, or the replicability of those results in clinical practice. But as pressures to control health care spending keep increasing, health-payers have pushed pharmaceutical manufacturers to decrease prices, potentially decreasing the incentives to invest in innovative treatments, and often resulting in the (temporary or definitive) absence of an agreement between both players at the loss of patient welfare and manufacturer's profits.

The United Kingdom's Department of Health has been one of the most innovative players regarding the relationship between the pharmaceutical manufacturers and health-payers.

Figure 2.2: Pharmaceutical expenditure growth per capita (2000 - 2008)



Source: 2012 OECD Health Data, http://www.oecd.org/health/healthpoliciesanddata/oecdhealthdata2012-frequentlyrequesteddata.htm

The Pharmaceutical Price and Regulation Schemes (PPRS) - a non-contractual agreement renegotiated every five years between the UK Department of Health and the Association of the British Pharmaceutical Industry (ABPI) - were first introduced in 1957 as a mechanism to ensure access to good quality branded medicines at reasonable prices to the National Health Service (NHS) and fair returns to the pharmaceutical industry, where the price is regulated by setting profit caps for pharmaceuticals. In 1999, NICE was established, playing a key advisory role in technology appraisal by quantifying benefits in a consistent and comparable way across the full range of health-related conditions and applying an Incremental Cost-effectiveness Ratio (ICER) methodology; it is important to note that NICE's recommendation is by law a sufficient but not necessary condition for inclusion of the drug in the NHS list. In the existing system based on the PPRS 2009, a standard willingness-to-pay (WTP) threshold is applied to all new products so that a drug will be recommended for inclusion in the NHS if the cost per QALY[3] achieved (*i.e.,* the incremental cost-effectiveness ratio) is less than £20,000, and analyzed on a case by case basis when the ICER is between £20,000 and £30,000 (*i.e.,* the WTP threshold). However, the recurrent rejection of new treatments, often on the grounds of lack of cost-effectiveness evidence, has motivated pharmaceutical manufacturers - particularly in the cardiovascular and oncology sectors - to explore more sophisticated agreements where risks can be more efficiently shared. Pressured by public and industry lobbying, and following the recommendations from the market

[3]NICE has used the Quality Adjusted Life Years (QALYs) - a QALY is the amount of health represented by a year of life at full health - to measure the benefits of an intervention.

study performed by the Office of Fair Trading (OFT, 2007), the NHS is expected to make a transition towards a value-based system of pricing medicines. Two temporary solutions are currently in practice: the Cancer Drug Fund which provides £200 million per year to fund cancer drug treatments not recommended by NICE that physicians deem appropriate for a particular patient; and Patient Access Schemes where manufacturer and payer agree on an evaluation time, a verifiable measurement, and a target, so that the manufacturer offers a discount or rebate to the NHS when the target is not reached. However, both programs are expected to be substituted in 2014 by value based pricing (VBP) which no longer considers a unique threshold across all interventions, but rather has a base threshold which is explicitly increased for high burden of illness, therapeutic innovation, and wider societal benefits, *i.e.,* the threshold may be different for different drugs and/or indications. The approach provides manufacturers with freedom to propose prices as long as the threshold is satisfied, and expects to increase transparency in the technology appraisal process and predictability for pharmaceutical manufacturers.

Other countries are following the trend of implementing mechanisms for using comparative clinical and cost-effectiveness to inform technology adoption and regulate selling prices. In Australia, the PBAC is responsible for recommending inclusion of drugs in the national formulary for reimbursement, using a variable cost-effectiveness threshold contingent on each drug's characteristics. As opposed to the UK, in Australia the PBAC recommendation is a necessary but not sufficient condition for the Minister for Health and Ageing to approve a drug's inclusion. Risk sharing contracts have included the use of rebates paid to the government when expenditures exceed an annual cap, a pooled annual sales cap for a group of drugs that treat a common condition, and price-volume agreements. The Life Saving Drugs Program operating outside the PBS has been set up to provide free access to certain expensive, life-saving drugs for rare, serious, life-threatening conditions, currently funding 8 drugs for almost 200 patients. Germany uses reference pricing groups to regulate prices

of drugs. Following the establishment in 2004 of the Institute for Quality and Efficiency in Health Care (IQWiG) - modeled after NICE -, since January 2011 reference pricing is only used for drugs that do not demonstrate additional benefits; for those that do, an agreement between the manufacturer and the national association of statutory health funds (SHI) must be reached within 6 months or else a central board of arbitration determines a rebate based on international prices. In the United States, the establishment of the Patient-Centered Outcomes Research Institute recognizes the relevance of evidence-based decision-making, and Medicare's Independent Payment Advisory Board is expected to motivate the use of price-negotiations and risk-sharing agreements. Some detailed, recent surveys on the use of risk sharing contracts across different countries include Pugatch, Healy and Chu (2010), and Espin, Rovira and Garcia (2011). However, as such papers comment, many of the so-called risk-sharing agreements seem to act as cost-containment mechanisms rather than a true re-assignment of risks and benefits to both players.

Motivated by the above trend, we understand that a health-payer must decide not only whether to accept a new drug under (partial or full) reimbursement for the patient population it serves, but also determine the volume purchased (how many patients are expected to be treated), access level (which patient groups will be serviced by the health-payer), and reimbursement conditions to the manufacturers (contract parameters). In our model, we explicitly acknowledge that each health-payer may have different priorities affected by the social and industry environment where it operates, and that the manufacturer may hold an information advantage about a drug's expected future value. Moreover, a health-payer's decision may be limited by absolute and relative expenditure restrictions. To exemplify the former, a prostate cancer drug that privately costs around £3,000 for a month's supply had been offered to the NHS at a discount, but NICE declared the number of men who need the drug would make it financially unworkable.[4] As for the latter, the ICER methodology which

[4]Retrieved Oct. 8, 2012 from http://www.savistamagazine.com/news/prostate-cancer-drug-provisionally-rejected-by-nhs

lies at the heart of any decision incorporating comparative clinical and cost-effectiveness, requires an upper bound or threshold that essentially makes the maximum allowable level of expenditures a function of the benefits.

As a result, the scope of this chapter is to analyze: a) the change in the system's optimal decisions as a function of the health-payer's decision-making priority, constraints, and the contract parameters using the newsvendor framework; b) the way in which double marginalization and asymmetric beliefs between the pharmaceutical manufacturer and the health-payer influence the mechanics; and c) the impact of all the latter on manufacturer's profits, health-payer's costs, and patient access and service levels. Our contribution can be summarized in three parts. First, we derive an efficient algorithm for determining the optimal access and service level policy, based on two very easy to calculate thresholds, in a setting based on the price and quantity newsvendor model but where the feasible prices are not continuous[5], the decision space is constrained, and the decision maker's priorities may vary. And second, we comment extensively on the situations - as a function of the combination of parameters - that are likely to induce full or restricted access, providing interesting insights for policy makers and pharmaceutical manufacturers.

The rest of the chapter is introduced with a literature review of relevant work. Section 2.3 introduces the model and solves the problem for a single decision maker under two types of patients. Section 2.4 expands the results to n number of patient types, and derives an algorithm to efficiently find the optimal solution without the need of full enumeration. When the pharmaceutical manufacturer and the health-payer act separately, section 2.5 expands the results to incorporate exogenous price-only contracts, setting the base model for the next Chapter. Conclusions and further research opportunities are discussed in §2.6.

[5]In our model, the additive part of the objective function is given by the health benefits obtained by those patients who receive the drug under analysis. As will be thoroughly explained in §2.3, there is a direct mapping between the access level chosen, and the average expected health benefits received by the patients, *i.e.,* the retail price in typical newsvendor models.

2.2 Literature Review

Ever since Arrow's (1963) seminal paper, the medical-care industry has been recognized as having particularities that set it apart from normal economic competitive models. Arrow (1963) sustains that the uncertainty in the incidence of disease (*i.e.,* the size of the demand) and in the efficacy of treatment (*i.e.,* the marginal revenue/benefits) causes adaptations that limit the descriptive power of the normal competitive model and its implications for economic efficiency. Acknowledging for these two sources of uncertainty, the chapter contributes to essentially two streams of literature: the health economics literature, by jointly analyzing the effects on manufacturers and payers of these two sources of uncertainty; and the price and quantity problem in the operations literature, by considering a discrete demand distribution and capturing the effect of different decision making criteria and constraints by the downstream player.

Health Economics

The main interest of the health-economics literature has not been directed at simultaneously considering randomness in the the size of patient demand and in the treatment outcome per patient. With regards to the former, the literature has focused on setting budget constraints on the payer or profit caps on the manufacturer, to analyze either the effect of manufacturer's private information or the manufacturer's reaction when she is able to influence the size of demand through detailing effort, *i.e.,* visits from pharmaceutical representatives to physicians. However these papers tend to ignore the costs associated with supply and demand mismatches. Zaric and O'Brien (2005) propose a model for financial risk sharing under demand uncertainty based on the payer's total budget in order to protect the payer from larger than expected demand that may increase costs dramatically. They let the manufacturer submit a budget impact analysis to achieve drug's approval and maximize

profits, assuming that the likelihood of the drug being accepted by the payer is decreasing in the budget statement. While the model proposed here is more limited in that the manufacturer is assumed to be able to observe the payer's maximum allowable budget, a layer of uncertainty is added by including randomness in the drug outcome. Also the present approach increases complexity in the decision-making interactions by modeling payers with multiple priorities and allowing the manufacturer to set a profit maximizing price in anticipation of health's actions. A closer paper is Zhang, Zaric and Huang (2011), who consider a setting where the market size is uncertain and the manufacturer may or may not have an informational advantage over the payer about the true demand's distribution. Similar to us, they include a cost-effectiveness constraint in order for trade to occur. They then let the cost-minimizing payer propose a price-volume agreement consisting of a unit price and a rebate rate in excess of the manufacturer's statement of projected future demand, and characterize the optimal contract parameters and the manufacturer's incentive compatible reaction. They find that there always exists a contract inducing the manufacturer to reveal his private information. We distinguish ourselves from this model in few but relevant ways. First in terms of the key modeling assumptions, we do not consider the case of asymmetric demand information; in our model it is the manufacturer, rather than the payer, who proposes the contract; and we consider payer heterogeneity by allowing his objective function to be maximizing either patient welfare, or the entire utility function (patient welfare minus purchasing costs). Second in terms of the design of the contract, the budget constraint in our model is fixed, while they use it essentially as a dynamic negotiation lever to extract information from the manufacturer and maintain cost-effectiveness. Furthermore, in their model the price is adjusted based on realized demand, while in our model the access level - which indirectly determines the "revenue" component in the objective function as will be explained in section 3 -, adjusts the expected demand, but the contract parameters will not change as a function of demand's realization.

With regards to access decisions due to patient heterogeneity, the drive to achieve treatment equality has raised interest in understanding the determinants of drug access for different medical indications, or even from patient subgroups within a drug's same indication. Zaric (2008) and Hawkins and Scott (2011) analyze the trade-off between increased access and reduced price under patient heterogeneity with respect to their response to a treatment. The former uses a Markov model of disease progression to find the optimal price and limited use conditions, both set by the manufacturer when seeking to achieve formulary listing. Hawkins and Scott (2008) compare a health-payer's reimbursement based on three different criteria: whole-population cost-effectiveness, stratified cost-effectiveness, and negotiated price and coverage. They allow the manufacturer to set the price of the drug to maximize revenue and through an example show how access, manufacturer's revenues and incremental net health benefits may all be increased through negotiation compared to stratified cost-effectiveness. Their whole-population and negotiation rationales are similar to the approach we use to determine cost-effectiveness since for any subset of patient groups, we consider the average incremental health benefits rather than doing the assessment independently for each patient group as stratified cost-effectiveness suggests. Furthermore, by allowing the health-payer in our model to be either a net utility or a social welfare maximizer, we indirectly capture the payer's willingness to subsidize the patient groups which would be rejected under stratified cost-effectiveness with the surplus achieved by the patient groups with higher incremental health benefits. Our analysis advances their work by incorporating demand and health outcome uncertainty into the model, in addition to including a budget constraint for the health-payer.

In the modeling of outcome uncertainty, So and Tang (2000) is worth mentioning as one of the earliest papers to recognize the need to study the joint impact of cost containment initiatives in health care on all relevant parties, taking a supply chain approach under the presence of uncertainty in a health care system. They model an outcome-oriented reim-

bursement policy aimed at discouraging excessive prescription behavior to determine the optimal prescription policy of a clinic when the effect of the drug on the patient is uncertain. Similar to our goal, they analyze the impact of the different parameters on the patient's well-being, the clinic's profitability, and the pharmaceutical firm's profitability. The main difference is that their analysis focuses on the dynamically made dosage quantity decision at the lowest echelon for all incoming patients, i.e., consumption is endogenously determined based on the information set available at the time of each patient's visit. On the contrary, our model considers a single-period and does not incorporate learning to determine the optimal consumption, but rather assumes the size of patients treated to be a random variable due to uncertainty in the size of the patients' pool. Further, rather than limiting the dose per patient, we allow the drug's use to be irreversibly restricted to only some categories of patients before the observation of patient arrivals, thus focusing jointly on the access and service level decisions.

Operations Management

The problem of determining access and service levels in our health care model has large similarities with the price and quantity problem studied in the operations literature. In our model, the access level decision shapes the distribution of patient demand, which is the key challenge in the price and quantity newsvendor model. This is because once the access level is endogenously determined in our model, an expected health benefit per patient treated can be obtained, i.e., expected health-benefits are stochastically decreasing as the access level becomes more inclusive. Since health-benefits are the source of "revenue" in our model, we can say that setting the optimal access level is equivalent to indirectly setting the expected health-benefit per patient. Despite the managerial attractiveness of this problem, the analysis of the price and quantity problem has received limited attention in the literature until

recent years, in part because of the complications to obtain general closed-form solutions. Some papers have considered additive demand models (e.g., Mills, 1959) where the demand is the sum of a deterministic downward-slope function of price - as in the classic marketing literature - and a second term incorporating the size randomness. However, we consider a multiplicative demand model where demand is given by a random number of incoming patients, of which only a fraction are expected to satisfy the endogenously set access level policy; for such reason we will focus only on multiplicative demand models. Karlin and Carr (1962) was the first work to consider the same demand components as Mills (1959), but assuming demand to be the product of those two terms, and find - as opposed to Mills (1959) -, that the optimal selling price is not lower than the riskless price, defined as the optimal price under no randomness in the demand function. Subsequent papers have extended the results within particular settings. Zabel (1970) shows the uniqueness of the stocking quantity solution under uniform and exponential distributions of demand when the penalty cost is zero. Nevins (1966) uses simulation to reach a similar conclusion when demand is normally distributed. Young (1978) explains how for the multiplicative demand case, the variance of demand is a decreasing function of price, while the coefficient of variation is independent of price.

More recently, a series of papers have tried to provide a more unifying framework. Petruzzi and Dada (1999) try to consolidate the previous results for additive and multiplicative forms of demand, arguing that pricing provides an opportunity to reduce the risks of overstocking and understocking, so that for the multiplicative case under isoelastic demand, it is possible to decrease demand variance without adversely affecting the coefficient of variation by choosing a higher price, i.e., in our model, this implies setting a more restrictive access policy. Also useful is their definition of a base price instead of a riskless price, where the former acknowledges that expected sales differ from expected demand, and define the optimal pricing strategy as the base price plus a premium; the intuition of this per-unit

premium is to recover on a per-sale basis, the total expected cost derived from the inventory that is used as a buffer against uncertainty in demand. Salinger and Ampudia (2011) explain how the price-setting newsvendor fits the Lerner relationship, linking profit-maximizing price with marginal costs and the elasticity of demand. They use the marginal cost of an expected unit sold instead of marginal cost, and elasticity of the average quantity sold with respect to price instead of the elasticity of demand with respect to price. They prove that in the multiplicative uncertainty case, the elasticity of the average quantity sold is constant for a mean-preserving spread in demand, therefore the mark-up factor is unaffected while the marginal cost of an expected unit sold increases due to a greater fraction of the marginal increase in production being unsold. These combined effects result in an increase in the optimal price as demand increases. Kocabiyikoglu and Popescu (2011) characterize more general models of stochastic demand, finding a series of necessary and sufficient conditions that guarantee uniqueness of the joint price-inventory solution. They do this by defining a *lost sales rate (LSR) elasticity*, described as the percentage change in the rate of lost sales with respect to the percentage change in price for a given quantity. They find that an increasing LSR elasticity is a sufficient condition for uniqueness of the optimal solution when the selling price observed by consumers and the order quantity decisions are coordinated. In terms of our model, the result which is shown in Corollary 2b of their paper, indicates that for the multiplicative demand (*i.e.,* our case), the LSR elasticity is increasing in the order quantity if and only if the distribution of the random variable that is independent of the health benefits is IGFR (Increased Generalized Failure Rate); and the LSR elasticity is decreasing in the access level (*i.e.,* increasing in the expected health benefits) if the access level dependent function of demand is elastic with respect to the access level. Since our model satisfies those conditions, we use their result as a starting point to incorporate particular characteristics of our problem.

There are four main distinctions between our model and the works discussed above. First,

we take a supply chain approach rather than focusing on a single echelon, by explicitly considering the role (and in Chapter 3, the best response function) of the manufacturer. Second, the "expected price", *i.e.,* the health benefit, is assumed to be discrete in our model, which complicates the analysis and requires a different treatment. Third, in addition to expanding the analysis for the traditional objective function in the price dependent newsvendor model, we consider a second alternative (maximizing social welfare, which will be formally defined below) motivated by the mission of some of the institutions, usually public, in charge of finding the optimal values for the model's decision variables. And fourth, we incorporate an absolute and a relative budget contraints, simultaneously, which has relevant implications on the feasible decision space, and therefore, on the solution process to find the optimal solution. In sum, our approach intends to aid public policy making and firm strategy by analytically showing the changing supply chain impact of different decision-making criteria by the payer under two fundamental constraints driven by the growing pressures surrounding health care costs, and in particular, pharmaceutical spending.

2.3 The Model

2.3.1 General Setup

Consider a health care supply chain where a risk-neutral pharmaceutical manufacturer, hereafter *Pharma* and denoted by subindex m, offers to sell a new (prescription) drug to a risk-neutral[6] central health care system (i.e., the health-payer and the health-provider are part of the same governing institution), hereafter *Health* and denoted by subindex h, through some take-it or leave-it contract agreement. Let N be a random variable representing the number

[6]This assumption, while inconsistent with the current observations regarding the introduction of new drugs aimed at treating chronic conditions will be relaxed in Chapter 3. As will be explained there, its relaxation under price-only contracts provides no additional insights and is therefore removed in the present chapter to avoid unnecessary noise.

of patients that arrive to receive treatment through Health within a single finite time period, where N follows a Poisson distribution with parameter λ. We assume heterogeneity within patients to exist so that upon arrival to Health, patients are categorized into 2 mutually exclusive types based on their physiological conditions and medical history.[7] Let $f(i)$ be the probability that an incoming patient is categorized as type i, $i = 1, 2$, and $F(i)$ be the probability that a patient is categorized as type i or lower, *e.g.,* $F(2) = f(1) + f(2)$, with $F(0) = 0$, $F(2) = 1$. Henceforth we assume $f(1) = \theta$, and $f(2) = (1 - \theta)$. Let b_i represent the multiplication of the incremental value of the health gains received by a patient categorized as type i, $i = 1, 2,$, and the health-payer's ceiling ratio; i.e., b_i is the incremental value derived from the drug's administration *times* the health-payer's maximum willingness-to-pay for that value. For simplicity of exposition, we will henceforth refer to b_i simply as *health benefits*. Without loss of generality, we organize patient categories such that $b_1 > b_2$. This implies that the perceived value to the health-payer from a drug being administered to a patient of category 1 versus to a patient of category 2, is $b_1 - b_2 > 0$; this can occur based on *(a)*superior clinical outcomes in patients of category 1; *(b)*on a higher ceiling ratio for category 1 originated by the drug's larger societal benefits, the disease severity, or the degree of innovation for that particular patient category; or *(c)*on a combination of both. The formulation therefore allows for a dynamic cost-effectiveness threshold, *e.g.,* it allows for the existence of different willingness-to-pay thresholds for each patient category, as is already the case in Australia and will be in the UK starting 2014.

We approximate the trading opportunities between Pharma and Health using the single period newsvendor framework where Pharma must commit production well in advance of receiving Health's order. This is based on *(a)*the long lead times in building production capacity, sourcing raw materials, manufacturing the drug and delivering it to Health; *(b)*the

[7]We assume that all incoming patients are diagnosed in order to learn their medical status, and as a result *even* in a fully inclusive policy, the examination cost is a constant which is not explicitly included in the model.

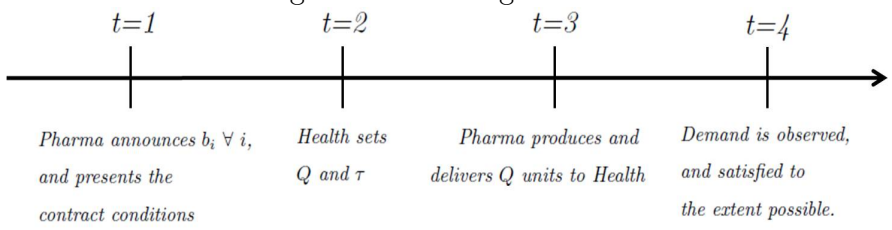
quantity discounts offered for large purchases which makes the partition of demand into long periods a reasonable assumption; and *(c)*the manufacturer's high utilization levels, given limited manufacturing capacity and possibly multiple clients, which reduces her ability to satisfy larger than expected demand in the short term.[8] Additionally, we consider that the newsvendor framework can be useful in representing the collateral costs associated with demand forecast mismatches exacerbated by the growing budget pressures in the health care sector. On one hand, overestimating demand may result in a portion of the health care budget being "trapped" in anticipation for more patients, possibly rejecting or delaying inclusion of other treatments into the formulary listings. On the other hand, underestimating demand may result in higher than expected costs, either in terms of health outcomes because patients are not treated in a timely manner with the best available treatment, or in monetary terms because the treatment is available but creates a budget deficit that negatively affects future introduction of innovative treatments.

The order of events in the model is depicted in Figure 2.3 and described next. (1)Pharma announces b_i for $i = 1, 2$, and announces the selling price for the drug, which may be either exogenously or endogenously determined as will be explained in the coming sections. (2)Health selects an order quantity Q of drugs and a prescription policy threshold τ so that an incoming patient of type i is prescribed the drug only if $i \leq \tau$; this means that the probability that an incoming patient is an eligible candidate for receiving the drug is $F(\tau)$. Mathematically, the conditional demand for the drug given some prescription policy threshold can be seen as a random number N of trials, each with a success probability $F(\tau)$; by the Poisson property, the *effective demand*, $D(\lambda, \tau)$, is also Poisson distributed with parameter $\lambda F(\tau)$. Let $p(x; \lambda F(\tau)) = \frac{(\lambda F(\tau))^x}{x!} e^{-(\lambda F(\tau))}$, be the probability that exactly x patient arrivals from the effective demand occur during the period; and let $P(x; \lambda F(\tau)) = \sum_{j=x}^{\infty} p(j; \lambda F(\tau))$ be the complement of the Poisson CDF. We can now define $B(\tau) \triangleq \sum_{i=1}^{\tau} \frac{b_i f(i)}{F(\tau)}$ to be the

[8]This assumption is relaxed in Chapter 3 through the introduction of a capacity buffer contract.

expected *health benefits* obtained by a patient belonging to the effective demand that is eligible to receive the drug treatment. (3)Pharma produces and delivers to Health Q units of the drug at marginal cost c. (4)Demand is realized. Excess drugs may be salvaged at a per unit value δ, which may be interpreted either as the opportunity cost or as a discounted sale to a secondary market; to avoid trivial problems, assume $\delta < c < b_1$. If $D(\lambda, F(\tau)) > Q$, a per unit cost, g, is accrued to Health for each patient arrival that satisfies the prescription policy threshold but does not receive the drug treatment due to a stock-out. To keep integrality, we will use $\lfloor x \rfloor$ and $\lceil x \rceil$ as the floor and ceiling functions, respectively. For the moment, all players are assumed to hold symmetric information about all functional forms and parameters.

Figure 2.3: Timing of Events



2.3.2 The decision making process

Define $A(Q,\tau) \triangleq \mathbb{E}\left[\min[Q\,,\, D(\lambda,\tau)]\right]$ to be the expected quantity of *administered* drug treatments; $\mathbb{E}[\max[0\,,\, Q - D(\lambda,\tau)]] = (Q - A(Q,\tau))$, to be the expected leftovers for Health; and $\mathbb{E}\left[\max[0\,,\, D(\lambda,\tau) - Q]\right] = \lambda F(\tau) - A(Q,\tau)$, to be the expected quantity of understocked units of the drug at the end of the period. Finally, let $T(\cdot)$ be the payment from Health to Pharma as a function of the contract parameters; it is well-known that when Pharma and Health act as a single decision-maker, it is optimal to set a transfer payment $T = cQ$ to prevent double-marginalization. Then, for a risk-neutral single decision-maker the social

welfare expected utility function is:

$$S(Q,\tau) \;=\; B(\tau)A(Q,\tau) + \delta(Q - A(Q,\tau)) - g\left(\lambda F(\tau) - A(Q,\tau)\right), \tag{2.3.1}$$

and the system's expected utility function is:

$$\begin{aligned} Z(Q,\tau) &= -cQ + B(\tau)A(Q,\tau) + \delta\left(Q - A(Q,\tau)\right) - g\left(\lambda F(\tau) - A(Q,\tau)\right) \\ &= -cQ + S(Q,\tau) \end{aligned} \tag{2.3.2}$$

Based on the latter, in order to better understand the impact of the different priorities used in practice by health care systems, we explicitly distinguish between two possible criteria for the decision-making process:

- *Maximize utility of social welfare,* $S(Q,\tau)$; this is the case when Health's priority are the recipients of the drug treatments, *i.e.,* the patients, and we consider it a more appropriate approach in settings where the health-payer is a non-for-profit institution as occurs in several national health systems. For example, in Germany a 2005 Court Decree establishes that treatment in the case of a life-threatening disease is an essential part of health care and statutory health funds must pay for it.

- *Maximize the decision maker's utility function,* $Z(Q,\tau)$; this is the case when the decision maker's priority is to make an efficient use of its resources and we consider it a more appropriate approach in settings where the health-payer's priority is to maximize the use of its resources or when the payer is a for-profit institution, *e.g.,* private insurance companies, as occurs in a large portion of the the United States market and some developing countries. For example, in Australia the PBAC recommendation may be accompanied by closely specified access restrictions due to a drug's lack of robust clinical evidence of a clinically important additional benefit, or because the incremental costs of obtaining those benefits mean that drugs are cost-effective in only a defined

group of patients. In other words, a drug's high cost-effectiveness for one of the indications cannot be considered to subsidize its use for an indication which benefits are on its own not cost-effective.

In addition, the reality is that for both decision-making criteria, the amount of resources is limited and the health-payer's decision space is typically bounded by a set of minimum conditions that should be satisfied in order for a drug treatment to be approved for a particular segment of the patient population. In response to these issues, two constraints are included in our analysis:

$$\text{a budget constraint: } T(\cdot) \leq \Gamma,$$

where Γ is an exogenous upper limit on Health's expenses for the drug under analysis;

$$\text{and a cost-effectiveness constraint: } Z(Q, \tau) \geq 0,$$

which makes sure that the expected net benefits derived from the drug's approval are above some minimum threshold. This approach to evaluating new technologies is known as net monetary benefits (NMBs), and as long as the terms in the calculations of incremental benefits and costs are the same, then positive NMBs are equivalent to the ICER being less than the willingness to pay in a cost-effectiveness analysis.

Next, we solve both the situations when an integrated decision-maker: maximizes expected social welfare utility (§3.3.1), and maximizes the system's expected utility function (§3.3.2).

2.3.3 The integrated channel

We refer to this setting with the symbol ς, where ς is used to denote the *single* decision-maker structure. Before going further, some useful structural properties of the model are presented.

Lemma 1: For a given τ, the social welfare function, $S(Q, \tau)$, is increasing and concave in Q.

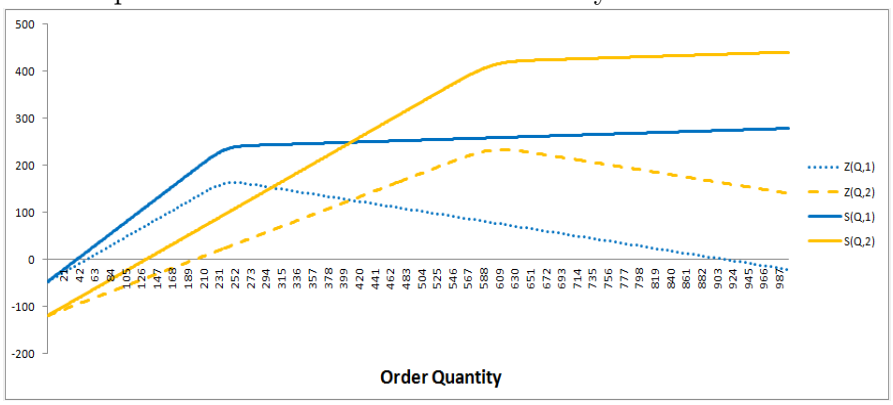
Lemma 2: For a given τ, the system's utility function, $Z(Q, \tau)$, is concave in Q.

Figures 2.4, 2.5, and 2.6 provide a graphical representation of Lemmas 1 and 2, and are useful in explaining the intuition created by our formulation. For increasing values of the order quantity, Q, the solid lines in Figures 2.4 and 2.5 plot the expected social welfare curves while the dashed lines plot the expected system's utility curves. Notice first that from the formulation, when the access level is restricted, the value of b_2 is irrelevant as only high health benefit patients are treated. Second, the reason for the steeper slope in the social welfare functions in Figure 2.4 versus 2.5 after reaching the maximum point in the expected utility functions is driven by the higher value of δ in Figure 2.4; this happens because as the incremental number of patients who are administered the drug goes to zero, the social welfare curve increases at a constant rate δ. The obvious consequence is that as the salvage

Figure 2.4: Expected social welfare and total utility functions with no intersection

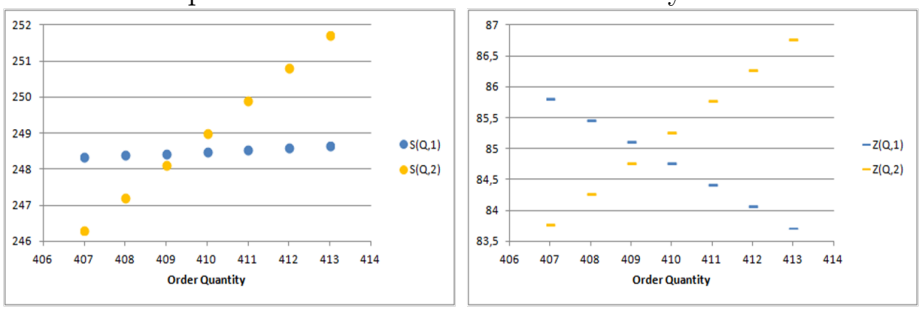
$\lambda = 600; \theta = 0.4; b_1 = 1; b_2 = 0.15; g = 0; \delta = 0.2; c = 0.3$

Figure 2.5: Expected social welfare and total utility functions with intersection



$\lambda = 600; \theta = 0.4; b_1 = 1; b_2 = 0.5; g = 0.2; \delta = 0.05; c = 0.3$

Figure 2.6: Zoom on expected social welfare and total utility functions with intersection



$\lambda = 600; \theta = 0.4; b_1 = 1; b_2 = 0.5; g = 0.2; \delta = 0.05; c = 0.3$

value is positive, the unconstrained social welfare maximizer will continue to order indefinitely. Third, when $g = 0$ both access levels provide a utility of zero when no drugs are ordered, while when $g > 0$, the full access policy is inferior to the restricted access policy when $Q = 0$. This occurs because under full access, more patients are expected to arrive, resulting in higher costs of understocking. Fourth, under restricted access, the curves initially grow at a faster rate due to the higher value of the average health benefits, but they do so for a shorter range of order quantities due to the decreasing probability that demand exceeds a particular order quantity. These effects are generated by the value of θ and the shape of the demand distribution. As θ increases, the slope of the expected utility under the restricted access policy will stay positive for higher values of Q; and also the initial increasing slope for both access level policies will get closer to each other as θ gets closer to 1. Finally, it is worth noting that in Figure 2.5 the access level policies cross paths, while in Figure 2.4 they do not. Figure 2.6 zooms in on the crossing point for the same parameter combination used

in Figure 2.5; since due to integrality there may not exist an integer order quantity where the functions are equal, henceforth when we speak of a crossing point or an intersection, we refer to a change in dominace of one access level curve versus the other. Proposition 1 provides the conditions for this crossing to occur, and explains the changes in the crossing point as a function of the problem's parameters.

Proposition 1: Let q be a positive order quantity such that $S(q,1) \geq S(q,2)$; and $S(q+1,2) > S(q+1,1)$.

a) If inequality (2.3.3) is satisfied, then q is unique and given by equation (2.3.4).

$$\theta \geq \left(\frac{1-P(Q;\lambda)}{1-P(Q;\lambda\theta)}\right)\left(\frac{P(Q+1;\lambda\theta)}{P(Q+1;\lambda)}\right) \tag{2.3.3}$$

$$q = max\left\{Q \;\middle|\; \left(\frac{g}{b_1-\delta+g}\right)\left(\frac{\lambda}{A(Q,2)}\right) + \left(\frac{b_1-b_2}{b_1-\delta+g}\right) \geq \left(\frac{A(Q,2)-A(Q,1)}{A(Q,2)}\right)\left(\frac{1}{1-\theta}\right)\right\} \tag{2.3.4}$$

b) $b_2 > \delta$ is a necessary and sufficient condition for q to exist.

c) If $b_2 < \delta$, then $S(Q,1) > S(Q,2),\ \forall\ Q > 0$.

d) If $b_2 = \delta$, then $S(Q,1) > S(Q,2)$, for some finite $Q > 0$, and $\lim_{P(Q,\lambda)\to 0} S(Q,1) - S(Q,2) = 0$.

e) The value of q is increasing in $b_1, g,\ \delta$, and decreasing in b_2. The change in q with respect to θ is ambiguous.

The existence of a unique q is critical in our analysis because it provides a threshold value for the dominance of either of the access level policies, both in terms of maximizing expected social welfare or expected system's utility - because the manufacturing cost is linear in the order quantity. In other words, q is independent of the transfer price and is then the same for the social welfare curves and for the system's utility curves. It is useful in representing the

essential tradeoff between choosing higher access levels versus higher service levels; notice that for any fixed order quantity, the service level will always be higher for lower levels of access. Consequently, q will be utilized below as a reference point to determine the optimal policy given the objective function and cost-effectiveness and budget constraints. Proposition 1a gives the conditions for finding q - equation (2.3.4) -, and for such q to be unique - equation (2.3.3) -, which through numerical experiments hasn't been found to be a very restrictive condition; being more specific, we haven't been able to find any combination of values for which the two access level curves cross more than once. Further, as is shown in the proofs in Appendix 1, if there exists at least one crossing point, then the number of crossing points must be odd. This implies that if q is not unique, then the curves from the two access level policies must cross at least three times. While imagining two concave curves that cross three times is not an impossible task, the shape required for such curves does not correspond with our observations. We believe there are two main situations where such complication could occur: First, when the average benefits for a given access level, and the shape of the effective demand distribution, are assumed to be independent; however, this assumption is essential to our model's formulation since the average benefit is given by the proportion of patients that belong to each type, and therefore does not concern the problem studied here. Second, when the demand distribution has very "fat tails", *e.g.,* when the expected system's utility function is relatively flat for a large range of values around the order quantity that maximizes the function. Such formulation could be appropriate if the size of the demand has high probabilities of being either very large or very small, and is left as an opportunity for further research.

As a result of the above discussion, we will henceforth use the following assumption:

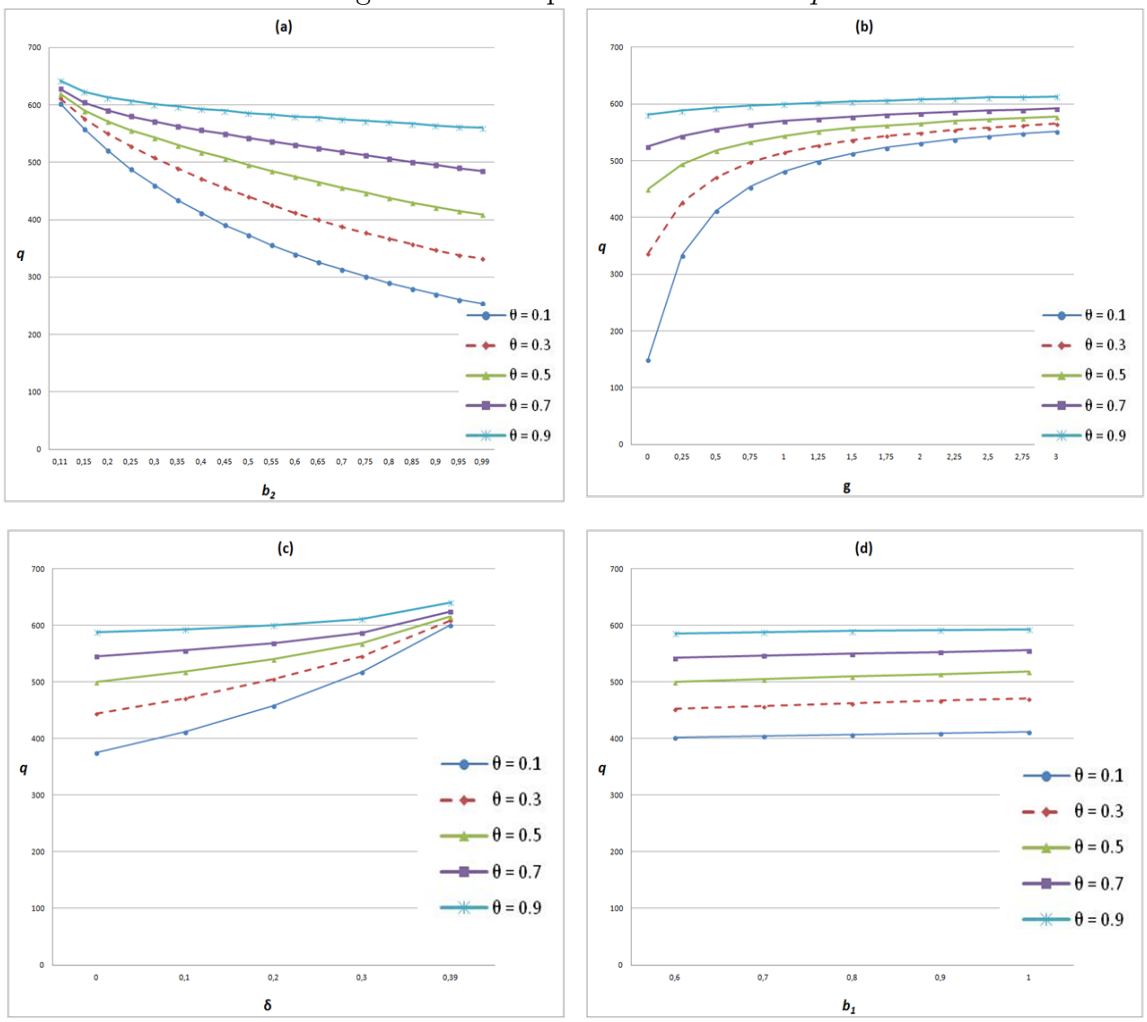
A1: *If q exists, it is unique.*

Proposition 1b, 1c, and 1d, provide the necessary and sufficient conditions for q to exist,

or to not exist. These results are very intuitive as when $\delta > b_2$, it is more efficient, regardless of the objective function, to get the salvage value of the drug than to provide the drug to type 2 patients. As a result, full access can only be superior to restricted access if $\delta < b_2$.

As for the comparative statics shown in Proposition 1e, the reader is invited to direct his attention to Figure 2.7 in order to get additional insights on the sensitivity of q with respect to changes in the problem's parameters. Notice first that for a given Q, limiting access is more likely to maximize social welfare when goodwill costs or salvage value are relatively high, and when the difference in the health benefits received by both patient groups is large. All these results are intuitive. First, as the goodwill cost - g -, increases, the cost of understocking increases; since the probability of understocks for a fixed Q increases in the access level, then there is an increase in the minimum required order quantity for full access to provide higher expected social welfare than restricted access. In terms of the salvage value - δ -, note that the overstocking cost decreases as δ increases; in the limit of the feasible values for q to exist, as δ approaches the expected benefit for patients of type 2, the value of providing full access decreases because large order quantities may be ordered and then salvaged at a high value if necessary. As for the health benefits, on one hand when b_1 increases, the slope of the expected utility curve under restricted access grows at a faster rate and for a larger range of order quantities; while the slope of the expected utility curve under full access will also increase, the effect is less because for any administered drug, the slope only increases with probability $(1 - \theta)$. On the other hand, when b_2 increases, the slope of the expected utility curve under restricted access is unaffected, while the slope of the expected utility curve under full access grows at a faster rate, causing the intersection point between the access level policies to take place at lower values of the order quantity. Analyzing it from a different perspective, as b_2 grows, the probability of full access to be preferred should (weakly) increase, meaning that q decreases in b_2.

Figure 2.7: Comparative statics for q



(a) $\lambda = 600; b_1 = 1; g = 0.5; \delta = 0.1$
(b) $\lambda = 600; b_1 = 1; b_2 = 0.4; \delta = 0.1$
(c) $\lambda = 600; b_1 = 1; b_2 = 0.4; g = 0.5$
(d) $\lambda = 600; b_2 = 0.4; g = 0.5; \delta = 0.1$

Last, it is worth paying special attention to the role of θ. Strictly speaking, q is increasing in θ if:

$$\lambda \left(P(Q;\lambda) - \theta P(Q;\lambda\theta) \right) > Q \left(P(Q+1;\lambda) - P(Q+1;\lambda\theta) \right) + \lambda(1-\theta) \left(P(Q;\lambda\theta) - p(Q+1;\lambda\theta) \right), \quad (2.3.5)$$

and q is decreasing in θ if:

$$\lambda\left(P(Q;\lambda)-\theta P(Q;\lambda\theta)\right)< \\ Q\left(P(Q+1;\lambda)-P(Q+1;\lambda\theta)\right)+\lambda(1-\theta)\left(P(Q;\lambda\theta)-p(Q+1;\lambda\theta)\right). \quad (2.3.6)$$

The details for these results can be observed in the proof of Proposition 1 provided in Appendix 1. Notice that both sides of the inequalities are positive, and while we are not able to provide intuition based on this expression, our numerical experiments always yield that q is increasing in θ. This implies that when θ increases, the effect of a larger expected demand - with mean $\lambda\theta$ - under restricted access is higher than the effect of a larger average benefit - $B(2)$ - under full access. Notice that as θ increases, the fraction of patients for which the average benefit increases under full access (*i.e.,* type 2 patients) decreases; this balancing effect may be the main element for which higher levels of θ require larger order quantities for full access to yield higher expected social welfare than restricted access. Figure 2.7 is also instructive in observing under which circumstances does θ play a more important role in the value of q. Looking at graph (a) in Figure 2.7, as b_2 approaches δ, the effect of θ is minimal since the order quantity needs to be very large for the administration of a drug to a type 2 patient to be more efficient than restricting access and selling such drug at salvage value. However as b_2 increases, the incentive to provide full access grows, and the impact is highest for low values of θ because the low levels of expected demand under restricted access are more rapidly overwhelmed by the full access policy. From graph (b), when g is very large, the risk of understocking is the main driver of the decision and θ becomes irrelevant. However as g decreases, the penalty associated with understocks drops, and the size of the expected demand becomes the main driver. Graph (c) in Figure 2.7 follows a similar intuition as that explained for b_2. As the salvage value - δ -, approaches the health benefit of type 2 patients, θ becomes almost irrelevant in determining q. However as δ goes to zero, the risk of overstocks increases and if θ is low, the minimal order quantity needed

for the full access policy to be above the restricted access policy decreases because a large demand is needed to justify larger order quantities. For graph (d), it is initially interesting to see that the effect of θ is relatively constant as b_1 oscillates between b_2 and 1 (due to normalization), and the increase in q with respect to b_1 is not too large. The intuition is that changes in b_1 affect both access level policies similarly because the average health benefits in both situations are contingent on b_1, and the proportion of the demand which receives b_1 under full access will be given by θ. Therefore, as b_1 changes, the main role played by θ is in determining the size of the demand, rather than the crossing point between the access level curves.

Now that the determination of q has been explained in detail, we will turn our attention to solving the risk-neutral single decision-maker's problem when the objective function is maximizing expected social welfare (§3.3.1) and maximizing expected system's utility (§3.3.2), in the presence of cost-effectiveness and budget constraints.

Case 1$^{\varsigma}$: Maximizing expected social welfare

In this section we solve the problem of maximizing expected social welfare as might be the case for a national health authority or a not-for-profit health organization. The problem faced by the single decision maker under this setting is:

$$\max_{(Q,\tau)} \quad S(Q,\tau) = (B(\tau) - \delta + g)A(Q,\tau) + \delta Q - g\lambda F(\tau)$$

subject to:

$$cQ \leq \Gamma$$

$$-cQ + B(\tau)A(Q,\tau) + \delta\left(Q - A(Q,\tau)\right) - g\left(\lambda F(\tau) - A(Q,\tau)\right) \geq 0 \quad (2.3.7)$$

We begin by defining the feasible area for trade to occur. Let $Q_{\Gamma}^{\varsigma} = \left\lfloor \frac{\Gamma}{c} \right\rfloor$, be the largest order quantity that satisfies the budget constraint, and let $\underline{Q}^{\varsigma} = \min\{\underline{Q}_1^{\varsigma}, \underline{Q}_2^{\varsigma}\}$, be the

minimum order quantity that allows the system's utility function to be nonnegative, where:

$$\underline{Q}_1^{\varsigma} = \min\left\{ Q \;\middle|\; Q \leq \frac{(B(1) - \delta + g)A(Q,1) - g\theta\lambda}{c - \delta}\,; Q \geq 0 \right\}, \text{ and}$$

$$\underline{Q}_2^{\varsigma} = \min\left\{ Q \;\middle|\; Q \leq \frac{(B(2) - \delta + g)A(Q,2) - g\lambda}{c - \delta}\,; Q \geq 0 \right\}.$$

Lemma 3: $Q_{\Gamma}^{\varsigma} \geq \underline{Q}^{\varsigma}$, is a necessary and sufficient condition for trade to occur.

Lemma 3 simply states that if the largest order quantity that satisfies the budget constraint is lower than the smallest order quantity that satisfies the cost-effectiveness constraint, then there is no feasible solution that results in a positive order quantity, *i.e.,* all patients are excluded from the access policy. This could happen when the goodwill cost is too large and the available budget is relatively low (*e.g.,* the case of the prostate cancer drug described in the introduction of this chapter).

Next, recall from Lemma 1 that the social welfare function is monotonically increasing in Q for a fixed prescription policy threshold. As a result, the optimal solution to the unconstrained problem is unbounded, which implies that either the budget constraint, the cost-effectiveness constraint, or both, must be binding. Define $Q_{S,\varsigma}^{*}$ and $\tau_{S,\varsigma}^{*}$ as the optimal order quantity and prescription policy threshold, respectively, when trade occurs. Let

$$\mathbb{Q}_S = \{Q_{\Gamma}^{\varsigma},\, \bar{Q}_1^{\varsigma},\, \bar{Q}_2^{\varsigma}\},$$

denote the set of possible optimal order quantities under this setting, where $\bar{Q}_{\tau}^{\varsigma}$, is the largest order quantity that satisfies the cost-effectiveness constraint for access level τ, *i.e.,*

$Z(\bar{Q}_\tau^\varsigma, \tau) \geq 0 > Z(\bar{Q}_\tau^\varsigma + 1, \tau)$.[9] By doing some algebra, we find

$$\bar{Q}_1^\varsigma = \max\left\{ Q \;\middle|\; Q \leq \frac{(B(1) - \delta + g)A(Q,1) - g\theta\lambda}{c - \delta}\,; Q > 0 \right\}, \text{ and}$$

$$\bar{Q}_2^\varsigma = \max\left\{ Q \;\middle|\; Q \leq \frac{(B(2) - \delta + g)A(Q,2) - g\lambda}{c - \delta}\,; Q > 0 \right\}.$$

From the above definitions, it should be evident that the existence of $\bar{Q}_\tau^\varsigma$ depends both on the combination of the marginal cost and benefit parameters and on the shape of the demand distribution. Still, Lemma 4 provides a key minimum condition that needs to be satisfied for $\bar{Q}_\tau^\varsigma > 0$.

Lemma 4: $c \leq B(\tau)$, is a necessary condition for $\bar{Q}_\tau^\varsigma$, $\tau = 1, 2$, to exist.

Notice that $c \leq B(\tau)$ is not a sufficient condition due to demand's uncertainty which may result in understocking and overstocking costs. At this point, it is also worth noting that even when the budget constraint is active, not necessarily all order quantities in the range $[\underline{Q}^\varsigma\,, Q_\Gamma^\varsigma]$ will be feasible, since when $\bar{Q}_1^\varsigma < q < \underline{Q}_2^\varsigma \leq Q_\Gamma^\varsigma$, all order quantities in the range $(\bar{Q}_1^\varsigma\,, \underline{Q}_2^\varsigma)$ do not satisfy the cost-effectiveness constraint. Lemma 5 further clarifies the ordering of the different reference values that have been introduced.

Lemma 5: Assume $\bar{Q}_1^\varsigma$ and $\bar{Q}_2^\varsigma$ exist.

a) Suppose q exists.

a1) If $\bar{Q}_1^\varsigma = q$, then the possible orderings are:

a1.1) $\underline{Q}_1^\varsigma \leq \bar{Q}_1^\varsigma = q = \underline{Q}_2^\varsigma \leq \bar{Q}_2^\varsigma$.

[9] Notice that the system's utility function may cross the zero-profit line at most at two positive values of Q.

a1.2) $\underline{Q}_1^{\varsigma} \leq \underline{Q}_2^{\varsigma} < \bar{Q}_1^{\varsigma} = q \leq \bar{Q}_2^{\varsigma}$.

a2) If $\bar{Q}_1^{\varsigma} > q$, then the possible orderings are:

a2.1) $\underline{Q}_1^{\varsigma} \leq \underline{Q}_2^{\varsigma} \leq q < \bar{Q}_1^{\varsigma} \leq \bar{Q}_2^{\varsigma}$.

a2.2) $q < \underline{Q}_2^{\varsigma} \leq \underline{Q}_1^{\varsigma} \leq \bar{Q}_1^{\varsigma} \leq \bar{Q}_2^{\varsigma}$.

a3) If $\bar{Q}_1^{\varsigma} < q$, then the possible orderings are:

a3.1) $\underline{Q}_1^{\varsigma} \leq \bar{Q}_1^{\varsigma} < q < \underline{Q}_2^{\varsigma} \leq \bar{Q}_2^{\varsigma}$.

a3.2) $\underline{Q}_1^{\varsigma} \leq \underline{Q}_2^{\varsigma} \leq \bar{Q}_2^{\varsigma} \leq \bar{Q}_1^{\varsigma} < q$.

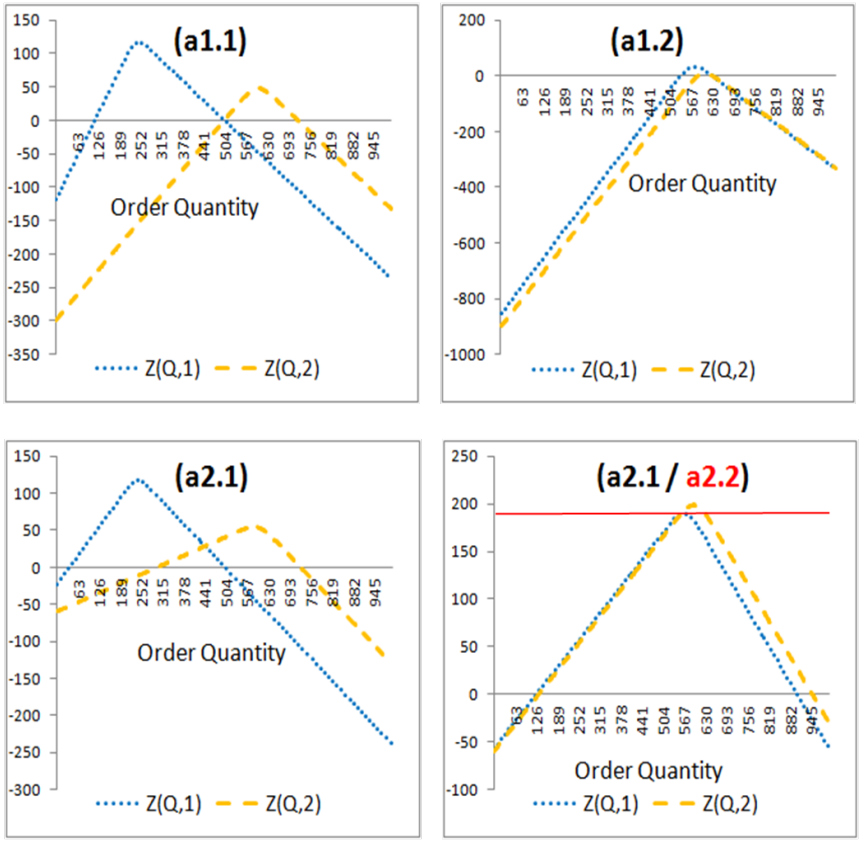
a3.3) $\underline{Q}_1^{\varsigma} < \bar{Q}_1^{\varsigma} = \underline{Q}_2^{\varsigma} = \bar{Q}_2^{\varsigma} < q$.

b) Suppose q does not exist. Then $\underline{Q}_1^{\varsigma} \leq \underline{Q}_2^{\varsigma} \leq \bar{Q}_2^{\varsigma} \leq \bar{Q}_1^{\varsigma}$.

There are a few interesting ideas on which it is worth elaborating based on Lemma 5; Figure 2.8 provides graphical examples to illustrate some of these different possible situations. Notice first that the crossing point between the two access level policies may occur: when the expected system's utility is still increasing in both access level policies; when the expected system's utility is already decreasing in both access level policies; or when the expected system's utility is decreasing under the restricted access policy and increasing in the the full access policy. Another observation is that a necessary condition for $\underline{Q}_{\tau}^{\varsigma} = \bar{Q}_{\tau}^{\varsigma} > 0$ is that $g > 0$; such equality would imply that under the access level policy τ, there exists only one order quantity that achieves cost-effectiveness.

Analyzing the Lemma by parts, for Lemma 5a1, as is formally shown in the proof, if $\bar{Q}_1^{\varsigma} = q$, then it is also true that the value of the curves is equal at an integer value. As result it is necessary that at least $q = \bar{Q}_2^{\varsigma}$ or $q = \underline{Q}_2^{\varsigma}$. Still, it is even possible that $\underline{Q}_1^{\varsigma} = \bar{Q}_1^{\varsigma} = q = \underline{Q}_2^{\varsigma} = \bar{Q}_2^{\varsigma}$; this situation requires two very specific circumstances to simultaneously occur. First, $q = \underline{Q}_2^{\varsigma} = \bar{Q}_2^{\varsigma}$ implies that under full access, the order quantity q maximizes the expected system's utility achieving a value of zero. Similarly, $\underline{Q}_1^{\varsigma} = \bar{Q}_1^{\varsigma} = q$ can only mean that the expected system's utility under the restricted access level policy

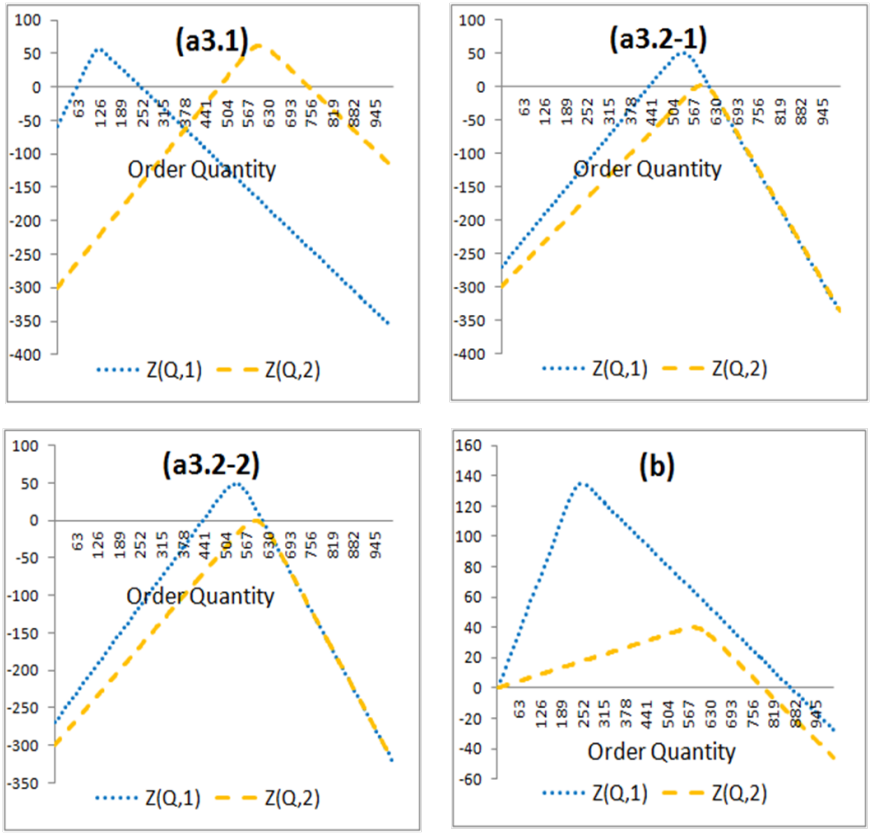
Figure 2.8: Ordering of the reference quantities (part 1)



(a1.1) $\lambda = 600; \theta = 0.4; b_1 = 1; b_2 = 0.29; g = 0.5; \delta = 0; c = 0.477$
$\underline{Q}_1^{\varsigma} = 118 < \bar{Q}_1^{\varsigma} = q = \underline{Q}_2^{\varsigma} = 503 < \bar{Q}_2^{\varsigma} = 722.$
(a1.2) $\lambda = 600; \theta = 0.95; b_1 = 1; b_2 = 0.12; g = 1.5; \delta = 0; c = 0.9055$
$\underline{Q}_1^{\varsigma} = 538 < \underline{Q}_2^{\varsigma} = 589 < \bar{Q}_1^{\varsigma} = q = \bar{Q}_2^{\varsigma} = 629.$
(a2.1) $\lambda = 600; \theta = 0.4; b_1 = 1; b_2 = 0.3; g = 0.1; \delta = 0; c = 0.4795$
$\underline{Q}_1^{\varsigma} = 39 < \underline{Q}_2^{\varsigma} = 300 < q = 442 < \bar{Q}_1^{\varsigma} = 500 < \bar{Q}_2^{\varsigma} = 725.$
(a2.2) $\lambda = 600; \theta = 0.95; b_1 = 1; b_2 = 0.96; g = 0.1; \delta = 0.05; c = 0.65$
$\underline{Q}_1^{\varsigma} = 127 < \underline{Q}_2^{\varsigma} = 134 < q = 557 < \bar{Q}_1^{\varsigma} = 902 < \bar{Q}_2^{\varsigma} = 947.$

crosses the zero utility line from below at order quantity $(q - \varepsilon)$, $\varepsilon \in (0, 1)$ and then crosses it again from above at order quantity q, - notice that $\lceil (q - \varepsilon) \rceil = q$. This case of full equality is highly unlikely and implies that for both access level policies, the only order quantity that achieves cost-effectiveness is q; it requires that $g > 0$, that θ approaches 1, and that b_2 approaches b_1. Further, taking into consideration that only a small subset of parameter combinations yield $\bar{Q}_1^{\varsigma} = q$, the relationship when the latter equality occurs is typically either: $\underline{Q}_1^{\varsigma} < \bar{Q}_1^{\varsigma} = q = \underline{Q}_2^{\varsigma} < \bar{Q}_2^{\varsigma}$, or $\underline{Q}_1^{\varsigma} < \underline{Q}_2^{\varsigma} < \bar{Q}_1^{\varsigma} = q = \bar{Q}_2^{\varsigma}$, as is shown in graphs (a1.1) and (a1.2) in Figure 2.8.

Figure 2.9: Ordering of the reference quantities (part 2)



(a3.1) $\lambda = 600; \theta = 0.2; b_1 = 1; b_2 = 0.5; g = 0.5; \delta = 0; c = 0.4795$
$\underline{Q}_1^s = 59 < \bar{Q}_1^s = 250 < q = 382 < \underline{Q}_2^s = 484 < \bar{Q}_2^s = 750.$
(a3.2-1) $\lambda = 600; \theta = 0.9; b_1 = 1; b_2 = 0.05; g = 0.5; \delta = 0; c = 0.88$
$\underline{Q}_1^s = 436 < \underline{Q}_2^s = 578 < \bar{Q}_2^s = 605 < \bar{Q}_1^s = 613 < q = 625.$
(a3.2-2) $\lambda = 600; \theta = 0.9; b_1 = 1; b_2 = 0.06; g = 0.5; \delta = 0; c = 0.88435$
$\underline{Q}_1^s = 439 < \underline{Q}_2^s = \bar{Q}_2^s = 592 < \bar{Q}_1^s = 610 < q = 622.$
(b) $\lambda = 600; \theta = 0.4; b_1 = 1; b_2 = 0.15; g = 0; \delta = 0.2; c = 0.42$
$\underline{Q}_1^s = \underline{Q}_2^s = 0 < \bar{Q}_2^s = 790 < \bar{Q}_1^s = 872.$

The main difference between parts a2 and a3 is that for the former, $Z(q, \tau) > 0$, while for the latter $Z(q, \tau) < 0$, which may result in the aforementioned discontinuous range of order quantities. In Figure 2.8, graphs (a2.1) and (a2.1/a2.2) show situations when the joint effect of demand distribution and the average expected benefits for a given access level are relatively high relative to the transfer cost, c. It should be mentioned that the situation from Lemma 5-a2.2 couldn't be exactly replicated numerically, but it is easy to show that if the cost-effectiveness threshold was positive and sufficiently large, instead of its currently assumed value of zero, then the shape of the graph (a2.1/a2.2) would fit into this case;

specifically if the cost-effectiveness threshold was raised to the horizontal red line, graph (a2.1/a2.2) corresponds to part a2.2 of Lemma 5. For part a3, graph (a3.1) shows the most common shape given that $\bar{Q}_1^c < q$. Graphs (a3.2-1) and (a3.2-2) both correspond to Lemma 5a3.2, the main distinction being that for the latter there is a single order quantity that satisfies cost-effectiveness under the full access policy. Part a3.3 requires a very specific combination of parameters, where there is a single order quantity that satisfies cost-effectiveness for full access, which coincides with the largest, but not unique, order quantity that satisfies cost-effectiveness under restricted access.

Another important observation arising from Lemma 5 is that of dominance, *i.e.,* the (weak) superiority of a particular access level policy in the expected social welfare utility - and consequently in the expected system's utility - for any order quantity that satisfies the cost-effectiveness constraint. On one hand, restricted access weakly dominates full access under the situations that satisfy parts a3.2, a3.3, and b, of Lemma 5. In the first two of these cases, the intersection between the two access level policies occurs at an order quantity that is higher than that which maximizes the expected system's utility under full access, and in such way that all order quantities above q yield a non cost-effective outcome. Through numerical experiments, we have observed that such situations are highly infrequent and occur when the fraction of type 1 patients is very high (above 0.9), the health benefit of type 2 patients is very low (below 0.1 b_1), and the transfer cost c is high. For the situation from Lemma 5b, the result follows from Proposition 1 and the only condition is for the value of the outside option, or salvage value, to be higher than the value of health benefits for type 2 patients. On the other hand, full access weakly dominates restricted access under the situations that satisfy Lemma 5a2.2. This means that the crossing point between the two access level policies occurs at an order quantity that is lower than that which maximizes the expected system's utility under restricted access. Again, through numerical experiments we have observed that this situation is not frequent either and requires the goodwill cost to be

positive and both the fraction of type 1 patients and the health benefits of type 2 patients to be very high, causing the two access level policies to almost overlap. For the rest of the combinations, there is no clear dominance and the optimal solution will depend on whether the inequalities in Lemma 5 are strong or weak, as well as on the value of the available budget.

Summarizing, and perhaps most importantly, even though Lemma 5 may appear to create complexity by identifying a large variety of possible orderings, it is very useful in finding a structure to understand the drivers of the optimal decision making process. From this result, and assuming that there exists at least one feasible solution that yields $Q^*_{S,\varsigma} > 0$, we are now able to reduce the analysis of the optimal decisions to the relationship between the budget constraint, the minimum and maximum feasible quantities for full access, $\underline{Q}^{\varsigma}_2$ and $\bar{Q}^{\varsigma}_2$, and the crossing point of the two access level policies, q.

Theorem 1: Assume $\underline{Q}^{\varsigma}_{\tau}$, $\tau = 1, 2$ exist, and let $Q^{\varsigma}_{\Gamma} \geq \underline{Q}^{\varsigma}$. When the integrated supply chain maximizes expected social welfare subject to budget and cost-effectiveness constraints, the optimal solution is as follows.

a) Suppose q exists.

a1) $\min\{\bar{Q}^{\varsigma}_2\,,\, Q^{\varsigma}_{\Gamma}\} \leq \max\{q\,,\, \underline{Q}^{\varsigma}_2\}$, is a necessary condition for $\tau^*_{S,\varsigma} = 1$ and $Q^*_{S,\varsigma} = \min\{Q^{\varsigma}_{\Gamma}\,,\, \bar{Q}^{\varsigma}_1\}$ to be an optimal solution.

a2) $\min\{\bar{Q}^{\varsigma}_2\,,\, Q^{\varsigma}_{\Gamma}\} < \max\{q\,,\, \underline{Q}^{\varsigma}_2\}$, is a necessary and sufficient condition for $\tau^*_{S,\varsigma} = 1$ and $Q^*_{S,\varsigma} = \min\{Q^{\varsigma}_{\Gamma}\,,\, \bar{Q}^{\varsigma}_1\}$ to be the unique optimal solution.

a3) $\min\{\bar{Q}^{\varsigma}_2\,,\, Q^{\varsigma}_{\Gamma}\} \geq \max\{q\,,\, \underline{Q}^{\varsigma}_2\}$ is a necessary condition for $\tau^*_{S,\varsigma} = 2$ and $Q^*_{S,\varsigma} = \min\{Q^{\varsigma}_{\Gamma}\,,\, \bar{Q}^{\varsigma}_2\}$ to be an optimal solution.

a4) Jointly satisfying $(\min\{\bar{Q}^{\varsigma}_2\,,\, Q^{\varsigma}_{\Gamma}\} > q)$ and $(\min\{\bar{Q}^{\varsigma}_2\,,\, Q^{\varsigma}_{\Gamma}\} \geq \underline{Q}^{\varsigma}_2)$, is a necessary and sufficient condition for $\tau^*_{S,\varsigma} = 2$ and $Q^*_{S,\varsigma} = \min\{Q^{\varsigma}_{\Gamma}\,,\, \bar{Q}^{\varsigma}_2\}$ to be the unique optimal solution.

a5) $\min\{\bar{Q}^{\varsigma}_2\,,\, Q^{\varsigma}_{\Gamma}\} = q \geq \underline{Q}^{\varsigma}_2$ is a necessary and sufficient condition for the decision-maker to be indifferent between $(\tau^*_{S,\varsigma} = 1$ and $Q^*_{S,\varsigma} = \min\{Q^{\varsigma}_{\Gamma}\,,\, \bar{Q}^{\varsigma}_1\})$ versus $(\tau^*_{S,\varsigma} = 2$ and

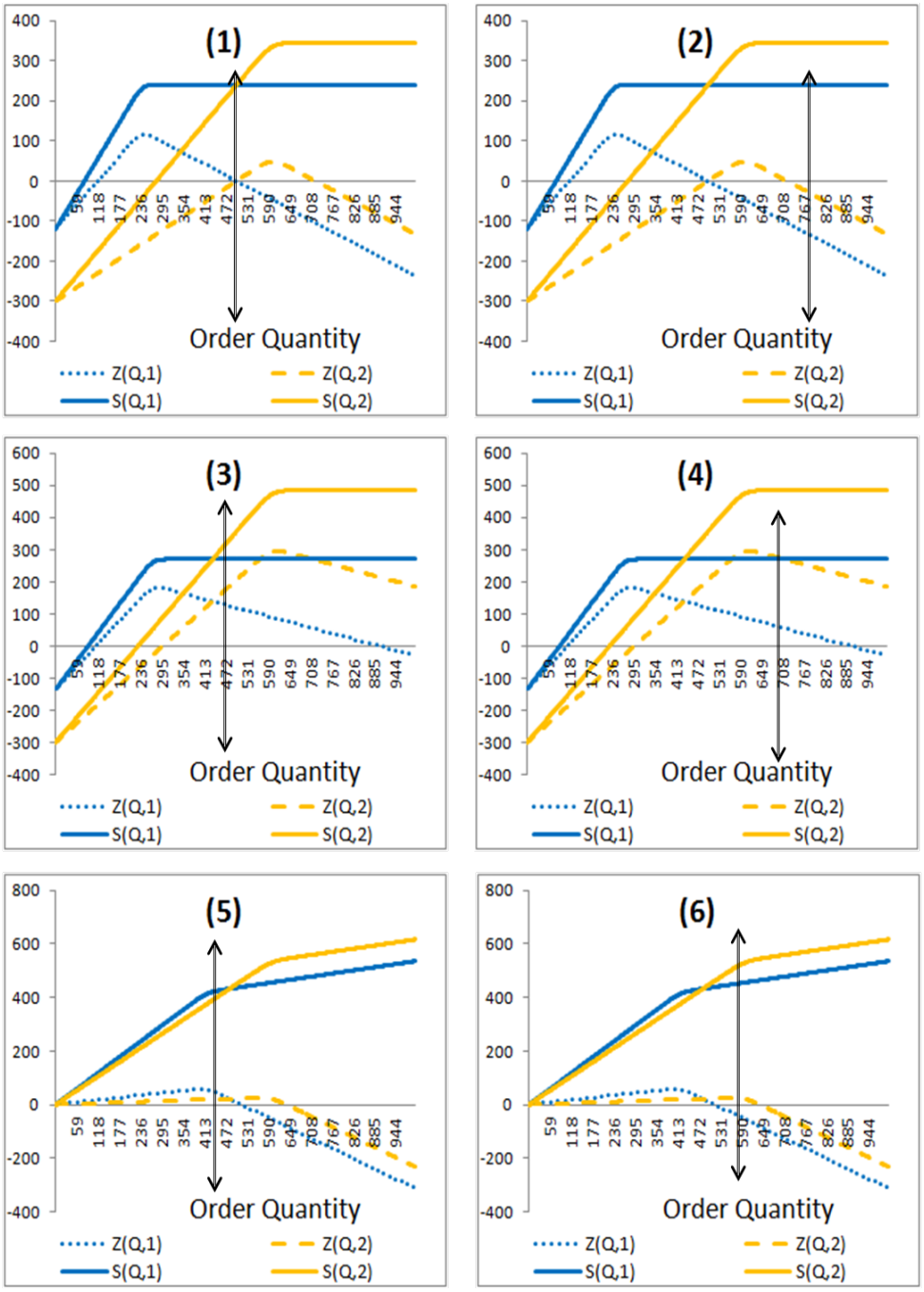
$Q^*_{S,\varsigma} = \min\{Q^\varsigma_\Gamma\,,\,\bar{Q}^\varsigma_2\}$).

b) Suppose q does not exist; then $\tau^*_{S,\varsigma} = 1$ and $Q^*_{S,\varsigma} = \min\{Q^\varsigma_\Gamma\,,\,\bar{Q}^\varsigma_1\}$.

Theorem 1 shows the conditions under which the optimal access level will be either restricted ($\tau^*_{S,\varsigma} = 1$), or fully inclusive ($\tau^*_{S,\varsigma} = 2$), and the corresponding optimal order quantity of drugs. Note first that Theorem 1-a1 and -a3 are the result of Theorem1-a2, -a4, and -a5. It is also easier to see the relevance of the (weak) dominance relationships we discussed after Lemma 5. Namely, Lemma 5-a1.2, -a3.2, -a3.3 correspond to Theorem 1-a1, and as a result also correspond to either Theorem 1-a2 or -a5. Similarly, Lemma 5-a2.2 corresponds to Theorem 1-a3, and as a result to either Theorem 1-a4, or -a5. Finally, Lemma 5-b corresponds to Theorem 1-b. Since the latter situations are less interesting from a joint decision making perspective, none of the cases where strong dominance exists is represented in Figures 2.10 and 2.11, but the interested reader can verify in the corresponding graphs shown in Figures 2.8 and 2.9 that regardless of the magnitude of the available budget, the access level will remain constant. Figures 2.10 and 2.11 is then used to assist the process of understanding how the solution changes in many of the remaining cases once the budget constraint is included.

One of the interesting observations is that when the budget is sufficiently high and the value of the outside option is lower than the expected health benefits for type 2 patients, full access level may be optimal for the social welfare maximizer even if the average health benefit received by patients of type 2 is lower than the cost of the drug, *e.g.,* Figure 2.10, graph (2); the explanation is that the expected health benefits achieved by patients of type 1 subsidize those belonging to the second type, making aggregate cost-effectiveness possible. In terms of indifference, graphs (1) and (12) show examples where the decision maker is indifferent between both choices of access level, but the drivers and implications are very different. On one hand, in graph (1) the budget constraint is the key driver of the decision

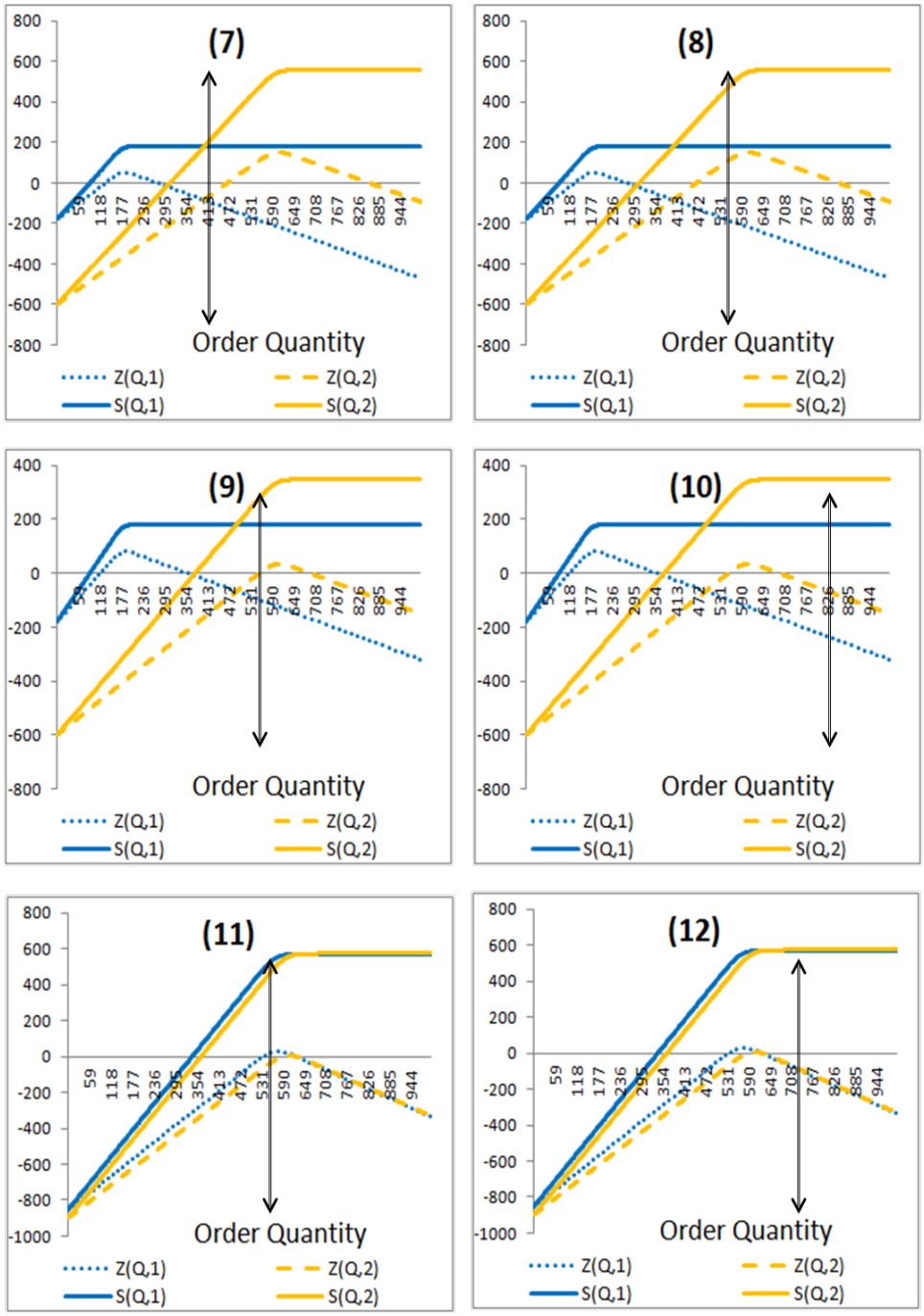
Figure 2.10: Optimal decision making under expected social welfare maximization (part 1)



$\updownarrow$ represents Q_Γ^ς

(1) $\lambda = 600; \theta = 0.4; b_1 = 1; b_2 = 0.29; g = 0.5; \delta = 0; c = 0.477;$ $\tau_{S,\varsigma}^*$ =1 or 2, and $Q_{S,\varsigma}^* = \bar{Q}_1^\varsigma = Q_\Gamma^\varsigma$
(2) $\lambda = 600; \theta = 0.4; b_1 = 1; b_2 = 0.29; g = 0.5; \delta = 0; c = 0.477;$ $\tau_{S,\varsigma}^* = 2$ and $Q_{S,\varsigma}^* = \bar{Q}_2^\varsigma$
(3) $\lambda = 600; \theta = 0.45; b_1 = 1; b_2 = 0.65; g = 0.5; \delta = 0; c = 0.3;$ $\tau_{S,\varsigma}^* = 2$ and $Q_{S,\varsigma}^* = Q_\Gamma^\varsigma$
(4) $\lambda = 600; \theta = 0.45; b_1 = 1; b_2 = 0.65; g = 0.5; \delta = 0; c = 0.3;$ $\tau_{S,\varsigma}^* = 2$ and $Q_{S,\varsigma}^* = Q_\Gamma^\varsigma$
(5) $\lambda = 600; \theta = 0.7; b_1 = 1; b_2 = 0.65; g = 0; \delta = 0.2; c = 0.85;$ $\tau_{S,\varsigma}^* = 1$ and $Q_{S,\varsigma}^* = Q_\Gamma^\varsigma$
(6) $\lambda = 600; \theta = 0.7; b_1 = 1; b_2 = 0.65; g = 0; \delta = 0.2; c = 0.85;$ $\tau_{S,\varsigma}^* = 2$ and $Q_{S,\varsigma}^* = Q_\Gamma^\varsigma$

Figure 2.11: Optimal decision making under expected social welfare maximization (part 2)



$\updownarrow$ represents Q_{Γ}^{ς}

(7) $\lambda = 600; \theta = 0.3; b_1 = 1; b_2 = 0.9; g = 1; \delta = 0; c = 0.65;$ $\tau_{S,\varsigma}^{*} = 1$ and $Q_{S,\varsigma}^{*} = \bar{Q}_1^{\varsigma}$
(8) $\lambda = 600; \theta = 0.3; b_1 = 1; b_2 = 0.9; g = 1; \delta = 0; c = 0.65;$ $\tau_{S,\varsigma}^{*} = 2$ and $Q_{S,\varsigma}^{*} = Q_{\Gamma}^{\varsigma}$
(9) $\lambda = 600; \theta = 0.3; b_1 = 1; b_2 = 0.4; g = 1; \delta = 0; c = 0.5;$ $\tau_{S,\varsigma}^{*} = 2$ and $Q_{S,\varsigma}^{*} = Q_{\Gamma}^{\varsigma}$
(10) $\lambda = 600; \theta = 0.3; b_1 = 1; b_2 = 0.4; g = 1; \delta = 0; c = 0.5;$ $\tau_{S,\varsigma}^{*} = 2$ and $Q_{S,\varsigma}^{*} = \bar{Q}_2^{\varsigma}$
(11) $\lambda = 600; \theta = 0.95; b_1 = 1; b_2 = 0.12; g = 1.5; \delta = 0; c = 0.9055;$ $\tau_{S,\varsigma}^{*} = 1$ and $Q_{S,\varsigma}^{*} = Q_{\Gamma}^{\varsigma}$
(12) $\lambda = 600; \theta = 0.95; b_1 = 1; b_2 = 0.12; g = 1.5; \delta = 0; c = 0.9055;$ $\tau_{S,\varsigma}^{*} = 1$ and $Q_{S,\varsigma}^{*} = \bar{Q}_1^{\varsigma} = \bar{Q}_2^{\varsigma}$

maker's indifference since moving it slightly in any direction would yield a unique solution. This situation tends to occur when the fraction of type 1 patients is medium to low; in such cases, the decision maker will need to choose between providing high service level to a limited fraction of the population, versus providing low service level to the whole patient population. On the other hand, in graph (12) the key driver of the decision maker's indifference is the cost-effectiveness constraint. This situation tends to occur when the fraction of type 1 patients is very high and the transaction cost, c, is also high; in such cases, there will be no major difference between restricting or not access to type 2 patients. Finally, note that by moving the budget constraint appropriately in some of the other graphs, such that $Q_{\gamma}^{\varsigma} = q$, the indifference issue would also arise; this brings the attention to the fact that under expected social welfare maximization, the decision maker may only be indifferent when $Z(q, \tau) \geq 0, \tau = 1, 2$, *i.e.,* for graphs (7)-(10) the optimal solution is unique regardless of the available budget because the crossing point of the curves occurs at a point which is not cost-effective. Finally, when only one access level satisfies the cost-effectiveness constraint, the solution is straightforward and given in Corollary 1.

Corollary 1: When the integrated supply chain maximizes social welfare subject to budget and cost-effectiveness constraints:

a) Assume $\underline{Q}_1^{\varsigma} \leq Q_{\Gamma}^{\varsigma}$ exists and $\underline{Q}_2^{\varsigma}$ does not exist; then $\tau_{S,\varsigma}^{*} = 1$ and $Q_{S,\varsigma}^{*} = \min\{Q_{\Gamma}^{\varsigma}\,,\, \bar{Q}_1^{\varsigma}\}$.

b) Assume $\underline{Q}_2^{\varsigma} \leq Q_{\Gamma}^{\varsigma}$ exists and $\underline{Q}_1^{\varsigma}$ does not exist; then $\tau_{S,\varsigma}^{*} = 2$ and $Q_{S,\varsigma}^{*} = \min\{Q_{\Gamma}^{\varsigma}\,,\, \bar{Q}_2^{\varsigma}\}$.

Case 2^{ς}: Maximizing the system's expected utility function

In this section we solve the problem of maximizing expected system's utility function as might be the case for a private insurance company. The problem faced by the single decision

maker under this setting is:

$$\max_{(Q,\tau)} \quad Z(Q,\tau) = (B(\tau) - \delta + g)A(Q,\tau) - (c-\delta)Q - g\lambda F(\tau)$$

subject to:

$$cQ \leq \Gamma$$

$$-cQ + B(\tau)A(Q,\tau) + \delta\left(Q - A(Q,\tau)\right) - g\left(\lambda F(\tau) - A(Q,\tau)\right) \geq 0 \quad (2.3.8)$$

Define $Q^*_{H,\varsigma}$ and $\tau^*_{H,\varsigma}$ to be the optimal order quantity and prescription policy threshold, respectively, when maximizing the system's expected utility function. Let

$$\mathbb{Q}^{\varsigma}_H = \{Q^{\varsigma}_{\Gamma},\, Q^{\varsigma}_1,\, Q^{\varsigma}_2\},$$

denote the set of possible optimal order quantities under this setting, recalling that $Q^{\varsigma}_{\Gamma} = \left\lfloor \frac{\Gamma}{c} \right\rfloor$, stands for the order quantity when the budget constraint is binding, while

$$Q^{\varsigma}_{\tau} \in \arg\max_{Q} \{Z(Q,\tau)\}, \qquad \tau = 1, 2,$$

is the order quantity that maximizes the system's utility function contingent on τ. From Lemma 2, the system's expected utility function is concave for a given τ, and we can use the method of finite differences to obtain:

$$Q^{\varsigma}_{\tau} = \max\left\{ Q \;\middle|\; P(Q; \lambda F(\tau)) \geq \frac{c-\delta}{B(\tau) - \delta + g} \right\}, \qquad \tau = 1, 2. \quad (2.3.9)$$

Before deriving the optimal access and service levels for this case, some intermediate results are needed.

Proposition 2: Define $\tilde{c} = \{c \mid Z(Q^{\varsigma}_1, 1) = Z(Q^{\varsigma}_2, 2)\}$.

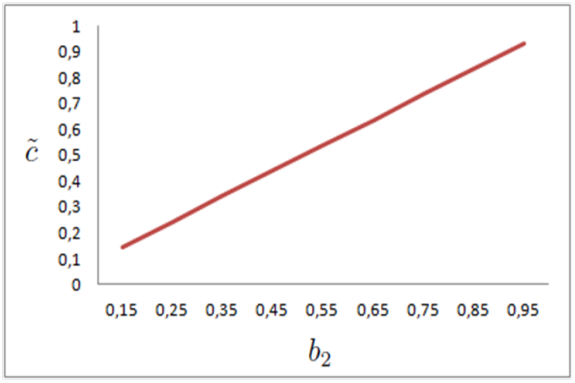
a) $\tilde{c}$ exists $\iff$ q exists.

b) $Z(Q_1^s, 1) > Z(Q_2^s, 2)$ for $c > \tilde{c}$, and $Z(Q_1^s, 1) < Z(Q_2^s, 2)$ for $c < \tilde{c}$.

c) $\delta < \tilde{c} < B(2)$.

The results from Proposition 2 allow us to set a threshold for our comparisons of the efficient decisions under the system's expected utility maximization. It is very interesting to note the connection between the existence of q and $\tilde{c}$, even though the value of q is independent of the drug's transfer price. The reason is that the existence of both q and $\tilde{c}$ depends on the same condition: $\delta < b_2$. While it has already been discussed why this is so for q, in terms of $\tilde{c}$ the intuition is that when $b_2 \leq \delta$, the utility under restricted access for any order quantity can be increased at least at the same rate (δ) as the utility under full access. When the latter is not the case, *i.e.,* when $\delta < b_2$, as the order quantity is increased and the probability of administering an additional unit of the drug decreases more rapidly for the restricted than for the full access policy then the value of $c > \delta$ can be reduced enough to achieve $Z(Q_1^s, 1) < Z(Q_2^s, 2)$.

Figure 2.12: Finding $\tilde{c}$ in relation to the health benefit

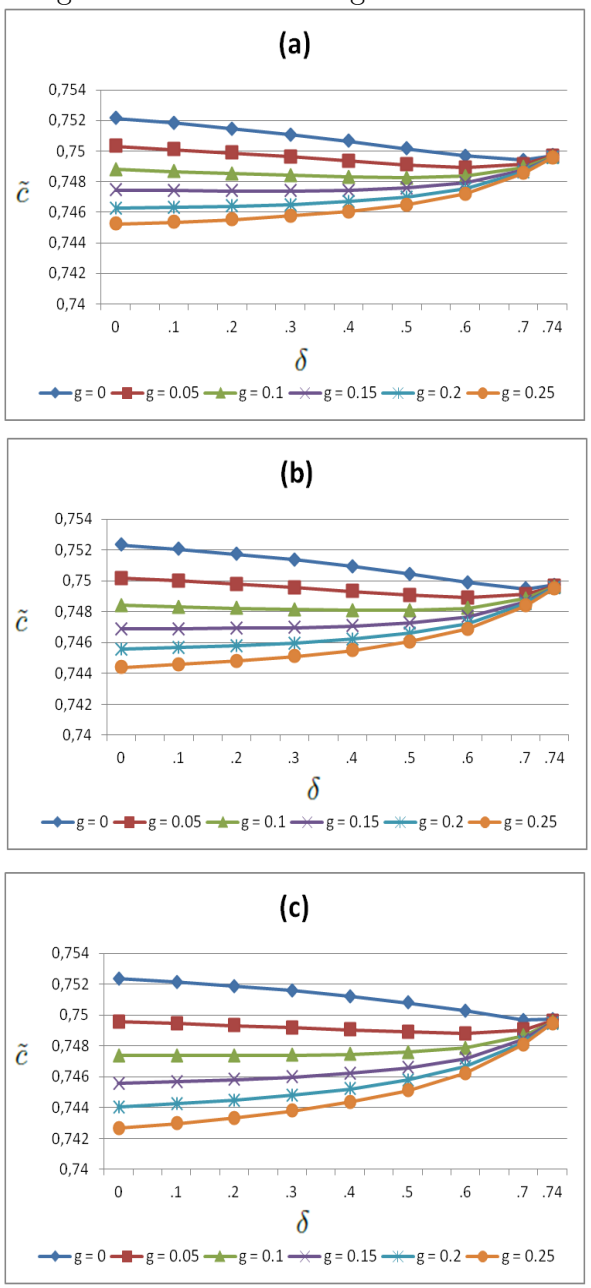


b_2	$\tilde{c}$
0,11	0,1093162
0,15	0,1474637
0,25	0,2444589
0,35	0,342392
0,45	0,440759
0,55	0,5393632
0,65	0,6380777
0,75	0,7368445
0,85	0,8356008
0,95	0,9339054
0,99	0,9737612

$\lambda = 600;\ \theta = 0.1;\ b_1 = 1;\ g = 0.5;\ \delta = 0.1$

Additionally, from numerical experiments we have observed that the value of $\tilde{c}$ is always in the neighborhood of b_2, as shown in Figure 2.12, typically approaching b_2 from below - even though Figure 2.13 shows the latter is not always true. It has also been observed that the change of $\tilde{c}$ with respect to the rest of the parameters is almost flat as shown in Figure

Figure 2.13: Changes in $\tilde{c}$ in relation to goodwill costs and salvage value



$\lambda = 600;\ b_1 = 1; b_2 = 0.75;\ (a)\theta = 0.8;\ (b)\theta = 0.5;\ (c)\theta = 0.2$

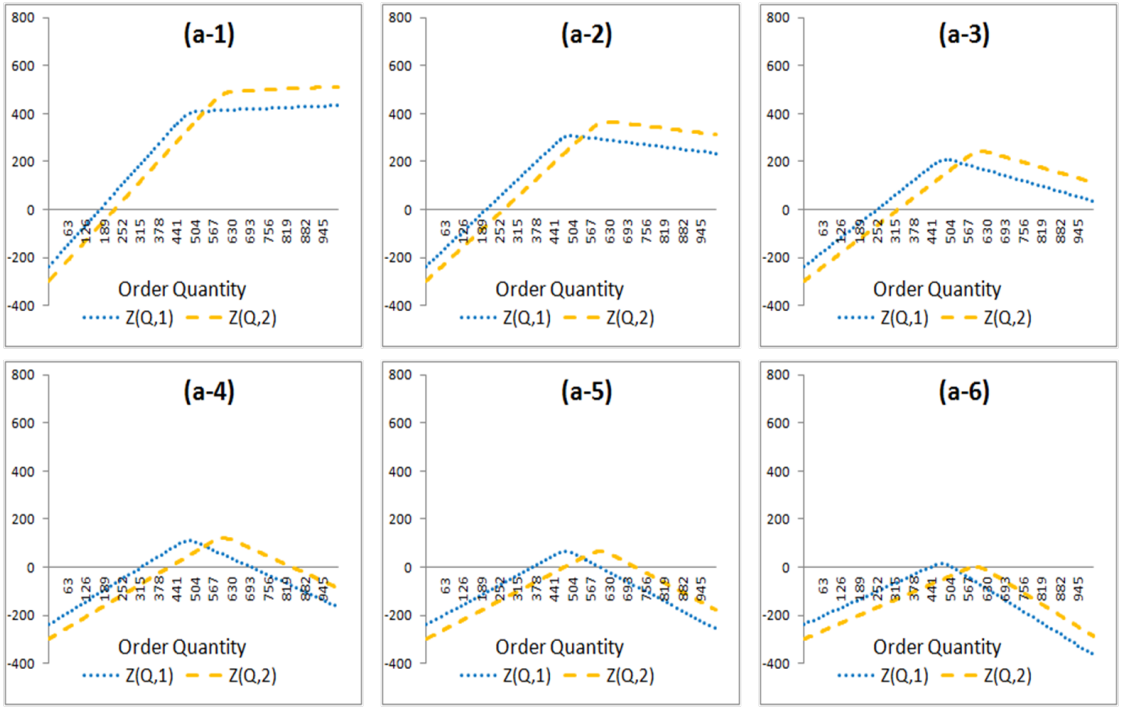
2.13. While initially surprising, the intuition is that when $c \simeq b_2$, then at order quantity Q_2^ς, a proportional inventory allocation would imply that the expected amount of drugs destined to type 1 patients is θQ_2^ς, which approaches Q_1^ς; notice that the $(1-\theta)Q_2^\varsigma$ remaining drugs which are expected to be destined to type 2 patients generate a utility of $(b_2 - c) \simeq 0$ if administered, and a cost of $(g + c - \delta) > 0$ when not administered. Therefore, having a stock of Q_1^ς drugs in a restricted access level policy, and having a stock of Q_2^ς drugs in a full

access level policy, when $c \simeq b_2$, yield very similar expected system's utilities because the type 2 patients barely affect the expected utility function at the optimal order quantity; in fact if demand were deterministic, then when $b_2 = c$, the effect of including the lower patient category would be null.

Figure 2.13 shows the changes in $\tilde{c}$ for a fixed b_2 with respect to the other parameters. It is observed from the graphs that $\tilde{c}$ is slightly decreasing in g, that the effect created by δ is ambiguous and depends on the value of g, and that for any given combination of g and δ, the range of values that $\tilde{c}$ can take increases as θ decreases. Our intuition is that the goodwill cost always affects the full access level policy more, and therefore the range of values of c for which restricted access is preferred increases as g increases. The effect of δ is ambiguous because when g is relatively small, then an increase in δ mainly favors the restricted access policy, and so the threshold decreases; but as g grows, then an increase in δ allows the full access level policy to compensate its increasing understocking cost with a decreasing overstocking cost, and therefore result in a larger value of $\tilde{c}$. Finally, regarding θ, as the proportion of patient population of type 1 increases, then the cost effect of understocking and overstocking under a full access level policy becomes smaller, and so $\tilde{c}$ becomes less responsive to changes in g and δ. To complement our intuition, Figures 2.14 and 2.15 show the change in the system's expected utility function as c grows. Also, they provides the value of $\tilde{c}$, if it exists, under different parameter combinations. As a consequence of the finding from Proposition 2, the remaining analysis is now greatly simplified.

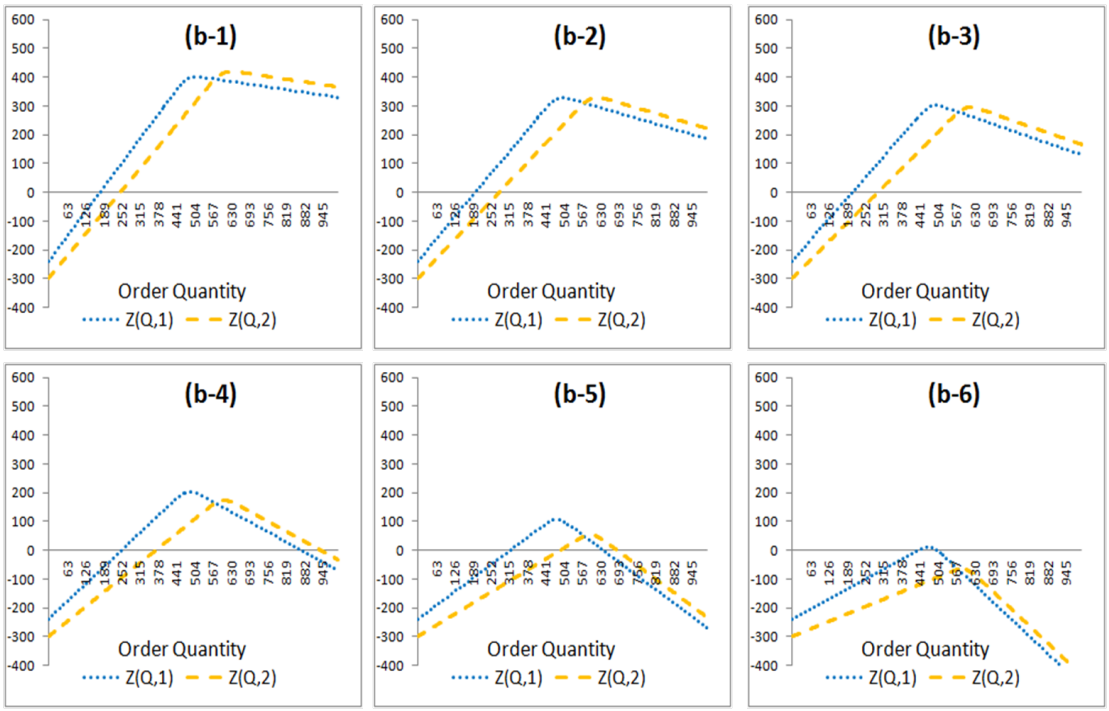
Proposition 3: Assume $\bar{Q}_\tau^\varsigma$, $\tau = 1, 2$, exist, and let $Q_\Gamma^\varsigma \geq \underline{Q}^\varsigma$. When $Q_\Gamma^\varsigma \geq \max\{q\,,\, \underline{Q}_2^\varsigma\}$, $c > \tilde{c}$,
a) is a sufficient condition for the optimal solution to be $\tau_{H,\varsigma}^* = 1$, and $Q_{H,\varsigma}^* = Q_1^\varsigma$, when $Q_\Gamma^\varsigma < \lfloor Q_2^\varsigma \rfloor$;
b) is a necessary and sufficient condition for the optimal solution to be $\tau_{H,\varsigma}^* = 1$, and $Q_{H,\varsigma}^* = Q_1^\varsigma$, when $Q_\Gamma^\varsigma \geq \lfloor Q_2^\varsigma \rfloor$.

Figure 2.14: Changing c (part 1)



$\lambda = 600;\ \theta = 0.8;\ b_1 = 1;\ b_2 = 0.85;\ g = 0.5;\ \delta = 0.2$

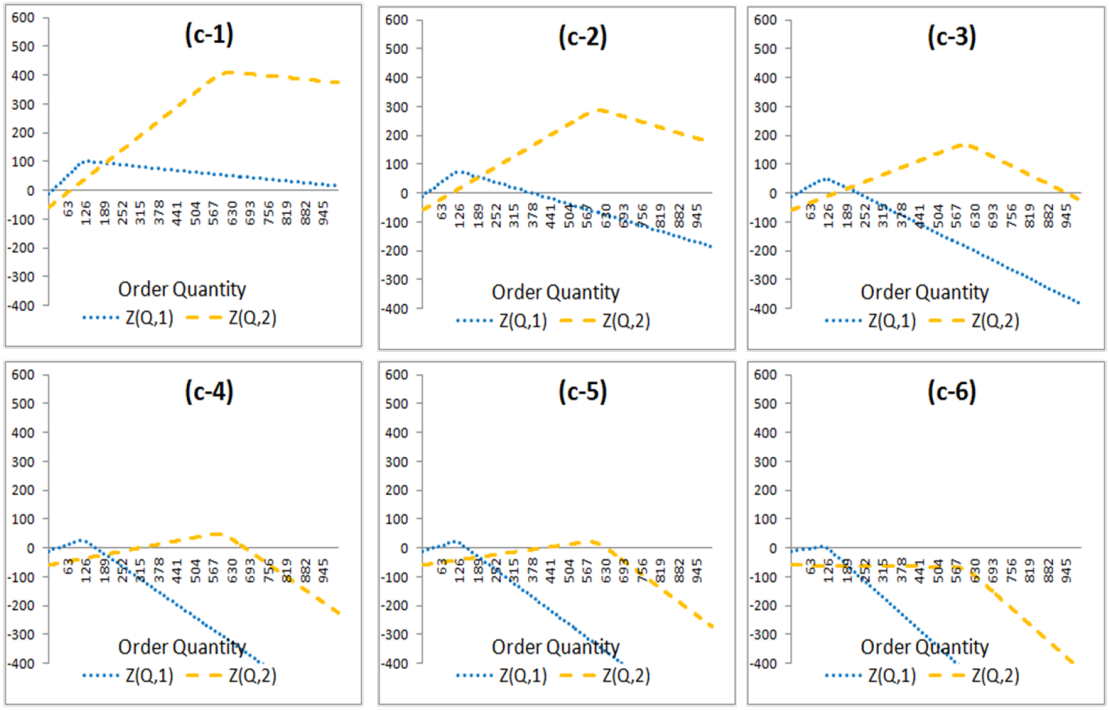
(a-1)$c = 0.15$; (a-2)$c = 0.35$; (a-3)$c = 0.55$; (a-4)$c = 0.75$; (a-5)$\tilde{c} = 0.841265$; (a-6)$c = 0.95$.



$\lambda = 600;\ \theta = 0.8;\ b_1 = 1;\ b_2 = 0.3;\ g = 0.5;\ \delta = 0$

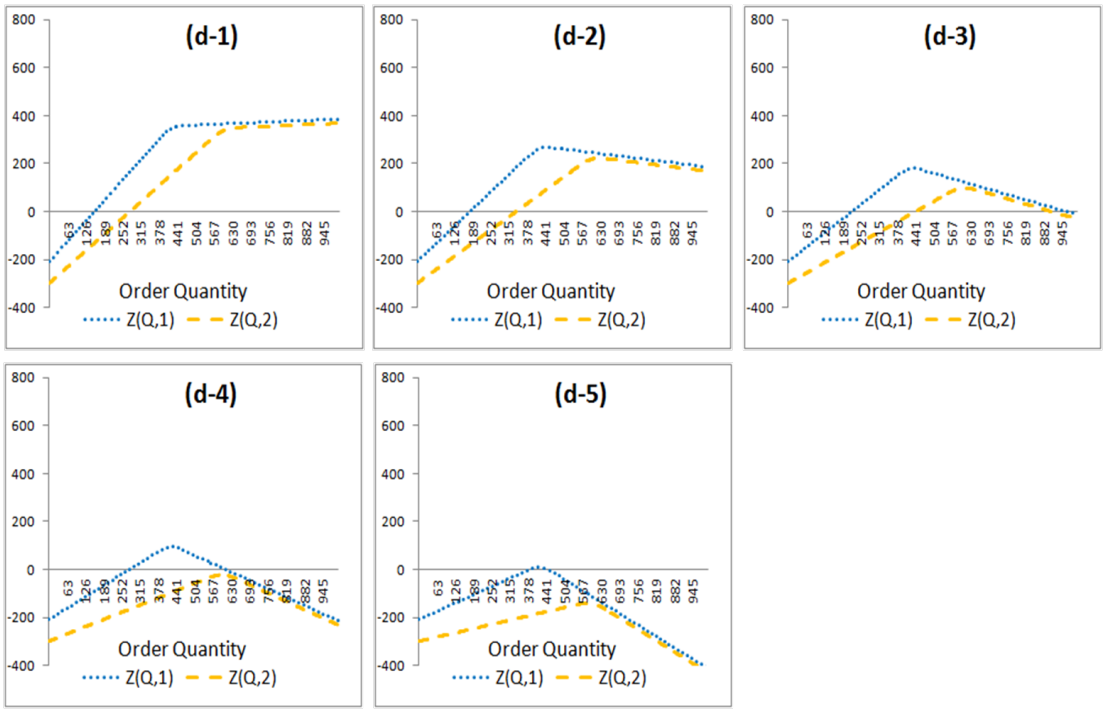
(b-1)$c = 0.15$; (b-2)$\tilde{c} = 0.29443$; (b-3)$c = 0.35$; (b-4)$c = 0.55$; (b-5)$\tilde{c} = 0.75$; (b-6)$c = 0.95$.

Figure 2.15: Changing c (part 2)



$\lambda = 600;\ \theta = 0.2;\ b_1 = 1;\ b_2 = 0.8;\ g = 0.05;\ \delta = 0.2$

(c-1)$c = 0.15$; (c-2)$c = 0.35$; (c-3)$c = 0.55$; (c-4)$c = 0.75$; (c-5)$\tilde{c} = 0.797108$; (c-6)$c = 0.95$.



$\lambda = 600;\ \theta = 0.7;\ b_1 = 1;\ b_2 = 0.1;\ g = 0.5;\ \delta = 0.2$

(d-1)$c = 0.15$; (d-2)$c = 0.35$; (d-3)$c = 0.55$; (d-4)$c = 0.75$; (d-5)$c = 0.95$; $\nexists \tilde{c}$

Proposition 3 incorporates the budget and the cost-effectiveness constraints to the results from Proposition 2. Note that the assumptions: $\bar{Q}^{\varsigma}_{\tau},\ \tau = 1, 2$, exist, and $Q^{\varsigma}_{\Gamma} \geq \underline{Q}^{\varsigma}$ are only there to restrict the attention to the (non-obvious) situations where there is in fact a trade-off because both access levels have at least one feasible solution. One major implication is that since it has been numerically observed that $\tilde{c}$ approaches b_2 from below, then $c > b_2$ will typically result in a restricted access policy. This contrasts with the result presented in Theorem 1 where patients whose health benefits were lower than the drug's cost could be subsidized; system's expected utility maximization is therefore more consistent with stratified cost-effectiveness policies.

Additionally, it is instructive to point out that the reason for which $c > \tilde{c}$ is not a necessary condition for restricted access to be the preferred policy is that the allowable budget may be sufficiently low to limit the implementation of the order quantity that maximizes the system's expected utility under full access. In other words, when $Q^{\varsigma}_{\Gamma} < \bar{Q}^{\varsigma}_{2}$, then the range of values for the cost c under which restricted access is preferred, (weakly) increases. Theorem 2 summarizes the conclusions from this section's analysis, and Corollary 2 points out the main consequences of maximizing the system's expected utility versus maximizing expected social welfare. A discussion of the results is provided below, along with a graphical representation (Figures 2.16 and 2.17) of the key findings.

Theorem 2: Let $Q^{\varsigma}_{\Gamma} \geq \underline{Q}^{\varsigma}$. When the integrated supply chain maximizes the system's expected utility subject to budget and cost-effectiveness constraints, the optimal solution is as follows.

a) Suppose q exists.

a1) $Q^{\varsigma}_{\Gamma} < \max\{q\,,\, \underline{Q}^{\varsigma}_{2}\}$, or $c > \tilde{c}$, are both sufficient conditions for $\tau^{*}_{H,\varsigma} = 1$ and $Q^{*}_{H,\varsigma} = \min\{Q^{\varsigma}_{\Gamma}\,,\, \lfloor Q^{\varsigma}_{1} \rfloor\}$ to be the unique optimal solution.

a2) $c < \tilde{c}$ is a necessary condition for $\tau^{*}_{H,\varsigma} = 2$ and $Q^{*}_{H,\varsigma} = \lfloor Q^{\varsigma}_{2} \rfloor$ to be the unique optimal

solution.

a3) Jointly satisfying $Q_{\Gamma}^{\varsigma} \geq \lfloor Q_2^{\varsigma} \rfloor$ and $c < \tilde{c}$ is a sufficient condition for $\tau_{H,\varsigma}^* = 2$ and $Q_{H,\varsigma}^* = \lfloor Q_2^{\varsigma} \rfloor$ to be the unique optimal solution.

a4) Jointly satisfying $(Q_{\Gamma}^{\varsigma} \geq \lfloor Q_2^{\varsigma} \rfloor)$ and $(c = \tilde{c})$, is a necessary and sufficient condition for the decision-maker to be indifferent between: $[\tau_{H,\varsigma}^* = 1$ and $Q_{H,\varsigma}^* = \lfloor Q_1^{\varsigma} \rfloor]$ and $[\tau_{H,\varsigma}^* = 2$ and $Q_{H,\varsigma}^* = \lfloor Q_2^{\varsigma} \rfloor]$.

a5) If $\lfloor Q_2^{\varsigma} \rfloor > Q_{\Gamma}^{\varsigma} \geq \max\{q\,,\, \underline{Q}_2^{\varsigma}\}$ and $c < \tilde{c}$, then the result is ambiguous.

b) Suppose q does not exist; then $\tau_{H,\varsigma}^* = 1$, and $Q_{H,\varsigma}^* = \min\{Q_{\Gamma}^{\varsigma}\,,\, \lfloor Q_1^{\varsigma} \rfloor\}$.

Corollary 2: When the integrated supply chain maximizes the system's expected utility subject to budget and cost-effectiveness constraints:

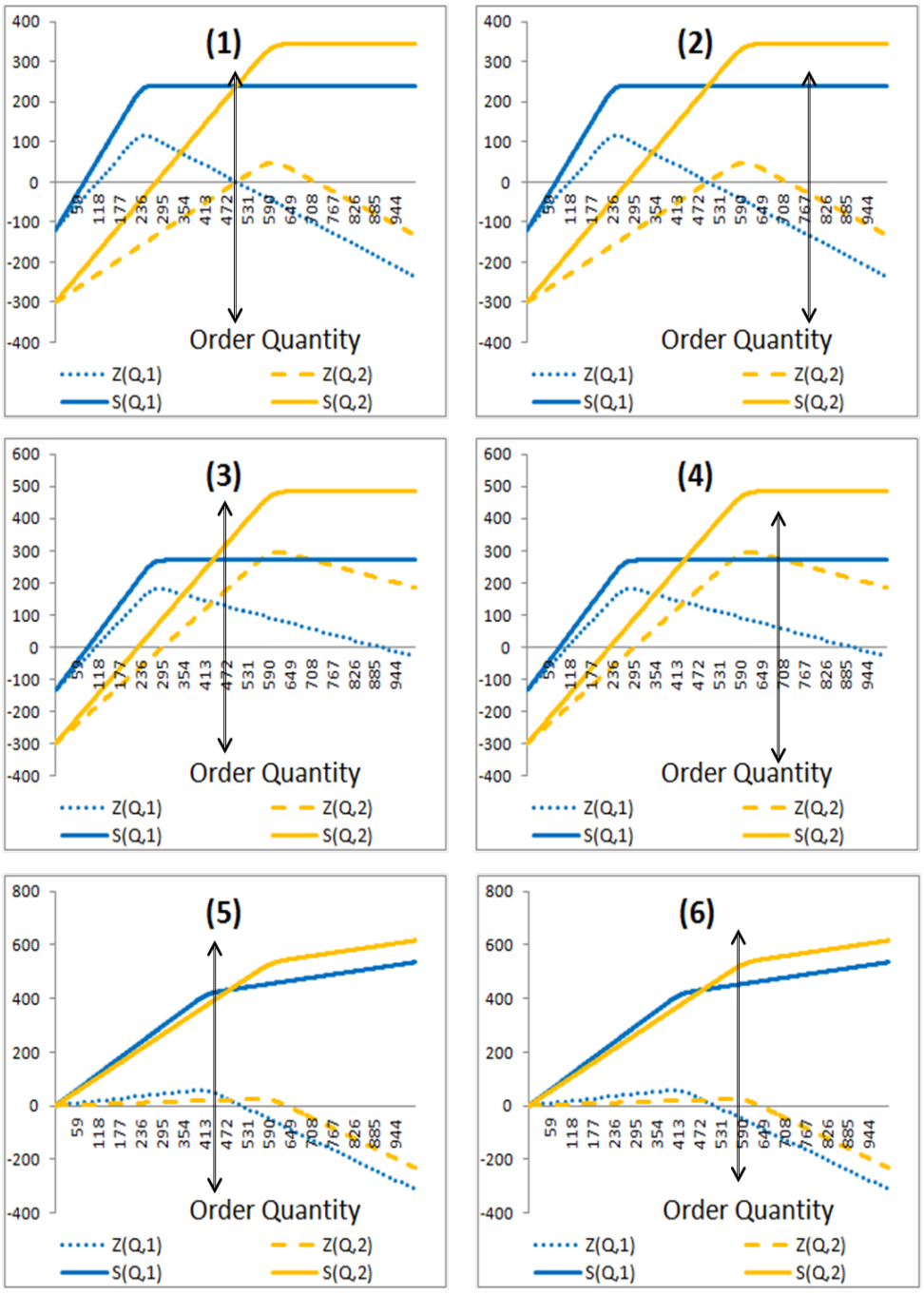
a) $\tau_{S,\varsigma}^* = 1$ is a sufficient, but not necessary, condition for $\tau_{H,\varsigma}^* = 1$.

b) $\tau_{S,\varsigma}^* = 2$ is a necessary, but not sufficient, condition for $\tau_{H,\varsigma}^* = 2$.

c) $Q_{H,\varsigma}^* \leq Q_{S,\varsigma}^*$.

One of the key implications from the above results is that the combination of parameters for which the full access policy is chosen decreases with respect to the social welfare maximization case. The result is obtained by comparing Theorem 2a2 versus Theorem 1a3. For example, under social welfare maximization, as the budget increased, the probability of choosing the full access policy was only limited by cost-effectiveness; however, under system's expected utility maximization, even as $\Gamma \to \infty$, full access policy necessarily requires $c \leq \tilde{c}$, including the indifference case. Figures 2.16 and 2.17 can facilitate the processing of the analytical results by observing how and when did the optimal decision change in relation to Figures 2.10 and 2.11. Graphs (2), (3), (6), (9), (10), and (12) all show strict decreases in the access level policy, while for graph (1), choosing restricted access has become the unique optimal solution; it's worth mentioning that from all the latter, only in graph (3) the budget acts as a limiting constraint. Going deeper, graphs (3) and (8) represent the situations

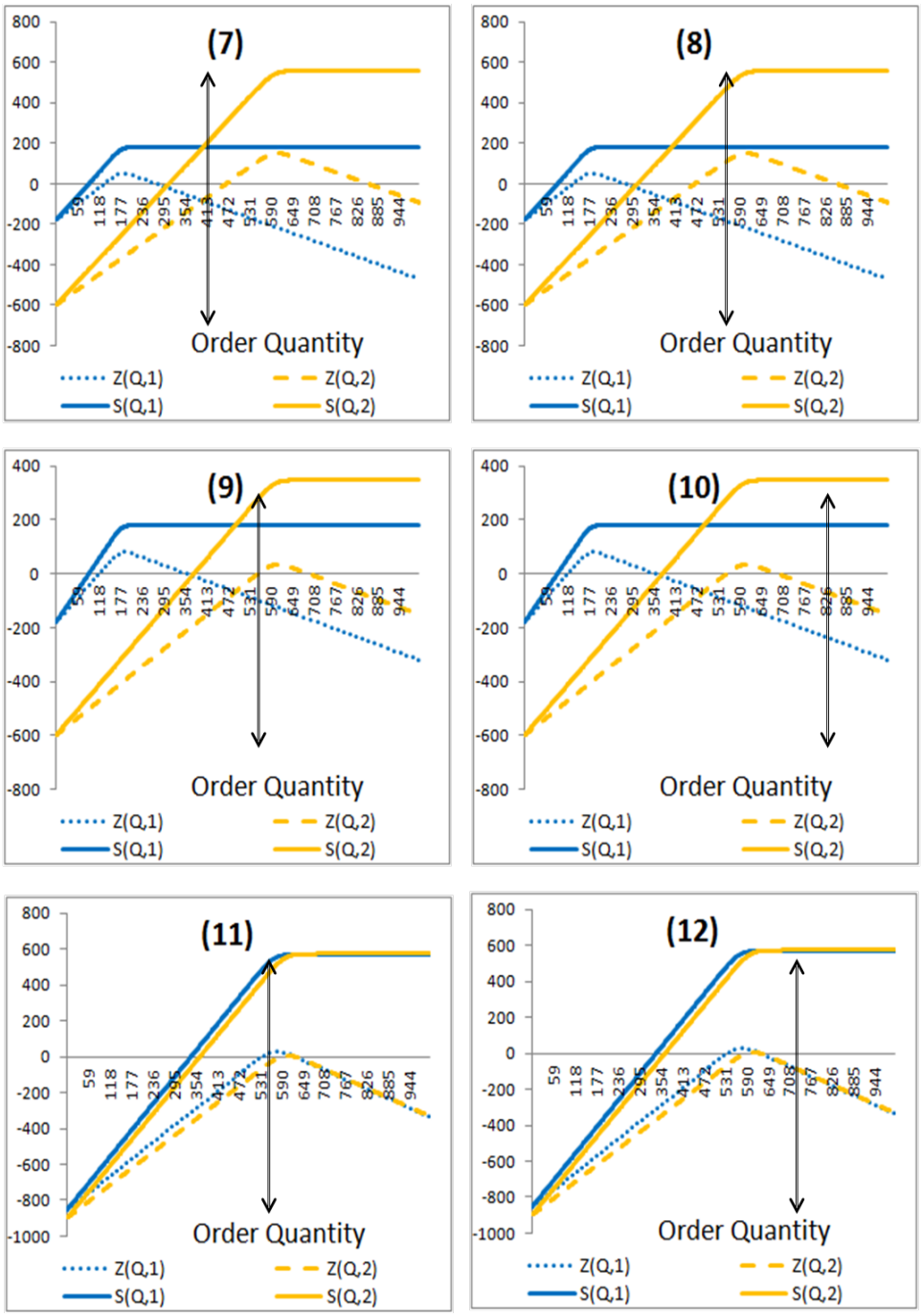
Figure 2.16: Optimal decision making under system's expected utility maximization (part 1)



$\updownarrow$ represents Q_{Γ}^{ς}

(1) $\lambda = 600; \theta = 0.4; b_1 = 1; b_2 = 0.29; g = 0.5; \delta = 0; c = 0.477;$ $\tau_{H,\varsigma}^{*}$ =1 and $Q_{H,\varsigma}^{*} = Q_{1}^{\varsigma}$
(2) $\lambda = 600; \theta = 0.4; b_1 = 1; b_2 = 0.29; g = 0.5; \delta = 0; c = 0.477;$ $\tau_{H,\varsigma}^{*} = 1$ and $Q_{H,\varsigma}^{*} = Q_{1}^{\varsigma}$
(3) $\lambda = 600; \theta = 0.45; b_1 = 1; b_2 = 0.65; g = 0.5; \delta = 0; c = 0.3;$ $\tau_{H,\varsigma}^{*} = 1$ and $Q_{H,\varsigma}^{*} = Q_{\Gamma}^{\varsigma}$
(4) $\lambda = 600; \theta = 0.45; b_1 = 1; b_2 = 0.65; g = 0.5; \delta = 0; c = 0.3;$ $\tau_{H,\varsigma}^{*} = 2$ and $Q_{H,\varsigma}^{*} = Q_{2}^{\varsigma}$
(5) $\lambda = 600; \theta = 0.7; b_1 = 1; b_2 = 0.65; g = 0; \delta = 0.2; c = 0.85;$ $\tau_{H,\varsigma}^{*} = 1$ and $Q_{H,\varsigma}^{*} = Q_{1}^{\varsigma}$
(6) $\lambda = 600; \theta = 0.7; b_1 = 1; b_2 = 0.65; g = 0; \delta = 0.2; c = 0.85;$ $\tau_{H,\varsigma}^{*} = 1$ and $Q_{H,\varsigma}^{*} = Q_{1}^{\varsigma}$

Figure 2.17: Optimal decision making under system's expected utility maximization (part 2)



$\updownarrow$ represents Q_{Γ}^{ς}

(7) $\lambda = 600; \theta = 0.3; b_1 = 1; b_2 = 0.9; g = 1; \delta = 0; c = 0.65;$ $\tau_{H,\varsigma}^{*} = 1$ and $Q_{H,\varsigma}^{*} = Q_{1}^{\varsigma}$
(8) $\lambda = 600; \theta = 0.3; b_1 = 1; b_2 = 0.9; g = 1; \delta = 0; c = 0.65;$ $\tau_{H,\varsigma}^{*} = 1$ and $Q_{H,\varsigma}^{*} = Q_{\Gamma}^{\varsigma}$
(9) $\lambda = 600; \theta = 0.3; b_1 = 1; b_2 = 0.4; g = 1; \delta = 0; c = 0.5;$ $\tau_{H,\varsigma}^{*} = 1$ and $Q_{H,\varsigma}^{*} = Q_{1}^{\varsigma}$
(10) $\lambda = 600; \theta = 0.3; b_1 = 1; b_2 = 0.4; g = 1; \delta = 0; c = 0.5;$ $\tau_{H,\varsigma}^{*} = 1$ and $Q_{H,\varsigma}^{*} = Q_{1}^{\varsigma}$
(11) $\lambda = 600; \theta = 0.95; b_1 = 1; b_2 = 0.12; g = 1.5; \delta = 0; c = 0.9055;$ $\tau_{H,\varsigma}^{*} = 1$ and $Q_{H,\varsigma}^{*} = Q_{1}^{\varsigma}$
(12) $\lambda = 600; \theta = 0.95; b_1 = 1; b_2 = 0.12; g = 1.5; \delta = 0; c = 0.9055;$ $\tau_{H,\varsigma}^{*} = 1$ and $Q_{H,\varsigma}^{*} = Q_{1}^{\varsigma}$

defined by Theorem 2a5. Intuitively, for Theorem 2a5 as c increases, Q^{ς}_{Γ} decreases, and c approaches $\tilde{c}$, increasing the the likelihood that access will be restricted; having a large gap between b_2 and c has higher likelihood of inducing full access. Taking a graphical approach based on Figures 2.16 and 2.17, it can be observed that while in (3) the budget is not large enough to justify full access, it is indeed so for graph (8) despite being forced to order less than the quantity that maximizes the unconstrained problem; the latter graph is also an appropriate example for why Theorem 2a3 does not provide a necessary condition.

The other change of interest is that of the order quantity. It is easy to see from Figures 2.16 and 2.17 how the optimal order quantity is (weakly) reduced in relation to the social welfare maximization case. The most frequent situation when $Q^{*}_{H,\varsigma} = Q^{*}_{S,\varsigma}$ given $\tau^{*}_{H,\varsigma} = 2$, is while being in the region defined by Theorem 2a5 such that $0 \leq Z(Q^{\varsigma}_{1}, 1) \leq Z(Q^{\varsigma}_{\Gamma}, 2) \leq Z(Q^{\varsigma}_{2}, 2)$; similarly, the most frequent situation when $Q^{*}_{H,\varsigma} = Q^{*}_{S,\varsigma}$ given $\tau^{*}_{H,\varsigma} = 1$, is when $0 \leq Z(Q^{\varsigma}_{\Gamma}, 2) \leq Z(Q^{\varsigma}_{1}, 1)$.

2.4 Extension for more than two types of patients

In this section, we describe a basic heuristic for solving the problem when there are $I > 2$ types of patients. The previous definitions are directly extended to this situation, *i.e.,* let b_i represent the health benefits for type i patients, and assume $b_1 < b_2 < \ldots b_I$. It is true that for most situations within the context described, the number of patient categories for which a drug can provide relevant health benefits is not expected to be very large. As a result, doing a sequential pairwise comparison wouldn't be a time consuming task. This would imply comparing types 1 and 2 and finding an optimal solution; then comparing the optimal access level between them against including type 3 patients, and so on. However, for the case when c is fixed, we are able to define more efficient algorithms contingent on the problem's structure. We will do so first for the social welfare maximizer, and then for the total utility maximizer.

2.4.1 Maximizing the expected social welfare given $I > 2$ types of patients

Based on the analysis from §2.3, we will use the additional definition for the crossing point between the expected social welfare curves of two access level policies.

Definition 1: Let $q_{i,j},\ 1 \leq i < j \leq I$, be a positive order quantity such that $S(q_{i,j}, i) \geq S(q_{i,j}, j)$; and $S(q_{i,j}+1, j) > S(q_{i,j}+1, i)$.

When the decision maker is trying to maximize the expected social welfare, Lemma 6 provides a key result which simplifies the calculations.

Lemma 6: For $1 \leq i \leq (I-2)$, $\min\{q_{i,i+1}, q_{i+1,i+2}\} \leq q_{i,i+2} \leq \max\{q_{i,i+1}, q_{i+1,i+2}\}$.

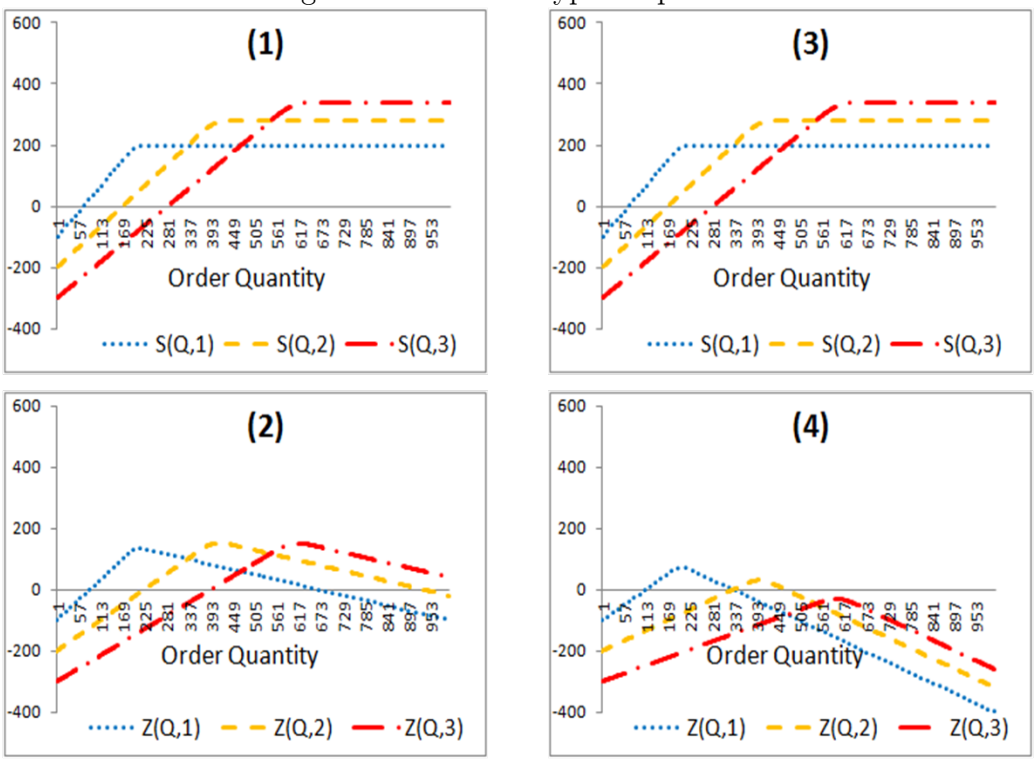
The usefulness of Lemma 6 is that not all the crossing points need to be calculated in order to determine the access level with the highest social welfare for any given order quantity. To be more specific, only the crossing points between consecutive access levels are necessary. Also, recall from Proposition 1 and Theorem 1, that when there is no crossing point between the curves of two policies, then the policy with greater access is dominated by the one with more restricted access. Since such result is now trivial, this section is only concerned with the situations where further analysis is necessary to determine the optimal solution. Next, Lemma 7 gives a condition which may further reduce the set of potential optimal solutions for the access level.

Lemma 7: Let $1 < i < I$.

a) if $q_{i,i+1} < q_{i-1,i+1} < q_{i-1,i}$, then the expected social welfare under access level i is dominated by either access levels $(i-1)$ or access level $(i+1)$, for any Q.

b) otherwise, if $q_{i-1,i} \leq q_{i-1,i+1} \leq q_{i,i+1}$, then no access level $(i-1)$, i, and $(i+1)$, is

Figure 2.18: Three Types of patients



(1) and (2) $\lambda = 600; f(1) = f(2) = f(3) = 1/3; b_1 = 1; b_2 = 0.4; b_3 = 0.3; g = 0.5; \delta = 0; c = 0.3$
(3) and (4) $\lambda = 600; f(1) = f(2) = f(3) = 1/3; b_1 = 1; b_2 = 0.4; b_3 = 0.3; g = 0.5; \delta = 0; c = 0.6$

dominated by the other two.

When Lemma 7a is satisfied for any i between 1 and I, then access level i can be discarded from the calculations. However, it must be noted that we were neither able to generate a combination of parameters that corresponds with this case, nor prove that this case can't occur; for this reason, it is included in the results, but will not be the focus of the analysis henceforth. In contrast, Lemma 7b represents the typical behavior of the expected social welfare and system's utility curves under different access level policies. Figure 2.18 provides a graphical representation for two different costs, c. Under this situation, before considering the cost-effectiveness and budget constraints, every access level policy remains a feasible candidate for optimality under expected social welfare maximization. As a result, we propose an algorithm to reach the optimal solution without having to solve completely for every access level policy. The main value of the algorithm is that it allows the results of §2.3 to

be extended to multiple types. While a trial and error approach would also be feasible, at this point it should be clear that the different ways in which the social welfare and total expected utility curves shift relative to changes in the parameters and decision variables makes it highly complicated to correctly anticipate the system's dynamics, and therefore its optimal solution, without full enumeration.

Proposition 4: Suppose $q_{i-1,i} \leq q_{i,i+1}$, for $1 < i < I$. The solution algorithm for maximizing expected social welfare is as follows.

1) Set $q_{k-1,k} = \hat{q}$, where $\hat{q}$ is given by equation (2.4.1).

2.1) If $q_{k-1,k} = \hat{q}$, calculate $\underline{Q}_k^{\varsigma}$, and $\bar{Q}_k^{\varsigma}$, and go to step 3.1.

2.2) Otherwise, calculate $q_{k-1,k}$, $\underline{Q}_k^{\varsigma}$, and $\bar{Q}_k^{\varsigma}$, and go to step 3.1.

3.1) If $\min\{\bar{Q}_k^{\varsigma}\,,\, Q_\Gamma^{\varsigma}\} \geq \max\{q_{k-1,k}\,,\, \underline{Q}_k^{\varsigma}\}$, then $\tau_{S,\varsigma}^* = k$ and $Q_{S,\varsigma}^* = \min\{Q_\Gamma^{\varsigma}\,,\, \bar{Q}_k^{\varsigma}\}$ is an optimal solution. Go to step 4.1.

3.2) Otherwise, set $k = (k-1)$, and go back to step 2.

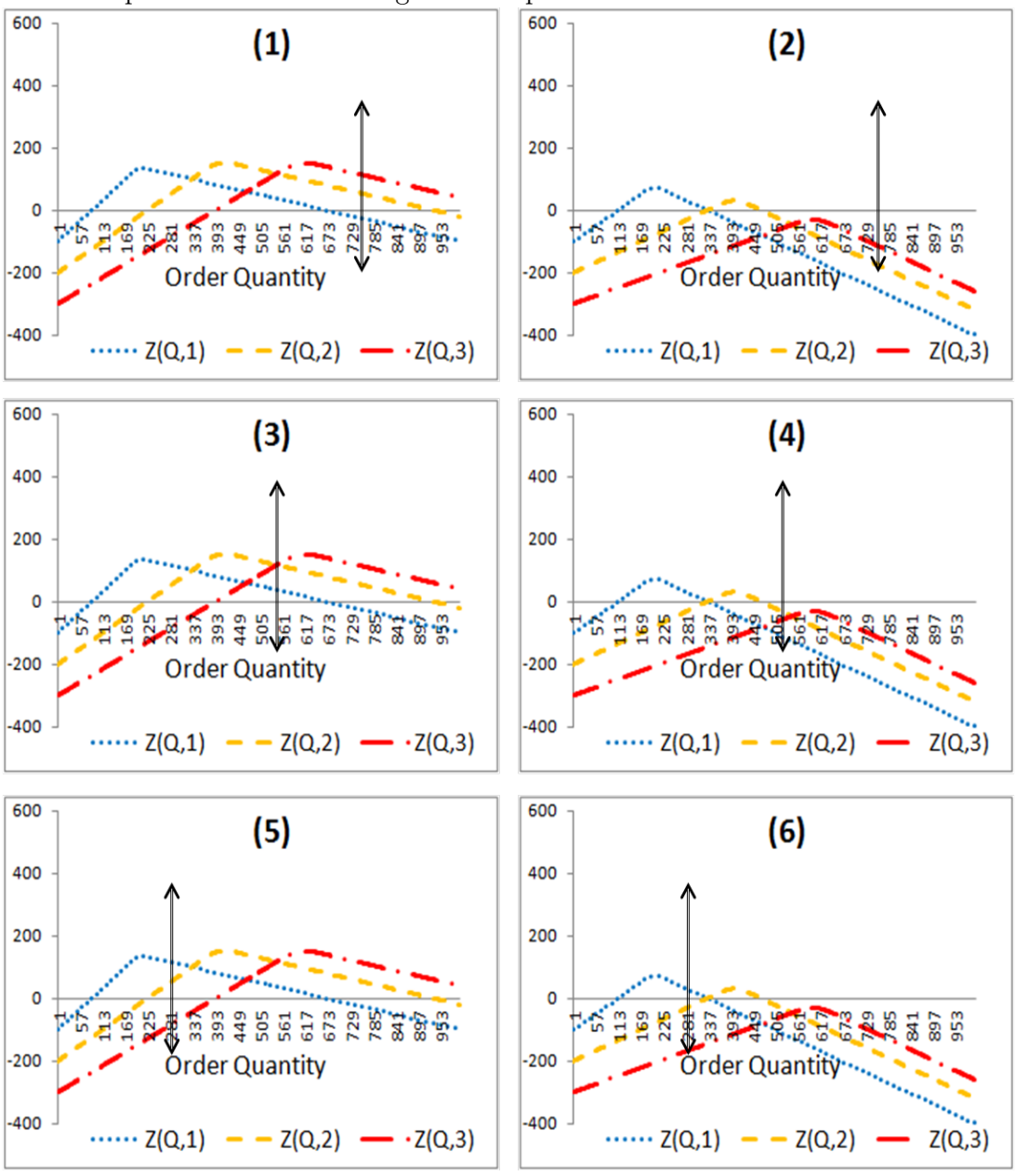
4.1) If $(\min\{\bar{Q}_k^{\varsigma}\,,\, Q_\Gamma^{\varsigma}\} > q_{k-1,k})$ and $(\min\{\bar{Q}_k^{\varsigma}\,,\, Q_\Gamma^{\varsigma}\} \geq \underline{Q}_k^{\varsigma})$, then $\tau_{S,\varsigma}^* = k$ and $Q_{S,\varsigma}^* = \min\{Q_\Gamma^{\varsigma}\,,\, \bar{Q}_k^{\varsigma}\}$ is the unique optimal solution. END.

4.2) Otherwise, the decision maker is indifferent between $(\tau_{S,\varsigma}^* = (k-1)$ and $Q_{S,\varsigma}^* = \min\{Q_\Gamma^{\varsigma}\,,\, \bar{Q}_{k-1}^{\varsigma}\})$ versus $(\tau_{S,\varsigma}^* = k$ and $Q_{S,\varsigma}^* = \min\{Q_\Gamma^{\varsigma}\,,\, \bar{Q}_k^{\varsigma}\})$. END.

$$\hat{q} \triangleq \begin{cases} \max\{\, q \in \{q_{1,2}, \ldots, q_{I-1,I}\} \,|\, q \leq Q_\Gamma^{\varsigma}\,\}, & \text{if } q_{1,2} \leq Q_\Gamma^{\varsigma} \\ q_{1,2} & \text{otherwise} \end{cases} \tag{2.4.1}$$

Proposition 4 integrates the results from Lemma 6 and Lemma 7 with the constraints faced by the single decision maker. Figure 2.19, is useful in observing how the algorithm works. Note that graphs (1), (3) and (5) only vary in the (decreasing) value of the budget constraint; the same applies to graphs (2), (4), and (6). For example, the algorithm path

Figure 2.19: Optimal decision making under expected social welfare maximization for $I = 3$



$\updownarrow$ represents Q^{ς}_{Γ}

(1), (3), and (5): $\lambda = 600; f(1) = f(2) = f(3) = 1/3; b_1 = 1; b_2 = 0.4; b_3 = 0.3; g = 0.5; \delta = 0; c = 0.3$
(2), (4), and (6): $\lambda = 600; f(1) = f(2) = f(3) = 1/3; b_1 = 1; b_2 = 0.4; b_3 = 0.3; g = 0.5; \delta = 0; c = 0.6$
(1) $\hat{q} = q_{2,3}$; $\tau^*_{S,\varsigma} = 3$ and $Q^*_{S,\varsigma} = Q^{\varsigma}_{\Gamma}$
(2) $\hat{q} = q_{2,3}$; $\tau^*_{S,\varsigma} = 2$ and $Q^*_{S,\varsigma} = \bar{Q}^{\varsigma}_{2}$
(3) $\hat{q} = q_{2,3}$; $\tau^*_{S,\varsigma} = 2$ or 3, and $Q^*_{S,\varsigma} = Q^{\varsigma}_{\Gamma}$
(4) $\hat{q} = q_{2,3}$; $\tau^*_{S,\varsigma} = 2$ and $Q^*_{S,\varsigma} = Q^{\varsigma}_{\Gamma}$
(5) $\hat{q} = q_{1,2}$; $\tau^*_{S,\varsigma} = 1$ or 2, and $Q^*_{S,\varsigma} = Q^{\varsigma}_{\Gamma}$
(6) $\hat{q} = q_{1,2}$; $\tau^*_{S,\varsigma} = 1$ or 2, and $Q^*_{S,\varsigma} = Q^{\varsigma}_{\Gamma}$

followed in graph (1) is steps: $1 \to 2 \to 3.1 \to 4.1$, which results in a unique solution. Instead, graph (2) follows the path: $1 \to 2.1 \to 3.2 \to 2.2 \to 3.1 \to 4.1$, which also results in a

unique solution; the difference is that the algorithm had to loop because the cost-effectiveness constraint was not satisfied for the feasible range of order quantities. The path followed in graph (4) is the same. Graph (3) shows the situation where the decision maker is indifferent between two access levels, following the path: $1 \rightarrow 2.1 \rightarrow 3.1 \rightarrow 4.2$. Finally, graphs (5) and (6) show the case where $q_{1,2} > Q_{\Gamma}^{\varsigma}$, and therefore $\hat{q} = q_{1,2}$; the algorithm then follows the path $1 \rightarrow 2.1 \rightarrow 3.1 \rightarrow 4.1$ which results in a unique solution.

2.4.2 Maximizing the system's expected utility function given $I > 2$ types of patients

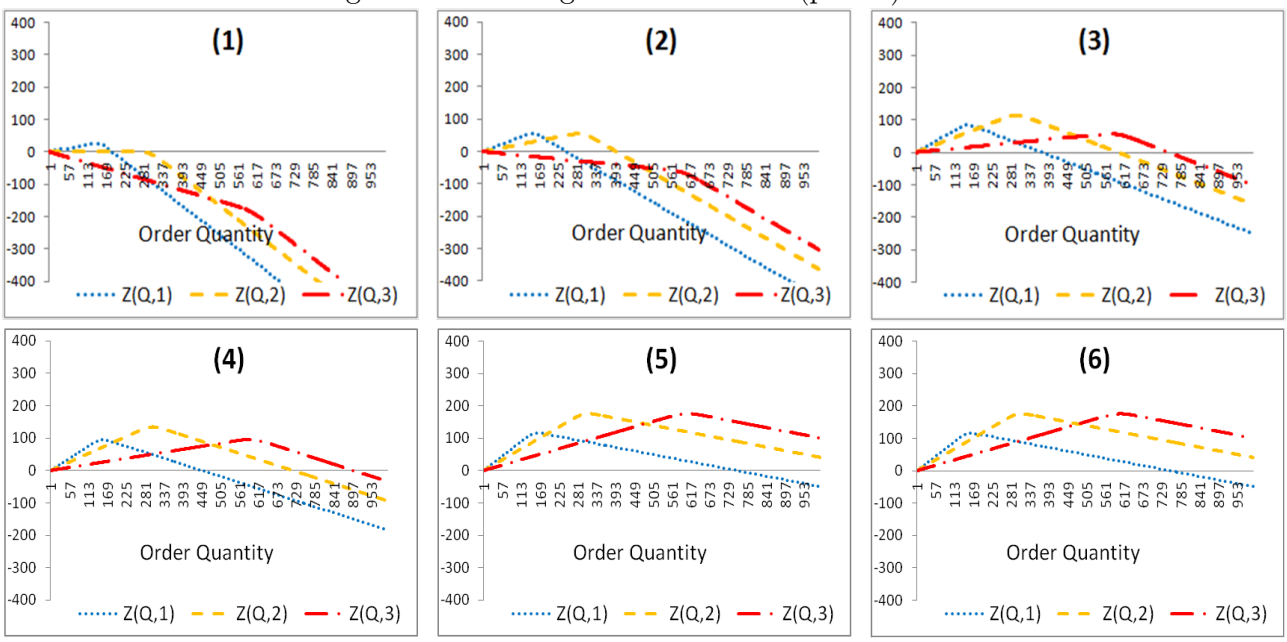
Now we look at the problem when the decision maker is concerned with maximizing the system's expected utility. Just like in §2.4.1 we had used the crossing point between the utility curves to simplify the analysis, in this part we use the key threshold that had been derived earlier in §2.3.3 to determine which access level policy would achieve a higher expected utility for the system.

Definition 2: Let $\tilde{c}_{i,j} = \left\{ c \;\middle|\; Z(Q_i^{\varsigma}, i) = Z(Q_j^{\varsigma}, j);\ 1 \leq i < j \leq I \right\}$.

Lemma 8: For $1 < i < I$, $\tilde{c}_{i-1,i} > \tilde{c}_{i,i+1}$.

Lemma 8 can be inferred from the previously stated observation that $\tilde{c}_{i-1,i} \rightarrow b_i$ for most parameter combinations; since b_i is decreasing in i, then the required selling price for equating the expected utility under two consecutive access level policies is also decreasing in i. The main implication is that if $c > \tilde{c}_{i,i+1}$, then $c > \tilde{c}_{j,j+1}$, $1 \leq i < j < I$, *i.e.,* when the selling price is sufficiently high for the more restricted policy to dominate in a consecutive pairwise combination, then that restricted policy also dominates the rest of the more inclusive access level policies. The result can be observed graphically in Figures 2.20 and 2.21, and is further

Figure 2.20: Changes in c for $I = 3$ (part 1)

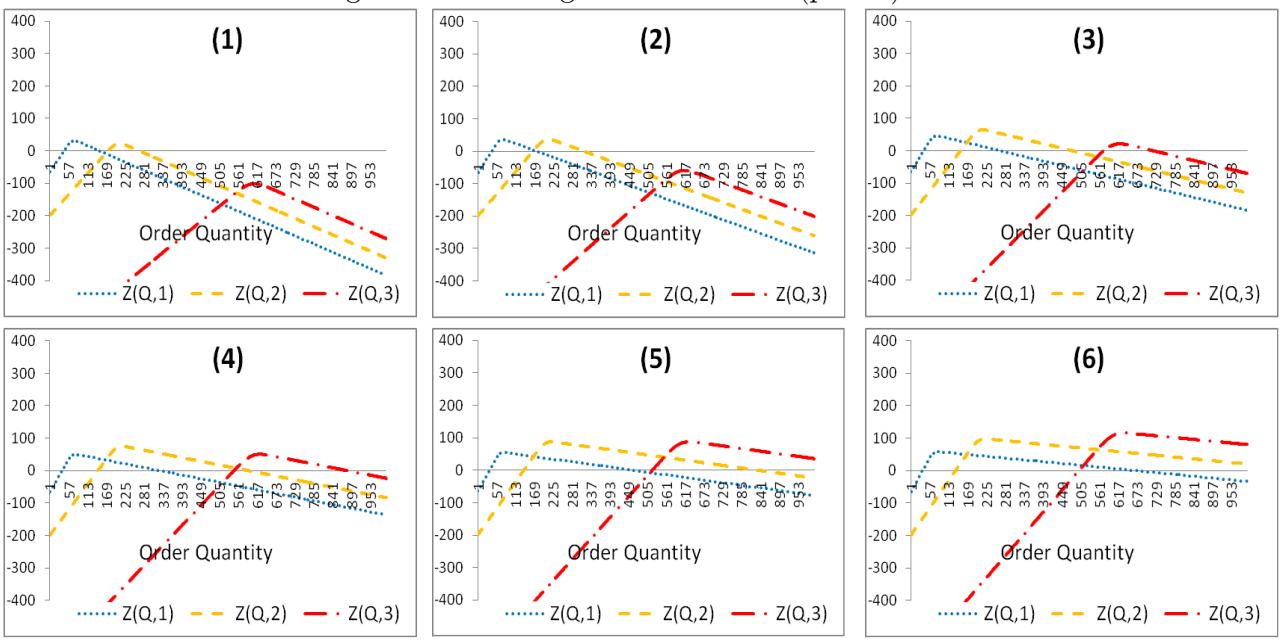


$\lambda = 600; f(1) = 0.25; f(2) = 0.25; f(3) = 0.5; b_1 = 1; b_2 = 0.6; b_3 = 0.2; g = 0; \delta = 0$

$(1)c = 0.8;\ (2)\tilde{c}_{1,2} = 0.60241;\ (3)c = 0.4;\ (4)\tilde{c}_{1,3} = 0.33342;\ (5)\tilde{c}_{2,3} = 0.19895;\ (6)c = 0.05.$

expanded in Lemma 9.

Figure 2.21: Changes in c for $I = 3$ (part 2)



$\lambda = 600; f(1) = 1/9; f(2) = 2/9; f(3) = 2/3; b_1 = 1; b_2 = 0.4; b_3 = 0.15; g = 1; \delta = 0$

$(1)c = 0.45;\ (2)\tilde{c}_{1,2} = 0.38090;\ (3)c = 0.25;\ (4)\tilde{c}_{1,3} = 0.20365;\ (5)\tilde{c}_{2,3} = 0.14416;\ (6)c = 0.1.$

Lemma 9: For $1 < i < I$, suppose $\tilde{c}_{i-1,i} > c > \tilde{c}_{i,i+1}$. Then $Z(Q_i^{\varsigma}, i) > Z(Q_j^{\varsigma}, j)$, $1 < j < I,\ j \neq i$.

In addition to verifying the dominance towards more inclusive access levels, Lemma 9 states the dominance over more restricted access level policies. To understand the relevance of the result, it is useful to note that in the unconstrained problem, $\tilde{c}_{i-1,i} > c > \tilde{c}_{i,i+1}$ implies that the system's expected policy is maximized at access level i and order quantity Q_i^{ς}. Furthermore, it is implied that when trade occurs, if the latter values for the access level and order quantity are not the solution to the constrained problem, then the limiting constraint must be the available budget. Clearly, if the unconstrained solution didn't satisfy the cost-effectiveness budget, then no solution would satisfy it. Leveraging on this intuition and previously derived results, Proposition 5 provides an algorithm to find the optimal solution under expected system's utility maximization.

Proposition 5: The solution algorithm for maximizing system's expected utility is as follows.

1) Set $k = \hat{\tau}$, where $\hat{\tau}$ is given by equation (2.4.2).

2) If $k = 1$, then go to step 3.1. Otherwise, go to step 3.2.

3.1) $\tau_{H,\varsigma}^* = 1$ and $Q_{H,\varsigma}^* = \min\{Q_{\Gamma}^{\varsigma}, \lfloor Q_1^{\varsigma} \rfloor\}$ is the unique optimal solution. END.

3.2) Calculate $\underline{Q}_k^{\varsigma}$ and $q_{k-1,k}$. Go to step 4.

4) If $Q_{\Gamma}^{\varsigma} < \max\{q_{k-1,k}\,,\, \underline{Q}_k^{\varsigma}\}$, then set $k = k - 1$, and go back to step 2; otherwise go to step 5.

5) Calculate Q_k^{ς}. If $Q_{\Gamma}^{\varsigma} \geq Q_k^{\varsigma}$, go to step 6.1; otherwise go to step 6.2.

6.1) The result is ambiguous between ($\tau_{H,\varsigma}^* = k$ and $Q_{H,\varsigma}^* = Q_{\Gamma}^{\varsigma}$) and ($\tau_{H,\varsigma}^* = k - 1$ and $Q_{H,\varsigma}^* = \lfloor Q_{k-1}^{\varsigma} \rfloor$). END.

6.2) If $c < \tilde{c}_{k-1,k}$, $\tau_{H,\varsigma}^* = k$ and $Q_{H,\varsigma}^* = \lfloor Q_k^{\varsigma} \rfloor$ is the unique optimal solution; and if $c = \tilde{c}_{k-1,k}$, then the decision maker is indifferent between ($\tau_{H,\varsigma}^* = k$ and $Q_{H,\varsigma}^* = \lfloor Q_k^{\varsigma} \rfloor$) versus ($\tau_{H,\varsigma}^* = k - 1$ and $Q_{H,\varsigma}^* = \lfloor Q_{k-1}^{\varsigma} \rfloor$). END.

$$\hat{\tau} \triangleq \begin{cases} \max\{\, \tau \in \{2, \dots I\} \,|\, c \leq \tilde{c}_{\tau-1,\tau}\} & \text{if } \tilde{c}_{I-1,I} < c \\ I & \text{otherwise} \end{cases} \tag{2.4.2}$$

Proposition 5 incorporates the constraints to the result from Lemma 9. The main advantage of the algorithm is that pairwise comparison will occur at most once.

2.5 Exogenous Price-only contracts

Finally, in this section we briefly show the relationship between the first-best solution and the case where the pharmaceutical manufacturer sells to a health-payer through an exogenous price-only contract. This situation is fairly common, and is a consequence of reference pricing. Essentially, reference pricing may work in two ways. On one hand, internal reference pricing sets the transfer price of the drug from Pharma to Health at a price that is comparable to an existing treatment that is considered equivalent in terms of the health benefits provided. On the other hand, external reference pricing implies that the transfer price will be set using the mean or median of the existing transfer price for the same drug between Pharma and other health-payers. To illustrate the existence of this method in practice, we may consider the related case of the United Kingdom, where a transfer price is negotiated between the pharmaceutical manufacturer and the NHS. Interestingly, even though the United Kingdom represents only about 3% of the global pharmaceutical market, the transfer price that is negotiated is used by many countries in their reference pricing calculations; as a result, such price directly affects more than a quarter of the global pharmaceutical market (OFT, 2007). The latter market is therefore the one being modeled in this section. Endogenous price agreements will be the focus of the next Chapter.

We next introduce the problem when Pharma and Health act as separate decision-makers and Health makes a transfer, $T(w, Q) = wQ$, to Pharma in exchange for a delivery of Q units of the drug. We denote this contract setting with the symbol χ for exogenous. Pharma's profit function is given by:

$$\begin{aligned} M^{\chi}(w; Q^*_{j,\chi}, \tau^*_{j,\chi}) &= T(w, Q^*_{j,\chi}) - cQ^*_{j,\chi}, \\ &= (w - c)Q^*_{j,\chi}. \end{aligned} \tag{2.5.1}$$

Health's expected utility function is defined as:

$$\begin{aligned} H(Q, \tau;\, w\,) &= -T(w, Q) + B_h(\tau)A(Q, \tau) + \delta(Q - A(Q, \tau)) - g\left(\lambda F(\tau) - A(Q, \tau)\right), \\ &= -wQ + (B_h(\tau) - \delta + g)A(Q, \tau) + \delta Q - g\lambda F(\tau). \end{aligned} \tag{2.5.2}$$

The social welfare's expected utility function from Health's perspective is:

$$\begin{aligned} S_h(Q, \tau) &= B_h(\tau)A(Q, \tau) + \delta(Q - A(Q, \tau)) - g\left(\lambda F(\tau) - A(Q, \tau)\right) \\ &= (B_h(\tau) - \delta + g)A(Q, \tau) + \delta Q - g\lambda F(\tau), \end{aligned} \tag{2.5.3}$$

And the cost-effectiveness constraint also needs to be appropriately modified such that

$$H(Q, \tau;\, w\,) \geq 0. \tag{2.5.4}$$

In the above formulations, we define $Q^*_{j,\chi}$, and $\tau^*_{j,\chi}$ as the optimal order quantity and access level policy, respectively, under exogenous price-only contracts, given the maximization of objective $j = (S)$ocial welfare, or (H)ealth's expected utility function. By comparing equations 2.5.2 and 2.5.3 to equations 2.3.1 and 2.3.2, we observe that there are two modifications. First, the production and delivery cost c has been replaced by the transfer price

w. By increasing the transfer price in this way, the well-known double marginalization effect creates a decrease in the optimal service level for a utility maximizing health-payer, leading to a decrease in the order quantity. An additional effect due to the price mark-up is that the range of order quantities for which full access is feasible decreases in the selling price, which not only decreases the order quantity under full access, but makes it more likely that Health chooses to restrict access, either due to the budget constraint being active and not allowing for a feasible solution under full access, or in the case of maximizing Health's utility, because the selling price is higher than the threshold

$$\tilde{w} = \{w \mid H(Q_1^\chi, 1;\, w) = H(Q_2^\chi, 2;\, w)\}\,, \tag{2.5.5}$$

where the latter is directly adapted from Proposition 2.

The second modification is the replacement of the average expected health benefits $B(\tau)$ with the value $B_h(\tau)$. As it was noted in the introduction, a drug's health benefits are often uncertain. Pharmaceutical manufacturers are already investing in targeted therapies such that patients with the same symptoms receive different treatments that are a function of their genetic characteristics (Roche, 2011). However, the current state of the art for most medical conditions is to provide a treatment based on the patients' symptoms, which may result in high degrees of uncertainty in the expected health benefits achieved by the drug. Additionally, even after a drug is approved by a health-payer to be included under its reimbursement scheme, the manufacturer may expand its clinical trials, or collect additional information from the drug's initial introduction, in order to increase the drug's expected health benefits or to reduce the uncertainty around them. These factors combined can be interpreted as a potential informational advantage by Pharma because Health may not fully observe Pharma's ongoing R&D efforts. In response to the latter, and the added increasing pressure to control spending, Health's risk aversion may come into play such that his certainty equivalent about the drug's expected health benefits may be lower than that claimed

by Pharma. As a result, we assume $B(\tau) \geq B_h(\tau) \triangleq \sum_{i=1}^{\tau} \frac{\beta_i f(i)}{F(\tau)}$, where β_i is Health's belief on the expected health benefits for patient type i; it is assumed $b_1 \geq \beta_1 > b_2 \geq \beta_2$. This asymmetry of beliefs implies that the slope of the expected utility functions under any access level will grow at a lower rate compared to the case where both players hold the same information and beliefs about the expected health benefits. Therefore, the range of order quantities that provide a feasible solution for any access level is reduced, plus there is a (weak) decrease in the order quantity that maximizes Health's utility function for any access level - the decrease is strong if the absolute budget constraint was not originally binding. The latter effects result not only in lower levels of social welfare - because the quantity ordered is weakly smaller -, but if Health's belief asymmetry is only with respect to the lower type patients, or even with respect to both patient categories but decreasing at least as much for the lower as for the higher type patients, then the probability of access being restricted increases, as mentioned in Proposition 6g. On the other hand, when Health's belief asymmetry is only with respect to the higher type of patients, it will actually result in a higher likelihood that full access will be preferred, as explained from Proposition 6f.

Proposition 6: When $I = 2$, let q^h be a positive order quantity such that $S_h^{\chi}(q^h, 1) \geq S_h^{\chi}(q^h, 2)$; and $S_h^{\chi}(q+1, 2) > S_h^{\chi}(q+1, 1)$.

a) If inequality (2.3.3) is satisfied, then q^h is unique and given by equation (2.5.6).

$$q^h = \max\left\{ Q \;\middle|\; \left(\frac{g}{\beta_1 - \delta + g}\right)\left(\frac{\lambda}{A(Q,2)}\right) + \left(\frac{\beta_1 - \beta_2}{\beta_1 - \delta + g}\right) \geq \left(\frac{A(Q,2) - A(Q,1)}{A(Q,2)}\right)\left(\frac{1}{1-\theta}\right) \right\} \tag{2.5.6}$$

b) $\beta_2 > \delta$ is a necessary and sufficient condition for q^h to exist.

c) If $\beta_2 < \delta$, then $S_h(Q,1) > S_h(Q,2),\ \forall\ Q > 0$.

d) If $\beta_2 = \delta$, then $S_h(Q,1) > S_h(Q,2)$, for some finite $Q > 0$, and $\lim_{P(Q,\lambda)\to 0} S_h(Q,1) - S_h(Q,2) = 0$.

e) The value of q^h is increasing in $\beta_1, g,\ \delta$, and decreasing in β_2. The change in q^h with respect to θ is ambiguous.

f) If $b_1 > \beta_1$ and $b_2 = \beta_2$, then $q^h < q$.

g) If $(b_1 - \beta_1) \leq (b_2 - \beta_2)$, then $q^h > q$.

h) If $(b_1 - \beta_1) > (b_2 - \beta_2) > 0$, then the relationship between q^h and q is ambiguous.

Despite these two modifications, it is important to stress that all previous results continue to hold for the exogenous price-only case, with the aforementioned adjustments in notation. In other words, here we have shown how the analyses presented in §2.3.3 and §2.4 collapse into the exogenous price-only contract by simply modifying the transfer price and expected health benefits parameters. Additionally, we have set the base model for the analysis of endogenously defined contracting mechanisms designed by Pharma, which is the focus of Chapter 3.

2.6 Conclusions

Chapter 2 has analyzed the joint access and coverage problem in the introduction process of a new drug with multiple indications, where access level is defined as the subset of patient categories that are eligible to receive the treatment under the health-payer's reimbursement scheme, and service level is defined as the probability that the order quantity purchased by the health-payer is enough to meet patient demand during a sales period. We formulated a model using the price and quantity newsvendor framework to understand how a health-payer defines his optimal policy as a function of his decision making priority and the existing absolute and relative cost constraints. The discussion has focused on the special situation when the supply chain is vertically integrated, or equivalently from an analytical standpoint, where the transaction price between the manufacturer and the health payer is exogenously deter-

mined; such assumption best reflects those situations where external reference pricing is used.

From a methodological perspective, the first contribution is the finding of a unique crossing point between the expected social welfare for the different levels of access that allows us to quickly determine the optimal order quantity and access level under expected social welfare maximization. Comparative statics and an extensive discussion has been included to explain the direction of such threshold as a function of the cost and benefit parameters. Second, for the case when the decision maker maximizes his expected net utility we achieve a similar result by finding a threshold transfer price such that any transaction price higher than the threshold will result in restricted access; through numerical experiments, we have observed this value to be very close to the marginal health benefit of the type of patients with lower health benefits. An additional, interesting property is that each of these thresholds exists if and only if the other threshold exists as well, despite the fact that social welfare does not depend on cost. Third, based on these two thresholds, we provide an efficient solution process which relates to the price and quantity newsvendor model studied in operations management; our contribution to earlier analyses is the determination of the optimal solution when the choice of the optimal access level (which serves the same purpose as the retail price in the operations literature) is discrete, the decision space is constrained by absolute and relative cost constraints, and when the objective function can be to either maximize the expected net utility (as traditional models in operations management do) or maximize social welfare (which is the mandate for many of the relevant players in our context). Finally, we have provided a heuristic for finding the optimal solution when there are more than two patient types under minor assumptions which represent most of the feasible space.

From a policy-making perspective, we first identify situations of strong dominance of a given access level policy, independent of the available budget. An important observation is that a social welfare maximizer is prone to subsidizing patients whose expected benefits

are lower than the transfer price as long as the available budget is sufficiently high, while maximizing the health payer's expected utility will not do so. This implies that markets where external reference pricing is used to determine prices will be highly dependent on their available budget (under social welfare maximization) and on relative health benefits, demand size, and demand uncertainty (under Health's utility maximization) to determine the optimal policy, since these were the main drivers of the results of the two cases, correspondingly, in §2.3.3. Also, we find that when access level is restricted under social welfare maximization, then it will be restricted as well under net utility maximization; however, the opposite is not necessarily true. Such situations occur due to either high transfer costs (higher than the threshold value) relative to the benefit of the lower type patients, or to relatively low budget constraints. Moreover, we find that the optimal order quantity is weakly reduced under expected net utility maximization, even in the situations when the budget is infinitely high or when the access decision remains unchanged relative to social welfare maximization.

An interesting and necessary extension relates to the response function from the manufacturer's perspective. Endogenously setting the price, or entering into some risk sharing agreement are the focus of Chapter 3. Additionally, the extension of our results to more general demand distributions would be a useful validation of the intuition here provided; however, we do expect the results to be qualitatively consistent for IGFR distributions.

Chapter 3

Analyzing the value of three endogenous contracting mechanisms in the joint access and coverage problem in health care

3.1 Introduction

In Chapter 2 we have analyzed the decision-making process of a health-payer under a fixed transfer price. However, pharmaceutical manufacturers also play an important role is setting the conditions under which transactions will occur between themselves and the health-payers. Motivated either by the increasing pressures held by the payers, or by selfish profit maximization, a variety of mechanisms have been attempted in order to modify the status quo. The financial risks associated to demand uncertainty and the asymmetry between the health benefits claimed by the pharmaceutical manufacturers and those acknowledged by the health-payers are two key reasons for the introduction of new drugs aimed at treating chronic conditions to be rejected, delayed, or accepted under terms that negatively impact the player with the lowest bargaining power. On top of the budget constraints that health-

payers may have, the health benefit value they use in their calculations of cost-effectiveness, expected social welfare, and expected total utility, has a high relevance on both the access and service levels under which the drug is commercialized (if at all). By delaying introduction, health-payers wish to either reduce their uncertainty about the drug's performance in clinical practice, or negotiate a lower selling price with the manufacturer. Said manufacturer, at the initial stage of negotiation has the option of either decreasing the selling price or accepting to commercialize a lower sales volume of the drug. If neither alternative is accepted by the manufacturer, then she will collect more evidence hoping to increase her bargaining power against health-payers in the future. Some of the main problems, however, are that until the drug is accepted for introduction, the manufacturer is losing revenues, the patent clock is ticking, and the patients are not able to receive what is supposed to be the most appropriate treatment.

As a result risk sharing contracts have received increasing attention. Particularly in the United Kingdom, the discussion in academic and political environments has been very active. Pouvourville (2006) discussed the attractiveness of risk sharing contracts in managing the uncertainty surrounding a product's performance in real life and the credibility of the claims by the manufacturers, while also providing some predictability for such manufacturers. Carapinha (2008) highlighted the importance of integrating clinical, quality of life, and financial outcome measures into a risk-sharing agreement as well as the challenges of patient compliance and inefficient delivery of health care services. He also comments on the inconclusive evidence on the impact of risk sharing agreements on an individual patient's clinical and quality of life outcomes, and on their effectiveness in containing pharmaceutical expenditure. This chapter attempts to add additional insight into the value and limitations of three contracting mechanisms proposed by the manufacturer.

First we analyze the impact of endogenizing the transaction cost of the drug, and find

situations when a health-payer maximizing his expected utility may be able to negotiate a lower price, and achieve higher social welfare, than a health-payer who maximizes expected social welfare. Second, we characterize the conditions under which Pharma is willing to build capacity above Health's initial commitment, and show that this type of contracts result in a weak increase in both access level and expected social welfare. Third, we propose a new performance-based mechanism that partially reduces the negative effects of asymmetric beliefs between Pharma and Health.

The rest of the chapter proceeds as follows. In §3.2, additional literature specific to the proposed contracts is briefly addressed to complement the one presented in the previous chapter. Section 3.3 solves the endogenous price-only contracts. Section 3.4 relaxes the single ordering assumption and solves the capacity buffer contract under an exogenously set transfer price. In section 3.5 Pharma offers a performance-based contract to Health to deal with belief asymmetry. Concluding remarks are offered in §3.6.

3.2 Literature Review

The search for alternatives to manage demand and health outcome uncertainty in order to make a better use of the available and continuously decreasing resources, has produced a large volume of work in recent years, both theoretical and applied. From an applied perspective, Pugatch, Healy, and Chu (2010) provide a survey of 27 agreements between manufacturers and health payers implemented over the last two decades across five countries (United Kingdom, Italy, Australia, Germany, and the United States) for drug treatments that would have otherwise been rejected. They identify 4 mechanisms: cost caps and rebates (which are driven by price), and patient monitoring and patient compliance (which are driven by performance). It is worth mentioning that 16 out of the 27 agreements included some form of rebate in the contract's conditions. Espin, Rovira and Garcia (2011) analyze risk sharing schemes in

Europe for oncology products, which they categorize as financially-based schemes, *i..e,* price-volume agreements with paybacks or price reductions, and outcome-based schemes. They find some form of risk sharing scheme in 7 countries: Portugal, France, United Kingdom, Italy, Slovenia, Germany, and Lithuania. The main objectives of the agreements observed were to control the budget, get additional data, or finance cost-effective medicines.

The contract proposed in §3.3 aims to represent the situations where reference pricing is not used to determine the transfer price between the manufacturer and the health-payer. We base our model on the analysis of price-only contracts by Lariviere and Porteus (2001) where the conditions for the manufacturer's objective function to be unimodal in the selling price are defined in Theorem 1 (p. 296). The key differences are that in their model, they consider the retail price to be fixed, and there are no absolute nor relative budget constraints for the downstream party (the retailer in their model, the health-payer in ours). As a result, our task is to develop a method for efficiently finding the optimal transfer price, incorporating the aforementioned factors. By doing this, we also expand the works of Salinger and Ampudia (2011) and Kocabiyikoglu and Popescu (2011), providing a supply chain perspective to the analysis of the price and quantity newsvendor model, where the the upstream party is able to determine the transfer price. The main contribution is the added visibility of the relationship between the transfer price-setting process and the downstream party's optimal decisions, as a function of the objective function and constraints faced by the latter.

The contract proposed in §3.4 goes back to the exogenous price assumption, but relaxes the single order opportunity constraint of the classic newsvendor setting by allowing the upstream party to overproduce, and therefore letting the downstream party to order above and beyond its initial order quantity. The contract has its roots in two well known contracts in the supply chain coordination literature. First, the quantity flexibility contracts (Tsay, 1999), where the buyer sends a purchasing signal of size q well before observing demand,

and the manufacturer builds a stock of $q(1+\alpha)$, the buyer is committed to purchase at least $q(1-w)$, and the contract parameters are the selling price and the sales range parameters $\alpha > -w$, and $w \in [0,1]$. Second, the buyback contract (Pasternack, 1985) where the manufacturer chooses selling price w and buyback rate b, which is the price paid by the manufacturer to the buyer for every unit of overstock at the end of the demand period. Both contracts coordinate the supply chain in a wide array of scenarios, but our contract is different both in its main objective and its decision variables. The proposed contract does not explicitly guarantee a minimum capacity on the part of the manufacturer because symmetric information is assumed regarding the manufacturer's production costs; therefore the buyer is able to anticipate the manufacturer's optimal capacity. Additionally, in our model the capacity is not necessarily bounded by the manufacturer's incentive compatibility constraint as in the quantity flexibility, but rather may be limited by Health's constraints. Finally, while selling K_T units to the buyer at price w and repurchasing the excess units at price b is similar to our approach of the buyer incurring a penalty for every unit ordered above its initial order quantity, it is worth noting that the only decision variable that the manufacturer has in our model is the total capacity, as price is considered to be exogenously determined. As a result, our contract is not directly aimed at coordinating the inventory decision, but rather seeks to understand the conditions under which the manufacturer will voluntarily build inventory above and beyond the health payer's initial order quantity when external reference pricing is used as a mechanism to determine transfer payments between the players.

Regarding the contract proposed in §3.5, which is a performance-based contract, Guajardo, Cohen, Kim and Netessine (2012) study the ability of performance based contracts in a general setting to increase product reliability, and find that such reliability is increased due to more frequent and more diligent maintenance activities induced by the optimal contract. Our model, rather than allowing the exertion of efforts that may affect the distribution of the realized health benefits, assumes that the manufacturer has private knowledge about the

expected performance of the drug, and therefore we focus not on the ability to modify the value of the product, but rather on the manufacturer's ability to signal the ability to the buyer. From the works of supply chain coordination using the newsvendor model, under price-dependent demand in a seller-buyer relationship, Emmons and Gilbert (1998) show that buy-back contracts with a fixed buy-back rate do not coordinate the chain; further, considering a fixed payment per unit sold, buy-back contracts coordinate the chain but allocate zero profit to the supplier (Marvel and Peck, 1995; Bernstein and Federgruen, 2005). Bernstein and Federgruen (2005) show that a buy-back contract coordinates the chain under arbitrary profit allocation only if the buy-back rate and the wholesale price are adjusted as a function of the retail price in what is referred to as "contingent buy-backs" or "discount pricing". Cachon and Lariviere (2005) show that revenue sharing may coordinate the chain when the buyer selects the retail price, but similar to buy-backs, an arbitrary allocation of profit requires the contract parameters to be contingent on the selling price. The main limitation of these models within our setting is that they assume the "selling price" to be a deterministic parameter while in our context after the prescription policy threshold is set, the resulting health benefits are a random variable. This would be equivalent to being able to select only the expected selling price, and setting the contract parameters accordingly. This distinction about a random selling price - health benefits in our model - implies that the player whose payoff function depends on the realized price holds a higher risk in the contract, and that the contract parameters would need to be modified after the true price is revealed. Needless to say, this raises concerns on how such a mechanism could be successfully implemented in our setting. By considering different objective functions for the downstream party - the health-payer - under a set of constraints relevant to the decision of introducing a new drug, we provide additional light into the theoretical reach and applicability of risk sharing contracts in health care.

This type of contracts has received increasing attention in recent years by the health

economics community. Barros (2011) looks at the relationship between a pharmaceutical manufacturer and a health-provider's prescription behavior assuming a binary health outcome and patient heterogeneity; he studies a risk sharing contract where the manufacturer is reimbursed for the drug only when the treatment is successful, which leads to high list prices and a higher than efficient prescription behavior, even though the latter effect may be alleviated by appropriately setting a revision cost. The main differences with our model is that Barros (2011) does not incorporate demand uncertainty and the prescriber experiences no risk in the contract. On a related paper, Zaric and Xie (2009) analyze two contracts in a two-period setting where the manufacturer sets the price for a drug seeking formulary listing and exerts promotional effort that deterministically shapes the demand curve: in one contract, the drug is listed in the payer's formulary during period 1 and delisted in the next period if cost-effectiveness is not achieved, and in the second contract, which is the most relevant to our work, the manufacturer pays a rebate to the health-payer in each period that cost-effectiveness is not achieved where the rebate amount is such that the payer's cost-effectiveness constraint binds. They find that no contract dominates, and provide a numerical analysis to observe the effects of uncertainty, the willingness to pay threshold, and the associated costs (or savings) derived from the drug's introduction. The first distinction in our model is that we allow the size of demand to be uncertain, which creates an inventory risk for the health-payer that becomes relevant for both his objective function and his cost-effectiveness constraint. Second, our model does not consider the manufacturer's ability to influence the size of total incoming demand. Third, we explicitly model information asymmetry with respect to the expected health outcome and allow the manufacturer to take advantage of its informational advantage through the parameters of the contract presented in §3.5. Fourth, the performance-based contract proposed here is similar in that a rebate is also offered, but in our case the rebate is a fixed amount given by a per unit rebate rate set by the manufacturer multiplied by the order quantity of the health-payer, while in Zaric and Xie (2009), the rebate is such that realized net monetary benefits are zero (*i.e.,*

if realized outcomes are below the minimum level of acceptance, the rebate is such that the cost-effectiveness constraint binds).

3.3 Endogenous price-only contracts

In this section we model the relationship between Pharma and Health when the former is able to endogenously select the transfer price, which is assumed to be the only contract parameter. This kind of contracts reflects more appropriately those cases where a negotiation process occurs between the manufacturer and the health-payer so that the former observes the parameters used by the latter in his calculations. We denote this setting with the symbol η for endogenous. We will continue to use the notation presented in Chapter 2, and introduce additional notation as needed. Pharma sells each unit of the drug to Health at endogenously selected price w, in order to maximize its utility function:

$$M^{\eta}(w; Q, \tau) = (w - c)Q, \tag{3.3.1}$$

It will be useful to henceforth use the notation $Q^*_{j,\eta\,(w)}$ and $\tau^*_{j,\eta\,(w)}$, $j = S, H$, to denote the optimal decisions by Health as a function of the selling price w, when maximizing (S)ocial welfare, or (H)ealth's utility function. Also, observe the following definitions adapted from Chapter 2 adjusting for this contracting scenario, as a function of the endogenously set w:

$$\bar{Q}^{\eta}_{\tau\,(w)} = \max\left\{ Q \;\middle|\; Q \leq \frac{(B_h(\tau) - \delta + g)A(Q,\tau) - g\lambda}{w - \delta}\,; Q > 0 \right\},$$

$$Q^{\eta}_{\tau\,(w)} = \max\left\{ Q \;\middle|\; P(Q; \lambda F(\tau)) \geq \frac{w - \delta}{B_h(\tau) - \delta + g} \right\}.$$

Since Health's problem remains unchanged with respect to Chapter 2, Pharma's optimal

decision satisfies:

$$
\begin{aligned}
w^*_{j,\eta} \in \arg\max_{(w)} \ & \{(w-c)Q^*_{j,\eta\,(w)}\} \\
& \text{subject to:} \\
& wQ^*_{j,\eta\,(w)} \leq \Gamma \\
& -wQ^*_{j,\eta\,(w)} + B_h(\tau^*_{j,\eta\,(w)})A(Q^*_{j,\eta\,(w)}, \tau^*_{j,\eta\,(w)}) + \delta\left(Q^*_{j,\eta\,(w)} - A(Q^*_{j,\eta\,(w)}, \tau^*_{j,\eta\,(w)})\right) \\
& \quad -g\left(\lambda F(\tau^*_{j,\eta\,(w)}) - A(Q^*_{j,\eta\,(w)}, \tau^*_{j,\eta\,(w)})\right) \geq 0 \qquad , \qquad j = S, H \\
Q^*_{j,\eta\,(w)}, \tau^*_{j,\eta\,(w)} \in \arg\max_{(Q,\tau)} & \begin{cases} S^\eta_h(Q,\tau;w) & \text{if } j = S, \\ H^\eta(Q,\tau;w) & \text{if } j = H, \end{cases}
\end{aligned}
\tag{3.3.2}
$$

Similar to what we have done earlier, we will present the solutions when expected social welfare and Health's expected utility are maximized, subject to the relevant constraints.

3.3.1 Case 1^η: Maximizing expected social welfare

First we solve the situation when Health maximizes social welfare under the endogenously determined price-only contract designed by Pharma. From Chapter 2, we know that Health will purchase the largest order quantity it can afford for a given τ, and by setting the selling price, Pharma is able to indirectly determine the access level and corresponding order quantity.

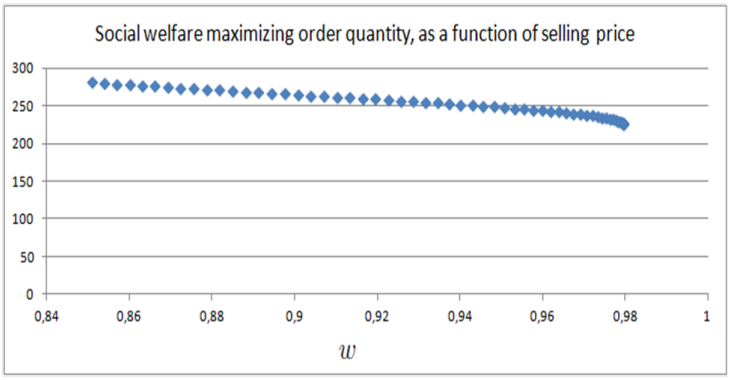
In order to reduce the search space for the optimal selling price, $w^*_{S,\eta}$, we define the following possible solutions:

$$
\bar{\bar{w}}^\eta_\tau = \max\{\, w \mid H(Q,\tau;\, w) \geq 0, \text{ for some } Q > 0\}\,; \tag{3.3.3}
$$

$$
\bar{w}^\eta_{\tau,S} \in \arg\max_{(w)} \ \{(w-c)Q^*_{S,\eta\,(w)};\ \Gamma \to \infty\}\,; \tag{3.3.4}
$$

$$
\underline{w}^\eta_\tau = \max\{\, w \mid H(Q,\tau;\, w) \geq 0\,;\ wQ = \Gamma, \text{ for some } Q > 0\}\,. \tag{3.3.5}
$$

Figure 3.1: Relationship between w and $\bar{Q}^{\eta}_{\tau\,(w)}$



w	$\bar{Q}^{\eta}_{\tau\,(w)}$	Marginal Revenue
0,955502	247	0,402494
0,95315	248	0,372206
0,95072	249	0,34808
0,9482	250	0,32072
0,9456	251	0,2956
0,94293	252	0,27276
0,94019	253	0,24971

$$\lambda = 600; \tau = 1; \beta_1 = 1; g = 0.2; \delta = 0; c = 0.3; \Gamma \to \infty$$
$$\bar{\bar{w}}^{\eta}_{1} = 0.97962;\ \bar{Q}^{\eta}_{1\,(0.97962)} = 226; \qquad \bar{w}^{\eta}_{1,S} = 0.94820;\ \bar{Q}^{\eta}_{1\,(0.94820)} = 250$$

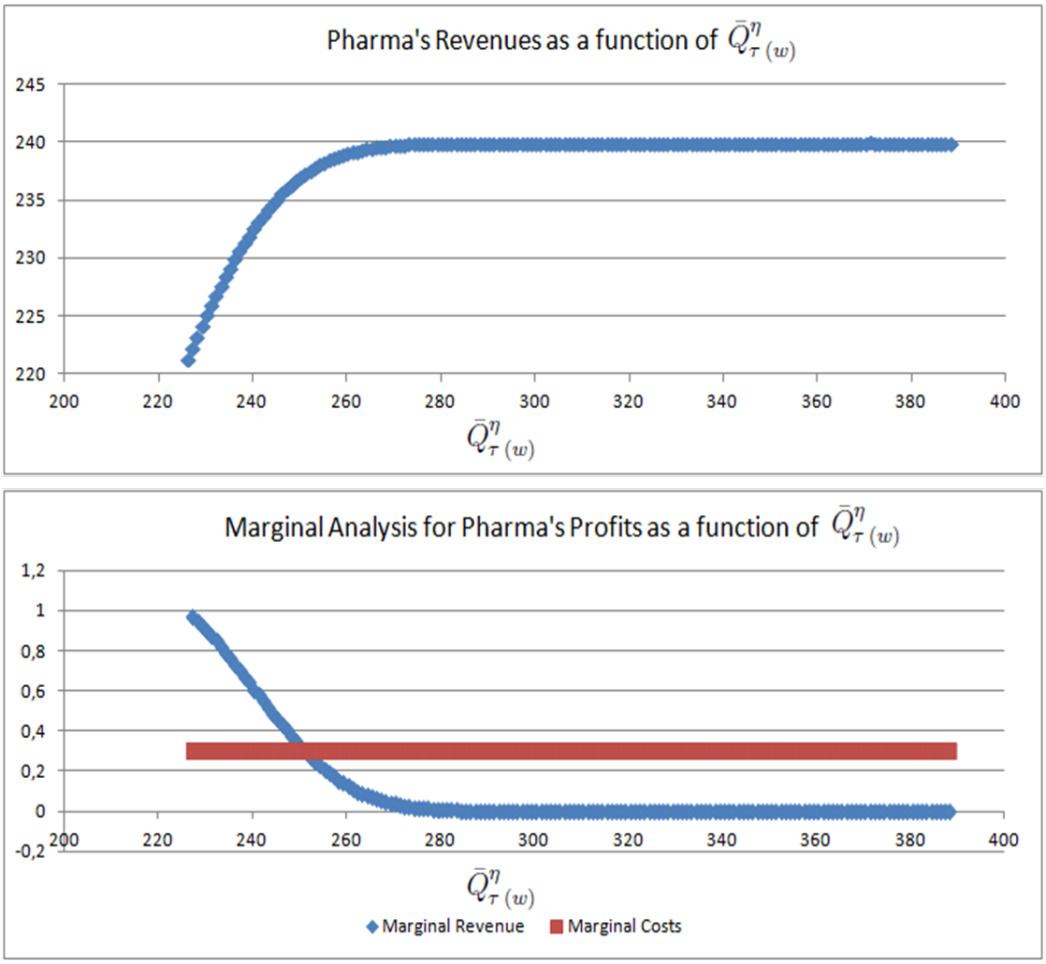
To develop some intuition on the latter definitions, observe firstly $\bar{\bar{w}}^{\eta}_{\tau}$, which acts as an upper bound on the optimal selling price. When the budget is infinitely high - or simply too high relative to the size of the population -, then the budget constraint becomes irrelevant and $\bar{\bar{w}}^{\eta}_{\tau}$ represents the largest selling price for which Health can achieve a non-negative expected utility function. Secondly, since Health is maximizing social welfare, Pharma may have an incentive to decrease the price[1] with respect to $\bar{\bar{w}}^{\eta}_{\tau}$, thus increasing Health's order quantity, as long as the increase in total revenue is higher than the increase in production costs. Mathematically, we have that $\bar{w}^{\eta}_{\tau,S}$ may be rewritten as the selling price that solves the following condition:

$$\bar{w}^{\eta}_{\tau,S} = \max\left\{ w \mid (w-c)(\bar{Q}^{\eta}_{\tau\,(w)}) > (w-\varepsilon_S-c)(\bar{Q}^{\eta}_{\tau\,(w-\varepsilon_S)}) \right\},$$
$$\text{where} \qquad \varepsilon_S = \min\left\{ \varepsilon \mid \bar{Q}^{\eta}_{\tau\,(w-\varepsilon)} = \bar{Q}^{\eta}_{\tau\,(w)} + 1 \right\}. \tag{3.3.6}$$

Figure 3.1 shows how the selling price changes in order to achieve higher order quantities. The definition of ε_S is simply the minimum decrease needed in the selling price in order to increase the order quantity by 1 unit. Figure 3.2 shows how Pharma's revenue changes as the selling price is gradually decreased from its upper bound, in the unconstrained budget situation. Notice how the marginal revenue quickly decreases until it converges to zero

[1]Note that any price higher than $\bar{\bar{w}}^{\eta}_{\tau}$ is infeasible.

Figure 3.2: Effect of w on Pharma's profits under social welfare maximization, as $\Gamma \to \infty$



$\lambda = 600; \tau = 1; \beta_1 = 1; g = 0.2; \delta = 0; c = 0.3$

marginal change. When the budget constraint is reachable, $\underline{w}^{\eta}_{\tau}$ gives the highest price that satisfies both constraints simultaneously, allowing Pharma to extract the maximum potential revenues Γ (or approaching Γ due to integrality of the order quantity) by selling as little units as possible. If $\underline{w}^{\eta}_{\tau}$ exists, by construction any price higher than $\underline{w}^{\eta}_{\tau}$ does not satisfy the cost-effectiveness constraint, and any price lower than $\underline{w}^{\eta}_{\tau}$ (weakly) reduces Pharma's utility because total revenues can't increase and production costs are nondecreasing due to (weakly) larger order quantities. Proposition 7 formally summarizes the implications on the optimal selling price offered by Pharma.

Proposition 7: Assume $\tau^*_{S,\eta} = \tau$ and $\bar{Q}^{\eta}_{\tau\,(w)}$ exists for some $w > c$.

Then $w^*_{S,\eta} = \begin{cases} \underline{w}^{\eta}_{\tau} & \text{if } \underline{w}^{\eta}_{\tau} \textit{ exists} \\ \bar{w}^{\eta}_{\tau,S} & \text{otherwise} \end{cases}$.

Note that $\bar{w}^{\eta}_{\tau,S} \leq \bar{\bar{w}}^{\eta}_{\tau}$ and $\underline{w}^{\eta}_{\tau} \leq \bar{\bar{w}}^{\eta}_{\tau}$; also, note that $\underline{w}^{\eta}_{\tau}$ and $\bar{w}^{\eta}_{\tau,S}$ can't simultaneously exist, and that the existence of $\bar{\bar{w}}^{\eta}_{\tau} > c$ is a necessary and sufficient condition for trade. While Proposition 7 explains Pharma's behavior for a given access level, recall that Pharma's choice of the selling price may modify the access and service levels selected by Health. In Theorem 1 we have defined the optimal order quantity for a given access level and selling price, which in combination with Proposition 7 implies that all we need to find before fully determining the solution is the optimal access level that Pharma wishes to induce in order to maximize her utility function. This is explored in Theorem 3.

Theorem 3:

a) When q^h exists, then:

a1) If $\underline{w}^{\eta}_{1}$ exists, then $\tau^*_{S,\eta} = 1$.

a2) If $\underline{w}^{\eta}_{1}$ does not exist and $\underline{w}^{\eta}_{2}$ exists, then $\tau^*_{S,\eta} = \begin{cases} 1 & \text{if } c > \frac{\underline{w}^{\eta}_{2} Q^*_{S,\eta\,(\underline{w}^{\eta}_{2})} - \bar{w}^{\eta}_{1,S} Q^*_{S,\eta,\ (\bar{w}^{\eta}_{1,S})}}{Q^*_{S,\eta\,(\underline{w}^{\eta}_{2})} - Q^*_{S,\eta\,(\bar{w}^{\eta}_{1,S})}} \\ 2 & \text{otherwise} \end{cases}$;

a3) If $\underline{w}^{\eta}_{1}$ and $\underline{w}^{\eta}_{2}$ do not exist, then $\tau^*_{S,\eta} = \begin{cases} 1 & \text{if } c > \frac{\bar{w}^{\eta}_{2,S} Q^*_{S,\eta\,(\bar{w}^{\eta}_{2,S})} - \bar{w}^{\eta}_{1,S} Q^*_{S,\eta\,(\bar{w}^{\eta}_{1,S})}}{Q^*_{S,\eta\,(\bar{w}^{\eta}_{2,S})} - Q^*_{S,\eta\,(\bar{w}^{\eta}_{1,S})}} \\ 2 & \text{otherwise} \end{cases}$.

b) When q^h does not exist, then $\tau^*_{S,\eta} = 1$.

The intuition, and true relevance, of Theorem 3.a2 is that for full access to be preferred, it must be not only that increasing access generates more revenues - i.e., $\bar{w}^{\eta}_{2,S} Q^*_{S,\eta\,(\bar{w}^{\eta}_{2,S})} > \bar{w}^{\eta}_{1,S} Q^*_{S,\eta\,(\bar{w}^{\eta}_{1,S})}$, but also that the increased revenue per incremental amount of drugs sold must be sufficiently large to justify the additional manufacturing cost; in other words, as c goes to zero, Pharma is more likely to induce higher levels of access; Theorem 3.a3 follows a

similar logic. Finally, notice that the results derived here allow for direct comparison when $\underline{w}_1^{\eta}$ doesn't exist, instead of doing a complete enumeration over a range of feasible price and quantity pairs.

3.3.2 Case 2^{η}: Maximizing Health's expected value function

In this subsection, we solve the situation where Health maximizes his expected value function given the endogenously selected price-only contract designed by Pharma. Recall that under endogenous price selection and social welfare maximization, Health chooses to increase the access level whenever it is cost-effective to do so. However, when Health maximizes its utility function, Pharma is no longer able to induce full access unless it allows Health to keep some of the surplus. Specifically, in order to induce full access, Pharma will have to set a selling price no larger than $\tilde{w}$, at which Health is indifferent between both access levels. We will need the following additional definition before following a similar approach to the previous subsection.[2]

$$\bar{w}_{\tau,H}^{\eta} \in \underset{(w)}{\arg\max} \{(w-c)Q_{H,\eta}^{*};\ \Gamma \to \infty\}\ ; \tag{3.3.7}$$

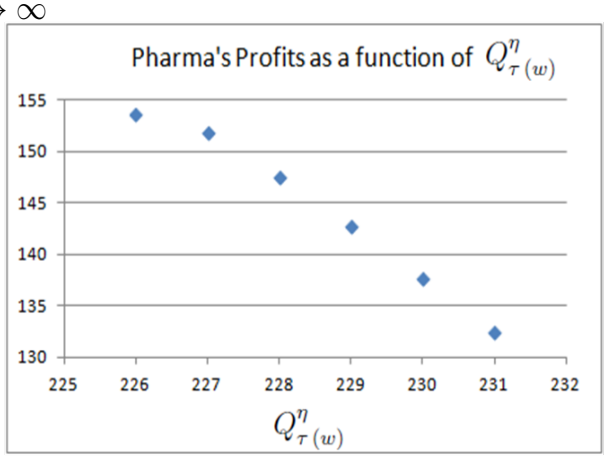
As before, it is possible that when the budget constraint is not active, Pharma will not charge the highest feasible price, and instead will set it satisfying:

$$\begin{aligned} \bar{w}_{\tau,H}^{\eta} &= \max\left\{ w \mid (w-c)(Q_{\tau(w)}^{\eta}) > (w-\varepsilon_H-c)(Q_{\tau(w-\varepsilon_H)}^{\eta}) \right\}, \\ \text{where} \qquad \varepsilon_H &= \min\left\{ \varepsilon \mid Q_{\tau(w-\varepsilon)}^{\eta} = Q_{\tau(w)}^{\eta} + 1 \right\}\ . \end{aligned} \tag{3.3.8}$$

Figure 3.3 shows how the selling price needs to change in order to increase the budget

[2]Lariviere and Porteus (2001) prove that the problem for a manufacturer selling to an unconstrained profit maximizing newsvendor is concave for continuous distributions that satisfy the IGFR property (see section 2.2 of their paper, pp. 295-296). Banciu and Mirchandani (2013) derive the conditions for this finding to be valid for some discrete distributions (see pp. 926-928 of their paper), among which is included the Poisson distribution assumed in our model.

Figure 3.3: Effect of w on Pharma's profits under Health's net utility maximization, as $\Gamma \to \infty$



w	$Q^{\eta}_{\tau(w)}$	Revenue	Profit
0,97962	226	221,3941	153,5941
0,96906	227	219,9766	151,8766
0,94687	228	215,8864	147,4864
0,92351	229	211,4838	142,7838
0,89903	230	206,7769	137,7769
0,87349	231	201,7762	132,4762

$\lambda = 600; \tau = 1; \beta_1 = 1; g = 0.2; \delta = 0; c = 0.3$

unconstrained order quantity, starting from the upper bound. It is interesting, although intuitive, to observe that for the combination of parameters used (which is the same as in Figures 3.1 and 3.2), $\bar{w}^{\eta}_{1,H} = \bar{\bar{w}}^{\eta}_{1} = 0.97962$. This means that Pharma chooses the largest possible price under which Health will not reject the contract. Although this is a common result, it need not be so in general, as the optimal selling price is given by the process described in equation (3.3.8). However, it does suggest that when $\tau = 1$, Pharma will extract (almost) all surplus from Health. More interestingly, Pharma may not be able to do so under restricted access, as is explained in Proposition 8.

Proposition 8.1: Assume $\tau^{*}_{H,\eta} = 1$. Then $w^{*}_{H,\eta} = \begin{cases} \underline{w}^{\eta}_{1} & \text{if } \underline{w}^{\eta}_{1} \textit{ exists} \\ \bar{w}^{\eta}_{1,H} & \text{otherwise} \end{cases}$

Proposition 8.2: Assume $\tau^{*}_{H,\eta} = 2$. Then: $w^{*}_{H,\eta} = \begin{cases} \min[\tilde{w}\,,\, \underline{w}^{\eta}_{2}] & \text{if } \underline{w}^{\eta}_{2} \text{ exists} \\ \min[\tilde{w}\,,\, \bar{w}^{\eta}_{2,H}] & \text{otherwise} \end{cases}$

The implications of the incentive problem are now evident, since Health indirectly uses his ability to define the access level to reduce Pharma's bargaining power if full access is to be selected. As a result, the conditions under which Pharma has an incentive to set a selling price that induces full access are now more restrictive compared to §3.3.1.

Theorem 4:

a) When q^h exists, then:

a1) If $\underline{w}_1^\eta$ exists, then $\tau_{H,\eta}^* = 1$.

a2) If $\underline{w}_1^\eta$ and $\underline{w}_2^\eta$ do not exist and $H^\eta(Q_{2\,(\tilde{w})}^\eta, 2;\, \tilde{w}) \leq 0$,

$$\text{then } \tau_{H,\eta}^* = \begin{cases} 1 & \text{if } c > \dfrac{\bar{w}_{2,H}^\eta Q_{H,\eta\,(\bar{w}_{2,H}^\eta)}^* - \bar{w}_{1,H}^\eta Q_{H,\eta\,(\bar{w}_{1,H}^\eta)}^*}{Q_{H,\eta\,(\bar{w}_{2,H}^\eta)}^* - Q_{H,\eta\,(\bar{w}_{2,H}^\eta)}^*} \\ 2 & \text{otherwise} \end{cases}$$

a3) If $\underline{w}_1^\eta$ does not exist, and $\underline{w}_2^\eta$ exists, and $\tilde{w}\underline{Q}_{2\,(\tilde{w})}^\eta > \Gamma$,

$$\text{then } \tau_{H,\eta}^* = \begin{cases} 1 & \text{if } c > \dfrac{\underline{w}_2^\eta Q_{H,\eta\,(\underline{w}_2^\eta)}^* - \bar{w}_{1,H}^\eta Q_{H,\eta\,(\bar{w}_{1,H}^\eta)}^*}{Q_{H,\eta\,(\underline{w}_2^\eta)}^* - Q_{H,\eta\,(\bar{w}_{1,H}^\eta)}^*} \\ 2 & \text{otherwise} \end{cases}$$

a4) Else, then $\tau_{H,\eta}^* = \begin{cases} 1 & \text{if } c > \dfrac{\tilde{w}Q_{H,\eta\,(\tilde{w})}^* - \bar{w}_{1,H}^\eta Q_{H,\eta\,\bar{w}_{1,H}}^*}{Q_{H,\eta\,(\tilde{w})}^* - Q_{H,\eta\,(w_1^\eta)}^*} \\ 2 & \text{otherwise} \end{cases}$.

b) When q^h does not exist, then $\tau_{H,\eta}^* = 1$.

The conclusions from Theorem 4 are very interesting. On one hand, and as anticipated, it is more likely that access will be restricted when Health maximizes his utility function versus when he maximizes social welfare. But on the other hand, we note that if Pharma selects price $\tilde{w}$, then Health's expected utility function will be positive and even the expected social welfare utility may be higher than when maximizing the latter was Health's objective. The reason is that the utility maximizing formulation creates an artificial incentive compatibility constraint in order for Health to prefer the full versus restricted access level, preventing Pharma from extracting all the surplus, achieving a lower selling price compared to social welfare maximization, and resulting in a larger order quantity and possibly larger expected social welfare. Some situations where the latter could occur include a) when the maximum allowable budget is large, the relative size of the patient population with lower health benefits is relatively large and the difference in expected health benefits between the two patient populations is low; or b) when capacity building and manufacturing costs are

low, so that Pharma finds a sufficient incentive in the incremental revenue from higher access levels.

3.4 Exogenous price contracts with capacity buffer allowed

So far we have assumed that Pharma has no excess capacity, and therefore Health may only order once per period. In this section we relax that assumption and we allow Pharma to build a capacity buffer, K, above Health's order quantity Q. In such a situation, when demand exceeds Q, then Health can purchase up to K additional units, paying the per unit selling price w to Pharma, and incurring a per unit penalty cost $p > 0$, which is interpreted as a penalty for increasing the initial order size or for delaying the patient's treatment. It is assumed that Pharma incurs an incremental cost p for delivering units above Q, and therefore is indifferent between selling a unit of the drug during the initial order, or at a later point in time. The rest of the parameters are consistent with Chapter 2. Health has a per unit salvage value δ for purchased units in excess of the realized demand, and when total capacity, defined $K_T \triangleq (Q + K)$, is exceeded by demand, a per unit goodwill cost g is incurred. The transfer payment from Health to Pharma is redefined as $T(w, Q, K) = wQ + (w + p)(\min(K\,,\,(D(\lambda, F(\tau)) - Q)^+))$. Pharma has no salvage value for unsold units.

Additionally, we make a weak assumption to guarantee that risklessly purchasing a drug for an incoming average patient is superior to incurring the goodwill cost of not meeting that

patient's demand. Mathematically, this is expressed as:

$$B_h(\tau) + g - w - p \geq 0. \tag{3.4.1}$$

Intuitively, by purchasing a drug from the capacity buffer and administering it to the patient, Health's (possibly negative) margin is $(B_h(\tau) - w - p)$, to which we add the 'saved' goodwill cost g.

Next we formulate the exogenous price contract when Pharma is willing to incur part of the inventory risk by being able to build excess inventory. We use the symbol 'κ' to denote exogenous price contracts *with* a positive capacity buffer allowed. Pharma's expected profit function is:

$$M^{\kappa}(K; Q, \tau) = (w - c)Q + w(A(Q + K, \tau) - Q)^{+} - cK \ , \tag{3.4.2}$$

which reduces to $M^{\kappa}(K; Q, \tau) = (w - c)(Q)$ when $K = 0$. The social welfare's expected utility function from Health's perspective is:

$$S_h^{\kappa}(Q, \tau; K) = (B_h(\tau) + g)A(Q + K, \tau) + \delta(Q - A(Q, \tau)) - g\lambda F(\tau) \tag{3.4.3}$$

and Health's expected utility function is:

$$\begin{aligned} H^{\kappa}(Q, \tau; K) = \quad & (B_h(\tau) + g - w - p)A(Q + K, \tau) + (p + w - \delta)A(Q, \tau) \\ & -(w - \delta)Q - g\lambda F(\tau) \end{aligned} \tag{3.4.4}$$

To explain the formulation, notice that the salvage value δ is only relevant for the first Q units. The health benefits and the goodwill costs only depend on K_T. The penalty cost p is only expected to be incurred for the difference between the expected administered drugs when K_T versus Q units are available. And the selling price w is deterministically incurred

for the first Q units, and is expected to be incurred for the difference between the expected administered drugs when K_T versus Q units are available.

For $j = S, H$, when the selling price is exogenously determined and Pharma is allowed to build a capacity buffer, define $K^*_{j,\kappa\,(Q,\tau)}$ as Pharma's optimal capacity buffer for Health's choice of Q and τ; and $Q^*_{j,\kappa\,(K)}$ and $\tau^*_{j,\kappa\,(K)}$ as Health's optimal order quantity and prescription policy threshold given Pharma's choice of capacity buffer K. We begin by solving Pharma's problem for any Q and τ, i.e., $K^*_{j,\kappa\,(Q,\tau)}, j = S, H$, and then proceed to find the equilibrium solution (or in some cases, solutions).

Pharma's Problem

In order to obtain a more intuitive characterization of the solution, we initially show some necessary conditions for $K^*_{j,\kappa\,(Q,\tau)} > 0$ and then find upper bounds on the feasible quantity that Health may purchase.

Lemma 10: For $j = S, H$, $Q^*_{j,\chi} < Q_\Gamma$ is a necessary condition for $M^\kappa(K^*_{j,\kappa}; Q^*_{j,\kappa}, \tau^*_{j,\kappa}) > M^\chi(Q^*_{j,\chi}, \tau^*_{j,\chi})$.

Lemma 10 marks an incentive compatibility constraint for Pharma, since it is only optimal to build a positive buffer if it leads to an increase in the expected profits relative to the exogenous price-only contract presented in Chapter 2. Therefore, in the rest of §3.4 we will assume that the condition from Lemma 10 is satisfied.

Lemma 11: Define $\bar{K}_{T\,(\tau)} = \underset{K \text{ int}}{\arg\max} \{wA(K,\tau) - cK\}$ to be the upper bound on Pharma's optimal total capacity. For a given access level τ, $\bar{K}_{T\,(\tau)} = \max\left\{K \,\middle|\, P(K, \lambda F(\tau)) > \frac{c}{w}\right\}$, and $\bar{K}_{(Q,\tau)} \triangleq \bar{K}_{T\,(\tau)} - Q$.

The notation $\bar{K}_{(Q,\tau)}$ defines Pharma's upper bound for the optimal capacity buffer as a function of the order quantity Q and access level τ. In other words, it shows the maximum level of inventory risk that Pharma is willing to accept by comparing the expected revenue of increasing the capacity buffer by 1 unit versus the corresponding (constant) deterministic cost c. While Lemma 12 gives the total capacity that Pharma would build in an unconstrained setting, two situations may occur. The first one is that Health chooses an order quantity larger than $\bar{K}_{T(\tau)}$ when Pharma's relative understocking costs are lower than those of Health; this will be formally shown when we solve Health's problem under each decision making criteria. The second situation is when Health's constraints do not justify building $\bar{K}_{T(\tau)}$, as is expressed below

Lemma 12: For any positive Q and K, let $w(Q+K)+pK$, be the largest possible *realized* expenses for Health. a) $K_{\Gamma(Q)} = \max\left\{ \lfloor K \rfloor \;\middle|\; K \leq \frac{\Gamma - wQ}{w+p} \right\}$ is the largest capacity buffer that Health will be able to utilize given the *budget* constraint Γ and Pharma's order quantity choice Q; b)$\frac{\partial((\Gamma - wQ)/(w+p))}{\partial Q} = -\frac{w}{w+p} \in (-1, 0)$; c) $\Delta \triangleq \lceil \frac{w+p}{p} \rceil$, is Health's minimum order quantity increase to trigger a 1 unit increase in the total quantity that satisfies the budget constraint.

Understanding the intuition from Lemma 12 is crucial to our analysis. First, it sets an upper bound on the feasible capacity buffer as a function of Q. Secondly, part b) explains that for an increase of 1 unit in Q, $K_{\Gamma(Q)}$ decreases in no more than 1 unit. This implies that the total capacity, $K_T = K + Q$, that satisfies the budget constraint is non-decreasing in Q; in fact increases in 1 unit for every Δ units that Q increases, as is explained in Lemma 12c. As such, $K_{\Gamma(Q)}$ represents Health's participation constraint with respect to its total budget. While for a fixed K_T larger values of Q are more likely to satisfy the budget constraint, the balancing effect is introduced in Lemma 13.

Lemma 13: $K_{E(Q,\tau)} = \min\left\{\lceil K\rceil \;\middle|\; A(Q+K,\tau) \geq \frac{(w-\delta)Q+g\lambda F(\tau)-(w+p-\delta)A(Q,\tau)}{B_h(\tau)-w-p+g}\right\}$ is the smallest capacity buffer that Pharma would need to provide in order for Health's *cost-effectiveness* constraint to be satisfied.

Lemma 13, while not a direct counterpart for Lemma 12, does contribute to a balancing effect in Health's choices. One way to interpret it is that as Q increases, the increase in Health's expected utility function (both from the increased health benefits and from the decreased goodwill costs) of having the capacity buffer must be sufficiently large to overcome the escalating expected overstocking costs. As Q continues to increase, the benefits achieved by the buffer may not be sufficient due to the low probability of high values of realized demand, or because the increase required in the buffer is larger than what the budget allows. Integrating the results from this subsection, Proposition 9 provides Pharma's best response for given Q and τ.

Proposition 9: For $j = S, H$, assume $Q^*_{j,\chi} < Q_\Gamma$. a) $0 < K_{E(Q,\tau)} \leq \min(\bar{K}_{(Q,\tau)}, K_{\Gamma(Q)})$, is a necessary condition for $K^*_{j,\kappa} > 0$. b) $K_{E(Q^*_{j,\chi},\tau^*_{j,\chi})} \leq \min(\bar{K}_{(Q^*_{j,\chi},\tau^*_{j,\chi})}, K_{\Gamma(Q^*_{j,\chi})})$, is a sufficient condition for $K^*_{j,\kappa} > 0$. c) For given Q and τ, Pharma's best response function when $K^*_{j,\kappa} > 0$, is delimited by: $K_{E(Q,\tau)} \leq K^*_{j,\kappa(Q,\tau)} = \min(\bar{K}_{(Q,\tau)}, K_{\Gamma(Q)})$.

Health's Problem

We now move to the analysis of Health's problem. Before finding Health's best response strategy under each of his decision-making criteria, we need to derive some additional results. In this respect, Lemma 14 finds the minimum order quantity necessary for Health's budget to not become a restriction on Pharma's desired capacity buffer based on her own critical fractile. Lemma 15 gives a sufficient condition for Health's budget to prevent Pharma's desired capacity buffer from being built.

Lemma 14: $\breve{Q}_{\tau}^{\kappa} = \min\left\{ \lceil Q \rceil \;\middle|\; P\left(\left\lfloor \frac{\Gamma + pQ}{w+p} \right\rfloor ; \lambda F(\tau)\right) < \frac{c}{w} ; Q < \frac{\Gamma}{w} \right\}$, is Health's minimum order quantity required for $K_{\Gamma(Q)} \geq \bar{K}_{(Q,\tau)}$.

Lemma 15: For $j = S, H$, if $P\left(\left\lfloor \frac{\Gamma}{w} \right\rfloor ; \lambda F(\tau)\right) > \frac{c}{w}$, then $K^*_{j,\kappa(Q,\tau)} \leq K_{\Gamma(Q)} < \bar{K}_{(Q,\tau)}$.

In other words, the two latter Lemmas provide a reference point for determining whether the budget constraint will be binding or not. On one hand, if the budget is large, the price is low, or Pharma's overstocking cost is relatively large, then Pharma's budget unconstrained maximization solution will yield the total capacity available. On the other hand, if the budget is low, the selling price and/or the penalty are high, or Pharma's incentive to overstock is high, then Health's budget will restrict the capacity built and the resulting drug amount available in the system.

3.4.1 Case 1^{κ}: Maximizing expected social welfare

In this subsection we solve the access and service level decisions when Health's objective is to maximize expected social welfare, and Pharma is allowed to create an excess capacity buffer. Recall that for $K^*_{S,\kappa} > 0$, it must be that $Q^*_{S,\chi} \in \{\bar{Q}_1^{\chi}, \bar{Q}_2^{\chi}\}$; otherwise Pharma has no incentive to provide the buffer. Consequently, Health's problem is expressed as follows:

$$\begin{aligned}
\max_{(Q,\tau)} \quad & (B_h(\tau) + g)A(Q + K^*_{S,\kappa(Q,\tau)}, \tau) + \delta(Q - A(Q,\tau)) - g\lambda F(\tau) \\
& \text{subject to:} \\
& K_{E(Q,\tau)} \leq K^*_{S,\kappa(Q,\tau)} = \min(\bar{K}_{(Q,\tau)}, K_{\Gamma(Q)}) \\
& T(w, Q, K^*_{S,\kappa(Q,\tau)}) \leq \Gamma \\
& (B_h(\tau) + g - w - p)A(Q + K^*_{S,\kappa(Q,\tau)}, \tau) + (w + p - \delta)A(Q,\tau) \\
& \quad -(w - \delta)Q - g\lambda F(\tau) \geq 0
\end{aligned} \tag{3.4.5}$$

Notice that the objective function is increasing in the total quantity of drugs available, which implies that Pharma's and Health's objectives are aligned in the same direction, even though their participation constraints are in general different. In other words, both players benefit from higher levels of available inventory (given the feasibility constraints). Therefore we begin our formal analysis by using $\bar{Q}_{\tau}^{\chi}$ as a reference point. Since we have assumed $B_h(\tau) + g > w + p$, then purchasing any order quantity K ex-post satisfies the cost-effectiveness constraint, and we only need to check the absolute budget constraint for feasibility. As was mentioned above, there exists the possibility that Pharma's optimal capacity buffer will be zero. Proposition 10 provides such situations.

Proposition 10: Set $\tau_{S,\kappa}^{*} = \tau$. There are three scenarios that will result in $K_{S,\kappa}^{*} = 0$.

i) If $\bar{Q}_{\tau}^{\chi} \geq Q_{\Gamma}^{\chi}$.

ii) If $\frac{\Gamma - w - p}{w} < \bar{Q}_{\tau}^{\chi} < Q_{\Gamma}^{\chi}$.

iii) If $\bar{K}_{T(\tau)} < \bar{Q}_{\tau}^{\chi} < Q_{\Gamma}^{\chi}$.

Proposition 10 allows us to further characterize the space for which the capacity buffer option is relevant. Conditions i) and ii) are related to the absolute budget constraint, so that any feasible combination of an initial order quantity and a positive capacity buffer will decrease social welfare when compared to the solution under the exogenous price-only contract. Condition iii) shows the case where Health is willing to accept a higher inventory risk than Pharma. This situation is more likely to occur as the selling price is relatively close to the production cost and relatively far from the benefit $B_h(\tau)$, or when the goodwill cost g is high for Health.

Next, we define limits on the feasible order quantity considering the possibility that Pharma keeps an excess inventory stock.

Definition 3:

a) Let $\bar{\bar{Q}}_\tau^\kappa = \max\left\{ \lfloor Q \rfloor \;\middle|\; (w-\delta)Q - (w+p-\delta)A(Q,\tau) < (B_h(\tau)+g-w-p)A(\bar{K}_{T(\tau)},\tau) \right.$
$\left. -(B_h(\tau)-\delta+g)A(\bar{Q}_\tau^\chi,\tau) + (w-\delta)\bar{Q}_\tau^\chi \right\}$, be Health's largest order quantity that satisfies the cost-effectiveness constraint for a fixed total capacity $\bar{K}_{T(\tau)}$.

b) Let $\bar{Q}_\tau^\kappa = \max\left\{ \lfloor Q \rfloor \;\middle|\; (w-\delta)Q - (w+p-\delta)A(Q,\tau) < (B_h(\tau)+g-w-p)A(Q+K_{\Gamma(Q,\tau)},\tau) \right.$
$\left. -(B_h(\tau)-\delta+g)A(\bar{Q}_\tau^\chi,\tau) + (w-\delta)\bar{Q}_\tau^\chi \right\}$, be Health's largest order quantity that satisfies both the cost-effectiveness constraint and the budget constraint.

Definition 3a provides the largest order quantity that Health is able to purchase in advance, conditioning on the fact that Pharma will stock $\bar{K}_{T(\tau)}$ units. It is known that such order quantity will be at least $\bar{Q}_\tau^\chi$; the larger $\bar{\bar{Q}}_\tau^\kappa$ is, the higher the probabilities that the budget constraint will not be binding. In a similar vein, Definition 3b defines the largest order quantity that Health is able to purchase without violating the constraints. Note that on one hand, such quantity may be larger than Pharma's choice of $\bar{K}_{T(\tau)}$, in which case Pharma would not build any excess stock. On the other hand, it is also possible that $\bar{Q}_\tau^\kappa < \bar{K}_{T(\tau)}$, which would imply that purchasing $\bar{K}_{T(\tau)}$ is not feasible, and therefore the budget constraint will limit the total inventory available in the system. Lemma 16 incorporates the previous results to explain the behavior of the expected social welfare function, and Proposition 11 finds the optimal order quantity and capacity buffer for a given access level.

Lemma 16:

a) For a fixed τ and K_T, $S_h^\kappa(Q,\tau;K)$ is weakly increasing in Q.

b) For a fixed τ and Q, $S_h^\kappa(Q,\tau;K)$ is increasing in K.

c) Assume $\breve{Q}_\tau^\kappa \leq \bar{\bar{Q}}_\tau^\kappa$. Then, for fixed τ, $S_h^\kappa(Q,\tau;K)$ is non-decreasing in Q for $Q \leq \bar{\bar{Q}}_\tau^\kappa$.

d) Assume $\breve{Q}_\tau^\kappa > \bar{\bar{Q}}_\tau^\kappa$. Then, for fixed τ, $S_h^\kappa(Q,\tau;K)$ is increasing in Q for $Q \leq \bar{Q}_\tau^\kappa$.

e) Assume $\breve{Q}_\tau^\kappa$ does not exist. Then, for fixed τ, $S_h^\kappa(Q,\tau;K)$ is increasing in Q for $Q \leq \bar{Q}_\tau^\kappa$.

Proposition 11: Fix $\tau^*_{S,\kappa} = \tau$.

a) If $\breve{Q}^\kappa_\tau$ exists and $\breve{Q}^\kappa_\tau < \bar{Q}^\kappa_\tau$, then: $Q^*_{S,\kappa} = \begin{cases} \bar{\bar{Q}}^\kappa_\tau & \text{if } \delta > 0 \\ \left[\breve{Q}^\kappa_\tau, \bar{\bar{Q}}^\kappa_\tau\right] & \text{otherwise} \end{cases}$;

and $K^*_{S,\kappa} = \bar{K}_{(Q,\tau)} \leq K_{\Gamma(Q)}$

b) Otherwise, $Q^*_{S,\kappa} = \bar{Q}^\kappa_\tau$ and $K^*_{S,\kappa} = K_{\Gamma(Q)} < \bar{K}_{(Q,\tau)}$.

It is interesting to note that when $\delta = 0$, there are multiple equilibrium between Health's order quantity and Pharma's capacity buffer K. However, this does not represent a problem since Pharma will build the same total capacity regardless of which equilibrium realizes, *i.e.,* since the total capacity for Pharma is constant, Pharma's capacity building choice is independent of Health's order quantity. The role of Health's order quantity, Q, will be therefore to allocate demand risk. As Q increases, Pharma's risk decreases, Health's risk of overstocking increases, and Health's total expenditures may either decrease or increase because as the initial order quantity decreases, the number of units for which Health expects to pay the penalty p increases. Also interesting is that when either $\delta > 0$, or $\breve{Q}^\kappa_\tau > \bar{Q}^\kappa_\tau$, or $\breve{Q}^\kappa_\tau$ does not exist, the equilibrium solution is unique for a given access level and the budget constraint is the limiting condition. To complete the analysis, we turn to the problem of finding the optimal access level.

Proposition 12: When $I = 2$, let q^κ be a positive order quantity such that $S^\kappa_h(q^\kappa, 1) \geq S^\kappa_h(q^\kappa, 2)$; and $S^\kappa_h(Q, 2) > S^\kappa_h(Q, 1),\ \forall\ Q > q^\kappa$.

a) $\beta_2 > \delta$ is a necessary and sufficient condition for q^κ to exist, and if equation (2.3.3) is satisfied and q^κ exists, it is unique and given by equation (3.4.6).

$$\left(\frac{g}{\beta_1 - \delta + g}\right)\left(\frac{\lambda}{A(Q+K,2)}\right) + \left(\frac{b_1 - b_2}{b_1 - \delta + g}\right) = \left(\frac{A(Q+K,2) - A(Q+K,1)}{A(Q+K,2)}\right)\left(\frac{1}{1-\theta}\right) \tag{3.4.6}$$

b) If $\beta_2 < \delta$, then $S^\kappa_h(Q, 1) > S^\kappa_h(Q, 2),\ \forall\ Q > 0$.

c) If $\beta_2 = \delta$, then $S_h^\kappa(Q,1) > S_h^\kappa(Q,2)$, for some finite $Q > 0$, and $\lim_{P(Q,\lambda)\to 0} S_h^\kappa(Q,1) - S_h^\kappa(Q,2) = 0$.

d) The value of q^κ is increasing in $\beta_1, g,\ \delta$, and decreasing in β_2. The change in q^κ with respect to θ is ambiguous.

e) If $\delta > 0$ and K_T is fixed, then the value of q^κ is decreasing in K; and for $K > 0$, $q^\kappa < q^h$.

f) If $\delta = 0$, then $q^\kappa = q^h$.

The key results from Proposition 12 are: that when $\delta = 0$, only K_T (and not the relative sizes of Q and K) is relevant in determining expected social welfare and the crossing point between the two access level options; and when $\delta > 0$, the total inventory available in the system that is required to achieve full access is lower when Pharma offers a capacity buffer in the contract, versus the situation of an exogenous price-only contract. This should not be confused with the result introduced in Lemma 16a, such that for a fixed available inventory level in the system, social welfare decreases as K increases. Instead, the intuition here is that for a fixed available inventory level in the system, as the capacity buffer increases, the crossing point between the two access level curves moves to the left because for a given service level the restricted access option benefits more from the salvage value than the full access option. Having this in mind, the choice of the optimal access level can now be expressed.

Proposition 13:

a) If $\tau_{S,\chi}^* = 1$, then $\begin{cases} \tau_{S,\kappa}^* = 2 & \text{if } \exists\ \left\{Q \,|\, q^\kappa \le Q + K_{E(Q,2)} \le Q + K_{\Gamma(Q)}\right\}, \\ \tau_{S,\kappa}^* = 1 & \text{otherwise ;} \end{cases}$

b) If $\tau_{S,\chi}^* = 2$, then $\tau_{S,\kappa}^* = 2$.

The first takeaway is that in the buffer capacity contract, access level can only increase relative to the exogenous price-only contract. The second takeaway involves the situations when the access level will be increased. We have already shown that when $\tau_{S,\chi}^* = 1$ and

$Q^*_{S,\chi} = \bar{Q}^{\chi}_1$, it must be that $S^{\chi}_h(q^{\chi}, 1) = S^{\chi}_h(q^{\chi}, 2) < 0$; this happened because either the budget constraint, Γ, didn't allow for $\underline{Q}^{\chi}_2$ to be purchased, or $\underline{Q}^{\chi}_2$ didn't exist. If we are in the first case, it follows that $Q + K_{\Gamma(Q)} < \underline{Q}^{\chi}_2$. This means that access is likely to be increased when: p is low such that the difference between Q^{χ}_{Γ} and $Q + K_{\Gamma(Q)}$ is not too large; and overstocking cost is relatively large and therefore Health can benefit from a low initial order quantity Q without incurring too many understocking costs by leveraging on $K_{\Gamma(Q)}$. If we are in the second case, there is no $\underline{Q}^{\chi}_2$ to use as a reference, but the intuition about Health ordering a quantity as low as possible ex-ante to avoid overstocking costs without drastically increasing its understocking costs, continues to hold.

3.4.2 Case 2^{κ}: Maximizing Health's expected utility function

In this subsection, we analyze the situation where Health maximizes his expected utility function under the exogenous price contract with capacity buffer. The problem Health solves is:

$$\max_{(Q,\tau)} \quad (B_h(\tau) + g - w - p)A(Q + K^*_{H,\kappa(Q,\tau)}, \tau) + (p + w - \delta)(A(Q, \tau)) - (w - \delta)Q - g\lambda F(\tau)$$

subject to:

$$K_{E(Q,\tau)} \leq K^*_{H,\kappa(Q,\tau)} = \min(\bar{K}_{(Q,\tau)}, K_{\Gamma(Q)})$$

$$T(w, Q, K^*_{H,\kappa(Q,\tau)}) \leq \Gamma$$

$$(B_h(\tau) + g - w - p)A(Q + K^*_{S,\kappa(Q,\tau)}, \tau) + (w + p - \delta)A(Q, \tau)$$

$$-(w - \delta)Q - g\lambda F(\tau) \geq 0 \tag{3.4.7}$$

When maximizing social welfare, we established that Health could be indifferent between choosing multiple order quantities as long as the total capacity available remain unchanged and $\delta = 0$. This can no longer be the case when Health maximizes its utility function, as is expressed in Lemma 17.

Lemma 17: a) $Q_\tau^\kappa = \max\left\{Q \;\middle|\; P(Q;\lambda F(\tau)) > \frac{w-\delta}{w+p-\delta}\right\}$ to be Health's optimal order quantity when $K \to \infty$ for a given τ. b) $Q_\tau^\kappa < Q_\tau^\chi$. c) $\lim_{(w+p)\to(B_h(\tau)+g)} Q_\tau^\kappa = Q_\tau^\chi$.

If $K \to \infty$, Lemma 17 describes Pharma's choice problem between purchasing units in advance assuming the inventory risk, versus purchasing units at a higher price after demand realization, in case it is needed. However, as it has been stated Pharma's optimal capacity buffer is bounded by her own incentive compatible critical fractile and by Health's budget and cost-effectiveness constraints.

Proposition 14: Fix $\tau_{H,\kappa}^* = \tau$.

a) $Q_{H,\kappa}^* = Q_\tau^\kappa + \bar{\alpha}\Delta$, if $\exists\ \bar{\alpha} \in \mathbb{N}_{>0}$, where

$$\bar{\alpha} = \max\left\{\alpha \;\middle|\; (B_h(\tau)+g-w-p)\sum_{x=Q_\tau^\kappa+1}^{Q_\tau^\kappa+\alpha} P(Q_\tau^\kappa+\alpha\Delta+K_{\Gamma(Q_\tau^\kappa+\alpha\Delta)};\lambda F(\tau)) > \right.$$

$$\left. \alpha\Delta(w-\delta)-(w+p-\delta)\sum_{x=Q_\tau^\kappa+1}^{Q_\tau^\kappa+\alpha\Delta} P(x;\lambda F(\tau))\ ;\ \ \bar{K}_{(Q_\tau^\kappa+\alpha\Delta,\tau)} \geq K_{\Gamma(Q_\tau^\kappa+\alpha\Delta)} \geq K_{E(Q_\tau^\kappa+\alpha\Delta,\tau)} \right\}.$$

b) Else if $\bar{\alpha}$ has no solution, then $Q_{H,\kappa}^* = \begin{cases} Q_\tau^\kappa & \text{if } K_{E(Q_\tau^\kappa,\tau)} \leq \min[\bar{K}_{(Q_\tau^\kappa,\tau)}\ ,\ K_{\Gamma(Q_\tau^\kappa)}], \\ Q_\tau^\kappa = 0 & \text{otherwise} \end{cases}$

Proposition 14 explains that Pharma's capacity buffer may be limited by the budget constraint, even though under an infinite supply assumption - and assuming a feasible solution exists -, Health would initially order Q_τ^κ units for access level τ. Still, as was proven in Lemma 12, Health can increase the total number of drugs that can be purchased given Γ by increasing his initial order quantity. In these situations, Proposition 14 explains that Health will only buy more than Q_τ^κ units, if the increase in K_T results in a larger expected utility. As before, the optimal capacity buffer will be bounded by the minimum buffer required to maintain cost-effectiveness, and the maximum buffer for which Pharma is willing to internalize the inventory risk.

For determining the optimal prescription policy threshold, we haven't been able to find a threshold value on the selling price as in the price-only contracts and therefore at this point direct comparison would be required. It is worth mentioning though that opposed to the case where Health is maximizing expected social welfare, here it is possible that the optimal decision is to restrict access for some parameter combinations that resulted in full access for the exogenous price-only contract. For instance, consider the situation where $w < \tilde{w}$, $\frac{\Gamma}{w}$ is slightly higher than Q_2^χ, c is low, and $B_h(2) + g \sim w + p$; then Health may receive little benefit from a buffer capacity contract under full access, but may increase its profits under restricted access by leveraging on Pharma's incentive to keep a high inventory availability. The situation where access level is increased under a capacity buffer contract can occur for instance when Γ and g are large, and τ is small, so that $B_h(2) + g \gg w + p$, and Health capitalizes on the low understocking probability.

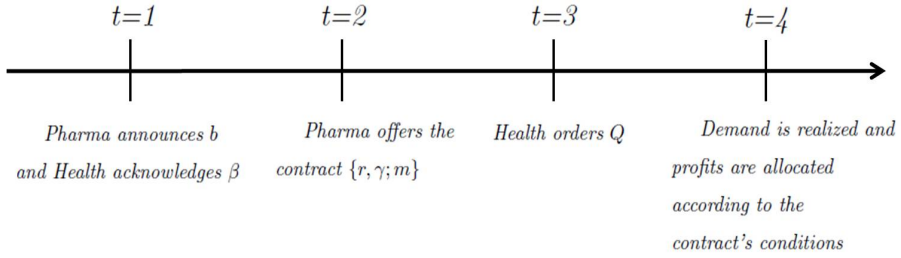
3.5 Performance-based contracts

In this section we analyze performance-based contracts as a mechanism aimed at reducing belief asymmetry between the players. We use a model very similar to the one used so far in Chapters 2 and 3, but instead of considering multiple patient categories, we consider a single category for which the health benefits are uncertain and explicitly model the information game between the players. It is important to understand that doing so does not imply that the access level decision is ignored for this contract; rather, and as it should be evident by the end of the section, Health's decision making criteria will remain unchanged in relation to Chapter 2, such that the only modification will be on the parameter he uses for the expected health benefits. Additionally, from an implementation perspective, we believe it is more sensible to provide a guarantee on the performance of the drug on a type by type basis, *i.e.,* Pharma may choose to offer a performance-based contract for patients of

type 1 and 3, but not for patients of type 2, even if the drug is administered to all three types.

In consequence, our previously used notation can be somewhat simplified, since the following analysis is applicable to any patient type. The setting consists of a pharmaceutical manufacturer, hereafter Pharma, who offers to sell a new drug to a health-payer at exogenously determined per unit price w. Let the demand, D, be a random variable following a Poisson distribution with parameter λ. Let $p(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$, be the probability that exactly x patient arrivals occur during the period; and let $P(x;\lambda) = \sum_{j=x}^{\infty} p(j;\lambda)$ be the complement of the Poisson CDF. Pharma announces the expected health benefit b for the drug, but Health chooses to use the certainty equivalent $\beta < b$ for his calculations. The reasons why Health would choose the value β rather than b include the uncertainty about the real present and future value of a drug's therapeutic innovation, the lack of solidity of the results presented by the manufacturer, or the replicability of those results in clinical practice. At this point we allow negotiation between the players to occur. We assume that Pharma holds an informational advantage about the drug's expected health benefits during clinical practice in the patient population serviced by Health. Specifically, Pharma holds a private belief π that the per patient realized health benefits will be b, and a belief $(1-\pi)$ that the realized health benefits will be β. An alternative formulation would assume Pharma's beliefs to be distributed over a continuous range. However, since the implementation of an insurance policy tends to be binary (see e.g., Pugatch et al., 2010) such that a rebate is paid when the health benefit is below a specified threshold value, then π would represent Pharma's belief that the minimum threshold value for a patient will be reached. For completeness, recall that the health benefit has been defined as the product of the incremental value of the health gains received by a patient and the health-payer's ceiling ratio; i.e., the incremental value derived from the drug's administration *times* the health-payer's maximum willingness-to-pay for that value.

Figure 3.4: Sequence of decisions and events



The order of events in the model is depicted in Figure 3.4 and described next. (1)Pharma announces health benefits b, and per unit selling price w. Health acknowledges health benefits β. (2)Pharma has the option to offer Health a performance-based contract $\{\gamma, r; m\}$ where: m is the exogenously set minimum number of administered drugs in order for the contract to be executed; γ is an endogenously set factor which serves as a guarantee that the average health benefits for the patient population who receive the drug (or a sample of the population) will be at least $b\gamma+\beta(1-\gamma)$; and r is the per unit rebate from Pharma to Health when the administered drugs are at least m and the average health benefit for the sampled patients is less than $\beta + (b - \beta)(\gamma)$. (3)Health selects order quantity Q of drugs which are then manufactured by Pharma at cost c.[3] (4)Demand is realized and the health benefits are measured for a sample of size m from the pool of patients who receive the drug treatment. Let $G(x)$ be the normal cumulative distribution function of the realized health benefits from the sample population, which has mean $m\pi$ and variance $m\pi(1-\pi)$, and let the fixed cost of verification to the manufacturer be $v(m)$, which is increasing in m. A rebate is paid if the corresponding contract's conditions are satisfied.[4] Excess drugs may be salvaged at a per unit value δ, which may be interpreted either as the opportunity cost or as a discounted

[3]Since we assume symmetric knowledge about the demand distribution and Health's decision making parameters, the assumption of the manufacturing and ordering decisions to be simultaneous versus sequential will not create a difference, and therefore we avoid the manufacturer's capacity term K_T to prevent unnecessary notation.

[4]It is assumed that $m\pi > 5$ and $m\pi(1-\pi) > 5$. This allows us to approximate the binomial distribution which would be used to calculate the probability of at least γm patients with a health outcome of b, through the normal distribution. For completeness, let it be known that using the binomial, the probability of the realized health benefit being higher than the threshold $(\beta + (b-\beta)\gamma)$ is expressed: $\sum_{x=0}^{m-\lfloor\gamma m\rfloor} \binom{m}{x} (\pi)^{m-x}(1-\pi)^{x}$.

sale to a secondary market; to avoid trivial problems, assume $\delta < c < \beta$. If $D(\lambda) > Q$, a per unit cost, g, is accrued to Health for each patient arrival which does not receive the drug treatment due to a stock-out. To keep integrality, we will use $\lfloor x \rfloor$ and $\lceil x \rceil$ as the floor and ceiling functions, respectively. Except where it has been otherwise specified, all players are assumed to hold symmetric information about functional forms and parameters.

Define $A(Q) \triangleq \mathbb{E}\left[\min[Q\,,\, D(\lambda)]\right]$ to be the expected quantity of *administered* drug treatments; $\mathbb{E}[\max[0\,,\, Q - D(\lambda)]] = (Q - A(Q))$, to be the expected leftovers for Health; and $\mathbb{E}\left[\max[0\,,\, D(\lambda) - Q]\right] = \lambda - A(Q)$, to be the expected quantity of understocked units of the drug at the end of the period. Also, let $T(w, \gamma, r)$ denote Health's transfer payment to Pharma as a function of the contract parameters. Pharma remains a profit maximizer, and Health's priority may be to either maximize expected social welfare, or maximize his entire expected utility function (*i.e.,* social welfare minus the transfer from Health to Pharma). However, we should observe that Health has no way of trusting Pharma. As a result we define Health's objective function, regardless of his priority, to be the smallest of the expected outcomes when the drug's performance is either as guaranteed by Pharma, or as originally assumed by Health. We denote this contract structure with the symbol ρ. The manufacturer's expected profit is:

$$M^{\rho}(\gamma, r;\, Q) \;=\; (w - c)Q - rQP(m; \lambda)G\left(\gamma m\right) - v(m), \tag{3.5.1}$$

where notice that $P(m; \lambda)G\left(\gamma m\right)$ represents the probability of rebate.

The social welfare expected utility function is:

$$S^{\rho}(Q;\,\gamma,r) = \min \begin{cases} S_{low}(Q;\,\gamma,r) = & (\beta-\delta+g)A(Q)+\delta Q-g\lambda, \\ & \text{if the health benefits are low,} \\ S_{high}(Q;\,\gamma,r) = & (\beta+(b-\beta)(\gamma)-\delta+g)A(Q)+\delta Q-g\lambda, \\ & \text{if the health benefits are high,} \end{cases}$$

Since $S_{high}(Q;\,\gamma,r) \geq S_{low}(Q;\,\gamma,r)$, then:

$$S^{\rho}(Q;\,\gamma,r) = (\beta-\delta+g)A(Q)+\delta Q-g\lambda$$

Health's expected utility function is:

$$H^{\rho}(Q;\,\gamma,r) = \min \begin{cases} H_{low}(Q;\,\gamma,r) = & (\beta-\delta+g)A(Q)-(w-\delta)Q-g\lambda\ +rQP(m;\lambda), \\ & \text{if the health benefits are low,} \\ H_{high}(Q;\,\gamma,r) = & (\beta+(b-\beta)(\gamma)-\delta+g)A(Q)-(w-\delta)Q-g\lambda, \\ & \text{if the health benefits are high,} \end{cases}$$

In addition, two types of constraints are included in our analysis: a budget constraint:

$$wQ \leq \Gamma, \tag{3.5.2}$$

where Γ is an exogenous upper limit on Health's expenses for the drug under analysis; and

the cost-effectiveness constraints:

$$H_{low}(Q;\, \gamma, r\,) \geq 0, \tag{3.5.3}$$

$$H_{high}(Q;\, \gamma, r\,) \geq 0 \tag{3.5.4}$$

which guarantee that the expected net benefits derived from the drug's approval are above some minimum threshold. Next, we map Health's optimal order quantity as a function of the contract's design, and analyze how it is affected by the contract parameters. Then we solve for Pharma's optimal contract.

3.5.1 Health's Problem

Case 1^{ρ}: Maximizing expected social welfare

In this subsection we adapt the previously obtained results from the exogenous price-only contract (or equivalently, from the integrated chain) to the performance-based contract when Health is maximizing the expected social welfare. In order to do so, we use the same mechanism that was derived in Chapter 2 to find the potential optimal quantities as a function of the contract parameters. First, Lemma 18 defines the possible optimal order quantities under social welfare maximization.

Lemma 18:

a) $\bar{Q}^{\rho}_{high} = \max\left\{ \lfloor Q \rfloor \;\middle|\; Q \leq \frac{(\beta+(b-\beta)\gamma-\delta+g)A(Q)-g\lambda}{w-\delta}\,; Q \geq 0 \right\}$.

b) $\bar{Q}^{\rho}_{low} = \max\left\{ \lfloor Q \rfloor \;\middle|\; Q \leq \frac{(\beta-\delta+g)A(Q)-g\lambda}{w-\delta-rP(m;\lambda)}\,; Q \geq 0 \right\}$.

c) $Q^{*}_{S,\rho} = \min\left[Q_{\Gamma},\, \bar{Q}^{\rho}_{low}\,,\, \bar{Q}^{\rho}_{high} \right]$.

As was the case under the simple contracts presented earlier, at optimality either the budget constraint or the cost-effectiveness constraint will be binding (or very close to binding,

because of integrality). Notice that when Health maximizes social welfare, our assumption of Health maximizing the minimum of the the two possible outcomes is irrelevant; this is because Health will purchase as many drugs as his constraints allow it to. This marks the difference here versus the analysis in Chapter 2, since Health's lack of trust in Pharma implies that the cost-effectiveness constraint must be satisfied both if the health benefits are as guaranteed by Pharma, and if the health-benefits are as Health had initially acknowledged. Recall that even when the latter occurs, the rebate is not guaranteed since demand must be at least m for the contract to be called upon. As a result, either r must be sufficiently large or m, which is exogenous, sufficiently small so that the largest feasible order quantity is increased in a significant way. The next results explains how the contract parameters affect the feasible region for Health's optimal order quantity.

Lemma 19: a) $\bar{Q}^{\rho}_{high}$ is weakly increasing in γ; b) $\bar{Q}^{\rho}_{high}$ is independent of r; c) $\bar{Q}^{\rho}_{low}$ is independent of γ; d) $\bar{Q}^{\rho}_{low}$ is weakly increasing in r. e) $Q^{*}_{S,\rho}$ is weakly increasing in γ and r.

Lemma 19 shows that as the value of the contract parameters increases, the largest feasible order quantity weakly increases. Intuitively, in the low benefits scenario as r increases and everything else is kept constant, then the expected lump sum transfer increases, which may allow for larger order quantities to be feasible. Similarly, in the high benefits scenario as γ increases and everything else is kept constant, the health benefit that Health expects to see for each unit of administered drugs increases; while for the same Q the utility function will increase, such difference may be large enough to allow the purchase of a larger quantity.

Case 2^{ρ}: Maximizing Health's expected utility

In this subsection we adapt the previously obtained results to the performance-based contract when Health is maximizing his expected utility function. To do so, we follow the same

mechanism as above, taking into account that there will be a limiting order quantity for each combination of parameters. Lemma 20 defines the feasible optimal order quantities under Health's expected utility maximization.

Lemma 20:

a) $Q^{\rho}_{high} = \max\left\{ \lfloor Q \rfloor \;\middle|\; P(Q;\lambda) > \frac{w-\delta}{\beta+(b-\beta)\gamma-\delta+g} \right\}$.

b) $Q^{\rho}_{low} = \max\left\{ \lfloor Q \rfloor \;\middle|\; P(Q;\lambda) > \frac{w-\delta-rP(m;\lambda)}{\beta-\delta+g} \right\}$.

c) $Q^{*}_{H,\rho} = \min\left[Q_{\Gamma},\, Q^{\rho}_{low},\, Q^{\rho}_{high} \right]$.

The first two parts of Lemma 20 yield the budget unconstrained order quantity that maximize Health's total utility function under the high and low scenarios, respectively. Lemma 20c then provides a parallel result to that of Chapter 2, since the optimal order quantity for a given access level will be given by the smallest of Q_{Γ}, and the order quantity that maximizes Health's objective function; the obvious difference is that by considering two possible scenarios, there are two (possibly overlapping) curves, each of them with its own maximum value. Lemma 21 explains how the order quantity that maximizes the curve under each scenario changes as a function of the contract parameters.

Lemma 21: a) Q^{ρ}_{high} increases in γ; b) Q^{ρ}_{high} is independent of r; c) Q^{ρ}_{low} is independent of γ; d) Q^{ρ}_{low} is increasing in r; e) $Q^{*}_{H,\rho}$ weakly increases in γ and r.

Now that Health's decision making criteria has been established, we proceed to analyze Pharma's problem of optimally designing the performance based contract.

3.5.2 Pharma's Problem

In this subsection we derive the optimal parameters of the performance-based contract from Pharma's perspective. Since the following results apply for both of Health's decision-making criteria, we use the letter $j = S, H$ to denote the expressions in a consistent manner without being repetitive.

Proposition 15: a) $r^*_{j,\rho\,(\gamma)} = \left(\frac{A(Q^*_{j,\rho})}{Q^*_{j,\rho}}\right)\left(\frac{b-\beta}{P(m;\lambda)}\right)(\gamma)$; b) $\gamma^*_{j,\rho\,(r)} = \left(\frac{Q^*_{j,\rho}}{A(Q^*_{j,\rho})}\right)\left(\frac{P(m;\lambda)}{b-\beta}\right)\left(\frac{1}{r}\right)$.

Proposition 15 states the optimal relationship between the contract parameters in order to avoid offering Health unnecessarily benevolent conditions in the contract that can't be compensated by larger order quantities. In other words, it implies that when the performance-based contract is implemented, then at optimality $\bar{Q}^{\rho}_{low} = \bar{Q}^{\rho}_{high}$, under social welfare maximization; and $Q^{\rho}_{low} = Q^{\rho}_{high}$, under Health's expect utility maximization. If that was not the situation, then one of the constraints would have a positive shadow price created by either a guarantee γ, that could be decreased (decreasing the probability of paying a rebate), or by a rebate rate r that could be decreased (decreasing the value of the rebate in case the guarantee is not satisfied), in both cases without inducing a decrease in Health's order quantity. Furthermore, this balancing relationship allows us to reformulate Pharma's problem as that of a single decision variable as is expressed next.

Lemma 22: At optimality, Pharma's problem under a performance based contract may be rewritten as a single variable problem, and is given by equation (3.5.6):

$$M(\gamma;\, Q^*_{j,\rho}) = (w-c)Q^*_{j,\rho} - A(Q^*_{j,\rho})(b-\beta)(\gamma)G(\gamma m) \tag{3.5.5}$$

Unfortunately we can't obtain the explicit solution for the parameters, but Proposition 15 gives the conditions that need to be satisfied by the optimal guarantee factor γ, and

this can be used to obtain the optimal rebate rate r. Similarly, we are unable to determine analytically whether Pharma has an incentive to truthfully reveal her private information. Intuitively though, as m becomes large, the sample mean will approach the true mean, and setting $\gamma = \pi$ would imply paying a rebate with approximately a 50% probability. As a result, we expect that under this contract's structure, Pharma will tend to understate its private information, or not provide any information at all. The latter situation occurs when the condition from Corollary 3 is not satisfied.

Corollary 3: Define $Q^*_{j,\chi}$ as Health's optimal order quantity under exogenous price-only contract with parameter w. Pharma can benefit from the performance based contract only if:

$$(w-c)\left(\frac{Q^*_{j,\rho}-Q^*_{j,\chi}}{Q^*_{j,\rho}}\right) > r^*_{j,\rho}G(\gamma^*_{j,\rho}m)P(m;\lambda)+\frac{v(m)}{Q^*_{j,\rho}}$$

.

Corollary 3 gives the necessary condition for the performance based contract to take place. If this is not met, then Pharma can simply set a rebate rate equal to zero, and the system will behave as in an exogenous price-only contract scenario.

3.6 Conclusions

Chapter has presented three mechanisms that the upstream player, Pharma, may use in order to increase her profits relative to the exogenous price-only contract presented in Chapter 2. These contracts are an enedogenous price-only contract; a capacity buffer under exogenous transfer price; and a performance-based contract under exogenous price.

For the endogenous price-only contract, we find a very interesting situation. The manu-

facturer has an incentive to increase the price and decrease order quantity as long as total profits keep increasing, and therefore the cost-effectiveness constraint tends to be binding or very close to binding under social welfare maximization. In fact, when the optimal access is limited, both social welfare and expected utility maximization yield extremely similar, or even equal, results; Pharma extracts all of Health's surplus. However, when the health payer maximizes his expected utility, the manufacturer is not able to induce full access and extract all of the health payer's surplus simultaneously because of the threshold price mentioned earlier. This implies that if the manufacturer wishes her product to be considered available to a larger fraction of the patient population, then she must reduce the price, which under some parameter combinations, may even result in a larger order quantity than that obtained by social welfare maximization because of the incentive compatibility constraint imposed by comparing the objective function under the different access levels. This result is a potential argument for why some markets allow the manufacturers to set prices freely and then act as profit maximizing entities. In short, according to the model, utility maximizing may be in some cases a more efficient tool for achieving social welfare than social welfare maximization itself.

After that, the buffer capacity contract was introduced to include Pharma's willingness to adopt some of the inventory risk. However, we have kept our distance from the infinite inventory assumption by considering Health's cost-effectiveness and budget constraints, which along with Pharma's incentive compatibility constraint, determine the optimal capacity buffer. We find that this contract is most useful to the manufacturer when the budget constraint under price-only contracts is not binding and the manufacturer's per unit overstocking/understocking cost ratio is much lower than that of the health payer. However, when Health's budget is large, the buffer capacity contract will result in a lower initial order quantity by Health compared to the exogenous price-only contract, implying that Pharma's certain revenue will be strictly lower if access level remains unchanged. For the health payer and the patients, the buffer capacity contract will be most useful when demand uncertainty

is high, goodwill cost is high, and the available budget is relatively large. Finally, we find that under social welfare maximization, the buffer capacity contract can't decrease access level nor available inventory relative to the exogenous-price only contract; however when the health payer maximizes its expected utility, access level and inventory available can either decrease or increase.

In the end, we have proposed a novel performance-based contract where the pharmaceutical manufacturer has the option to increase the health-payer's willingness to pay for a given drug by offering a partial guarantee on the average realized health benefits. While other similar contracts have been studied and implemented, the particularities of the approach here developed is that given the manufacturer's informational advantage, the health-payer must base its decisions on the possibility that every contracting scenario may realize. As a result, if Pharma sends a signal indicating a high level of trust on her originally announced health-benefit, then she will also need to offer Health a large rebate as collateral. From this perspective, and while we cannot prove it analytically, the manufacturer is not expected to send a reliability signal that is higher than her privately held knowledge when the variance of Pharma's private information is high, because doing so would increase the probability that the rebate will need to be paid. In short, the health payer can only calculate his expected utility based on his own perception of the health benefits and the manufacturer's signals, and bases his decisions on the scenario that imposes the tightest constraints; but the manufacturer has visibility on the probability that each scenario may occur so that if she were completely confident that her announced health benefit will be equal to the realized benefit, then she could set an extremely large rebate without any negative consequences. Another interesting property of this contract is its attempt to distance itself from the so-called cost-containment initiatives and become a truly risk sharing mechanism. For example, without the need to alter the initial selling price, both the manufacturer and the health payer benefit from a successful outcome: the payment to the manufacturer increases due to the possibility of charging the initially negotiated price and selling a larger than initially negotiated

order quantity, cost-effectiveness is maintained, and drug availability increases. In terms of the downside risks for the players, the manufacturer assumes most of the risk from lower than expected health benefits, while the payer assumes the risk relative to the size of the demand. It is worth discussing the implementability of this contract when multiple patient categories can be treated by the same drug, and explain why the access level decision was not included in this model. The situations where it is not expected to add value are when the budget surplus from Health is low so that the manufacturer's risk acceptance can't be rewarded with a larger revenue stream, and when the cost of verification is large. As for the exclusion of the access level decision, the main reason is that a fair assessment of the drug's performance would require the proportion of patients from each category within the sample to be the same as in the population. This situation is on one part harder to control, and additionally would lay an additional layer of risk on the manufacturer's utility function, as it would be dependent on the realized distribution of patient arrivals which is prone to being manipulated by an unethical health-payer. Instead, the belief is that this contract design should be implemented for a single category of patients per drug (even if, and perhaps specially when, the drug is administered to multiple categories), and that this category should be the one where the health-payer's trust in the manufacturer's announced health benefits is low and the manufacturer's level of confidence is high. In such situations, the contracting mechanism can allow for increases in access and service level by increasing the range of order quantities for which cost-effectiveness is achieved, and in general increasing the health-payer's expected utility function. Furthermore, since manufacturers tend to lose patients' and prescribers' goodwill following negative health outcomes, it is considered unlikely for this type of contracts to be viciously implemented.

Last, it is true that a policy-maker or a pharmaceutical manufacturer may implement the contracts here analyzed in a simultaneous manner. Our analysis does not show any evidence against doing so. Rather, we have attempted to identify the virtues and shortcomings

of each contract, and we do not expect that these results will change qualitatively when interaction exists. However, analytically tracking the simultaneous implementation of these policies does not appear to result in any clear intuition. Taking a simulation approach to test the theories here presented under contract interaction may prove to be an interesting research opportunity.

The final Chapter will depart from the decision of whether to serve a fraction or the entirety of the population, and instead will focus on whether service will be provided to different patient categories through a single product, or through differentiated products.

[page left intentionally blank]

Chapter 4

Pulling, pooling, and contracting in the presence of heterogeneous consumers: implications of supply chain design on innovation, coverage and profits

4.1 Introduction

Consumer heterogeneity, which has been at the center of our discussions, can often be measured either subjectively (*e.g.,* an individual's favorite color, movie, or tourist destination) or objectively (*e.g.,* a bidders' maximum willingness to pay in a Vickrey auction or the health outcome of patients after receiving a drug treatment with multiple therapeutic applications). Organizations have tried to somehow incorporate the value that their clients (*i.e.,* their target population) perceive from the products and services these organizations provide. However, since value is often not easily observable or accurately captured, much of the operations literature's focus has been on maximizing profits by determining optimal

levels of supply typically considering demand to be stochastic, and sometimes even price dependent, but not explicitly considering the specific preferences of individuals. The economics literature has tried to acknowledge such heterogeneity in consumers' valuations by measuring the resulting social surplus created by the demand and supply equilibrium, however it often ignores demand stochasticity and the corresponding inventory risk. The recent theory of revenue management is an attempt to efficiently allocate supply taking into account the heterogeneity in consumers' valuations, for example, by implementing reservation values for each consumer category. However, there are some settings where first-come first-serve is the only implementable allocation policy due to operational, technological, or ethical reasons. As a result, it is not clear what the optimal supply levels should be when demand for a given product is partitioned into consumer segments with distinct observable valuations of the good and where both the demand's size and order of arrivals are stochastic. To the best of our knowledge, this is the first attempt to model the mentioned conditions simultaneously to obtain operational results.

The scope of the chapter is therefore on a good that is either sold at different prices to different consumers depending on the consumers' category (*e.g.,* medical attention provided to incoming insured and uninsured patients); or equivalently, where different consumer categories pay the same price for the good but value it differently and the player who delivers the good observes and internalizes consumer surplus (*e.g.,* fully subsidized drugs that can treat both terminal and stage 2 cancer patients, providing different QALY's to each of these two patient categories). It is important to consider that the distinction in valuations need not be caused by differences in the taste or willingness to pay of the consumers; but it may also originate from a good that has multiple applications, each intended for a separate consumer category and providing a different outcome for each category even when demand is inelastic with respect to price.

In order to be consistent with the rest of the dissertation, the model is motivated within a healthcare setting, but the conclusions of the paper can be adapted to more general contexts as is illustrated in the last section, as long as the following assumptions are met:

A1: There is heterogeneity in the consumers' valuations of the good (*i.e.,* different types of consumers exist), and both the type of consumer and the consumer's valuation are observable.

A2: The player who decides the inventory stock (or capacity level) of the good either is able to price-discriminate between consumer types (*e.g.,* through membership cards), or sells at a unique price to all types of consumers but observes and internalizes their surplus.

A3: The exact order of arrivals of the consumers is *ex-ante* unknown, but there is some probabilistic knowledge about it.

A4: No inventory (or capacity) can be reserved and the goods are provided to consumers on a first-come first-serve basis (*i.e.,* if an incoming consumer demands the good and the good is available, then the consumer's demand must be satisfied).

In a newsvendor-type demand process with two observable types of consumers, we use a separable demand function composed by two categories of patients who consume the same drug: an initial category (incumbent) of stochastic size and a second category (entrant) of either deterministic or stochastic size which may or may not realize as a function of R&D efforts. When the efforts are successful and the entrant category realizes, we show that accounting for the heterogeneity in consumers' types in combination with the stochastic order of arrivals creates an increase in the optimal order quantity that is either larger or smaller than the size of the newly created demand, but never the same even if the size of the additional demand is deterministic. Intuitively, when the size of the entrant category is deterministic, the variance of the demand's size is unaffected (hence, there is no inventory pooling effect), but the possibility that the incumbent category consumes drugs that are intended for the entrant category modifies the understocking cost, affecting the order quantity

for the stochastic part of the demand; we refer to this phenomena as the *inventory pulling effect* since the category of patients that composes the stochastic portion of the demand pulls inventory from a stockpile that was intended for a different category. In this setting, the chapter's first key contribution is that depending on which category of consumers (the stochastic or the deterministic) obtains higher benefits from the drug, then either having a single common channel that treats multiple patient categories, or having separate channels exclusively dedicated to each patient category, maximize total welfare, create the maximum incentives to exert innovation effort, and reduce total inventory levels compared to the inefficient selection in the event that the innovation efforts are successful; comparative statics are presented to show the determinants of the distortion in the order quantity due to the pulling effect. The second contribution is to analyze the incentive misalignment in a vertically disintegrated supply chain under the model's stated assumptions, and show that under exogenous price-only contracts, the direction of the incentives for innovation efforts changes and the manufacturer always prefers the inefficient channel structure; a contract design is proposed in the extensions to deal with this issue. Thirdly, we formulate and implicitly solve the case where the inventory pooling and pulling effects interact, and the effects on inventory and system's welfare. Since closed-form solutions are not reachable, numerical experiments are presented to further our intuition.

The chapter proceeds as follows. We first present a review of relevant literature. Section 4.3 introduces the model in the context of marginal innovations of a drug in a vertically integrated supply chain involving a pharmaceutical manufacturer and a health-payer, where the latter delivers the drug to the patients. §4.3.2 solves the case where the two demand categories are independently served through dedicated channels and in §4.3.3 we introduce the pulling effect by forcing a single inventory stock to serve the entire market. Section 4.4 departs from vertical integration and allows exogenously determined price-only contracts to be signed between the manufacturer and the health-payer. Section 4.5 presents some numeri-

cal results. Extensions to the basic model are included in §4.6, and the interaction of pooling and pulling effects is modeled in Section 4.7. §4.8 concludes with managerial implications and brief descriptions of the application of the model into other contexts.

4.2 Literature Review

Products that have multiple applications or that are valued differently by separate consumer segments are not uncommon but have been narrowly studied in the operations management literature, particularly when reservation is not allowed.[1] While the literature on optimal effort levels under uncertain conditions is extensive both in the economics and operations fields, we distinguish that given our approach of the innovation effort in the presence of heterogenous consumers, our model is closest to Glass (2001). She examines quality improvement efforts in a competitive environment when consumers value quality differently. By analyzing the interaction between quality and price-setting, she finds that allowing the firm to select price results in price-discrimination with higher quality-adjusted prices for consumers than those achieved through minimum quality intervention. However, such intervention reduces the firm's incentives to invest in quality. We observe a similar behavior in the vertically separated chain such that the design option selected by the manufacturer is not efficient from a global system perspective, but imposing the efficient design reduces the manufacturer's innovation efforts. Additionally, in our model we account for demand stochasticity to capture the inventory implications of each effort decision.

With respect to such inventory decisions, the immediate reference is the literature around pooling effects and consolidation. In his seminal work, Eppen (1979) shows that consolidating supply for multiple demand sources that follow a normal distribution always leads to a

[1]There is extensive work on revenue management and the calculation of reservation levels as a function of demand's arrival rate and willingness to pay. However, the assumptions required in those models are in conflict with this chapter's motivation, eliminating their applicability, and are therefore not discussed here.

decrease in inventory costs relative to a decentralized setting where each demand source is satisfied by a dedicated inventory stock. This happens due to the firms ability to aggregate uncertainty and more efficiently match supply and demand. The results from Eppen have been expanded to cover more general distributions, as well as being the inspiration for empirical work incorporating behavioral factors (see Gerchack and He, 2003, and Alfaro and Corbett, 2003, for reviews).

Our model shares some commonalities with the models of the competitive (e.g., Lippman and McCardle, 1994) and substitution (e.g., Parlar, 1988) newsvendor in that there are two sources of demand that can consume the good from two separate stockpiles. However, we depart from such works in two fundamental ways. First, a common assumption in those models is that the selling price of the good is independent of the source of demand. Second, in both models customers will consume their own stockpile before they attempt to consume the competitor's stockpile. In our first setting where one of the sources of demand is of deterministic size, consuming from the competing stockpile would only occur if the stochastic source of demand occurred before realization of the deterministic demand. As a result, while in Lippman et al., 1994, perfect substitutability never leads to a decrease in total inventory, in our model this may happen when the demand source with highest uncertainty generates a higher revenue from the good's consumption.

There are some recent works that consider customer heterogeneity in the inventory decisions but do not satisfy our allocation rule. Deshpande, Cohen and Donohue (2003) study the optimal inventory rationing policy in a continuous replenishment setting for two demand classes that differ in their arrival rates and shortage costs. Under their allocation scheme, the first come first serve allocation rule is assumed only until a threshold level of on-hand inventory is reached, at which point demand for the lower class is backordered. As is explained in this chapter, excluding demand classes after they have been considered eligible is

not always a feasible option, which is the key assumption that distinguishes our model. In other words, while their model keeps the threshold as a decision variable, in our model we fix a threshold of minus infinity as a restriction, so that first-come first-serve is always the allocation mechanism. Alptekinoglu, Banerjee, Paul and Jain (2012) also have a motivation very close to ours but follow a different solution approach. In their model, they recognize the existence of customer heterogeneity, but rather than assuming a different marginal revenue from selling to different customers, they solve the problem when a pool of inventory (equivalent to our formulation of a single channel) is used to serve customers with varying service level requirements, in order to find the minimum inventory level and the optimal allocation policy that would satisfy all customers. Their model is more extensive in that they consider multiple allocation policies while we assume that customers are provided the good on a first come first serve basis, which is the constraint that drives our counterintuitive results. Also, instead of focusing on achieving a minimum service level, the goal of this chapter is to provide insights on the drivers and service level of the optimal supply chain design, as opposed to limiting it to be a single a pool.

Finally, Swinney (2012) explores the effects of pooling on customer purchasing behavior where forward looking consumers anticipate end-of-season clearance sales and the firm chooses to sell in two markets through a separated selling strategy or a pooled selling strategy, which is an approach almost parallel to ours, but with a different focus. He assumes that prices are time-dependent, rather than demand-dependent, and concentrates the analysis on the change in consumer's optimal time to purchase as a function of the selling strategy. The operational benefits of pooling in his model are consistent with the literature, but finds that when margins are low and demand is positively correlated, having a separated selling strategy may be optimal, and that when consumers are strategic a pooling strategy may decrease consumer welfare by increasing competition during high price periods and increasing the probability of understocks during the clearance period. While his results are driven by the

consumers' expectation on the available inventory at different times, our model also finds that a pooling strategy may hurt social welfare, and that pooling is not necessarily on the manufacturer's nor the system's best interest.

4.3 The Model

4.3.1 General Set-up

We model a two-stage, single period supply chain where a profit-maximizing pharmaceutical manufacturer, hereafter *Pharma*, offers to sell a new (prescription) drug to a utility-maximizer health-payer, hereafter *Health*, through some take-it or leave-it contract arrangement. We consider the health-payer to be interested in the health benefits obtained by the recipients (*i.e.,* patients) of the new drug, and assume that if a contract is signed between Pharma and Health, then both players must honor the contract.

Our focus on healthcare, and more specifically the context of prescription drugs which treat chronic diseases, stems both from its fit with the proposed scope and from its empirical testability in future research. Drugs often apply to more than one category of patients or have more than one indication, and all the potential "consumers" of the drug "compete" to gain access to the same inventory stock. The earlier stated assumptions are met: (1) patients from different categories (*e.g.,* stage 2 and terminal stage cancer patients) are expected to obtain a different and measurable expected health-benefit (*i.e.,* value) by consuming the same drug; (2) regardless of the category of patient that consumes the drug, the health-payer pays a unique unit price to the pharmaceutical manufacturer and if there is a patient co-payment, it is also independent of the patient's category; moreover, for moral, political or economic reasons, health-payers want to make sure that their health investments generate a positive health-benefit on the intended population so that the inclusion of the drug can

be considered cost-effective; (3) the frequency and category of patient arrivals are stochastic, but historical data can be used to develop a probabilistic estimation; and (4) because of ethical reasons, an individual that qualifies for a drug's treatment (given a pre-defined prescription policy threshold) can't be denied the treatment on the basis of the possible future arrival of a patient with higher expected health-benefits derived from the same drug treatment; the contracts with the insurer (public or private) only discuss eligibility, which is independent of available stock.

In terms of the empirical testability of the propositions here stated, note that a patient's change in health as a result of a given medical treatment can be, at least to some extent, measured and translated into economic terms[2]. An a priori estimation of the benefits from following a medical treatment contingent on the patient's conditions can also be estimated, which means both that a cost-effectiveness analysis can be performed to decide if the patient should be treated with the drug, and that a patient's lost surplus from taking a late treatment or alternative option as opposed to the primary drug treatment can be calculated. All this implies that the parameters of the model can be empirically estimated enhancing the applicability of the results.

In stage 1, let $e \geq 0$ be the effort exerted by Pharma directly aimed at increasing demand for its drug in an additive form, *i.e.,* by investing in R&D activities to achieve marginal innovations so that the drug's therapeutical applications are expanded to a new patient category. The cost to Pharma of exerting e is $C(e)$, which is assumed to be increasing and convex in the effort level, and let $C(0) = 0$. In addition to being a common assumption in newsvendor models, a convex cost of effort may represent the greater financial risks, or the increased administration and coordination complexity, associated with higher levels of effort. Before the beginning of stage 2, the players observe whether the innovation efforts were successful

[2]QALY's and DALY's are typical units of measurement. The appropriateness of each is beyond this paper's scope. The interested reader is referred to Airoldi and Morton (2009) for a recent and rich discussion.

or not. Let x be a binary random variable that takes the value 1 if the innovation efforts, e, accomplish their goal, which occurs with probability $g(e)$, and 0 otherwise, with probability $(1-g(e))$. Assume $g(0) = 0,\ g(1) \leq 1,\ g'(e) \geq 0$, and $g''(e) \leq 0$. Assuming the probability of a successful effort to be non-decreasing and concave in the effort level allows us to guarantee concavity of the objective function. It is more than intuitive that increasing effort will not decrease the probability of the effort being successful; and regarding concavity, it represents how initial efforts can make a large impact, but the marginal impact becomes smaller and smaller as the effort level increases.

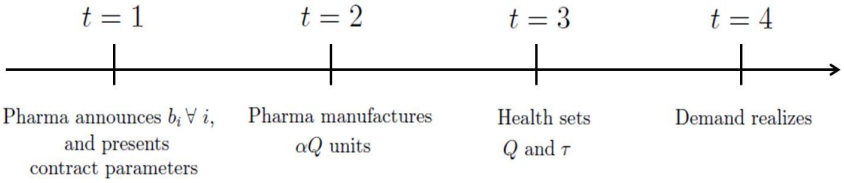
In stage 2, after observing x, Pharma offers a contract to Health.[3] If accepted, and before demand for the drug is realized, Health orders quantity Q of the drug which immediately after is produced by Pharma at constant marginal cost c. Let $D(\varepsilon, e) = \varepsilon + Nx$, be the total quantity of incoming patients (each representing a unit demand) who show-up in order to get treated, either with the drug or with an outside option. ε is a random variable that lies within the interval $[0, \lambda]$ with probability distribution $\Psi(\cdot)$ and density $\psi(\cdot)$, where $\Psi(\cdot)$ is IGFR and is exogenous to efforts;[4] and N is the deterministic number of additional patients that could benefit from taking the drug if the innovation efforts are successful.[5] To differentiate the patient categories and facilitate notation later on, let S_a be the subset of patients from the total demand population, D, for whom the drug can provide positive health benefits, denoted β_a, *before* e is exerted (*i.e.,* the "incumbent" category of patients). Similarly, let S_b be the subset of patients from the total demand population, D, for whom the drug can provide positive health benefits, denoted β_b, *only after, and if, the innovation efforts are successful* (*i.e.,* the "entrant" category of patients that realizes only when $x = 1$). $S_a \cap S_b = 0$, and $|S_a| + |S_b| = D$, where $|y|$ denotes the cardinality of y. For simplicity in

[3]Note that even under advance contract committments, Health would not be bound to his committed order quantity unless $x = 1$; therefore Pharma has no possibility of misrepresenting the value of x.

[4]IGFR: increasing generalized failure rate. The generalized failure rate is the product of a continuous random variable and its hazard rate, *i.e.,* $\frac{x\psi(x)}{1-\Psi(x)}$ is weakly increasing in $x|\Psi(x) < 1$.

[5]In §4.7 we allow for the size of the entrant category of patients to be stochastic.

Figure 4.1: Sequence of decisions and events



the analysis we also define Q_a (Q_b) as the order quantity that corresponds to the stochastic (deterministic) part of the demand; and $Q = Q_a + Q_b$. When both patient categories are served by the same channel and there is excess demand, then patients gain access to the drug on a first-come first-serve basis. The timing is depicted in Figure 4.1. There are no holding costs, and any excess inventory has no salvage value nor incurs any additional cost of disposal. Common knowledge is assumed among the players regarding all functional forms and cost parameters.

In what follows we will characterize the solution to the model introduced above under two vertical arrangements: Integrated chain, and disintegrated chain with an eXogenous price-only contract; and with two different supply chain designs: Multiple dedicated distribution channels, and a Single distribution channel. When used as a superindex on the decision variables, they will denote the optimal solution for that particular scenario.

4.3.2 Multiple channels under vertical integration

We first analyze the situation where each patient category is served through a different channel. This can be interpreted as a drug that is marketed under a different brand name, packaging, or delivery method for each consumer category, or as perfectly separated channels where each patient category is served only at its corresponding dedicated channel. In either case, it is assumed that the drugs in each channel have an identical cost structure, and that there is zero substitutability across channels. The problem corresponding to the patients

belonging to S_a is:

$$Z_{S_a}^{MI}(Q_a) = -cQ_a + \beta_a \left(\int_0^{Q_a} \xi\psi(\xi)\ d\xi + \int_{Q_a}^{\lambda} Q_a\psi(\xi)\ d\xi \right) \tag{4.3.1}$$

The first term is the cost of producing the drug, and the second term is the expected health-benefits achieved through the drugs that are given to patients. Since the problem is concave, taking the first-order condition we obtain:[6]

$$\begin{aligned} &(\beta_a)\left[1 - \Psi\left(Q_a^{MI}\right)\right] - c = 0 \\ &\Rightarrow Q_a^{MI} = \Psi^{-1}\left(1 - \frac{c}{\beta_a}\right) \end{aligned} \tag{4.3.2}$$

For the patients belonging to S_b, the problem in the second stage is deterministic:

$$Z_{S_b,2}^{MI}(Q_b) = \beta_b \min[Q_b, Nx] - cQ_b\ , \tag{4.3.3}$$

and clearly:

$$Q_b^{MI} \ = \ \begin{cases} N & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.3.4}$$

Next, recall that $\mathbb{E}[x] = g(e)$; therefore $\mathbb{E}[Q_b^{MI}] = Ng(e)$, and the problem in the first stage is:

$$Z_{S_b,1}^{MI}(e;\, Q_b^{MI}) = -C(e) + (\beta_b - c)\, Ng(e) \tag{4.3.5}$$

The first term is the cost of exerting effort and the second term is the expected benefit

[6] This is the basic newsvendor formulation. The second order condition is $\frac{\partial^2}{\partial Q^2} Z_{S_a}^{MI}(Q_a) = -\beta_a \Psi'\left(Q_a^{MI}\right) \leq 0$, which guarantees concavity.

minus the production cost generated from the drugs that are given to patients. The objective is clearly a concave function as it is, by definition, the sum of two concave functions. Taking the first-order condition with respect to the effort level we obtain:

$$C'(e^{MI}) \;=\; (\beta_b - c)\, N g'(e^{MI}) \tag{4.3.6}$$

The results presented so far are standard knowledge in the operations management literature. Summarizing, the expected quantity of drugs available is: $Q_a^{MI} + Ng(e^{MI})$, the actual quantity of drugs available after observing the outcome of the exerted effort will be: $Q_{MI} = Q_a^{MI} + Q_b^{MI}$.

4.3.3 Single channel under vertical integration

Now consider the case of a single decision-maker when both patient categories compete on a first-come first-serve basis to gain access to the same inventory of the drug, *i.e.,* when there is a single channel that serves both patient categories. As usual, we solve the problem by backwards induction. Instead of solving for Q, we can solve for Q_a, taking into account that $Q_a^{SI} = Q^{SI} - xN$.[7] In the second stage, after the realization of x which is a function of the efforts e, the single decision-maker solves:

$$\begin{aligned} Z_2^{SI}(Q) = -cQ &+ \int_0^{Q_a} \left[\beta_a \xi + \beta_b x N\right] \psi(\xi)\; d\xi \\ &+ Q \int_{Q_a}^{\lambda} \left[\beta_a \left(\frac{\xi}{\xi + xN}\right) + \beta_b \left(\frac{xN}{\xi + xN}\right)\right] \psi(\xi)\; d\xi \end{aligned} \tag{4.3.7}$$

The first term in the objective function is the cost of producing the drug's order quantity. The second term is the health-benefits obtained by patients belonging to S_a and S_b when the available inventory is sufficient to cover all the demand. The third term captures the

[7] Recall that $x = 0$ if the innovation is unsuccessful, and $x = 1$ if the innovation is successful.

situation when the stochastic part of the demand (*i.e.,* the inflow of patients belonging to S_a) exceeds the drug quantity Q_a intended to cover that demand; as a result, there is a positive probability that patients from S_a will consume the drugs intended to satisfy patients from S_b; this probability depends on the relative sizes of the demand coming from patients of each category. We introduce the definition of such phenomena as an *inventory pulling effect* since the category of patients that composes the stochastic portion of the demand *pulls* inventory from a stockpile that was intended for the category of deterministic size.

To provide additional intuition about the last term in (4.3.7), consider the following. Since there is no information about the order of arrivals, then for a given realized demand, the order of arrivals is completely random. For example, if $x = 1$ and $D = N + d > Q$, then the expected fraction of patients who receive the drug that belong to type b (*i.e.,* the fraction of demand which belongs to S_b and arrives within the first Q patients) is $(N/(N + d))$, and the expected number of type b patients who receive the drug is the number of drugs available *times* the fraction of drugs allocated to S_b patients: $Q\,(N/(N + d))$. Since the value of d is stochastic, the formulation needs to integrate over all values of $d \geq Q$.

Define $Q^{SI} \in \arg\max\ Z_2^{SI}(Q)$. Therefore:

$$Q^{SI} \;=\; \begin{cases} Q_{a,1}^{SI} + N & \text{if } x = 1 \\ Q_{a,0}^{SI} & \text{otherwise} \end{cases} \tag{4.3.8}$$

Using backwards induction, the first order condition for the optimal order quantity is as follows:

When $x = 1$:

$$(\beta_a)\left[1 - \Psi\left(Q_{a,1}^{SI}\right)\right] - c + (\beta_b - \beta_a)\int_{Q_{a,1}^{SI}}^{\lambda} \frac{xN}{\xi + xN}\psi(\xi)d\xi = 0 \tag{4.3.9}$$

When $x = 0$:

$$(\beta_a)\left[1 - \Psi\left(Q_{a,0}^{SI}\right)\right] - c = 0 \tag{4.3.10}$$

The proof for equations (4.3.9) and (4.3.10) is provided in the Appendix. A straightforward, yet important to notice, result is that $Q_{a,0}^{SI} = Q_a^{MI}$, which means that when effort is unsuccessful the order quantity in the system is independent of the channel structure; this follows by direct comparison of (4.3.2) and (4.3.10). A second interesting observation that is also consistent with previous results in the literature is that if $\beta_a = \beta_b$, then the optimal order quantity is exactly the same as in the case of multiple channels; this follows because the last term in (4.3.9) becomes zero, and the equation collapses into (4.3.2). While we cannot get a closed form solution for $Q_{a,1}^{SI}$ when $\beta_a \neq \beta_b$, we can take advantage of the fact that the problem is concave in the order quantity. Namely, we can observe whether the total order quantity increases or decreases with respect to the dedicated channels case by observing the sign of the slope when we replace $Q_{a,1}^{SI}$ with Q_a^{MI} into equation (4.3.9). These results are summarized in Proposition 16.

Proposition 16: a) When $x = 0$, $Q^{SI} = Q^{MI}$. b) When $x = 1$ and $\beta_a = \beta_b$, then $Q^{SI} = Q^{MI}$. c) When $x = 1$ and $\beta_a > \beta_b$, then $Q^{SI} < Q^{MI}$. d) When $x = 1$ and $\beta_a < \beta_b$, then $Q^{SI} > Q^{MI}$.

Proposition 16a simply states that when there is only one category of consumers, the optimal order quantity is the same as in the multiple channels structure, as was expected. Similarly, Proposition 16b states that even if two patient categories exist, when the benefit received by members of both populations are equivalent (or considered as equivalent), then

the order quantity is also the same as in the multiple channels case. More interestingly, Proposition 16c says that when a second patient category exists and the benefit obtained by a member belonging to the second category is lower than that obtained by a member from the first category, the buyer will increase its inventory in a proportion lower than the increase in expected demand (*i.e.,* $(Q^{SI} < Q_a^{MI} + xN)$), for reasons other than the pooling effect. The intuition is that the pulling effect creates an externality on the consumption of the drug for the patients belonging to S_b, which is ignored in the case of independent monopolies. As a result, the relative weight that the decision-maker allocates to understocking for the stochastic part of the demand (*i.e.,* patients $\in S_a$) decreases because in case of excess demand, members from the first category may pull inventory from the second category, thus resulting in a lower total order quantity. Finally, Proposition 16d states the opposite scenario, so that when the benefits obtained by a member from the second category are greater than those of the first category, then the relative cost of excess demand is larger because of the same pulling effect, thus resulting in a larger total quantity, even if both demands are managed by the same profit maximizer. For the decision-maker, the pulling effect implies that the quantity of drugs manufactured and the resulting expected quantity of patients treated will be higher (respectively, lower) for a single drug that treats multiple patient categories than for multiple category-specific drugs when the category of patients with less demand stochasticity receives higher (respectively, lower) benefits from the drug. While the relevance of this observation will become more apparent in following sections, for the moment it further adds to the discussion between marginal innovations of existing compounds (*e.g.,* a single product, or product line extensions) versus new drugs development (*e.g.,* multiple products or product differentiation).

A crucial aspect to notice at this point is that the changes in the optimal order quantities at stage 2 do not depend on how much effort was exerted at stage 1; they only depend on whether the effort was successful or not (i.e., $x = 0$ or $x = 1$). Having made this point

explicit, let us look at how the demand increasing efforts are affected by the pulling effect. In the first stage, the problem that the single decision-maker solves is:

$$\begin{aligned} Z_1^{SI}(e;\, Q^{SI}) = & -C(e) - c[g(e)(Q_{a,1}^{SI} + N) + (1 - g(e))Q_{a,0}^{SI}] \\ & + g(e) \left[\int_0^{Q_{a,1}^{SI}} [\beta_a \xi + \beta_b N]\, \psi(\xi)\, d\xi \right. \\ & \left. + (Q_{a,1}^{SI} + N) \int_{Q_{a,1}^{SI}}^{\lambda} \left[\beta_a \left(\frac{\xi}{\xi + N} \right) + \beta_b \left(\frac{N}{\xi + N} \right) \right] \psi(\xi)\, d\xi \right] \\ & + (1 - g(e)) \left[\int_0^{Q_{a,0}^{SI}} \beta_a \xi \psi(\xi)\, d\xi \; + \; Q_{a,0}^{SI} \int_{Q_{a,0}^{SI}}^{\lambda} \beta_a \psi(\xi)\, d\xi \right] \end{aligned}$$

subject to:

$$Q^{SI} \in \arg\max \quad Z_2^{SI}(Q) \tag{4.3.11}$$

The only variations in the objective function of the first stage relative to the second stage are that: (1) the innovation effort cost is now included (in stage 2, this was considered a sunk cost); and (2) x is still a random variable and dependent on the decision-maker's choice of e, which impacts the expected order quantity. Although closed form solutions for the level of effort can't be obtained either - because the optimal order quantity can't be plugged in equation (4.3.11) -, the first order condition can be expressed as follows (*if needed, see details in the proof of Proposition 17 in the Appendix*):

$$\frac{\partial Z_1^{SI}(e;\, Q^{SI})}{\partial e} = -C'(e) - g'(e) \left(Z_2^{SI}(Q_{a,0}^{SI}) \right)\Big|_{x=0} + g'(e) \left(Z_2^{SI}(Q_{a,1}^{SI} + N) \right)\Big|_{x=1} = 0 \tag{4.3.12}$$

This structure is sufficient to determine whether the incentives to exert effort have increased or decreased relative to the multiple channels structure, based on the change in profit generated when the efforts are successful. This first result is expressed in Lemma 23.

Lemma 23:

a)Suppose x=0. Then $Z_2^{SI}(Q^{SI}) = Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(Q_b^{MI})$.

b) Suppose x=1. Then:

b1) if $\beta_a = \beta_b$, then $Z_2^{SI}(Q^{SI}) = Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(Q_b^{MI})$;

b2) if $\beta_a > \beta_b$, then $Z_2^{SI}(Q^{SI}) > Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(Q_b^{MI})$;

b3) if $\beta_a < \beta_b$, then $Z_2^{SI}(Q^{SI}) < Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(Q_b^{MI})$.

Part a of Lemma 23 is trivial because when $x = 0$, then $Z_{S_b,2}^{MI}(Q_b^{MI}) = 0$, and from Proposition 16, the order quantity in the system will be the same under any structure because there will be only one patient population. Part b, however, does benefit from a more elaborate explanation. Observe that the total health benefits derived from the deterministic portion of the demand can't be higher than $\beta_b N$, and by definition, Q_a^{MI} minimizes the cost of uncertainty in the multiple channel scenario for patients of type a. Notice then that there are two ways in which the value of the objective function can increase. One is to reduce inventory costs, and the other is to increase the health benefits achieved by the patients. If under a single channel structure, the inventory was set equal to Q^{MI}, then patients from S_a are guaranteed at least a private stock of Q_a^{MI}, because those from S_b will not consume more than N drugs. Therefore setting $Q^{SI} < Q^{MI}$ (alternatively, $Q^{SI} > Q^{MI}$) because $\beta_a > \beta_b$ (alternatively, $\beta_a < \beta_b$) is intended to take advantage from (alternatively, restrict) the pulling effect. Another way to see this is that given excess demand under the single channel structure, then for every patient belonging to S_b who does not get the drug, which happens when a patient from S_a arrives first and receives the drug instead, a net margin of $(\beta_a - \beta_b)$ is generated. On one hand, when the margin is positive, i.e., $\beta_a > \beta_b$, expected health benefits for the same inventory level increase, also increasing net utility; since it has been shown that $Q^{SI} < Q^{MI}$ under these circumstances, the explanation is that the total inventory is lowered compared to the dedicated channels up to the point where the decrease in purchasing costs is no longer justified by the health benefits achieved thanks to the pulling effect. On the

other hand, when the margin is negative, i.e., $\beta_a < \beta_b$, expected health benefits for the same inventory level decrease, also decreasing net utility; lowering the total inventory would increase the negative consequences of the pulling effect, and since it has been shown that $Q^{SI} > Q^{MI}$ under these circumstances, it must be that both inventory costs are rising and health benefits are dropping compared to the dedicated channels case. The inventory then is increased up to the point where the increase in purchasing costs is no longer justified by the savings achieved by limiting the pulling effect. In summary, since profits always increase (respectively, decrease) in SI with respect to MI when $\beta_a > \beta_b$, (respectively, $\beta_b > \beta_a$), then the incentives to exert innovation effort are higher (respectively, lower) in SI. The result is summarized in Proposition 17.

Proposition 17: a) When $\beta_a = \beta_b$, then $e^{SI} = e^{MI}$. b) When $\beta_a > \beta_b$, then $e^{SI} > e^{MI}$. c) When $\beta_a < \beta_b$, then $e^{SI} < e^{MI}$.

Proposition 17 provides interesting managerial implications. First, Proposition 17a says that when the benefits are equivalent to both populations, then the channel structure has no effect on the effort level. However, when $\beta_a > \beta_b$, (alternatively $\beta_a < \beta_b$), Proposition 17b (alternatively, Proposition 17c) states that innovation efforts will increase (decrease) despite the fact that the total inventory decreases (increases). This is because of the decreased (increased) inventory cost and the reallocation of the available inventory created by the pulling effect.

Theorem 5: (a) if $\beta_a > \beta_b$, having a single sales channel for all patient categories maximizes expected total welfare; (b) if $\beta_a < \beta_b$, having separate sales channels for each consumer category maximizes expected total welfare; and (c) in both cases, the efficient channel structure maximizes innovation efforts and in the event of a successful innovation, results in (weakly) less coverage with respect to the one that would be achieved under the inefficient channel

structure.

Theorem 5 summarizes the chapter's first key contribution. The first takeaway is that the design of the efficient channel structure depends on which patient category obtains higher benefits from the drug. Theorem 5a implies that when the marginal benefits for the entrant population of consumers are lower than those of the incumbent consumers' population, then it is more efficient to serve both populations with a unique multi-purpose product rather than serving each population with a specific product. Another implication is that even if a single channel is used to distribute the drug to both patient categories, if the order quantity is chosen based on independent forecasts and then put together, then the single channel will order a higher than efficient order quantity. Theorem 5b describes the opposite case, such that when the entrant population has a higher valuation of the product than the incumbent population it is efficient that each population manages its own inventory stock (or equivalently, that a different product serves each population), rather than keeping a common inventory (equivalently, a single product). Finally Theorem 5c states that choosing the most efficient channel structure increases the probability that a drug that treats the second category of patients exists, but when the efforts are successful and the drug does exist, then total quantity of drugs available in the system will be (weakly) lower than under the inefficient channel structure.[8]

[8] Although it may be obvious to the reader, it is worth mentioning that even though expected utility will be higher given the choice of the efficient channel, the realized utility may not be so. The most explicit example is when the effort is unsuccessful. Since the level of effort exerted is higher under the efficient channel, then so is the cost of such effort. Therefore, if despite higher effort levels, $x = 0$, then the utility generated at stage 2 will be the same for both structures (Lemma 23a), but the cost of effort incurred at stage 1 will be higher under the efficient channel design (Proposition 17), resulting in a lower total utility.

4.4 Exogenous price-only contracts

In this section, assume that Pharma sells the drug to Health at some exogenously determined price w, where $\min[\beta_a, \beta_b] > w > c$. As before, we will first analyze the problem under multiple channels where each channel is used to satisfy a different patient category, and then we'll extend the analysis to the single channel case.

4.4.1 Multiple Channels with Exogenous Price (MX)

The problem for Health's channel that is in charge of treating the incumbent population is:

$$H_{S_a}^{MX}(Q_a) = -wQ_a + \beta_a \left(\int_0^{Q_a} \xi\psi(\xi)\, d\xi + \int_{Q_a}^{\lambda} Q_a\psi(\xi)\, d\xi \right) \tag{4.4.1}$$

Compared to the integrated supply chain arrangement, Health now replaces the manufacturing cost c with the transfer price w in its objective function. Taking the first-order condition:

$$Q_a^{MX} = \Psi^{-1}\left(1 - \frac{w}{\beta_a}\right) \tag{4.4.2}$$

By direct comparison, we can observe that $Q_a^{MX} < Q_a^{MI}$, which is also a known result in the literature due to double marginalization. As for Health's channel serving the patients belonging to S_b, the problem is:

$$H_{S_b,2}^{MX}(Q_b) = \beta_b \min[Q_b, Nx] - wQ_b \tag{4.4.3}$$

where clearly:

$$Q_b^{MX} = \begin{cases} N & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.4.4}$$

In the first stage, the problem for Pharma is:

$$\begin{aligned} M^{MX}(e) &= -C(e) + (w-c)(Q_a^{MX} + Ng(e)) \\ &\text{subject to:} \\ & Q_a^{MX} = \Psi^{-1}\left(1 - \frac{w}{\beta_a}\right) \end{aligned} \tag{4.4.5}$$

The first term is the cost of exerting effort and the second term is the marginal profit generated from the drugs that are sold to Health. Taking the first-order condition we obtain:

$$C'(e^{MX}) = (w-c)\, Ng'(e^{MX}) \tag{4.4.6}$$

In summary, the only difference is the double-marginalization effect, which will decrease the order quantity and the effort exerted, relative to the vertically integrated chain.

4.4.2 Single Channel with Exogenous Price (SX)

The problem for Health is:

$$\begin{aligned} H^{SX}(Q) = & -wQ + \int_0^{Q_a} [\beta_a \xi + \beta_b x N]\, \psi(\xi)\, d\xi \\ & + Q \int_{Q_a}^{\lambda} \left[\beta_a \left(\frac{\xi}{\xi + xN}\right) + \beta_b \left(\frac{xN}{\xi + xN}\right)\right] \psi(\xi)\, d\xi \end{aligned} \tag{4.4.7}$$

As before, we solve for Q_a instead of Q, recalling that $Q_a^{SX} = Q^{SX} - xN$. Define

$Q^{SX} \in \arg\max\ H^{SX}(Q)$. Therefore:

$$Q^{SX} = \begin{cases} Q^{SX}_{a,1} + N & \text{if } x = 1 \\ Q^{SX}_{a,0} & \text{otherwise} \end{cases} \tag{4.4.8}$$

Using backwards induction, the first order condition for the optimal order quantity is as follows:

When $x = 1$:

$$(\beta_a)\left[1 - \Psi\left(Q^{SX}_{a,1}\right)\right] - w + (\beta_b - \beta_a)\int_{Q^{SX}_{a,1}}^{\lambda} \frac{xN}{\xi + xN}\psi(\xi)d\xi = 0 \tag{4.4.9}$$

When $x = 0$:

$$(\beta_a)\left[1 - \Psi\left(Q^{SX}_{a,0}\right)\right] - w = 0 \tag{4.4.10}$$

Once more, it is straightforward to observe that given $w > c$, then both $Q^{SX}_{a,1} < Q^{SI}_{a,1}$, and $Q^{SX}_{a,0} < Q^{SI}_{a,0}$. The problem for Pharma is:

$$M^{SX}(e;\, Q^{SX}) = -C(e) + (w - c)\left(g(e)(Q^{SX}_{a,1} + N) + (1 - g(e))Q^{SX}_{a,0}\right)$$

subject to:

$$Q^{SX} \in \arg\max\ \ H^{SX}(Q) \tag{4.4.11}$$

Taking the first-order condition we obtain:

$$C'(e^{SX}) = (w - c)\left(g'(e^{SX})(N + Q^{SX}_{a,1} - Q^{SX}_{a,0}\right) \tag{4.4.12}$$

We now begin our analysis of the supply chain's design incentive misalignment under the vertically separated chain.

Lemma 24: a) When $x = 0$, $Q^{SX} = Q^{MX}$. b) When $x = 1$ and $\beta_a = \beta_b$, then $Q^{SX} = Q^{MX}$. c) When $x = 1$ and $\beta_a > \beta_b$, then $Q^{SX} < Q^{MX}$. d) When $x = 1$ and $\beta_a < \beta_b$, then $Q^{SX} > Q^{MX}$.

Lemma 24 shows that the results from the previous section regarding the impact of the pulling effect on the optimal order quantity carry on to the case when there is vertical separation and the players contract through exogenously determined price-only contracts. However, the direction of the results in the first stage are no longer consistent with §4.3, as is expressed in Proposition 18.

Proposition 18: a) If $\beta_a = \beta_b$, then $e^{SX} = e^{MX}$; b) If $\beta_a > \beta_b$, then $e^{SX} < e^{MX}$; c) If $\beta_a < \beta_b$, then $e^{SX} > e^{MX}$.

Proposition 18 shows an important difference when the players act separately. In the vertically integrated setting from the previous subsection, the effort decision was made based on the system's profits. In the case when the innovation effort was successful, the incremental profits were higher (respectively, lower) when $\beta_a > \beta_b$, (respectively, $\beta_b > \beta_a$), recalling that this occurred as a consequence of lower (respectively, higher) total order quantities. However, in the current contracting setting, Pharma makes her effort decision based only on the quantity of drugs that she sells and not necessarily on the benefits that are transferred to the population of patients. Comparing between the channel structures, Proposition 18b and Proposition 18c say that the intensity of Pharma's effort are not aligned with the net utility consequences. Namely, Proposition 18b, (respectively, Proposition 18c) says that when $\beta_a > \beta_b$, (respectively, $\beta_a < \beta_b$), Pharma will have less (respectively, more) incentives to exert innovation effort when there is a single channel versus when there are multiple channels; but

using the results from section 4.3, Pharma's optimal decisions result in a higher probability of the innovation being successful if the inefficient channel structure is in place. A second way of interpreting the results is that if the efficient channel structure is imposed by some external agent, then Pharma's efforts will be lower than if she were allowed to operate under an inefficient channel structure. The conclusions from this section are written in Theorem 6 and serve as the motivation to explore a risk sharing contract that links Pharma's payoff to the realized benefits of the drug, which is the topic of section 4.6.

Theorem 6: Under exogenous price-only contracts, (a) the pharmaceutical manufacturer benefits, on expectation, from selecting the inefficient channel structure; and (b) selecting the inefficient channel structure provides the pharmaceutical manufacturer with higher incentives to exert innovation effort and increases the probability that a second patient category exists.

4.5 Numerical Studies

To gain additional intuition, we have run numerical studies, for which some graphs are presented next. In addition to confirming the direction of the pulling effect which was analytically proven, we are able to gain additional insights on the key drivers that determine the required adjustment in the optimal order quantity as a result of the aforementioned pulling effect. First, from Figure 4.2, we find that the required adjustment on the order quantity, whether it is upwards or downwards, increases in the size of the population with deterministic demand, N.[9] The intuition is that a higher relative size of N increases the ability to pull drugs from the deterministic portion of the stock and therefore: when $\beta_a > \beta_b$, the pulling effect reduces the inventory stocking needs (so the decrease in the order quantity is greater as N increases). Alternatively, when $\beta_a < \beta_b$, there is a need to restrict the pulling effect,

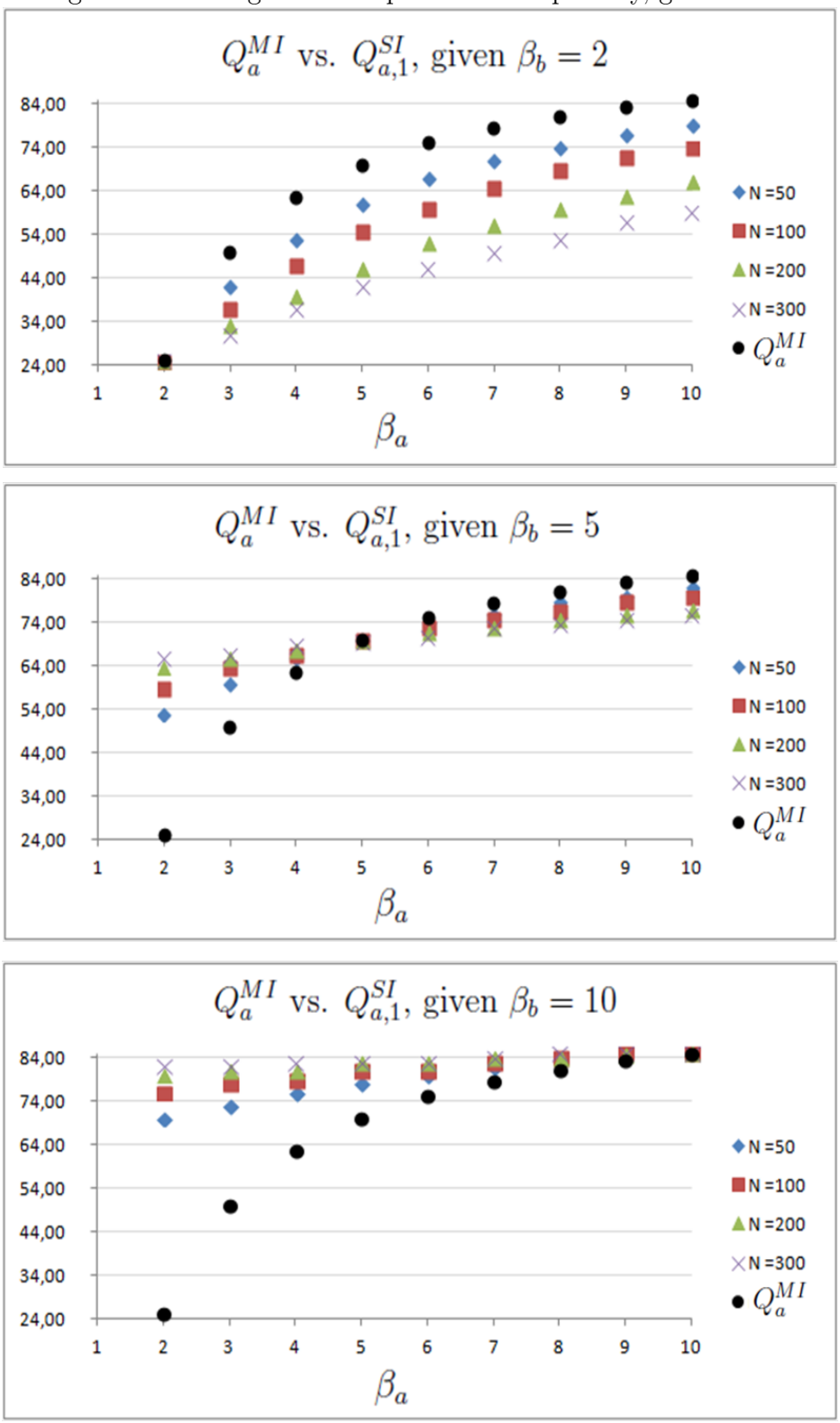
[9]Recall that when $\beta_a = \beta_b$, or when $x = 0$, it has been shown that no adjustment is made.

which is achieved by increasing the inventory stock; in such situation, the increase in the order quantity is greater as N increases, *i.e.,* as the severity of the pulling effect increases. Secondly, from Figure 4.3 we confirm that the change in the optimal net utility also increases as the size of N increases. The explanation is a consequence of the changes in the optimal order quantity; the larger the profit impact of moving from a dedicated to a single channel arrangement, the larger the correction in the order quantity will tend to be.

Third, Figures 4.4 and 4.5 show (from top to bottom) the relative change in utility, the relative change in the order quantity, and the absolute change in the order quantity. We can observe that when $\beta_b > \beta_a$, the required adjustment (which is upwards) increases as β_b increases. The logic is that the higher the relative value of β_b, the more important it becomes to prevent the pulling effect, and therefore a larger adjustment is needed. On the complementary part, when $\beta_b < \beta_a$, the required adjustment (which is downwards) decreases as β_b increases. This occurs because when β_b is very small, the desired pulling effect will be large, but as β_b grows, the deterministic benefit becomes more desirable, and it is no longer optimal to reduce so much the total order quantity, *i.e.,* the adjustment becomes smaller. A summarizing way to interpret the relationship between β_a and β_b is that for fixed β_a, when the difference in the expected health benefits from patients belonging to S_a and S_b is low, then the relevance of the pulling effect decreases, reducing the need to distort the order quantity. This is consistent with equation 4.3.10 which defines the optimal order quantity under a single integrated channel, Q_a^{SI}, since the magnitude of the slope (the first order derivative) increases with the difference between the benefits from each category.

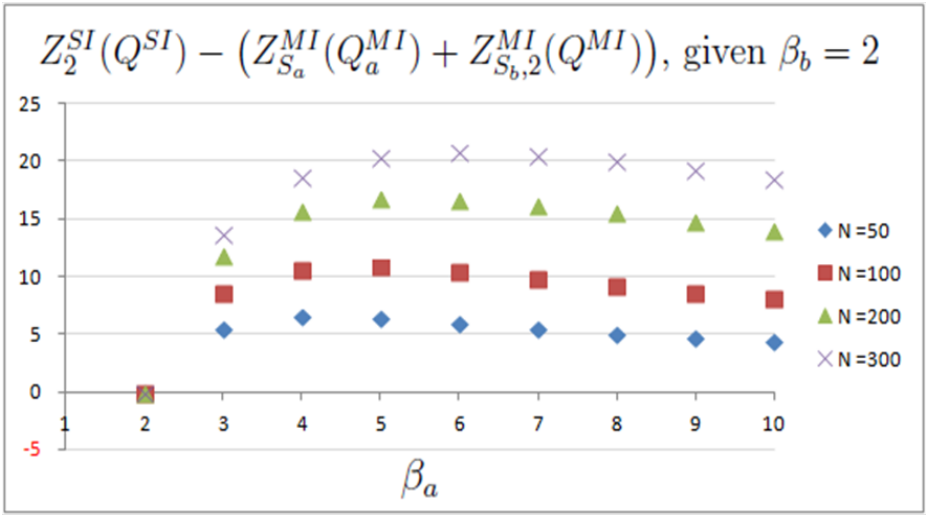
Fourth, Figures 4.6 and 4.7 provide the relationship between the effort levels exerted at stage 1, depending on whether the design is that of a single or dedicated channels. These results are traced back to the previous figures which showed if the change in utility when $x = 1$ is positive or negative relative to the dedicated channels structure. Since the latter

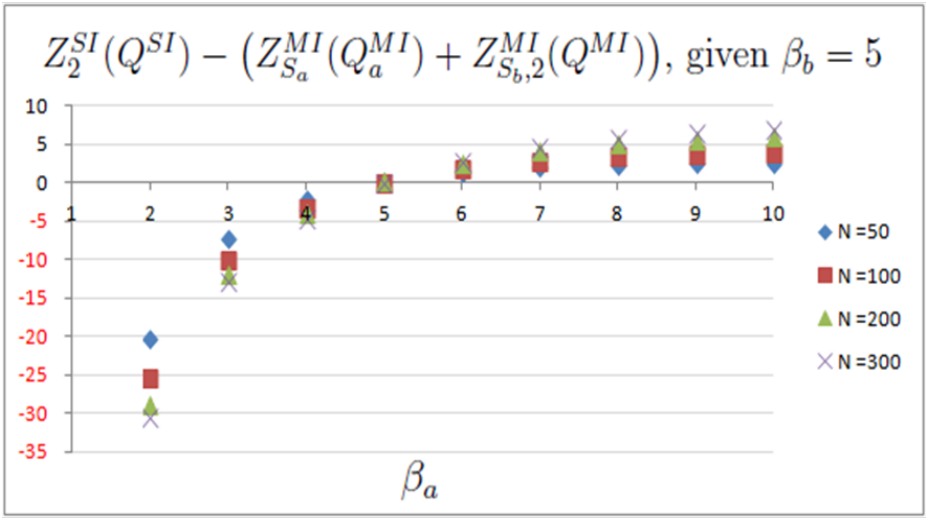
Figure 4.2: Changes in the optimal order quantity, given $x = 1$

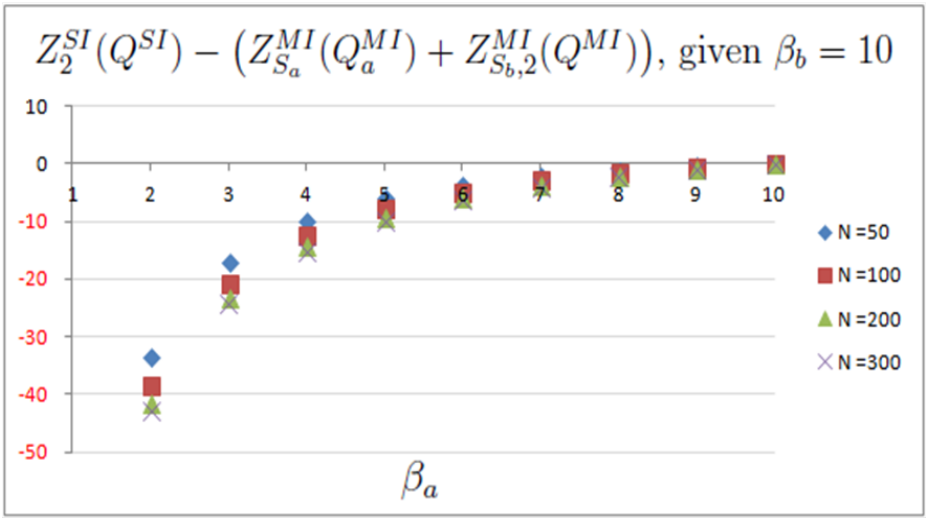


$\varepsilon \sim$ Uniform $[0, 100]; c = 1.5$

Figure 4.3: Changes in optimal net utility for different values of N, given $x = 1$

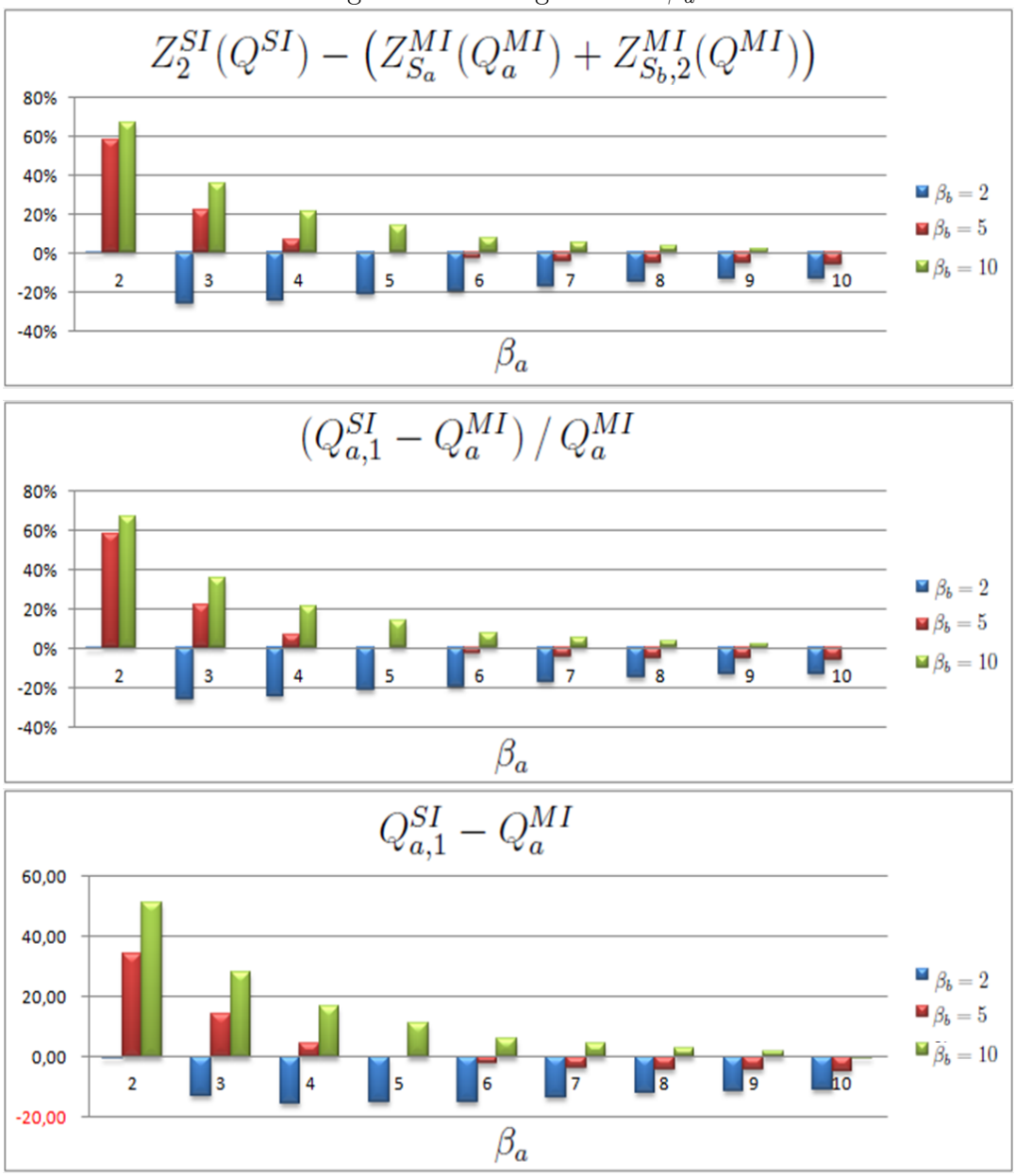






$\varepsilon \sim$ Uniform $[0, 100]; c = 1.5$

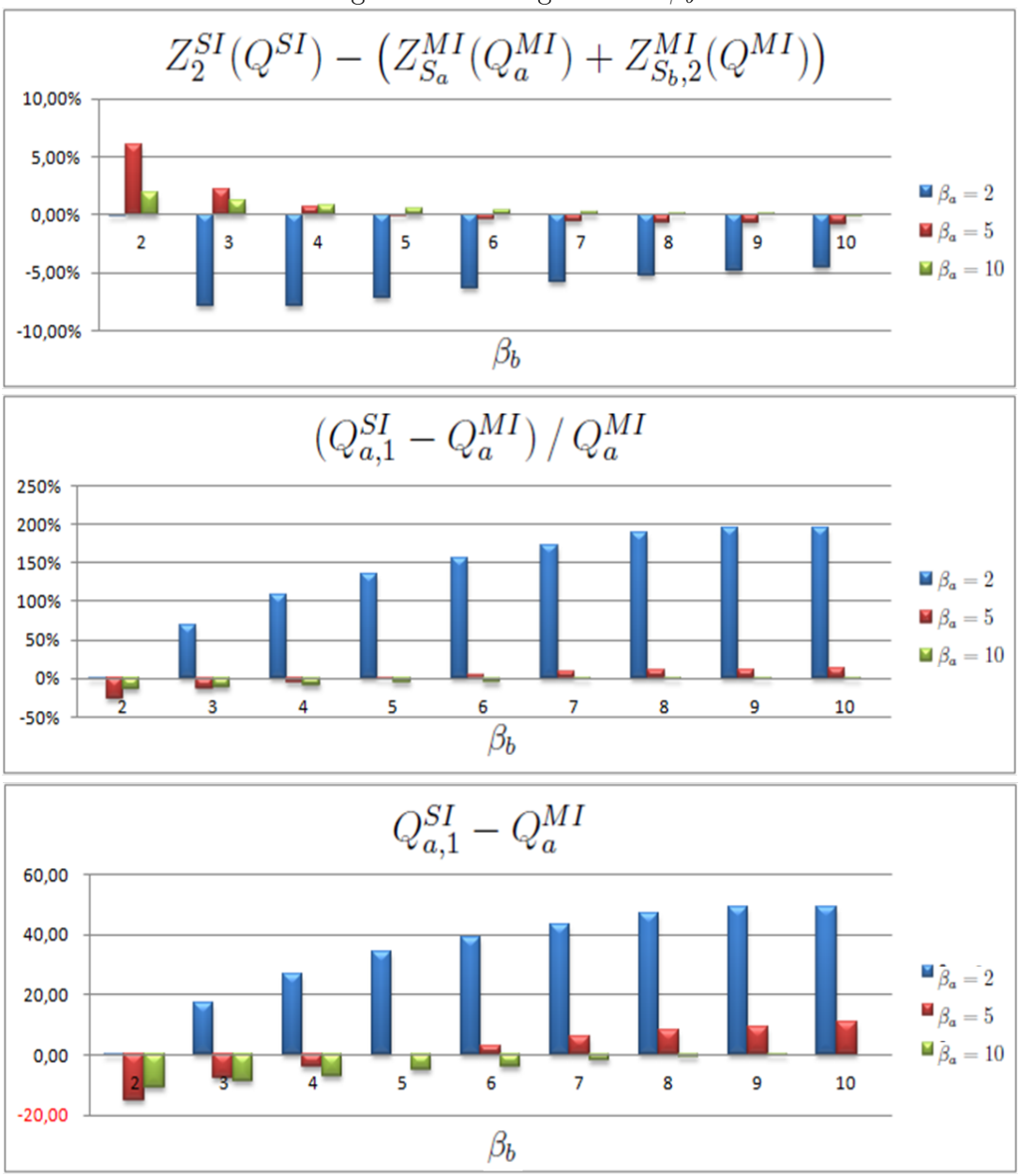
Figure 4.4: Changes across β_a



$\varepsilon \sim$ Uniform $[0, 100]; N = 100; c = 1.5$

relationship depends only on whether $beta_a \gtrless \beta_b$, then the effort levels are driven by the same mechanics. As the net increase in the single channel versus the dedicated channels design, derived from $x = 1$, is positive then $e^{SI}/e^{MI} > 1 \Rightarrow e^{SI} > e^{MI}$. Figures 4.6 and 4.7 describe this behavior for two specific functional forms of $g(e)$ and $C(e)$. However, the results evidently hold as long as the conditions described in the model setup hold.

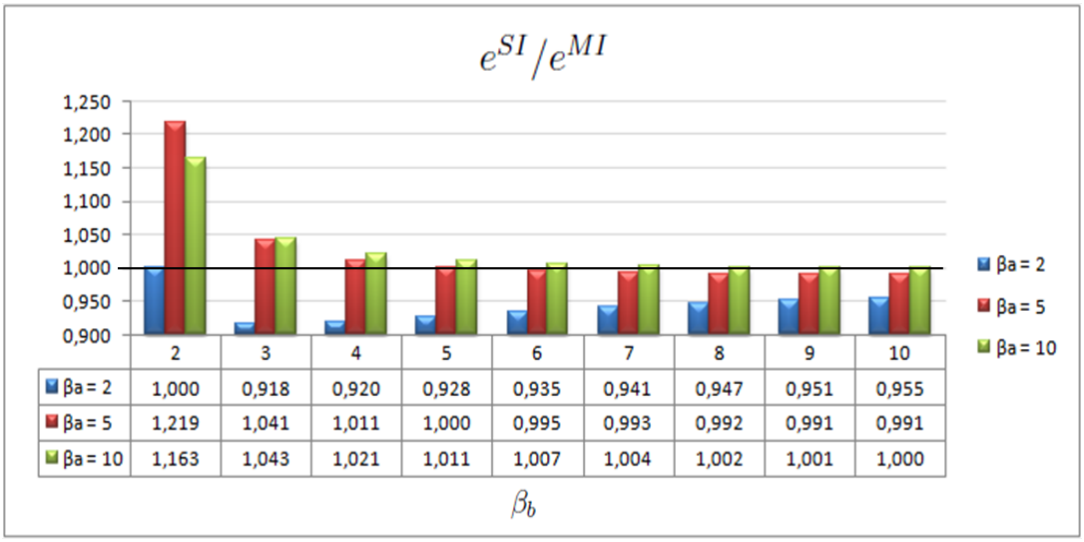
Figure 4.5: Changes across β_b



$\varepsilon \sim$ Uniform $[0, 100]; N = 100; c = 1.5$

Finally, it's worth mentioning that as can be observed from the figures, the relative adjustment in terms of quantity, and specially in terms of expected utility, may be quite considerable. Therefore, the optimal choice of the design of the supply chain is likely to have a relevant impact both on consumers, and on the player selecting the design and supplying the good to be consumed. Motivated by this, the next section proposes some alternatives to correct the incentive misalignments arising in the vertically disintegrated chain, as well as

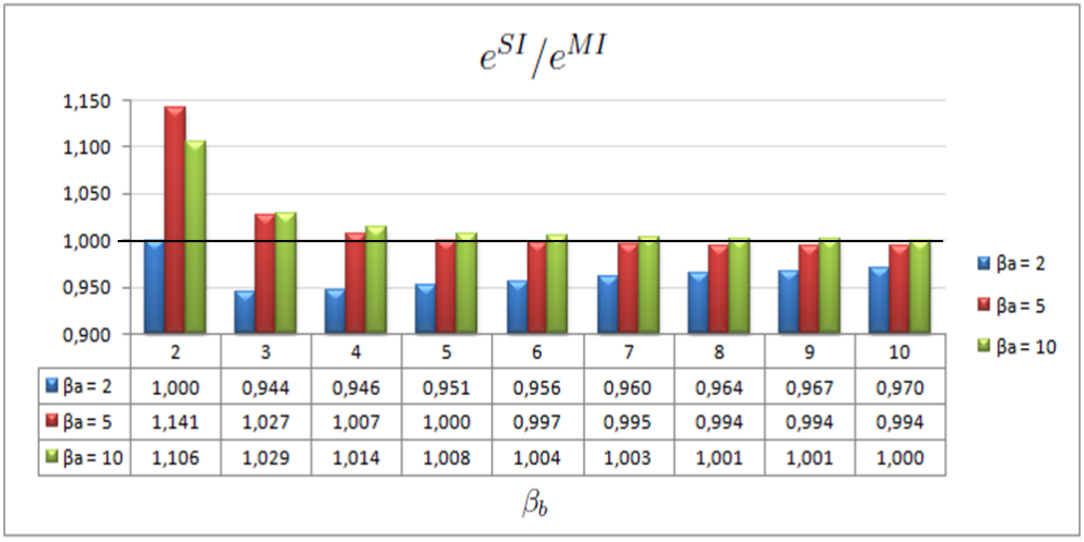
Figure 4.6: Changes in the relative effort levels across β_b (example 1)



	2	3	4	5	6	7	8	9	10
βa = 2	1,000	0,918	0,920	0,928	0,935	0,941	0,947	0,951	0,955
βa = 5	1,219	1,041	1,011	1,000	0,995	0,993	0,992	0,991	0,991
βa = 10	1,163	1,043	1,021	1,011	1,007	1,004	1,002	1,001	1,000

$\varepsilon \sim$ Uniform $[0, 100]; N = 100; c = 1.5; g(e) = e; C(e) = e^2/2$

Figure 4.7: Changes in the relative effort levels across β_b (example 2)



	2	3	4	5	6	7	8	9	10
βa = 2	1,000	0,944	0,946	0,951	0,956	0,960	0,964	0,967	0,970
βa = 5	1,141	1,027	1,007	1,000	0,997	0,995	0,994	0,994	0,994
βa = 10	1,106	1,029	1,014	1,008	1,004	1,003	1,001	1,001	1,000

$\varepsilon \sim$ Uniform $[0, 100]; N = 100; c = 1.5; g(e) = e^{1/2}; C(e) = e^2$

extending the applicability of the results by relaxing some of the previous assumptions.

4.6 Extensions to the Model

4.6.1 Coordinating the supply chain design

At the end of the exogenous price analysis, it was noted that a social planner would be faced with a complicated decision in terms of choosing the optimal supply chain design. For example, if $\beta_a > \beta_b$, then it has been established that the optimal design is to have a single channel that can benefit from the pulling effect. However, the probability of the second patient category materializing is contingent on Pharma's efforts, but Pharma has no incentive neither to choose the efficient supply chain design, and even if it is imposed, to exert the level of effort that would benefit the supply chain. One option, while not a very advisable one, is for the social planner to allow the establishment of the inefficient supply chain design in order to provide the manufacturer with higher incentives to exert innovation effort, and therefore increase the probabilities of achieving a higher access level, even if it comes at a larger than optimal cost. A second option aimed at providing the manufacturer with the right incentives to innovate without sacrificing the efficient choice of the supply chain design is provided next based on a simple lump sum fee.

A social planner is offered with two basic options in order to modify the manufacturer's incentives: impose a tax T, or offer a reward R. Suppose $\beta_a > \beta_b$; then Pharma prefers to have multiple channels, which would waste the benefits from the pulling effect. The first option is to tax the manufacturer a lump sum amount T_m^* for every additional channel used (e.g., a fixed fee for every new brand registry). The alternative option is to offer the manufacturer a reward R_s^* for every additional application of a single existing good (*e.g.,* rewarding marginal innovations in health care, or offering a discount on the registration process for additional therapeutic applications of a drug).

Proposition 19:

Assume $\beta_a > \beta_b$. Define $T_m^* = M^{MX}(e^{MX}; Q^{MX}) - M^{SX}(e^{SX}; Q^{SX}) = R_s^* > 0$.

a) Both taxing T_m^*, or rewarding R_s^*, coordinate the supply chain design decision.

b) By taxing T_m^*, then $(e^{MX}|T = T_m^*) = (e^{SX}|T = 0) < (e^{MX}|T = 0)$.

c) By rewarding R_s^*, then $(e^{SX}|R = R_s^*) = (e^{MX}|R = 0) > (e^{SX}|R = 0)$.

The main takeaway from Proposition 19 is that while there is more than one simple, implementable alternative to coordinate the supply chain design, each alternative can be used to affect the level of efforts exerted. Notice that the conditions in Proposition 19b coordinate the chain, but induce the inefficient effort level; the tax only allows Health to take advantage of the pulling effect in case the efforts are successful, but the probability of those efforts being successful is lower than the case where no tax exists and Pharma chooses a multiple channels design. In Proposition 19c, both the supply chain decision and the effort level are chosen efficiently; this occurs because by offering a fixed reward for marginal innovations of the same product, then the expected benefit for Pharma derived from a successful innovation effort is the same regardless of the design choice.

Additionally, the contract could include a revenue sharing scheme or any other form of inventory coordinating mechanism in order to avoid the presence of double-marginalization. Such problem is vastly studied in the literature (see Cachon, 2003) and their inclusion here would add no theoretical value. For complementarity, Proposition 20 looks at the situation where $\beta_a < \beta_b$. In such case, Pharma prefers a single channel structure, while the social planner prefers to have multiple channels to avoid the pulling effect. Under such a case, the social planner could establish a tax, T_s^*, for multiple applications of the same product, or offer a reward, R_m^*, for the introduction of new, dedicated or differentiated products.

Proposition 20:

Assume $\beta_a < \beta_b$. Define $T_s^* = M^{SX}(e^{SX}; Q^{SX}) - M^{MX}(e^{MX}; Q^{MX}) = R_m^* > 0$.

a) Taxing T_s^* and rewarding R_m^* both coordinate the supply chain design decision.

b) By taxing T_s^*, then $(e^{SX}|T = T_s^*) = (e^{MX}|T = 0) < (e^{SX}|T = 0)$.

c) By rewarding R_m^*, then $(e^{MX}|R = R_m^*) = (e^{SX}|R = 0) > (e^{MX}|R = 0)$.

4.6.2 Design Dependent Cost Functions

An assumption of the model is that the cost of innovation is independent of the supply chain design, which we relax to some extent here. On one hand, economies of scale could suggest lower costs of innovation for single channel designs, in addition to the set-up cost for introducing new products in a multiple channels case. On the other hand, innovation effort in a single channel design may be considered more expensive due to the lower probabilities of finding additional applications for a single product, and having multiple dedicated channels may reduce the uncertainty in the effort's efficacy or can increase the possibility of finding subsidies. Proposition 21 presents the scenarios for which the results still apply, even under different cost functions.

Proposition 21: Define $C_M(\cdot)$ as the cost of exerting innovation effort in a multiple channels design. Define $C_S(\cdot)$ as the cost of exerting innovation effort in a single channel design.

a) If $C_M(e) > C_S(e)$, and $\beta_a > \beta_b$, the efficient design remains a single channel design.

b) If $C_M(e) < C_S(e)$, and $\beta_a < \beta_b$, the efficient design remains a multiple channels design.

c) If $C_M(e) = C_S(e) + K$, where K is a constant, then the optimal level of effort is the same under both designs, but the efficient supply chain design may change.

4.6.3 Binary success probabilities

In the model we have defined the probability of success to be non-decreasing and concave in the level of effort. It is possible that the innovation effort is of the "win-all lose-all" type, *i.e,* a threshold level of effort needs to be incurred in order for the second patient category to realize, so that the probability of success is 0 for efforts below the threshold, and x for any effort level equal to or above the threshold, where $x \in (0, 1)$. Based on the continued interest on this situation, it was considered important to point out that the formulation used allows for such circumstances, and that in those cases the optimal effort will be either equal to zero (if at the threshold the marginal costs exceed marginal benefits) or equal to the threshold (otherwise).

4.7 When Pulling meets Pooling

Throughout the analysis it has been assumed that the additional demand created by the innovation efforts is deterministic. This was done in order to cleanly observe the impact of the inventory pulling effect on the optimal order quantity and the innovation effort. In order to observe the combined effects of pooling and pulling, for $j = a, b$, let ε_j be a normal distributed random variable with mean μ_j and variance σ_j^2, where $\Psi_j(\cdot)$ and $\psi_j(\cdot)$ are used to denote the distribution and probability density functions, respectively. Assume ε_a and ε_b are independent and uncorrelated, and that the coefficient of variation is sufficiently small so that $\Psi_j(0) = 0,\ j = a, b$. Recall that the demand for type b patients is contingent on e and is only realized when $x = 1$. We also define the random variable $\Theta = \varepsilon_1 + x\varepsilon_2$, which is conditional on the outcome of the innovation effort and has a conditional normal distribution function $\Phi(\cdot \mid 1)$ and density $\phi(\cdot \mid 1)$ when $x = 1$, and conditional normal distribution function $\Phi(\cdot \mid 0) \triangleq \Psi_a(\cdot)$, and density $\phi(\cdot \mid 0) \triangleq \psi_a(\cdot)$, when $x = 0$. The goal is then to observe the impact on the total drug order quantity and the incentives for innovation

effort. We do so by following the procedure shown in section 4.3 (*i.e.,* we assume vertical integration).

4.7.1 Multiple Channels with stochastic demands

We use superindex (MS), referring to Multiple channels with Stochastic size demands, to denote the optimal solutions in this subsection. Notice first that by providing different markets with different goods (*i.e.,* the equivalent setting to §4.3.2) both the pulling and the pooling effects are suppressed and we obtain a solution slightly different to that presented earlier, where the only adjustment is that a critical fractile must be calculated for the new category of patients. For $j = a, b$, let Q_j be the drug order quantity to satisfy demand from patients $\in S_j$.

The problem for satisfying the demand from patients belonging to S_a is:

$$Z^{MS}_{S_a}(Q_a) \;=\; -cQ_a + \beta_a \left(\int_0^{Q_a} \xi \psi_a(\xi)\, d\xi + \int_{Q_a}^{\infty} Q_a \psi_a(\xi)\, d\xi \right) \tag{4.7.1}$$

Taking the first order condition:

$$Q^{MS}_a \;=\; \Psi_a^{-1}\left(1 - \frac{c}{\beta_a}\right) \tag{4.7.2}$$

It is clear that when $x = 0$, then $Q^{MS}_b = 0$, Then, the problem in stage 2 is only relevant when $x = 1$, in which case the decision-maker solves:

$$Z^{MS}_{S_b,2}(Q_b) \;=\; -cQ_b + \beta_b \left(\int_0^{Q_b} \xi \psi_b(\xi)\, d\xi + \int_{Q_b}^{\infty} Q_b \psi_b(\xi)\, d\xi \right) x \tag{4.7.3}$$

Taking the first order condition:

$$Q_b^{MS} = \begin{cases} \Psi_b^{-1}\left(1-\frac{c}{\beta_b}\right) & \text{if } x=1 \\ 0 & \text{otherwise} \end{cases} \tag{4.7.4}$$

As a result, the problem in stage 1 for the decision-maker planning to serve patients belonging to category S_b is:

$$\begin{aligned} Z_{S_b,1}^{MS}(e;\, Q_b) &= -C(e) + (\beta_b - c)\mathbb{E}[Q_b^{MS}] \\ &\text{subject to:} \\ &\mathbb{E}[Q_b^{MS}] = g(e)\Phi^{-1}\left(1-\frac{c}{\beta_b}\right) \end{aligned} \tag{4.7.5}$$

The first term is the cost of exerting effort and the second term is the expected benefit minus the production cost generated from the drugs that are given to patients. Notice that we take into account the order quantity that realizes when $x = 1$ *times* the probability that $x = 1$. Taking the first-order condition we obtain:

$$C'(e^{MS}) = (\beta_b - c)\,\Psi_b^{-1}\left(1-\frac{c}{\beta_b}\right) g'(e^{MS}) \tag{4.7.6}$$

4.7.2 Single channel with stochastic demands

If a single channel is used to satisfy both categories, then both pooling and pulling will occur and the decision-maker solves the following problem which we denote with the superindex (SS), referring to Single channel with Stochastic size demands. Again, we solve by backwards

induction. In the second stage, after the realization of x, the cost of effort is sunk and the single decision-maker solves:

$$\begin{aligned} Z_2^{SS}(Q) = -cQ &+ \int_{\varepsilon_a=0}^{Q} \int_{0}^{Q-\xi} (\beta_a \xi + \beta_b x\eta)\, \psi_a(\xi)\psi_b(\eta)\, d\eta\, d\xi \\ &+ \int_{\varepsilon_a=0}^{Q} \int_{Q-\xi}^{\infty} Q\left(\beta_a\left(\frac{\xi}{\xi+x\eta}\right) + \beta_b\left(\frac{x\eta}{\xi+x\eta}\right)\right) \psi_a(\xi)\psi_b(\eta)\, d\eta\, d\xi \\ &+ \int_{\varepsilon_a=Q}^{\infty} \int_{0}^{\infty} Q\left(\beta_a\left(\frac{\xi}{\xi+x\eta}\right) + \beta_b\left(\frac{x\eta}{\xi+x\eta}\right)\right) \psi_a(\xi)\psi_b(\eta)\, d\eta\, d\xi \end{aligned} \tag{4.7.7}$$

Breaking down equation (4.7.7), the first line includes the cost of purchasing Q units, and the revenue when the total demand doesn't exceed Q. The second and third lines calculate the revenue captured when demand exceeds the available inventory. On the second line, the demand derived from type a patients is not enough to exceed the available inventory on its own, and so the second integral considers those cases where demand of type b patients is sufficiently large to generate a stockout. On the third line, the demand derived from type a patients is enough to exceed capacity, and therefore all possible realizations of patients of type b are considered. For visibility purposes, the second and third line can be merged as is shown in equation (4.7.8). Notice that taking ξ as the realized value of the type a patients, when $(Q-\xi) > 0$, the second line in (4.7.8) captures the second line in (4.7.7); and when $(Q-\xi) \leq 0$, the second line in (4.7.8) captures the third line in (4.7.7).

$$\begin{aligned} Z_2^{SS}(Q) = -cQ &+ \int_{\varepsilon_a=0}^{Q} \int_{0}^{Q-\xi} (\beta_a \xi + \beta_b x\eta)\, \psi_a(\xi)\psi_b(\eta)\, d\eta\, d\xi \\ &+ \int_{\varepsilon_a=0}^{\infty} \int_{\max[0,Q-\xi]}^{\infty} Q\left(\beta_a\left(\frac{\xi}{\xi+x\eta}\right) + \beta_b\left(\frac{x\eta}{\xi+x\eta}\right)\right) \psi_a(\xi)\psi_b(\eta)\, d\eta\, d\xi \end{aligned} \tag{4.7.8}$$

We believe that (4.7.7) and (4.7.8) are useful for illustration purposes. However, from this point forward, it becomes easier to continue the analysis by working with the joint probability

distribution, which is conditional on the value of x. We then rewrite the objective function at stage 2 as follows:

$$\begin{aligned} Z_2^{SS}(Q) = -cQ + (1-x)(\beta_a)\left(\int_0^Q \theta\phi(\theta \mid 0)\, d\xi + \int_Q^\infty Q\phi(\theta \mid 0)\, d\theta\right) \\ +x\left(\int_{\Theta=0}^{Q} \left(\beta_a \mathbb{E}[\varepsilon_a \mid \Theta=\theta,\, x=1] + \beta_b \mathbb{E}[\varepsilon_b \mid \Theta=\theta,\, x=1]\right)\phi(\theta \mid 1)\, d\theta \right. \\ \left. + \int_{\Theta=Q}^{\infty} Q\left(\frac{\beta_a \mathbb{E}[\varepsilon_a \mid \Theta=\theta,\, x=1] + \beta_b\left(\mathbb{E}[\varepsilon_b \mid \Theta=\theta,\, x=1]\right)}{\theta}\right)\phi(\theta \mid 1)\, d\theta\right) \quad (4.7.9) \end{aligned}$$

where

$$\mathbb{E}[\varepsilon_a \mid \Theta=\theta,\, x=1] = \int_{\varepsilon_a=0}^{\theta} \xi\psi_b(\theta-\xi)\psi_a(\xi)\, d\xi\,, \quad (4.7.10)$$

$$\mathbb{E}[\varepsilon_b \mid \Theta=\theta,\, x=1] = \theta - \mathbb{E}[\varepsilon_a \mid \Theta=\theta,\, x=1] = \int_{\varepsilon_b=0}^{\theta} \eta\psi_a(\theta-\eta)\psi_b(\eta)\, d\eta\,. \quad (4.7.11)$$

Notice that when $x = 0$, the second and third lines of (4.7.9) are equal to zero, and when $x = 1$, the value of the first line is simply $(-cQ)$. Additionally, it is instructive to point out that in (4.7.9), the third line (which captures the situations where total demand exceeds Q) includes a denominator θ inside the integral, unlike the equation's second line (which captures the situations where total demand is lower than Q). The reason is simple. On one hand, the realized health benefits in the demand scenarios captured in the second line is independent of the order of arrivals because there are no stockouts; on the other hand, the realized health benefits in the scenarios captured in the third line do depend on the order of arrivals, and therefore the total arrivals must be normalized in order to determine the expected allocation of the Q drugs to type a and type b patients.

Define $Q^{SS} \in \arg\max_{(Q)} \; Z_2^{SS}(Q)$. Therefore:

$$Q^{SS} \;=\; \begin{cases} Q_{a,b}^{SS} & \text{if } x = 1 \\ Q_{a,0}^{SS} & \text{otherwise} \end{cases} \tag{4.7.12}$$

Taking the first order condition of (4.7.9) conditional on $x = 0$ yields[10]:

$$Q_{a,0}^{SS} = \Phi^{-1}\left(1 - \frac{c}{\beta_a}\middle| 0\right) \tag{4.7.13}$$

Lemma 25: $Q_{a,0}^{SS} = Q_a^{MS}$.

Lemma 25 states that when $x = 0$, the optimal order quantity is independent of whether the system's design has a single channel or dedicated channels. As expected, this is consistent with the results from §4.3 given that $x = 0$ even though the second category is now stochastic.

Taking the first order condition of (4.7.9), conditional on $x = 1$ yields[11]:

$$\int_{\Theta=Q_{a,b}^{SS}}^{\infty} \left(\frac{\beta_a \mathbb{E}[\varepsilon_a \mid \Theta = \theta,\, x = 1] + \beta_b \left(\mathbb{E}[\varepsilon_b \mid \Theta = \theta,\, x = 1]\right)}{\theta} \right) \phi(\theta \mid 1)\, d\theta - c = 0 \tag{4.7.14}$$

[10]The proof for equation (4.7.13) is provided in the Appendix.
[11]The proof for equation (4.7.14) is provided in the Appendix.

Replacing (4.7.10) and (4.7.11) into (4.7.14), we simplify the expression as follows:

$$\int_{\Theta=Q_{a,b}^{SS}}^{\infty} \left(\frac{(\beta_b - \beta_a)\left(\int_{\varepsilon_b=0}^{\theta} \xi\psi_a(\theta-\xi)\psi_b(\xi)\,d\xi\right)}{\theta} + \beta_a \right) \phi(\theta \mid 1)\,d\theta = c$$

$$\beta_a\left(1-\Phi\left(Q_{a,b}^{SS} \mid 1\right)\right) + \int_{\Theta=Q_{a,b}^{SS}}^{\infty} \left(\frac{(\beta_b - \beta_a)\left(\int_{\varepsilon_b=0}^{\theta} \xi\psi_a(\theta-\xi)\psi_b(\xi)\,d\xi\right)}{\theta} \right) \phi(\theta \mid 1)\,d\theta = c$$

$$\beta_a\left(1-\Phi\left(Q_{a,b}^{SS} \mid 1\right)\right) + (\beta_b - \beta_a) \int_{\Theta=Q_{a,b}^{SS}}^{\infty} \left(\frac{(\mathbb{E}[\varepsilon_b \mid \Theta=\theta,\, x=1])}{\theta} \right) \phi(\theta \mid 1)\,d\theta = c \quad (4.7.15)$$

or alternatively:

$$\beta_b\left(1-\Phi\left(Q_{a,b}^{SS} \mid 1\right)\right) + (\beta_a - \beta_b) \int_{\Theta=Q_{a,b}^{SS}}^{\infty} \left(\frac{(\mathbb{E}[\varepsilon_a \mid \Theta=\theta,\, x=1])}{\theta} \right) \phi(\theta \mid 1)\,d\theta = c \quad (4.7.16)$$

A remarkable finding is the similarity in the structure between equations (4.7.15) and (4.7.16), and equation (4.3.9). Once again, as we did in §4.3, we observe that when $\beta_a = \beta_b$, the second term in (4.7.15) and (4.7.16) becomes zero and our model collapses into the familiar newsvendor result from the operations management literature. In fact, by assuming demand for group S_b to be deterministic, equation (4.7.10) also collapses into (4.3.9). Note then that the only relevant distinction with respect to §4.3 is that now both populations are stochastic, and from the pooling effect literature, we know that the variance of the distribution of total demand $\Phi(\cdot \mid 1)$ is less than the sum of the variances of the distributions $\Psi_a(\cdot)$ and $\Psi_b(\cdot)$; it is equally well known that the pooling effect pushes the optimal order quantity closer to the total mean. Our key contribution is that when $\beta_a \neq \beta_b$, there is a second effect captured by the second term in either equation (4.7.15) or (4.7.16); we have coined this behavior as the *pulling effect*, and its impact on the optimal order quantity is consistent with the results from §4.3. This is summarized in Lemma 26.

Lemma 26: Fix the value of β_a. a) When $\beta_b = \beta_a$, then $Q_{a,b}^{SS} = \Phi^{-1}\left(1 - \frac{c}{\beta_a}\right)$. b) When $\beta_a > \beta_b$, then $\Phi^{-1}\left(1 - \frac{c}{\beta_b}\right) < Q_{a,b}^{SS} < \Phi^{-1}\left(1 - \frac{c}{\beta_a}\right)$. c) When $\beta_b > \beta_a$, then $\Phi^{-1}\left(1 - \frac{c}{\beta_a}\right) < Q_{a,b}^{SS} < \Phi^{-1}\left(1 - \frac{c}{\beta_b}\right)$.

Lemma 26 provides intuitive bounds on the optimal order quantity. As an illustration, consider the case where $\beta_a > \beta_b$. (4.7.15) explains that the optimal order quantity will be lower than $\Phi^{-1}\left(1 - \frac{c}{\beta_a}\middle| 1\right)$, because the type a patients benefit from stochastically pulling from the additional inventory that is created for type b patients. Similarly, (4.7.16) explains that the optimal order quantity will be higher than $\Phi^{-1}\left(1 - \frac{c}{\beta_b}\middle| 1\right)$, because the pulling effect of type b patients taking inventory from type a patients wants to be restricted.

Lemma 26 uses therefore the same logic as Proposition 16 in order to analyze the pulling effect in isolation. Unlike §4.3, here we cannot say whether the optimal order quantity increases or decreases as we move from multiple dedicated channels to a single channel design. The reason is that the pulling and pooling effects may push the optimal order quantity in different directions. However, the truly relevant conclusion we are able to make is that when $\beta_a \neq \beta_b$, then the pulling and pooling effects are two distinct forces which act simultaneously to affect the order quantity. This also means that contrary to "conventional wisdom", having a single channel (or pooled system) may not be optimal compared to dedicated channels, even in a basic setting as the one we propose where there are no incremental costs to pooling demand.

In the first stage, the decision-maker solves:

$$
\begin{aligned}
Z_1^{SS} = \max_{(e)} \quad & -C(e) - c\left\{(1-g(e))Q_{a,0}^{SS} + g(e)(Q_{a,b}^{SS})\right\} \\
& +(1-g(e))\beta_a \left\{ \int_0^{Q_{a,0}^{SS}} \xi\psi(\xi)\, d\xi + \int_{Q_{a,0}^{SS}}^{\lambda} Q_{a,0}^{SS}\psi(\xi)\, d\xi \right\} \\
& +g(e) \left\{ \int_{\Theta=0}^{Q} \left(\beta_a \mathbb{E}[\varepsilon_a \mid \Theta=\theta] + \beta_b \mathbb{E}[\varepsilon_b \mid \Theta=\theta]\right) \phi(\theta)\, d\theta \right. \\
& \qquad \left. + \int_{\Theta=Q_{a,b}^{SS}}^{\infty} Q_{a,b}^{SS} \left(\frac{\beta_a \mathbb{E}[\varepsilon_a \mid \Theta=\theta] + \beta_b \left(\mathbb{E}[\varepsilon_b \mid \Theta=\theta]\right)}{\theta} \right) \phi(\theta)\, d\theta \ , \right\}
\end{aligned}
$$

subject to:

$$
\begin{aligned}
Q_{a,0}^{SS} &\in \arg\max_{(Q)} \ Z_2^{SS} \mid_{x=0} \\
Q_{a,b}^{SS} &\in \arg\max_{(Q)} \ Z_2^{SS} \mid_{x=1}
\end{aligned}
\tag{4.7.17}
$$

Again, we can rewrite as follows:

$$
\begin{aligned}
Z_1^{SS} = \max_{(e)} \quad & -C(e) - c\left\{(1-g(e))Q_{a,0}^{SS} + g(e)(Q_{a,b}^{SS})\right\} \\
& +(1-g(e))\beta_a \left\{ \int_0^{Q_{a,0}^{SS}} \xi\psi(\xi)\, d\xi + \int_{Q_{a,0}^{SS}}^{\lambda} Q_{a,0}^{SS}\psi(\xi)\, d\xi \right\} \\
& +g(e) \left\{ \int_{\Theta=0}^{Q_{a,b}^{SS}} \left(\beta_a \mathbb{E}[\varepsilon_a \mid \Theta=\theta,\, x=1] + \beta_b \mathbb{E}[\varepsilon_b \mid \Theta=\theta,\, x=1]\right) \phi(\theta \mid 1)\, d\theta \right. \\
& \qquad \left. + \int_{\Theta=Q_{a,b}^{SS}}^{\infty} Q_{a,b}^{SS} \left(\frac{\beta_a \mathbb{E}[\varepsilon_a \mid \Theta=\theta,\, x=1] + \beta_b \left(\mathbb{E}[\varepsilon_b \mid \Theta=\theta,\, x=1]\right)}{\theta} \right) \phi(\theta \mid 1)\, d\theta \right\}
\end{aligned}
$$

subject to:

$$
\begin{aligned}
Q_{a,0}^{SS} &\in \arg\max_{(Q)} \ Z_2^{SS} \mid_{x=0} \\
Q_{a,b}^{SS} &\in \arg\max_{(Q)} \ Z_2^{SS} \mid_{x=1}
\end{aligned}
\tag{4.7.18}
$$

The first two terms represent the costs of exerting innovation effort and producing the drug's order quantity, respectively. The third term is the health-benefits obtained by patients belonging to S_a when the innovation efforts are unsuccessful. The last terms represent

the case when the innovation efforts are successful and so both pulling and pooling occur: the first double integral considers the probability that total demand doesn't exceed total supply, and the second double integral considers the cases where total demand exceeds total supply and shows the corresponding split of the available inventory in response to the first-come first-serve rule; The last two double integrals represent the health benefits obtained by patients belonging to S_b with an interpretation of the terms inside the brackets analogous to the ones for patients belonging to S_a.

4.8 Conclusions

In this final part of the analysis, our focus has shifted away from the design of the contract in an attempt to understand the effect of patient heterogeneity on the optimal supply chain design strategy. An important contribution is not only the ability to determine the efficient supply chain design simply in terms of the relative health benefits of the patient categories (or from a more general perspective, based on the relative valuations), but also the operational implications of having a single common stock versus multiple dedicated stocks to serve heterogenous demand. Even though the perspective of a social welfare maximizer was not explicitly modeled, the results implicitly provide the necessary grounds for discussion; in other words, the operational tradeoff between efficiency and availability allows us to get additional insights from a strategic perspective as a function of the decision-maker's priority. As a result, to close this chapter and the analytical component of the dissertation, other situations to which this chapter's results can be relevant are briefly sketched.

Case 1: Turning "lemon" markets into "cherries"

Consider a pharmaceutical manufacturer that sells its drug through a distributor in country Alpha and is evaluating whether to also introduce the drug in country Beta. Doing so

would require the manufacturer to go through a costly evaluation process in order to get its drug approved for consumption in country Beta, whose willingness to pay for the drug is different from Alpha's. There is a probability that the drug doesn't get approved in Beta, but if it does, then the demand forecast at that location is expected to be extremely accurate (*i.e.,* demand variance $\simeq 0$). Based on our results, if the pharmaceutical manufacturer sells the drugs through a unique distributor to both countries who supplied them on a first-come first-serve basis, then on top of the incremental order quantity corresponding to country Beta's demand, the order quantity that the distributor will order to satisfy demand in country Alpha will increase in relation to the quantity ordered when the drug would sell only in country Alpha. As a result, the pharmaceutical company has more incentives to invest in getting its drug accepted for sale in country Beta when there exists a unique distributor because that will result in a larger quantity of drugs being sold compared to the case where there was a different distributor in each country (assuming the distributors do not trade among them, or do not consider the possibility of mutual trade when they decide the order quantity). We can even take our conclusions one echelon upstream and consider two pharmaceutical manufacturers (namely, P_x and P_y) with drugs equivalent in their cost structure and expected health benefits (*i.e.,* valuation) and that are distributed through a single health-payer (*e.g.,* a single NGO that purchases malaria drugs for all African countries); we can then say that the total quantity of drugs produced and sold will be higher under the single-manufacturer single-distributor configuration than if P_x had a monopoly in country Alpha and P_y had a monopoly in country Beta; further, the first manufacturer (incumbent) to sell in any of the differentiated countries has the largest incentive to get its drug also accepted in an additional country because of the positive externality created on its initial market(s) resulting from acceptance into a new market. This suggests important benefits for the manufacturers of dealing with suppliers who do aggregate demand forecasts, and also important benefits for consumers in less attractive markets from market consolidation both at the distributor and the manufacturer level. We also believe that our results moti-

vate and should be incorporated into the design of donor coordination mechanisms aimed at guaranteeing sufficient installed production capacity by drug manufacturers and increasing coverage in developing countries.

Case 2: Consolidation of health services providers

Consider a health insurance company that is vertically integrated so that it is also a health services provider. If this company is in charge of serving both its insured patients (who are charged for only a fraction of the received treatment cost) and incoming patients who have no insurance (who are charged for the full cost of the received treatment cost), then it will build more capacity (*i.e.,* in terms of manpower, have a larger staff of physicians and nurses; in terms of installations, have more beds, special equipment, operating rooms; in terms of inventory, have a larger stock of drugs in its pharmacy) than if two exclusive health service providers existed: one that treated the insured population, and one that treated the uninsured population. Further, the single provider will have a larger incentive to invest in marketing activities to create a base of insured members than the incentives that a service provider that treated only its own patients would have. This example raises questions on the ways in which the mix of public-private health providers should be established, particularly in developing countries where access is a big issue.

Case 3: Coupon distribution and inventory planning

Consider a product sold at retail outlets for which demand is relatively stable. The product's manufacturer wishes to reach a new customer segment and distributes a large number of discount coupons at the local colleges. The number of students who will attempt to use their coupons is not known, and the manufacturer must choose whether to limit the exchange of the coupons to a few retail outlets, or to allow the coupons to be exchanged at

any location. On one hand, if the goodwill cost of losing new potential customers if ignored, then the efficient design choice is to limit the exchange to specific (almost dedicated) outlets. On the other hand, if product availability is preferred to revenues, and the goodwill cost of understocking for a client belonging to the stable demand is ignored, then allowing the coupons to be exchanged at any retail outlet is the best option.

[page left intentionally blank]

Chapter 5

Final Comments

This dissertation has analytically studied the introduction process of a new drug sold by a pharmaceutical manufacturer to a health-payer. It has looked into the decisions of these two players considering the setting's particular characteristics, of which the most relevant have been found to be: i) the heterogeneity in the health benefits received by different patient categories who are eligible to receive the drug; ii) the heterogeneity in the health-payer's decision-making priority and constraints; iii) the uncertainty and information asymmetry with respect to the expected health benefits; and iv) the uncertainty about the size of the demand and the order of arrivals between patient categories.

The focus of the thesis has been two-fold. On the first part, Chapters 2 and 3 pay special attention to the effect of the contract design on the players' decisions and the resulting levels of profit, expenses, and drug availability for the manufacturer, health-payer, and the patients, respectively. The main findings are the requirements for achieving higher access levels depending on the health-payer's priority, the ability that the manufacturer has to extract a health-payer's available budget without necessarily increasing social welfare when maximizing the latter is the payer's main goal, and the (incidental) upper bound on the selling price that is created when the payer's priority is to maximize its own value function and the manufacturer is interested in achieving higher levels of access. These findings suggest

that the willingness to subsidize patient categories with low efficacy levels through the high efficacy levels of other patients consuming the same drug may result in high per-treatment expenditures for the payer, low service levels, and higher profitability for the manufacturers. Two more contracts are analyzed where the selling price is not a decision variable. The capacity buffer contract is appropriate to represent the circumstances where the manufacturer is willing to hold an important portion of the inventory risk. We find that this contract tends to increase access level and is most beneficial when the manufacturer has a relatively low overstocking cost - as opposed to the health-payer -, the cost of understocking for the health-payer is large, and the available budget is sufficient to cover the increased costs of smaller, more frequent inventory orders. The last contract analyzed is a performance based agreement between the players which has highest value when the health payer has little trust in the manufacturer's announced health benefits, and the manufacturer holds private information that increases her reliability on her own product.

On the second part of the dissertation, the main objective is to determine if the existence of patient heterogeneity can have an influence on the supply chain design strategy, and more specifically on whether a pooled versus a separated stocking strategy is more convenient. In a two-patient categories environment, where one category is of stochastic size and the other is deterministic, we find that: i) the optimal design depends exclusively on the relative health benefits obtained by each patient category; ii) the difference between the optimal levels of inventory for each design option is a function of what has been coined here as a pulling effect; iii) in a vertically integrated chain, the efficient supply chain design always has less inventory than the inefficient design and provides higher incentives for innovation; and iv) in a vertically separated chain, the upstream player (the manufacturer) prefers the inefficient design, and the efficient design provides the manufacturer with less incentives to innovate than the inefficient design. Last, an attempt to model two stochastic categories was made - in addition to exploring other simple extensions to the basic model -, where even though the

determination of the optimal solution was considered analytically intractable, the analysis revealed a structure similar to the one causing the pulling effect in the basic model, thus allowing for limited findings when the pulling and pooling effects interact.

Finally, it is important to mention other research avenues that can be extended as a result of this work, and for that purpose, it may be appropriate to divide them into empirical and analytical projects. For the first category, a logical next step is to try and obtain data to validate the models, obtain additional insights, and be able to make better grounded recommendations to public policy-makers. The new pharmaceutical price regulation scheme that will become active in 2014 in the United Kingdom represents a great opportunity to run pilot studies that can allow us to determine which factors are truly relevant in the decision-making process. Similarly, risk sharing agreements based on the capacity buffer and the performance based contracts could be developed in real life. On an idea unrelated to the health care sector, and leveraging on the preliminary findings for the interaction of pulling and pooling effects, the availability of data related to the distribution of discount coupons and the associated capacity utilization/inventory availability could open new areas of research.

As for modeling, the options available are large. A first approach would be the expansion of the analysis presented here by combining different contracting scenarios. For instance, allowing the selling price to be endogenously determined in the capacity buffer and performance-based contract could offer interesting results. A key assumption in this work is the distribution of demand which was assumed to be Poisson distributed for most of the contracting scenarios due to the properties of a split Poisson process that allowed for better tractability of the demand distribution for different access levels. Expanding the results to more general distributions would be a valuable contribution. Another assumption that can be relaxed in order to obtain more general results is that of having only two patient

categories. Preliminary work around this issue suggests that a heuristic based on bundled-pairwise comparison and stopping rules can be developed. While in the setting described in the dissertation, the problem is not critical as the number of patient categories per drug treatment is usually sufficiently small, the application of the model in other contexts (such as the price and quantity newsvendor) might benefit from such developments. Finally, more general extensions include the presence of competition at different echelons of the supply chain and the incorporation of the health-provider and his role in the supply chain.

References

Alfaro, J. and C. Corbett. 2003. The value of SKU rationalization in practice (The pooling effect under suboptimal inventory policies and nonnormal demand). *Production and Operations Management.* 12(1): 12-29.

Alptekinoglu, A., A. Banerjee, A. Paul, and N. Jain. 2013. Inventory pooling to deliver differentiated service. *Manufacturing and Services Operations Management.* 15(1): 33-44.

Arrow, K. 1963. Uncertainty and the welfare economics of medical care. *The American Economic Review* 53(5): 941-973.

Banciu, M. and P. Mirchandani. 2013. Technical Note - New results concerning probability distributions with increasing generalized failure rates. *Operations Research.* 61(4): 925-931.

Barros, P. P. 2011. The simple economics of risk-sharing agreements between the NHS and the pharmaceutical industry. *Health Economics* 20: 461-470.

Bell, D. E. 2003. Incorporating the customer's perspective into the newsvendor problem. *Working Paper, Harvard Business School.*

Bernstein, F., A. Federgruen. 2005. Decentralized Supply Chains with Competing Re-

tailers Under Demand Uncertainty. 2005. 51(1): 18-29.

Cachon, G. 2003. Supply chain coordination with contracts. Graves, S., T. de Kok (Eds.). *Handbooks in Operations Research and Management Science, Supply Chain Manan-agement.* North-Holland.

Cachon, G. and M. Lariviere. 2005. Supply Chain Coordination with Revenue-Sharing Contracts: Strengths and Limitations. *Management Science* 51(1): 30-44.

Carapinha, J.L. 2008. Setting the Stage for Risk-Sharing Agreements: International Experiences and Outcomes-based Reimbursement. *South African Family Practice.* 50(4)

Chalkidou K., R. Lopert, and A. Gerber. Paying for "End-of-Life" Drugs in Australia, Germany, and the United Kingdom: Balancing Policy, Pragmatism, and Societal Values. The Commonwealth Fund, January 2012.

Deshpande, V., M. Cohen, and K. Donohue. 2003. A Threshold Inventory Rationing Policy for Service-Differentiated Demand Classes. *Management Science.* 49(6): 683-703.

Emmons, H., S. Gilbert. 1998. Returns policies in pricing and inventory decisions for catalogue goods. *Management Science.* 44(2): 276-83.

Eppen, G. D. 1979. Effects of centralization on expected costs in a multi-location news-boy problem. *Management Science.* 25(5), 498-501.

Espin, J., J, Rovira, and L. Garcia. Experiences and Impact of European Risk-Sharing Schemes Focusing on Oncology Medicines. Andalusian School of Public Health, January

2011.

Geljins, A., J. Zivin, R. Nelson. 2001. Uncertainty and technological change in medicine. *Journal of Health Politics, Policy and Law* 26(5): 913-924.

Glass, A.J. 2001. Price discrimination and quality improvement. *Canadian Journal of Economics*. 34(2): 549-569.

Guajardo, M. A. Cohen, S. Kim, and S. Netessine. 2012. Impact of Performance-Based Contracting on Product Reliability: An Empirical Analysis. *Management Science*. 1110.1465; published online before print February 10, 2012.

Hadley, G. and T.M. Whitin. Analysis of Inventory Systems. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963.

Hawkins, N. and D. A. Scott. 2011. Reimbursement and value-based pricing: stratified cost-effectiveness analysis may not be the last word. *Health Economics* 20(6): 688-98.

Jelovac, I. and I. Macho-Stadler. 2002. Comparing Organizational Structures in Health Services. *Journal of Economic Behavior and Organization* 49 (4): 69-90.

Karlin, S., C. R. Carr. 1963. Prices and optimal inventory policy. K.J. Arrow, S. Karlin, and H. Scarf (Eds.) *Studies in Applied Probability and Management Science*. Stanford University Press, Stanford, CA, 159-172.

Kraiselburd, S., V.G. Narayanan, A. Raman. 2004. Contracting for inventory in a distribution channel with stochastic demand and substitute products. *Production and Operations*

Management. 13(1): 46-62.

Kocabiyikoglu, A. and I. Popescu. 2011. An elasticity approach to the newsvendor with price-sensitive demand. *Operations Research.* 59(2): 301-312.

Lariviere, M., Porteus, E. 2001. Selling to the Newsvendor: an analysis of price-only contracts. *Manufacturing and Service Operations Management* 3(4): 293-305.

Lippman, S. and K. McCardle. 1997. The comptetitive newsboy. *Operations Research.* 45(1): 54-65.

Marvel, H., J. Peck. 1995. Demand uncertainty and return policies. *International Economic Review.* 36(3): 691-714.

Mills, E. S. 1959. Uncertainty and price theory. *Quartely Journal of Economics.* 73: 116-130.

Nevins, A.J. 1966. Some effects of uncertainty: simulation of a model of price. *Quarterly Journal of Economics.* 80: 73-87.

OFT - Office of Fair Trading. 2007. The pharmaceutical price regulating scheme: an OFT market study. Office of Fair Trading, London.

Pasternack, B. 1985. Optimal pricing and returns policies for perishable commodities. *Marketing Science* 4:166-176.

Pasternack, B. 1999. Using revenue sharing to achieve channel coordination for a news-

boytype inventory model. *CSU Fullerton working paper.*

Parlar, M. 1998. Game theoretic analysis of the substitutable product inventory problem with random demands. *Naval Research Logistics.* 35: 397-409.

Pearson S. D. and M. D. Rawlins. 2005. Quality, Innovation, and Value for Money: NICE and the British National Health Service. *Journal of the American Medical Association* 294(20): 2618-22.

Petruzzi, N., M. Dada. 1999. Pricing and the newsvendor problem: a review with extensions. *Operations Research.* 47(2): 183-194.

Pouvourville, G. 2006. Risk-sharing agreements for innovative drugs: a new solution to old problems? *European Journal of Health Economics* 7: 155-7.

Pugatch, M., P. Healy, and R. Chu. Sharing the burden: could risk-sharing change the way we pay for healthcare? The Stockholm Network, 2010.

Roche. *Roche Personalized Healthcare. Small differences, big effects.* F. Hoffman-La Roche Ltd, Group Communications. 4070 Basel, Switzerland, October 2011.

Salinger, M. and M. Ampudia. 2011. Simple economics of the price-setting newsvendor problem. *Management Science.* 57(11): 1996-1998.

Scherer, F.M. The Pharmaceutical Industry - Prices and Progress. *New England Journal of Medicine* 351: 927-932.

So,K., C. Tang. 2000. Modeling the impact of an outcome-oriented reimbursement policy on clinic, patients, and pharmaceutical firms. *Management Science* 46(7): 875-892.

Swinney, R. 2012. Inventory pooling with strategic consumers: operational and behavioral benefits. Graduate School of Business, Stanford University. *Working Paper.*

Taylor, T. 2002. Supply Chain Coordination Under Channel Rebates with Sales Effort Effects. *Management Science.* 48(8): 992-1007.

Tsay, A. 1999. Quantity-flexibility contract and supplier-customer incentives. *Management Science* 45:1339-1358.

Young, L. 1978. Price, inventory and the structure of uncertain demand. *New Zealand Operations Research.* 6: 157-177.

Zabel, E. 1970. Monopoly and uncertainty. *Review of Economic Studies.* 37: 205-219.

Zaric, G. S. 2008. Optimal drug pricing, limited use conditions and stratified net benefits for markov models of disease progression. *Health Economics* 17: 1277-94.

Zaric, G. S. and B.J. O'Brien. 2005. Analysis of a pharmaceutical risk sharing agreement based on the purchaser's total budget. *Health Economics* 14: 793-803.

Zaric, G. S., B. Xie. 2009. The impact of two pharmaceutical risk-sharing agreements on pricing, promotion, and net health benefits. *Value in Health.* 12(5): 838-845.

Zhang, H., G.S. Zaric, and T. Huang. 2011. Optimal design of a pharmaceutical price-

volume agreement under asymmetric information about expected market size. *Production and Operations Management* 20(3): 334-346.

[page left intentionally blank]

Appendix 1: Proofs for Chapter 2

Proof of Lemma 1:

Recall $S^{\varsigma}(Q,\tau) = (B(\tau) - \delta + g)A(Q,\tau) + \delta Q - g\lambda F(\tau)$. The first term is nondecreasing $\forall\ Q$, and increasing for at least some $Q > 0$; the second term is strictly increasing in Q; and the third term is independent of Q. Therefore the function is increasing in Q.
About concavity, note that $A(Q,\tau) = \sum_{x=1}^{Q} P(x;\lambda F(\tau))$. Therefore,

$$A(Q+1,\tau) - A(Q,\tau) = P(Q+1;\lambda F(\tau)) < P(Q;\lambda F(\tau)) = A(Q,\tau) - A(Q-1,\tau),\ \ \forall\ Q \geq 1,$$

which implies that the first term in the social welfare function has diminishing returns from increasing Q. Since the second and third terms have no second-order effects, concavity is guaranteed. □

Proof of Lemma 2:

Follows directly from Lemma 1 since the extra term is linear in Q. □

Proof of Proposition 1:

First, note the following limits:

$$\lim_{Q\to 0} S(Q,\tau) = -g\lambda F(\tau)$$
$$\lim_{Q\to\infty} S(Q,\tau) = [B(\tau)-\delta][\lambda F(\tau)] + \delta Q$$

Therefore:

$$\lim_{Q\to 0} S(Q,1) - S(Q,2) = g\lambda(1-\theta) \geq 0$$
$$\lim_{Q\to\infty} S(Q,1) - S(Q,2) = [B(1)-\delta][\lambda\theta] - [B(2)-\delta][\lambda] \gtrless 0$$

As a result, for $S(Q,2) > S(Q,1)$, the curves must cross for at least one value of $Q > 0$. Suppose $S(Q,1) = S(Q,2)$, for some $Q > 0$. We then obtain the following equality which results in equation (2.3.4):

$$[B(1)+g-\delta][A(Q,1)] + \delta Q - g\lambda F(1) = [B(2)+g-\delta][A(Q,2)] + \delta Q - g\lambda F(2)$$
$$[B(1)+g-\delta][A(Q,1)] - g\lambda\theta = [B(2)+g-\delta][A(Q,2)] - g\lambda$$
$$g\lambda[1-\theta] - g[A(Q,2)-A(Q,1)] =$$
$$[b_1\theta + b_2(1-\theta)][A(Q,2)] - b_1A(Q,1) - \delta[A(Q,2)-A(Q,1)]$$
$$g\lambda[1-\theta] - g[A(Q,2)-A(Q,1)] =$$
$$b_1[A(Q,2)-A(Q,1)] - [b_1-b_2][1-\theta][A(Q,2)] - \delta[A(Q,2)-A(Q,1)]$$
$$[1-\theta][g\lambda + (b_1-b_2)A(Q,2)] = [b_1-\delta+g][A(Q,2)-A(Q,1)]$$
$$\left(\frac{g}{b_1-\delta+g}\right)\left(\frac{\lambda}{A(Q,2)}\right) + \left(\frac{b_1-b_2}{b_1-\delta+g}\right) = \left(\frac{A(Q,2)-A(Q,1)}{A(Q,2)}\right)\left(\frac{1}{1-\theta}\right)$$

To generalize the latter, suppose there is no integer quantity that satisfies the latter equation. Then:

$$q = max\left\{Q \,\middle|\, \left(\frac{g}{b_1-\delta+g}\right)\left(\frac{\lambda}{A(Q,2)}\right) + \left(\frac{b_1-b_2}{b_1-\delta+g}\right) \geq \left(\frac{A(Q,2)-A(Q,1)}{A(Q,2)}\right)\left(\frac{1}{1-\theta}\right)\right\}$$

Next, since $A(Q,2)$ is increasing in Q, the left-hand side of the inequality inside (2.3.4) is decreasing in Q, $\forall Q > 0$. Consider the following limits:

$$\lim_{Q\to 0}\left(\frac{A(Q,2)-A(Q,1)}{A(Q,2)}\right)\left(\frac{1}{1-\theta}\right)=0,$$

$$\lim_{Q\to \infty}\left(\frac{A(Q,2)-A(Q,1)}{A(Q,2)}\right)\left(\frac{1}{1-\theta}\right)=1.$$

These limits imply that if there exists at least one crossing point, then the number of crossing points (or points where the dominance of social welfare curves between access levels changes) must be odd; moving forward, changes in dominance are referred to as crossing points. Next, notice $\frac{A(Q,2)-A(Q,1)}{A(Q,2)} = 1 - \frac{A(Q,1)}{A(Q,2)}$. To show that the crossing point is unique, it suffices to show that the right-hand side in (2.3.4) is monotonic in Q, i.e., it must be that: $\frac{A(Q,1)}{A(Q,2)}$, is non-increasing in Q. First we show that $A(Q+1,1) - A(Q,1) \leq A(Q+1,2) - A(Q,2)$:

$$\begin{aligned}
A(Q+1,1) - A(Q,1) &= \sum_{x=1}^{Q+1} P(x;\lambda\theta) - \sum_{x=1}^{Q} P(x;\lambda\theta) \\
&= P(Q+1;\lambda\theta) \\
&\leq P(Q+1;\lambda) \\
&= \sum_{x=1}^{Q+1} P(x;\lambda) - \sum_{x=1}^{Q} P(x;\lambda) \\
&= A(Q+1,2) - A(Q,2)
\end{aligned}$$

Let $x = P(Q+1;\lambda\theta)$ and $y = P(Q+1;\lambda) - x$. We need that $\frac{A(Q,1)}{A(Q,2)} \geq \frac{A(Q,1)+x}{A(Q,2)+x+y} = \frac{A(Q+1,1)}{A(Q+1,2)}$. Doing some algebra:

$$\begin{aligned}
A(Q,1)A(Q,2) + xA(Q,1) + yA(Q,1) &\geq A(Q,1)A(Q,2) + xA(Q,2) \\
y &\geq x\left(\frac{A(Q,2)-A(Q,1)}{A(Q,1)}\right)
\end{aligned}$$

Replacing x and y, we obtain $P(Q+1;\lambda) - P(Q+1;\lambda\theta) \geq \left(\frac{A(Q,2)-A(Q,1)}{A(Q,1)}\right) P(Q+$

$1;\lambda\theta)$. Doing some more algebra, and replacing $A(Q,\tau) = \lambda F(\tau) - \lambda F(\tau)P(Q;\lambda F(\tau)) + QP(Q+1;\lambda F(\tau))$ (from: Hadley and Whitin (1963), Appendix 3, equation 6), we obtain the following inequality expressed in equation (2.3.3): $\theta \geq \left(\frac{1-P(Q;\lambda)}{1-P(Q;\lambda\theta)}\right)\left(\frac{P(Q+1;\lambda\theta)}{P(Q+1;\lambda)}\right)$. When (2.3.3) is satisfied, it follows that $S(\cdot,1)$ and $S(\cdot,2)$, $Q > 0$, can cross at most once. The necessary and sufficient condition for the crossing to occur is given by the following limit: $\lim_{Q\to\infty} \quad S(Q,1) - S(Q,2) < 0 \Rightarrow [B(1)-\delta][\lambda\theta] - [B(2)-\delta][\lambda] < 0 \Rightarrow \delta < b_2$. It follows that for part c), if $\delta > b_2$, $S(Q,1) > S(Q,2)$, $\forall\ Q > 0$ and setting a tighter prescription policy threshold is a dominant strategy for any positive order quantity. Part d) is simply the boundary condition.

Finally, for part e), notice the following derivatives. For the goodwill cost,

$$\frac{\partial}{\partial g}\left(\left(\frac{g}{b_1-\delta+g}\right)\left(\frac{\lambda}{A(Q,2)}\right) + \left(\frac{b_1-b_2}{b_1-\delta+g}\right)\right) = \left(\frac{b_1-\delta}{(b_1-\delta+g)^2}\right)\left(\frac{\lambda}{A(Q,2)}\right) - \left(\frac{b_1-b_2}{(b_1-\delta+g)^2}\right) > 0,$$

because $b_2 > \delta$ when q exists, and $\left(\frac{\lambda}{A(Q,2)}\right) > 1$, therefore the first term is larger than the second term. Since the left-hand side of (2.3.4) increases in g, then the order quantity must be increased to increase the right-hand side and keep the equality. Following similar arguments: for the salvage value,

$$\frac{\partial}{\partial \delta}\left(\left(\frac{g}{b_1-\delta+g}\right)\left(\frac{\lambda}{A(Q,2)}\right) + \left(\frac{b_1-b_2}{b_1-\delta+g}\right)\right) = \left(\frac{g}{(b_1-\delta+g)^2}\right)\left(\frac{\lambda}{A(Q,2)}\right) + \left(\frac{b_1-b_2}{(b_1-\delta+g)^2}\right) > 0.$$

For the relative benefit of the lower type category,

$$\frac{\partial}{\partial b_2}\left(\left(\frac{g}{b_1-\delta+g}\right)\left(\frac{\lambda}{A(Q,2)}\right) + \left(\frac{b_1-b_2}{b_1-\delta+g}\right)\right) = -\frac{1}{b_1-\delta+g} < 0.$$

For the relative benefit of the higher type category, we rearrange the initial inequality so that

$$\left(\frac{g\lambda}{A(Q,2)}\right)+(b_1-b_2)=\left(\frac{A(Q,2)-A(Q,1)}{A(Q,2)}\right)\left(\frac{1}{1-\theta}\right)(b_1-\delta+g)$$

Taking derivatives:

$$\frac{\partial}{\partial b_1}\left(\frac{g\lambda}{A(Q,2)}+(b_1-b_2)\right)=1.$$
$$\frac{\partial}{\partial b_1}\left(\frac{A(Q,2)-A(Q,1)}{A(Q,2)}\right)\left(\frac{b_1-\delta+g}{1-\theta}\right)=\left(\frac{A(Q,2)-A(Q,1)}{A(Q,2)}\right)\left(\frac{1}{1-\theta}\right).$$

Recall that $\left(\frac{A(Q,2)-A(Q,1)}{A(Q,2)}\right)\left(\frac{1}{1-\theta}\right)$ is less or equal to 1, and increasing in Q. Therefore, to balance the increase of 1 unit in the left-hand side of the equation, Q must be increased by more than 1, i.e., q increases in b_1.

For the proportion of the patient population that is of type $i=1$, using equation 6 from Appendix 3 in Hadley and Whitin (1963):

$$\begin{aligned}\frac{\partial}{\partial\theta}\left(\frac{A(Q,2)-A(Q,1)}{A(Q,2)}\right)\left(\frac{1}{1-\theta}\right)&=\left(\frac{1}{1-\theta}\right)\left[\left(1-\frac{A(Q,1)}{A(Q,2)}\right)\left(\frac{1}{1-\theta}\right)-\frac{\partial}{\partial\theta}\left(\frac{A(Q,1)}{A(Q,2)}\right)\right]\\&=\left[\frac{1}{(1-\theta)A(Q,2)}\right]\left[\left(\frac{A(Q,2)-A(Q,1)}{1-\theta}\right)-\frac{\partial A(Q,1)}{\partial\theta}\right]\\&=\left[\frac{1}{(1-\theta)^2A(Q,2)}\right][-\lambda\left(P(Q;\lambda)-\theta P(Q;\lambda\theta)\right)+Q\left(P(Q+1;\lambda)-P(Q+1;\lambda\theta)\right)\\&\quad+\lambda(1-\theta)\left(P(Q;\lambda\theta)-p(Q+1;\lambda\theta)\right)]\gtrless 0.\quad\square\end{aligned}$$

Proof of Lemma 3:

By definition, if $Q_\Gamma^\varsigma\geq\underline{Q}^\varsigma$, then $\underline{Q}^\varsigma$ satisfies both constraints and therefore at least one

feasible solution exists. To complete the proof, if $Q_\Gamma^\varsigma < \underline{Q}^\varsigma$, then $Z(Q,\tau) < 0$ for $Q \leq Q_\Gamma^\varsigma$, *i.e.,* all order quantities (if any) that satisfy the cost-effectiveness constraint do not satisfy the budget constraint. □

Proof of Lemma 4:

We can rearrange the system's expected utility function as: $Z(Q,\tau) = (B(\tau)-\delta)A(Q,\tau) - g(\lambda F(\tau) - A(Q,\tau)) - (c-\delta)Q$. Since $\lambda F(\tau) \geq A(Q,\tau)$, a necessary condition for $Z(Q,\tau) \geq 0$ is $(B(\tau)-\delta)A(Q,\tau) \geq (c-\delta)Q$, and since $Q \geq A(Q,\tau)$, then $B(\tau) \geq c$ is a necessary condition. □

Proof of Lemma 5:

First, recall that by definition: $\underline{Q}_\tau^\varsigma \leq \bar{Q}_\tau^\varsigma,\ \tau = 1,2$.

For part a1.1, suppose $\bar{Q}_1^\varsigma = q$, and $S(q,1) = S(q,2)$. Then $Z(\bar{Q}_1^\varsigma, 1) = Z(q,1) = Z(q,2) \geq 0 \rightarrow \underline{Q}_2^\varsigma \leq q$. But if $\underline{Q}_2^\varsigma < q$, then $S(q,1) > S(q,2)$, which is a contradiction. Therefore, $\bar{Q}_1^\varsigma = q = \underline{Q}_2^\varsigma$. By definition, $\underline{Q}_1^\varsigma \leq \bar{Q}_1^\varsigma$ and $\underline{Q}_2^\varsigma \leq \bar{Q}_2^\varsigma$, which completes the first ordering possibility.

For part a1.2, suppose $\bar{Q}_1^\varsigma = q$, but $S(q,1) > S(q,2)$. Then $Z(\bar{Q}_1^\varsigma, 1) = Z(q,1) > Z(q,2) \geq 0 \rightarrow \underline{Q}_2^\varsigma \leq q$. Since the case of $\underline{Q}_2^\varsigma = q$ is covered in a1.1, then the other possibility is $\underline{Q}_2^\varsigma < \bar{Q}_1^\varsigma = q$. Next suppose that, $\underline{Q}_1^\varsigma > \underline{Q}_2^\varsigma$; this is a contradiction because $S(0,1) \geq S(0,2)$, and $\underline{Q}_2^\varsigma < \underline{Q}_1^\varsigma$ only if $q \leq \underline{Q}_2^\varsigma$, which is a contradiction. Finally, $\bar{Q}_2^\varsigma < \bar{Q}_1^\varsigma$ would be a contradiction as it would require a second crossing point. Therefore the other possibility is the ordering: $\underline{Q}_1^\varsigma \leq \underline{Q}_2^\varsigma < \bar{Q}_1^\varsigma = q \leq \bar{Q}_2^\varsigma$.

For part a2, if $\bar{Q}_1^\varsigma > q$, then $Z(\bar{Q}_1^\varsigma, 1) < Z(\bar{Q}_1^\varsigma, 2) \Rightarrow \bar{Q}_1^\varsigma \leq \bar{Q}_2^\varsigma$. For a2.1, if $\underline{Q}_1^\varsigma < q$, then $Z(\underline{Q}_1^\varsigma, 1) > Z(\underline{Q}_1^\varsigma, 2)$, and due to (A1), $\underline{Q}_1^\varsigma \leq \underline{Q}_2^\varsigma$. $\underline{Q}_2^\varsigma = q$ is clearly feasible, and it would automatically imply $\underline{Q}_1^\varsigma = \underline{Q}_2^\varsigma = q$. For a2.2, if $q = \underline{Q}_2^\varsigma$, then it must also be that $\underline{Q}_1^\varsigma = q = \underline{Q}_2^\varsigma$, which is already included in a2.1. Therefore, suppose $q < \underline{Q}_2^\varsigma$; since $\bar{Q}_1^\varsigma > q$, the only possibility is that $Z(\underline{Q}_1^\varsigma, 1) < Z(\underline{Q}_1^\varsigma, 2) \Rightarrow \underline{Q}_1^\varsigma \geq \underline{Q}_2^\varsigma$.

For part a3, if $\bar{Q}_1^\varsigma < q$, then $0 > Z(q, 1) \geq Z(q, 2) < 0 \Rightarrow \underline{Q}_1^\varsigma \neq q \neq \bar{Q}_2^\varsigma$. For part a3.1, if $\underline{Q}_2^\varsigma > q$, then it has been established that $\underline{Q}_\tau^\varsigma \leq \bar{Q}_\tau^\varsigma,\ \tau = 1, 2$. For part a3.2, if $\bar{Q}_2^\varsigma \leq \bar{Q}_1^\varsigma < q$, then $Z(Q, 1) > Z(Q, 2)\ \forall\ Q \leq \bar{Q}_2^\varsigma \Rightarrow \underline{Q}_1^\varsigma \leq \underline{Q}_2^\varsigma \leq \bar{Q}_2^\varsigma$. Finally, for a3.3 there is the possibility that $\underline{Q}_2^\varsigma = \bar{Q}_2^\varsigma < q$. If $\bar{Q}_1^\varsigma < \underline{Q}_2^\varsigma$, then $q > \underline{Q}_2^\varsigma$ is a contradiction.

For part b, when a finite q does not exist, $Z(Q, 1) > Z(Q, 2)\ \forall Q > 0$. Since $\bar{Q}_1^\varsigma < \bar{Q}_2^\varsigma$ would require a crossing point, it must be that $\underline{Q}_1^\varsigma \leq \underline{Q}_2^\varsigma \leq \bar{Q}_2^\varsigma \leq \bar{Q}_1^\varsigma$. □

Proof of Theorem 1:

Theorem 1-a assumes that q exists. We will prove Theorem 1-a2, -a4, and -a5, and then Theorem 1-a1, and -a3 follow by complementarity.

For Theorem 1-a2, first suppose $Q_\Gamma^\varsigma \geq \bar{Q}_2^\varsigma$, and recall that $\bar{Q}_2^\varsigma \geq \underline{Q}_2^\varsigma$.
On one hand, consider $q \leq \underline{Q}_2^\varsigma$; then $\bar{Q}_2^\varsigma < q$ is a contradiction, and when $Q_\Gamma^\varsigma \geq \bar{Q}_2^\varsigma \geq q$, $\exists\ \{Q \,|\, Z(Q, 2) \geq 0;\ Z(Q, 2) \geq Z(Q, 1);\ Q_\Gamma^\varsigma \geq Q \geq q\}$ under which $\tau = 2$ is weakly preferred.
On the other hand, consider $q > \underline{Q}_2^\varsigma$; then if $q \leq \bar{Q}_2^\varsigma$, then $\exists\ \{Q \,|\, Z(Q, 2) \geq 0;\ Z(Q, 2) \geq Z(Q, 1);\ Q_\Gamma^\varsigma \geq Q \geq q\}$ and $\tau = 2$ would be preferred, which proves that $\bar{Q}_2^\varsigma < q$ is a necessary condition; further, when $q > \bar{Q}_2^\varsigma$, from Proposition 1, $\nexists\ \{Q \,|\, Z(Q, 2) \geq 0;\ Z(Q, 2) > Z(Q, 1);\ Q < q\}$, proving the sufficient condition. Recall from Lemma 1 that $S(Q, \tau)$ is

increasing in Q; then $\tau^*_{S,\varsigma} = 1$ is strictly preferred, and $Q^*_{S,\varsigma} = \min\{Q^\varsigma_\Gamma\,,\,\bar{Q}^\varsigma_1\}$ is the unique order quantity.

Next for Theorem 1-a3, suppose that $Q^\varsigma_\Gamma < \bar{Q}^\varsigma_2$.

On one hand, consider $q \leq \underline{Q}^\varsigma_2$; the argument is the same as above.

On the other hand, consider $q > \underline{Q}^\varsigma_2$; then if $q \leq Q^\varsigma_\Gamma$, then $\exists\ \{Q \,|\, Z(Q,2) \geq 0;\ Z(Q,2) \geq Z(Q,1); Q^\varsigma_\Gamma \geq Q \geq q\}$ and $\tau = 2$ would be preferred, which proves that $Q^\varsigma_\Gamma < q$ is a necessary condition; further, when $q > Q^\varsigma_\Gamma$, from Proposition 1, $\nexists\ \{Q \,|\, Z(Q,2) \geq 0;\ Z(Q,2) > Z(Q,1);\ Q < q\}$, proving the sufficient condition. Since $S(Q,\tau)$ is increasing in Q, then $\tau^*_{S,\varsigma} = 1$ is strictly preferred, and $Q^*_{S,\varsigma} = \min\{Q^\varsigma_\Gamma\,,\,\bar{Q}^\varsigma_1\}$ is the unique order quantity.

The proof for Theorem 1-a4 is very similar. We split the analysis into two possibilities: either $(\min\{\bar{Q}^\varsigma_2\,,\,Q^\varsigma_\Gamma\} \geq \underline{Q}^\varsigma_2 > q)$, or $(\min\{\bar{Q}^\varsigma_2\,,\,Q^\varsigma_\Gamma\} > q \geq \underline{Q}^\varsigma_2)$.

First suppose that $Q^\varsigma_\Gamma \geq \bar{Q}^\varsigma_2$.

On one hand, consider $\underline{Q}^\varsigma_2 > q$; then $\bar{Q}^\varsigma_2 \leq q$ is a contradiction. Alternatively, when $Q^\varsigma_\Gamma \geq \bar{Q}^\varsigma_2 \geq \underline{Q}^\varsigma_2 > q$, $\exists\ \{Q \,|\, Z(Q,2) \geq 0;\ Z(Q,2) > Z(Q,1);\ Q^\varsigma_\Gamma \geq Q > q\}$ under which $\tau = 2$ is preferred, satisfying the necessary condition for an optimal solution. To complete the proof, notice from Proposition 1 that $\nexists\ \{Q \,|\, Z(Q,1) \geq Z(Q,2);\ Q > q\}$. Using Lemma 1, $\tau^*_{S,\varsigma} = 2$ is strictly preferred, and $Q^*_{S,\varsigma} = \bar{Q}^\varsigma_2$ is the unique order quantity.

On the other hand, consider $q \geq \underline{Q}^\varsigma_2$; then if $q \geq \bar{Q}^\varsigma_2 \geq \underline{Q}^\varsigma_2$, then $\nexists\ \{Q \,|\, Z(Q,2) \geq 0;\ Z(Q,2) > Z(Q,1), Q^\varsigma_\Gamma \geq Q \geq q\}$ and $\tau = 1$ would be weakly preferred, which proves that $\bar{Q}^\varsigma_2 > q$ is a necessary condition; further, when $q < \bar{Q}^\varsigma_2$, from Proposition 1, $\nexists\ \{Q \,|\, Z(Q,1) \geq Z(Q,2);\ Q > q\}$, and $\exists\ \{Q \,|\, Z(Q,2) \geq 0;\ Z(Q,1) < Z(Q,2);\ Q^\varsigma_\Gamma \geq Q > q\}$, proving the sufficient condition. By use of Lemma 1, $\tau^*_{S,\varsigma} = 2$ is strictly preferred, and $Q^*_{S,\varsigma} = \bar{Q}^\varsigma_2$ is the unique order quantity.

Next suppose that $Q^{\varsigma}_{\Gamma} < \bar{Q}^{\varsigma}_{2}$.

On one hand, consider $\underline{Q}^{\varsigma}_{2} > q$; then if $Q^{\varsigma}_{\Gamma} < \underline{Q}_{2}$, $\nexists\{Q \,|\, Z(Q,2) \geq 0;\, Q \leq Q^{\varsigma}_{\Gamma}\}$; this implies that $Q^{\varsigma}_{\Gamma} \geq \underline{Q}_{2}$ is a necessary condition. To complete the proof, notice that when $\bar{Q}^{\varsigma}_{2} > Q^{\varsigma}_{\Gamma} \geq \underline{Q}^{\varsigma}_{2} > q$, then $Z(q,2) \geq 0$, and therefore $\exists\ \{Q \,|\, Z(Q,2) \geq 0;\, Z(Q,2) > Z(Q,1);\, Q^{\varsigma}_{\Gamma} \geq Q > q\}$, but $\nexists\ \{Q \,|\, Z(Q,1) \geq Z(Q,2);\, Q > q\}$. Using Lemma 1, $\tau^{*}_{S,\varsigma} = 2$ is strictly preferred, and $Q^{*}_{S,\varsigma} = Q^{\varsigma}_{\Gamma}$ is the unique order quantity.

On the other hand, consider $q \geq \underline{Q}^{\varsigma}_{2}$; then if $q \geq Q^{\varsigma}_{\Gamma}$, then from Proposition 1, $\nexists\ \{Q \,|\, Z(Q,2) > Z(Q,1); Q^{\varsigma}_{\Gamma} \geq Q\}$ which proves the necessary condition that $Q^{\varsigma}_{\Gamma} > q$; further, when $q < Q^{\varsigma}_{\Gamma}$, $\exists\ \{Q \,|\, Z(Q,2) \geq 0;\, Z(Q,2) > Z(Q,1);\, Q^{\varsigma}_{\Gamma} \geq Q > q\}$, but $\nexists\ \{Q \,|\, Z(Q,1) \geq Z(Q,2);\, Q > q\}$, proving the sufficient condition. Using Lemma 1, $\tau^{*}_{S,\varsigma} = 2$ is strictly preferred, and $Q^{*}_{S,\varsigma} = Q^{\varsigma}_{\Gamma}$ is the unique order quantity.

For Theorem 1-a5, we need to analyze the remaining cases, *i.e.,* $\min\{\bar{Q}^{\varsigma}_{2}\,,\, Q^{\varsigma}_{\Gamma}\} = q \geq \underline{Q}^{\varsigma}_{2}$. The condition is necessary because when q exists, all cases not involving $(\min\{\bar{Q}^{\varsigma}_{2}\,,\, Q^{\varsigma}_{\Gamma}\} = q \geq \underline{Q}^{\varsigma}_{2})$ have already been proven to have a unique solution. To prove sufficiency, suppose $Q^{\varsigma}_{\Gamma} \geq \bar{Q}^{\varsigma}_{2} = q \geq \underline{Q}^{\varsigma}_{2}$. Clearly $Z(\bar{Q}^{\varsigma}_{2}, 2) = Z(q,2) = Z(q,1)$ and for $x > 0$, $0 > Z(q+x,2) > Z(q+x,1)$, therefore $S(q,\tau) > S(Q,\tau),\ 0 < Q \neq q$. Alternatively, suppose $\bar{Q}^{\varsigma}_{2} \geq Q^{\varsigma}_{\Gamma} = q \geq \underline{Q}^{\varsigma}_{2}$. Then $S(\underline{Q}^{\varsigma}_{2}, 2) \leq S(q,2) \leq S(\bar{Q}^{\varsigma}_{2}) \geq 0$, and for $x > 0,\ S(q,1) > S(q-x,1)$.

Theorem 1-b follows directly from Proposition 1. □

Proof of Corollary 1:

Follows directly from previous results when only one access level policy is considered. □

Proof of equation (2.3.9):

We find the largest order quantity for which the system's expected utility function is increasing, that is:

$$\begin{aligned} 0 \leq Z(Q,\tau) - Z(Q-1,\tau) &= -(c-\delta) + (B(\tau) - \delta + g)\,(A(Q,\tau) - A(Q-1,\tau)) \\ &= -(c-\delta) + (B(\tau) - \delta + g)\,(P(Q;\lambda F(\tau))) \\ &\Rightarrow \quad P(Q_\tau^\varsigma;\lambda F(\tau)) \geq \frac{c-\delta}{B(\tau)-\delta+g} > P(Q_\tau^\varsigma + 1;\lambda F(\tau)) \end{aligned}$$

□

Proof of Proposition 2:

Recall that $Z(Q,\tau) = S(Q,\tau) - cQ$, and that $\delta < b_2$ is a sufficient condition for q to exist. We will make use of limits to show the connection between the crossing point q and $\tilde{c}$. First we show that if there is a feasible solution, then for $c > B(2),\ Z(Q_1^\varsigma, 1) > Z(Q_2^\varsigma, 2)$.

$$\begin{aligned} \lim_{c \to B(2)^+} Z(Q_2^\varsigma, 2) &= -(c - B(2))Q_2^\varsigma - (B(2) - \delta)\,(Q_2^\varsigma - A(Q_2^\varsigma, 2)) \\ &\quad -g\,(\lambda - A(Q_2^\varsigma, 2)) < 0, \\ \lim_{c \to B(2)^+ | Z(Q_1^\varsigma,1) \geq 0} (Z(Q_1^\varsigma, 1) - Z(Q_2^\varsigma, 2)) &> 0, \end{aligned}$$

The condition that $Z(Q_1^\varsigma, 1) \geq 0$ is necessary because when $Z(Q_1^\varsigma, 1) < 0$ and $c > B(2)$ simultaneously, there doesn't exist a feasible solution.

At this point we state $b_2 > \delta$ as a necessary condition for $\tilde{c}$ to exist. Recall from Proposition 1 that when $b_2 \leq \delta$, there is no crossing point between the expected social welfare curves; since the system's expected utility curves are equal to the expected social welfare curves minus a linear cost, then $b_2 \leq \delta$ would imply that restricted access dominates, i.e., that $Z(Q_1^\varsigma, 1) > Z(Q_2^\varsigma, 2)$. Therefore, the existence of $\tilde{c}$ implies the existence of q.

Next, we show that when $c \to \delta^+$, the overstocking cost approaches zero and $Z(Q_1^\varsigma, 1) < Z(Q_2^\varsigma, 2)$.

$$\begin{aligned}
\lim_{c\to\delta^+|\delta<b_2} (Q_1^\varsigma) &= \infty \\
\lim_{c\to\delta^+|\delta<b_2} (Q_2^\varsigma) &= \infty \\
\lim_{c\to\delta^+|\delta<b_2} (Z(Q_1^\varsigma, 1) - Z(Q_2^\varsigma, 2)) &= (B(1)-\delta)(\lambda\theta) - (B(2)-\delta)(\lambda) \\
&= -\lambda(1-\theta)(b_2-\delta) \\
&< 0,
\end{aligned}$$

This means that as long as $b_2 > \delta$, i.e., as long as q exists, then there is at least one value of c for which $Z(Q_1^\varsigma, 1) = Z(Q_2^\varsigma, 2)$, and that such value will lie in the range $(\delta\,,\, B(2))$. This proves that the existence of q implies the existence of $\tilde{c}$.

Finally, note that the order quantity Q_τ^ς decreases in the magnitude of the critical fractile, and that for a unit change in the critical fractile, the change in the order quantity increases in the parameter of the Poisson distribution. Therefore we analyze the change of the critical fractile that determines the unconstrained optimal order quantity. Observe that $\frac{\partial(c-\delta)/(B(1)-\delta+g)}{\partial c} = \frac{1}{B(1)-\delta+g} < \frac{1}{B(2)-\delta+g} = \frac{\partial(c-\delta)/(B(2)-\delta+g)}{\partial c}$. Therefore Q_2^ς is more sensitive to changes in c than Q_1^ς. Since the margin of an administered unit of a drug is fixed for a given access level policy, then $\frac{\partial(Z(Q_1^\varsigma,1)-Z(Q_2^\varsigma,2))}{\partial c} < 0$ is sufficient to prove uniqueness of $\tilde{c}$. □

Proof of Proposition 3:

When $Q_\Gamma^\varsigma \geq \max\{q\,,\, \underline{Q}_2^\varsigma\}$, then $\tau = 2$ is a feasible solution. Therefore, if $Q_\Gamma^\varsigma \geq Q_2^\varsigma$ and $c < \tilde{c}$, then from Proposition 2, it would be optimal to set $\tau = 2$; therefore $c > \tilde{c}$ is both a necessary and sufficient condition for $\tau_{H,\varsigma}^* = 1$. If $\max\{q\,,\, \underline{Q}_2^\varsigma\} < Q_\Gamma^\varsigma < Q_2^\varsigma$, then Q_Γ^ς maximizes the system's expected utility given $\tau = 2$, and such utility, $Z(Q_\Gamma^\varsigma, 2) \gtrless Z(Q_1^\varsigma, 1)$;

therefore $c > \tilde{c}$ is a sufficient condition to guarantee restricted access, but it may or may not be necessary. □

Proof of Theorem 2:

Theorem 2 follows directly from the results shown in Propositions 1, 2 and 3.

For part a1, if $Q_\Gamma^\varsigma < q \leq \underline{Q}_2^\varsigma$, then $\tau = 2$ is feasible but is dominated by $\tau = 1$, by Proposition 1. If $Q_\Gamma^\varsigma < \underline{Q}_2^\varsigma < q$, then $\tau = 2$ is not feasible. If $c > \tilde{c}$, then restricted access dominates, by Proposition 2.

Part a2 is given by the complement of Proposition 3-a, and part a3 follows directly from Proposition 2-b.

For part a4, recall from Proposition 2 that $c = \tilde{c} \iff Z(Q_1^\varsigma, 1) = Z(Q_2^\varsigma, 2)$. Therefore, in order to be indifferent, it must be that $\lfloor Q_2^\varsigma \rfloor$ is a feasible solution, $\Rightarrow \lfloor Q_2^\varsigma \rfloor \leq Q_\Gamma^\varsigma$.

For part a5, when $\lfloor Q_2^\varsigma \rfloor > Q_\Gamma^\varsigma \geq \max\{q\,,\, \underline{Q}_2^\varsigma\}$ and $c < \tilde{c}$, then $Z(Q_1^\varsigma, 1) \gtrless Z(Q_\Gamma^\varsigma, 2) < Z(Q_2^\varsigma, 2)$, therefore the result is ambiguous.

Part b follows directly from Proposition 1, since $Z(Q_1^\varsigma, 1) = S(Q_1^\varsigma, 1) - cQ_1^\varsigma > S(Q, 2) - cQ$, for any $Q > 0$. □

Proof of Corollary 2:

For parts a and b, we show that Theorem 2-a2 is a subset of Theorem 1-a3. From Theorem 1-a3, $\min\{\bar{Q}_2^\varsigma\,,\, Q_\Gamma^\varsigma\} \geq \max\{q\,,\, \underline{Q}_2^\varsigma\}$ is a necessary condition for $\tau_{S,\varsigma}^* = 2$. First, if $\bar{Q}_2^\varsigma \leq Q_\Gamma^\varsigma$, then $\bar{Q}_2^\varsigma \geq \max\{q\,,\, \underline{Q}_2^\varsigma \Rightarrow Q_\Gamma^\varsigma \geq \max\{q\,,\, \underline{Q}_2^\varsigma$. Therefore, we just need to prove that the set of parameters that jointly satisfy $(Q_\Gamma^\varsigma \geq Q_2^\varsigma)$, and $(c < \tilde{c})$, is a subset of the set of parameters that satisfy $(Q_\Gamma^\varsigma \geq \max\{q\,,\, \underline{Q}_2^\varsigma\})$. Suppose $Q_2^\varsigma < \underline{Q}_2^\varsigma$; since $Z(Q_2^\varsigma, 2) \geq Z(Q, 2)$ for any $Q \neq Q_2^\varsigma$ and $Z(Q, 2)$ is concave in Q, this is a contradiction; therefore $Q_2^\varsigma \geq \underline{Q}_2^\varsigma$. Next, suppose $Q_2^\varsigma < q$; then $Z(Q_2^\varsigma, 1) > Z(Q_2^\varsigma, 2) \Rightarrow Z(Q_1^\varsigma, 1) > Z(Q_2^\varsigma, 2) \Rightarrow c > \tilde{c}$. Therefore, it

must be that either $Q_2^{\varsigma} \geq q$ or that $c > \tilde{c}$. Therefore, $(Q_\Gamma^{\varsigma} \geq Q_2^{\varsigma}) \cup (c < \tilde{c}) \Rightarrow \min\{\bar{Q}_2^{\varsigma}\,,\, Q_\Gamma^{\varsigma}\} \geq \max\{q\,,\, \underline{Q}_2^{\varsigma}\}$, but not the other way around. Additionally, the conditions from Theorem 2-a3 which create indifference in the decision maker with respect to the access level policy were included in Theorem 1-a4, where the full access policy was the unique solution; also the set of parameters that satisfy Theorem 2-a4 and result in $\tau_{H,\varsigma}^* = 1$ due to the budget constraint further reduce the set of parameters that achieve $\tau_{H,\varsigma}^*$.

For part c, consider the case when $\tau_{S,\varsigma}^* = 2$ and $\tau_{H,\varsigma}^* = 1$, then $Q_{H,\varsigma}^* \leq Q_{S,\varsigma}^*$, where $Q_{H,\varsigma}^* = Q_{S,\varsigma}^*$ only if $Q_1^{\varsigma} = \underline{Q}^{\varsigma} = Q_\Gamma^{\varsigma}$. Next consider the case where $\tau_{S,\varsigma}^* = \tau_{H,\varsigma}^*$. Then, for $\tau = 1, 2$, if $Q_\Gamma^{\varsigma} \leq Q_\tau^{\varsigma} \Rightarrow Q_{S,\varsigma}^* = Q_{H,\varsigma}^* = Q_\Gamma^{\varsigma}$; alternatively, if $Q_\Gamma^{\varsigma} > Q_\tau^{\varsigma} \Rightarrow Q_{S,\varsigma}^* = \min\{Q_\Gamma^{\varsigma}, \bar{Q}_\tau^{\varsigma}\} > Q_\tau^{\varsigma} = Q_{H,\varsigma}^*$. □

Proof of Lemma 6:

Let $I = 3$, and first assume that $q_{1,2} \leq q_{2,3}$. There are two associated possibilities which would contradict the Lemma. Suppose for the first option, that $q_{1,2} \leq q_{2,3} < q_{1,3}$, which would imply $S(q_{1,2}+1, 3) < S(q_{1,2}+1, 1) < S(q_{1,2}+1, 2)$; since $S(\cdot, 1)$ and $S(\cdot, 2)$ can't cross more than once, then the access level curve with $\tau = 3$ can't cross the access level curve with $\tau = 2$ before crossing the access level curve with $\tau = 1$; hence $q_{1,3} < q_{2,3}$. Suppose for the second option that $q_{1,3} < q_{1,2} \leq q_{2,3}$, and recall that $S(0, 3) \leq S(0, 2) \leq S(0, 1)$, therefore the access level curve with $\tau = 3$ can't cross the access level curve with $\tau = 1$ before either crossing the access level curve with $\tau = 2$, or $q_{1,2}$ existing before. Therefore, if $q_{1,2} \leq q_{2,3}$, it must be that $q_{1,2} \leq q_{1,3} \leq q_{2,3}$.

The argument is parallel when it is assumed that $q_{1,2} > q_{2,3}$. The proof for any three consecutive access levels then follows by induction. □

Proof of Lemma 7:

Let $I = 3$. For part a, suppose $q_{2,3} < q_{1,3} < q_{1,2}$. Then for any $x > 0, x < q_{1,2}$, $S(q_{1,2} - x, 1) > S(q_{1,2} - x, 2)$, and $S(q_{1,3} + x, 3) > S(q_{1,3} + x, 1)$. Since $q_{1,3} < q_{1,2}$, then $S(Q, 2) < \max\{S(Q, 1), S(Q, 3)\}$, for any $Q > 0$.

For part b), suppose $q_{1,2} < q_{1,3} < q_{2,3}$. Then $S(q_{1,2}-x, 1) > S(q_{1,2}-x, 2) > S(q_{1,2}-x, 3)$; also, for small $\varepsilon > 0$, $S(q_{1,2}+\varepsilon, 2) > S(q_{1,2}+\varepsilon, 1) > S(q_{1,2}+\varepsilon, 3)$; next, $S(q_{1,3}+\varepsilon, 2) > S(q_{1,3}+\varepsilon, 3) > S(q_{1,3}+\varepsilon, 1)$; and finally for any $y > 0$, $S(q_{2,3} + y, 3) > S(q_{2,3} + y, 2) > S(q_{2,3} + y, 1)$.

The proof for any three consecutive access levels follows by induction. □

Proof of Proposition 4:

We only need to prove that if $q_{i,i+1} \leq q_{i+1,i+2}$, for $i = 1, \ldots, I-2$, then if $\hat{q} = q_{i,i+1}$, then $(i+2)$ cannot be an optimal solution. Note that for $x > 0, x < Q^{\varsigma}_{\Gamma}$, $S(Q^{\varsigma}_{\Gamma} - x, i+1) > S(Q^{\varsigma}_{\Gamma} - x, i+2)$, and $Q > Q^{\varsigma}_{\Gamma}$ is not a feasible solution. The algorithm then extends directly from Theorem 1 and Lemma 7-b. □

Proof of Lemma 8:

First, note that $\tilde{c}_{i-1,i}$ is increasing in b_i. Therefore, since $b_i < b_{i+1}, i \geq 1$, then $\tilde{c}_{i-1,i} > \tilde{c}_{i-1,i+1}$. Second, by the same principle, $\tilde{c}_{i-1,i}$ is increasing in b_{i-1}. Therefore, since $b_{i-1} > b_i, i \geq 2$, then $\tilde{c}_{i-1,i+1} > \tilde{c}_{i,i+1}$. Putting the two results together, $\tilde{c}_{i-1,i} > \tilde{c}_{i-1,i+1} > \tilde{c}_{i,i+1}$. □

Proof of Lemma 9:

Follows directly from Proposition 2b. □

Proof of Proposition 5:

Follows directly from Lemma 9 and Theorem 2. □

Proof of Proposition 6:

The parts a) - e) are fully equivalent to Proposition 1, where only b_1 is replaced with β_1, and b_2 is replaced with β_2.

Part f) follows directly from the fact that q is increasing in b_1 and decreasing in b_2. Therefore, when $\beta_1 < b_1$ is the only change, the crossing point decreases.

For part g), if $\beta_2 < b_2$ is the only change, clearly the crossing point increases due to Proposition 1e, and Proposition 6e. Even if both categories decrease their expected health benefits by the same amount, then the effect caused by the decrease in category 1 is lower than that caused by the decrease in category 2. To see this, replace $\beta_1 = b_1 - (b_2 - \beta_2) + x, x \in [0, b_2 - \beta_2)$ into the left hand side of the inequality in (2.3.4), which is:

$$\left(\frac{g}{b_1 - \delta + g}\right)\left(\frac{\lambda}{A(Q,2)}\right) + \left(\frac{b_1 - b_2}{b_1 - \delta + g}\right)$$

When $x = 0$, we obtain:

$$\left(\frac{g}{b_1 - (b_2 - \beta_2) - \delta + g}\right)\left(\frac{\lambda}{A(Q,2)}\right) + \left(\frac{b_1 - b_2}{b_1 - (b_2 - \beta_2) - \delta + g}\right)$$

where the numerators are the same as those in (2.3.4), and the denominators are lower, therefore requiring an increase in the order quantity to bring balance back to the equation. When $b_2 - \beta_2 > x > 0$, the denominator is still decreased because $(b_2 - \beta_2) - x > 0$, and the numerator in the second term is increased by x, therefore the logic remains, completing the

proof for part g).

$$\left(\frac{g}{b_1-(b_2-\beta_2)+x-\delta+g}\right)\left(\frac{\lambda}{A(Q,2)}\right)+\left(\frac{b_1-b_2+x}{b_1-(b_2-\beta_2)+x-\delta+g}\right)$$

Finally, when $b_2-\beta_2<x$, the direction of the change is ambiguous because the numerator in the second term would decrease relative to (2.3.4), resulting in part h). □

Appendix 2: Proofs for Chapter 3

Proof of equation (3.3.3)

For a fixed τ and Q, $\frac{\partial H(Q,\tau;w)}{\partial w} = -Q < 0$. And for $\varepsilon > 0$, $\underline{Q}^{\eta}_{\tau(w)} \leq \underline{Q}^{\eta}_{\tau,(w+\varepsilon)}$ and $\bar{Q}^{\eta}_{\tau(w)} \geq \bar{Q}^{\eta}_{\tau(w+\varepsilon)}$ Therefore, for fixed τ, $\exists\ \bar{\bar{w}}^{\eta}_{\tau} = \{w \mid H(Q,\tau;w) \geq 0$ for at least one value of Q; and $H(Q,\tau;w+\varepsilon) < 0, \forall\ Q$. □

Proof of equation (3.3.4)

For a fixed τ and Q, $\frac{\partial H(Q,\tau;w)}{\partial w} = -Q < 0$. And for $\varepsilon > 0$, $\underline{Q}^{\eta}_{\tau(w)} \leq \underline{Q}^{\eta}_{\tau,(w+\varepsilon)}$ and $\bar{Q}^{\eta}_{\tau(w)} \geq \bar{Q}^{\eta}_{\tau(w+\varepsilon)}$ Therefore, for fixed τ, $\exists\ \bar{\bar{w}}^{\eta}_{\tau} = \{w \mid H(Q,\tau;w) \geq 0$ for at least one value of Q; and $H(Q,\tau;w+\varepsilon) < 0, \forall\ Q$. □

Proof of equation (3.3.5)

For a fixed τ and Q, $\frac{\partial H(Q,\tau;w)}{\partial w} = -Q < 0$. And for $\varepsilon > 0$, $\underline{Q}^{\eta}_{\tau(w)} \leq \underline{Q}^{\eta}_{\tau,(w+\varepsilon)}$ and $\bar{Q}^{\eta}_{\tau(w)} \geq \bar{Q}^{\eta}_{\tau(w+\varepsilon)}$ Therefore, for fixed τ, $\exists\ \bar{\bar{w}}^{\eta}_{\tau} = \{w \mid H(Q,\tau;w) \geq 0$ for at least one value of Q; and $H(Q,\tau;w+\varepsilon) < 0, \forall\ Q$.

.

Proof of Proposition 7:

Notice that the existence of $\underline{w}^{\eta}_{\tau}$ depends entirely on the budget constraint being rele-

vant or not to Health under a given access level. The proof is then given by equations (3.3.3) - (3.3.6). First, if $\underline{w}_\tau^\eta$ exists, then $T(\underline{w}_\tau^\eta, Q^*_{S,\eta\,(\underline{w}_\tau^\eta)} + 1) > \Gamma \geq T(\underline{w}_\tau^\eta, Q^*_{S,\eta\,(\underline{w}_\tau^\eta)})$, and $T(w, Q^*_{S,\eta\,(\underline{w}_\tau^\eta)}) > \Gamma \geq T(\underline{w}_\tau^\eta, Q^*_{S,\eta\,(\underline{w}_\tau^\eta)})$ for any $w > \underline{w}_\tau^\eta$. Therefore revenues for Pharma can't increase, and manufacturing costs can only be reduced by increasing the selling price which has been shown to be infeasible. Second, if $\underline{w}_\tau^\eta$ does not exist, then by definition $\bar{w}_{\tau,S}^\eta$ satisfies both constraints and maximizes Pharma's utility function for a given $\tau^*_{S,\eta}$. □

Proof of Theorem 3:

For part a1), if $\underline{w}_1^\eta$ exists, then inducing full access would require a price $w < \underline{w}_1^\eta$, which would result in a larger order quantity and therefore larger costs; since $\underline{w}_1^\eta$ already achieves the highest revenue Γ, Pharma's profits strictly decrease by inducing full access.

For parts a2) and a3), the optimal selling price for a given access level and a set of parameters is given from Proposition 7; so we must compare Pharma's profits for each induced access level. For example, in part a2), restricted access is preferred iff $(\bar{w}_{1,S}^\eta - c)Q^*_{S,\eta\,(\bar{w}_{1,S}^\eta)} > (\underline{w}_2^\eta - c)Q^*_{S,\eta\,(\underline{w}_2^\eta)}$. The same direct comparison applies for a3).

For part b), it is certain from Proposition 1 that Health will choose to restrict access regardless of Pharma's choice of the selling price. □

Proof of Proposition 8:

This proof makes use of equation (2.5.5), which defines $\tilde{w}$. See Chapter 2 for further references on it.

For Proposition 8.1, i.e., when the access level is most restrictive, the mechanics are exactly the same as in Proposition 7.

For Proposition 8.2, i.e., when $\tau = 2$ is a feasible solution for some w, then $(H^\eta(Q^\eta_{2\,(w)}, 2;\, w) < H^\eta(Q^\eta_{1\,(w)}, 1;\, w))\mid_{w>\tilde{w}}$, and the only way to induce Health to increase access is by setting $w \leq \tilde{w}$, such that $(H^\eta(Q^\eta_{2\,(w)}, 2;\, w) \geq H^\eta(Q^\eta_{1\,(w)}, 1;\, w))\mid_{w\leq\tilde{w}}$.

When $\underline{w}^\eta_2$ exists we can consider two possibilities. If $\tilde{w}\underline{Q}^\eta_{2\,(\tilde{w})} > \Gamma$, then $\tilde{w}$ is not a feasible selling price, and $\tilde{w} > \underline{w}^\eta_2 = w^*_{H,\eta}$, given full access. Else if $\tilde{w}\underline{Q}^\eta_{2\,(\tilde{w})} \leq \Gamma$, then $\underline{w}^\eta_2 \geq \tilde{w} = w^*_{H,\eta}$.

When $\underline{w}^\eta_2$ does not exist, recall $H^\eta(Q^\eta_2(\bar{w}^\eta_{2,H}), 2;\, \bar{w}^\eta_{2,H}) \geq 0$; so if $H^\eta(Q^\eta_{1\,(\tilde{w})}, 1;\, \tilde{w}) = H^\eta(Q^\eta_{2\,(\tilde{w})}, 2;\, \tilde{w}) < 0$, then $\tilde{w} > \bar{w}^\eta_{2,H}$ and Pharma does not need to artificially reduce its selling price to induce higher access. Else if $H^\eta(Q^\eta_{1\,(\tilde{w})}, 1;\, \tilde{w}) = H^\eta(Q^\eta_{2\,(\tilde{w})}, 2;\, \tilde{w}) \geq 0$, then $\tilde{w} \leq \bar{w}^\eta_{2,H}$, and Pharma loses some potential profit due to Health's ability to restrict access in order to maximize the expected utility function. □

Proof of Theorem 4:

The proof follows directly from the proof in Proposition 8 and the argument is parallel to that of Theorem 3. □

Proof of Lemma 10:

From (3.4.2) , if $Q^*_{j,\chi} = Q_\Gamma$, i.e., if the budget constraint is binding in the exogenous price-only contract, then $(w-c)Q^*_{j,\chi} > (w-c)Q,\ \forall\ Q \neq Q^*_{j,\chi}$, i.e., Pharma can't increase its revenues nor reduce its costs by setting a positive buffer. □

Proof of Lemma 11:

Using the condition expressed in Lemma 10, for $j = S, H$, given that $K^*_{j,\kappa} > 0$, Pharma's

objective function can be rewritten as $M^{\kappa}(K^*_{j,\kappa}; Q^*_{j,\kappa}, \tau^*_{j,\kappa}) = wA(K_T, \tau) - cK_T - w(Q^*_{j,\kappa} - A(Q^*_{j,\kappa}, \tau^*_{j,\kappa}))$. Following the proof from Lemma 1, the last two terms of the latter equation are independent of K_T, the function is concave in K_T, and taking first order differences yields that $P(\bar{K}_T, \lambda F(\tau)) > \frac{c}{w} > P(\bar{K}_T + 1, \lambda F(\tau))$. □

Proof of Lemma 12:

Note that $T(w, Q, K) = wQ + (w + p)(\min(K\,,\, (D(\lambda, \tau) - Q)^+) \leq \Gamma$. Therefore, for a choice of Q and τ, and a demand realization, it must be that $wQ + (w + p)K \leq \Gamma \Rightarrow K \leq \frac{\Gamma - wQ}{w+p}$.

For the second part, by taking the partial derivative we learn that by increasing Q in 1 unit, $K_{\Gamma(Q)}$ decreases in $\frac{w}{w+p}$ units, and K_T increases in $\frac{p}{w+p}$ units. Therefore for the total quantity available, given the budget constraint, to increase by 1 unit, it must be that $\frac{p\Delta}{w+p} \geq 1 \Rightarrow \Delta \geq \frac{w+p}{p}$, where Δ represents how many incremental units of the initial order quantity are required to increase the total quantity. The ceiling function is used to keep integrality. □

Proof of Lemma 13:

We simply rearrange $H^{\kappa}(Q, \tau; K) = (B_h(\tau) + g - w - p)A(Q + K, \tau) + (p + w - \delta)A(Q, \tau) - (w - \delta)Q - g\lambda F(\tau) \geq 0 \Rightarrow A(Q + K, \tau) \geq \frac{(w-\delta)Q + g\lambda F(\tau) - (w+p-\delta)A(Q,\tau)}{B(\tau) - w - p + g}$. Since $A(Q + K, \tau)$ is non-decreasing in K, then all capacity buffers larger (respectively, smaller) than $K_{E(Q,\tau)}$ will (respectively, will not) satisfy the cost-effectiveness constraint. □

Proof of Proposition 9:

For part a), Lemmas 11 and 12 provide upper bounds, while Lemma 13 provides a lower bound. Then, when $K_{E(Q,\tau)} > \min(\bar{K}_{(Q,\tau)}, K_{\Gamma(Q)})$, there is no feasible solution involving a positive capacity buffer.

For part b), $(Q^*_{j,\chi}, \tau^*_{j,\chi})$ is a feasible solution when $K^*_{j,\chi} = 0$. Therefore, if $0 < K_{E(Q^*_{j,\chi},\tau^*_{j,\chi})} \leq \min(\bar{K}_{(Q^*_{j,\chi},\tau^*_{j,\chi})}, K_{\Gamma(Q^*_{j,\chi})})$, a positive capacity buffer is a feasible solution. Since Health's expected utility and social welfare are non-decreasing in K (because Health can choose not to purchase K), then $K^*_{j,\kappa} > 0$ is guaranteed.

For part c), since Pharma's profits are increasing in K_T up to the value $\bar{K}_T$, then the optimal solution binds at the upper bound given by the smallest of $\bar{K}_{(Q,\tau)}$ and $K_{\Gamma(Q)}$. □

Proof of Lemma 14:

Recall that $\frac{\Gamma - wQ}{w+p} \geq K_{\Gamma(Q)} > \frac{\Gamma - wQ}{w+p} - 1$, and $P\left(\bar{K}_{T(\tau)}; \lambda F(\tau)\right) > \frac{c}{w}$. Notice that for $K_{\Gamma(Q)} > 0$, $P\left(Q + K_{\Gamma(Q)}; \lambda F(\tau)\right) \geq P\left(\left\lfloor Q + \frac{\Gamma - wQ}{w+p} \right\rfloor; \lambda F(\tau)\right) = P\left(\left\lfloor \frac{\Gamma + pQ}{w+p} \right\rfloor; \lambda F(\tau)\right)$. Then if $P\left(\left\lfloor \frac{\Gamma + pQ}{w+p} \right\rfloor; \lambda F(\tau)\right) < \frac{c}{w} \Rightarrow Q + K_{\Gamma(Q)} \geq Q + \bar{K}_{(Q,\tau)} \Rightarrow K_{\Gamma(Q)} \geq \bar{K}_{(Q,\tau)}$. □

Proof of Lemma 15:

Recall that $K_{\Gamma(Q)}$ is decreasing in p. As $p \to 0^+$, $T(w, Q, K) \to w(Q + K)$, and $K_{\Gamma(Q)} \to \left\lfloor \frac{\Gamma}{w} \right\rfloor - Q$, *i.e.,* $K_{\Gamma(Q)} + Q \to \left\lfloor \frac{\Gamma}{w} \right\rfloor$. Therefore, if $P\left(Q + K_{\Gamma(Q)}; \lambda F(\tau)\right) > P\left(\left\lfloor \frac{\Gamma}{w} \right\rfloor; \lambda F(\tau)\right) > \frac{c}{w}$, then $K_{\Gamma(Q)} < \bar{K}_{(Q,\tau)}$. □

Proof of Proposition 10:

In condition i), clearly $Q^*_{S,\kappa} \leq Q^{\chi}_{\Gamma}$. Then for integer $0 < x < Q^{\chi}_{\Gamma}$, from Lemma 12b)

it follows that $K_{\Gamma(Q_\Gamma^\chi - x)} + Q_\Gamma^\chi - x < Q_\Gamma^\chi$, which creates a decrease in social welfare due to the decrease in the total inventory available in the system, and is therefore not incentive compatible for Health.

In condition ii), $\bar{Q}_\tau^\chi$ is a feasible solution, but $w\bar{Q}_\tau^\chi + (w+p)x > \Gamma$, for $x \geq 1$. Therefore purchasing any positive capacity buffer above $\bar{Q}_\tau^\chi$ violates the budget constraint, and from condition i), reducing the initial order quantity can only reduce social welfare.

In condition iii), $\bar{Q}_\tau^\chi$ is a feasible solution, but $P(\bar{Q}_\tau^\chi; \lambda F(\tau)) < \frac{c}{w} < P(\bar{K}_{T(\tau)}; \lambda F(\tau))$, i.e., $\bar{K}_{T(\tau)} < \bar{Q}_\tau^\chi$. □

Proof of Definition 3:

Recall that $H(\bar{Q}_\tau^\chi, \tau) \geq 0 > H(\bar{Q}_\tau^\chi + 1, \tau)$. Therefore, to maintain cost-effectiveness for $Q > \bar{Q}_\tau^\chi$, it must be that the incremental costs: $(B_h(\tau) - \delta + g)(A(Q,\tau) - A(Q_\tau^\chi, \tau)) - (w - \delta)(Q - Q_\tau^\chi) < 0$, are lower than the incremental benefits: $(B_h(\tau) + g - w - p)(A(K_{T(\tau)}, \tau) - A(Q, \tau)) \geq 0$. Part a) replaces $K_{T(\tau)}$ with the fixed value $\bar{K}_{T(\tau)}$. Part b) replaces $K_{T(\tau)}$ with the maximum capacity, as a combination of Q and K, that satisfies the budget constraint. □

Proof of Lemma 16:

For part a), given a fixed K_τ, $S_h^\kappa(Q,\tau;K) - S_h^\kappa(Q-1,\tau;K) = \delta(1 - A(Q,\tau) + A(Q-1,\tau)) \geq 0$, for $Q > 0$.

For part b), given a fixed Q, $S_h^\kappa(Q,\tau;K) - S_h^\kappa(Q,\tau;K-1) = (B_h(\tau) + g)(A(Q+K,\tau) + A(Q+K-1,\tau)) > 0$, for $K > 0$.

For part c), given $\breve{Q}_\tau^\kappa \leq \bar{\bar{Q}}_\tau^\kappa$, $K_{S,\kappa}^*(Q,\tau) = \bar{K}_{(Q,\tau)}$, for $Q \in [\breve{Q}_\tau^\kappa\,,\, \bar{\bar{Q}}_\tau^\kappa]$. Therefore K_T is fixed, and from part a), social welfare weakly increases in Q.

For part d), given $\breve{Q}_\tau^\kappa > \bar{Q}_\tau^\kappa$, $K_{T\,(Q,\tau)}$ increases in Q from Lemma 12 as long as cost-effectiveness is satisfied. From Lemma 13, increasing Q also increases the required K for cost-effectiveness. Therefore increasing Q increases K_T up to the point where satisfying cost-effectiveness would require a total budget higher than Γ; i.e., social welfare increases for $Q \in [\breve{Q}_\tau^\kappa\,,\, \bar{Q}_\tau^\kappa]$.

Part e) is simply a special case of part d), as $\bar{K}_{T(\tau)}$ is not a feasible solution due to the budget constraint. □

Proof of Proposition 11:

For part a), from Lemma 12a, any order quantity in the range $\left[\breve{Q}_\tau^\kappa\,,\, \bar{\bar{Q}}_\tau^\kappa\right]$ is feasible contingent on $K_T = \bar{K}_{T\,(\tau)}$. Therefore Pharma's total capacity choice is independent of Health's constraints. To find the order quantity, from Lemma 16c, social welfare is increasing in the order quantity up to when $\delta > 0$, and constant when $\delta = 0$ (because only K_T is relevant).

For part b), from Definition 3b, the budget constraint will be binding for Health, and Pharma's total capacity is bounded by Γ and the order quantity. From Lemma 16d and Lemma 16e, social welfare is increasing in the order quantity up to $\bar{Q}_\tau^\kappa$, and Pharma will make its total capacity choice knowing Health's optimal order quantity. □

Proof of Proposition 12:

Parts a) - d) are equivalent to Propositions 1 and 3.

For part e), suppose $S_h^\kappa(Q+K,1) = S_h^\kappa(Q+K,2)$, for some $Q+K>0$. We then obtain the following equality:

$$(B_h(1)+g)A(Q+K,1)+\delta(Q-A(Q,1))-g\lambda F(1) =$$
$$(B_h(2)+g)A(Q+K,2)+\delta(Q-A(Q,2))-g\lambda F(2)$$
$$(B_h(1)+g-\delta)A(Q+K,1)-g\lambda F(1)+\delta(A(Q+K,1)-A(Q,1)) =$$
$$(B_h(2)+g-\delta)A(Q+K,2)-g\lambda F(2)+\delta(A(Q+K,2)-A(Q,2))$$

Notice that when $K=0$, the expression is reduced to $(B_h(1)+g-\delta)A(Q,1)-g\lambda F(1) = (B_h(2)+g-\delta)A(Q,2)-g\lambda F(2)$, and recall $A(Q,2)-A(Q,1)$ is non-decreasing in Q, which implies that for $K>0$: $A(Q+K,2)-A(Q+K,1) \geq A(Q,2)-A(Q,1) \Rightarrow A(Q+K,2)-A(Q,2) \geq A(Q+K,1)-A(Q,1)$. Therefore if $q^h = Q+K$, then $S_h^\kappa(Q+K,1) < S_h^\kappa(Q+K,2)$ for $K>0$, which implies that $q^h > q^\kappa$.

For part f), when $\delta=0$, $S_h^\kappa(Q+K,1) = (B_h(1)+g)A(Q+K,1)-g\lambda F(1) = (B_h(2)+g)A(Q+K,2)-g\lambda F(2) = S_h^\chi(Q+K,2)$, which is the same relationship from equation (2.5.6). □

Proof of Proposition 13:

For part a), when $\tau_{S,\chi}^* = 1$ and $Q_{S,\chi}^* = \bar{Q}_1^\chi$, it must be that $S_h^\chi(q,1) = S_h^\chi(q,2) < 0$. Therefore the condition on the total capacity $K_T > q^\kappa$ is necessary but not sufficient. Therefore to satisfy both constraints, it must be that $q^\kappa < K_T = Q + K_{E(Q,2)} \leq Q + K_{\Gamma(Q)}$.

For part b), when $\tau_{S,\chi}^* = 2$, then the same order quantity is feasible under the capacity buffer contract, and since $q^\kappa \leq q^\chi$, then full access will continue to be optimal. □

Proof of Lemma 17:

For part a), on one hand the expected profit of purchasing an additional unit in advanced is: $B_h(\tau)P(Q;\lambda F(\tau)) + \delta(1 - P(Q;\lambda F(\tau))) - w$. On the other hand, the expected profit of delaying the purchase is: $(B_h(\tau) - w - p)P(Q;\lambda F(\tau))$. Therefore, purchasing in advanced is profitable only if: $P(Q;\lambda F(\tau)) > \frac{w-\delta}{w+p-\delta}$. Since $P(Q;\lambda F(\tau))$ is decreasing in Q, it must be that $P(Q^\kappa_\tau;\lambda F(\tau)) > \frac{w-\delta}{w+p-\delta} > P(Q^\kappa_\tau + x;\lambda F(\tau))$ for any $x \geq 1$.

For part b), recall $P(Q^\chi_\tau;\lambda F(\tau)) > \frac{w-\delta}{B_h(\tau)+g-\delta}$. Since we have assumed $B_h(\tau)+g > w+p$, then $\frac{w-\delta}{B_h(\tau)+g-\delta} < \frac{w-\delta}{w+p-\delta} < P(Q^\kappa_\tau;\lambda F(\tau))$.

For part c), $\lim_{(w+p)\to(B_h(\tau)+g)} \frac{w-\delta}{B_h(\tau)+g-\delta} - \frac{w-\delta}{w+p-\delta} = 0.$ □

Proof of Proposition 14:

For part a), from Lemma 8c), $Q+\Delta+K_{\Gamma(Q+\Delta)} = Q+K_{\Gamma(Q)}+1$, and $Q+\alpha\Delta+K_{\Gamma(Q+\Delta)} = Q + K_{\Gamma(Q)} + \alpha$. Recall $A(Q,\tau) = \sum_{x=1}^{Q} P(x;\lambda F(\theta))$. Then it must be that

$$
\begin{aligned}
0 < H^\kappa(Q^\kappa_\tau + \alpha\Delta, \tau; K_{\Gamma(Q^\kappa_\tau+\alpha\Delta)}) - H^\kappa(Q^\kappa_\tau, \tau; K_{\Gamma(Q^\kappa_\tau)}) = \\
(B_h(\tau) + g - w - p) \sum_{x=Q^\kappa_\tau+1}^{Q^\kappa_\tau+\alpha} P(Q^\kappa_\tau + \alpha\Delta + K_{\Gamma(Q^\kappa_\tau+\alpha\Delta)};\lambda F(\tau)) - \alpha\Delta(w-\delta) \\
+(w+p-\delta) \sum_{x=Q^\kappa_\tau+1}^{Q^\kappa_\tau+\alpha\Delta} P(x;\lambda F(\tau))
\end{aligned}
$$

For part b), it must be that Health's cost-effectiveness constraint is satisfied, using the definition of Lemma 9 and the result in Proposition 6. □

Proof of Lemma 18:

Follows directly from the solution process in Chapter 2, section 2.3.3 □

Proof of Lemma 19:

Suppose Q is fixed. Then, for part a), notice that $\frac{\partial H_{high}(\gamma,r;Q)}{\partial \gamma} = (b-\beta)A(Q) > 0$. Therefore, when Q is at the upper bound of $H_{high}(\gamma, r; Q)$, the increased profit gap created by an increase in γ may be compensated by increasing Q.

Parts b) and c) are straightforward since the terms are not present in the corresponding objective function.

For part d), notice that $\frac{\partial H_{low}(\gamma,r;Q)}{\partial r} = QP(m;\lambda) > 0$. The argument then follows as for part a).

For part e), Q_Γ is independent of γ and r; $\bar{Q}^\rho_{low}$ weakly increases in r; and $\bar{Q}^\rho_{high}$ weakly increases in γ. Therefore, $Q^*_{S,\rho}$ weakly increases in γ and r. □

Proof of Lemma 20:

Follows directly from the solution process in Chapter 2, section 2.3.3. □

Proof of Lemma 21:

Observe the following derivatives of the critical fractiles from Lemma 20.

$\frac{\partial \frac{w-\delta}{\beta+(b-\beta)\gamma-\delta+g}}{\partial \gamma} = -\frac{(w-\delta)(b-\beta)}{(\beta+(b-\beta)\gamma-\delta+g)^2} < 0.$

$\frac{\partial \frac{w-\delta}{\beta+(b-\beta)\gamma-\delta+g}}{\partial r} = 0$

$\frac{\partial \frac{w-\delta-rP(m;\lambda)}{\beta-\delta+g}}{\partial \gamma} = 0$

$\frac{\partial \frac{w-\delta-rP(m;\lambda)}{\beta-\delta+g}}{\partial r} = -\frac{P(m;\lambda)}{\beta-\delta+g} < 0.$

When the critical fractiles decrease, the order quantity increases, proving parts a-d. For part e), Q_Γ is independent of γ and r; Q^ρ_{low} weakly increases in r; and Q^ρ_{high} weakly increases in γ. Therefore, $Q^*_{H,\rho}$ weakly increases in γ and r. □

Proof of Proposition 15:

From Lemmas 18 and 19, if $\bar{Q}^\rho_{high} > \bar{Q}^\rho_{low}$, then γ can be decreased without affecting the optimal order quantity; if $\bar{Q}^\rho_{high} < \bar{Q}^\rho_{low}$, then r can be decreased without affecting the optimal order quantity. Lemmas 20 and 21 give an equivalent result for Q^ρ_{high} and Q^ρ_{low}. Notice that $H_{high}(Q;\, \gamma, r) - H_{low}(Q;\, \gamma, r) = \gamma(b-\beta)A(Q) - rQP(m;\lambda)$. By setting the difference to zero, it follows that: $\frac{r}{\gamma} = \left(\frac{A(Q)}{Q}\right)\left(\frac{b-\beta}{P(m;\lambda)}\right)$. Since this must hold for any Q chosen by Health, it must hold for $Q^*_{j,\rho}$. □

Proof of Lemma 22:

It follows directly by replacing $r^*_{j,\rho\,(\gamma)} = \left(\frac{A(Q^*_{j,\rho})}{Q^*_{j,\rho}}\right)\left(\frac{b-\beta}{P(m;\lambda)}\right)(\gamma)$ into Pharma's objective function: $M(\gamma,\, r^*_{j,\rho(\gamma)};\, Q^*_{j,\rho}) = M(\gamma;\, Q^*_{j,\rho}) = (w-c)Q^*_{j,\rho} - A(Q^*_{j,\rho})(b-\beta)(\gamma)G(\gamma m)$. □

Proof of Corollary 3:

Given Pharma's private knowledge, the expected transfer from Pharma to Health is: $r^*_{j,\rho}Q^*_{j,\rho}G(\gamma^*_{j,\rho}m)P(m;\lambda)$. Additionaly, Pharma incurs the verification cost $v(m)$. As a result the incremental revenue achieved by offering the performance based contract, must be larger than the incremental costs; that is: $(w-c)\left(Q^*_{j,\rho} - Q^*_{j,\chi}\right) > r^*_{j,\rho}Q^*_{j,\rho}G(\gamma^*_{j,\rho}m)P(m;\lambda) + v(m) \Rightarrow$ $(w-c)\left(\frac{Q^*_{j,\rho}-Q^*_{j,\chi}}{Q^*_{j,\rho}}\right) > r^*_{j,\rho}G(\gamma^*_{j,\rho}m)P(m;\lambda) + \frac{v(m)}{Q^*_{j,\rho}}$. □

[page left intentionally blank]

Appendix 3: Proofs for Chapter 4

Proof of equation (4.3.9) and (4.3.10):

To avoid unnecessary repetition, the notation Q_a is used as a general order quantity, contingent on the value of x (instead of separating for $Q_{a,1}$ and $Q_{a,0}$). By use of Leibniz Rule:

$$\begin{aligned}\frac{\partial Z_2^{SI}(Q)}{\partial Q_a} &= -c + (\beta_a Q_a + \beta_b xN)\psi(Q_a) - (Q_a + xN)\left[\beta_a\left(\frac{Q_a}{Q_a + xN}\right) + \beta_b\left(\frac{xN}{Q_a + xN}\right)\right]\psi(Q_a) \\ &\quad + \int_{Q_a}^{\lambda}\left[\beta_a\left(\frac{\xi}{\xi + xN}\right) + \beta_b\left(\frac{xN}{\xi + xN}\right)\right]\psi(\xi)\,d\xi = 0\end{aligned}$$

Rearranging terms:

$$\begin{aligned}&-c + \int_{Q_a}^{\lambda}\left[\beta_a\left(\frac{\xi}{\xi + xN}\right) + \beta_b\left(\frac{xN}{\xi + xN}\right)\right]\psi(\xi)\,d\xi = 0 \\ &c = \int_{Q_a}^{\lambda}\left[\beta_a\left(1 - \frac{xN}{\xi + xN}\right) + \beta_b\left(\frac{xN}{\xi + xN}\right)\right]\psi(\xi)\,d\xi \\ &c = \beta_a\int_{Q_a}^{\lambda}\psi(\xi)\,d\xi + (\beta_b - \beta_a)\int_{Q_a}^{\lambda}\left(\frac{xN}{\xi + xN}\right)\psi(\xi)\,d\xi \\ &c = \beta_a\left(1 - \Psi(Q_a)\right) + (\beta_b - \beta_a)\int_{Q_a}^{\lambda}\left(\frac{xN}{\xi + xN}\right)\psi(\xi)\,d\xi\end{aligned}$$

□

Proof of Proposition 16:

By replacing $Q_{a,1}$ with Q_a^{MI} in equation (4.3.9):

$$\left\{(\beta_a)\left[1-\Psi\left(Q_a^{MI}\right)\right]-c\right\}+(\beta_b-\beta_a)\int_{Q_a^{MI}}^{\lambda}\frac{xN}{\xi+xN}\psi(\xi)d\xi$$

From equation 4.3.2, $(\beta_a)\left(1-\Psi\left(Q_a^{MI}\right)\right)-c=0$, and therefore the sign of the slope (i.e., the sign of equation 4.3.9) will be positive (alternatively, negative) when $\beta_b > \beta_a$ (alternatively, $\beta_a > \beta_b$), which implies that the optimal order quantity lies to the right (alternatively, left) of the order quantity under dedicated channels. □

Proof of Lemma 23:

The case when $x = 0$ is trivial because $Q_b^{MI} = Q_b^{SI} = 0$, and therefore $Z_2^{SI}(Q+0) = Z_{S_a}^{MI}(Q) + Z_{S_b,2}^{MI}(0)$, for any Q.

The case when $x = 1$ and $\beta_a = \beta_b$ is also trivial since equation (4.3.9) collapses into (4.3.2), and it is irrelevant which patient receives the drugs.

Next, suppose $x = 1$ and $\beta_a > \beta_b$. If $D(\varepsilon, e) = \varepsilon + N \leq Q_a^{MI} + N$, then clearly $Z_2^{SI}(Q_a^{MI}+N) = Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(N)$. However, if $D(\varepsilon, e) = \varepsilon + N > Q_a^{MI} + N$, then $Z_2^{SI}(Q_a^{MI}+N) > Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(N)$, because for every expected stockout from a patient of S_b under SI, a health benefit of $(\beta_a - \beta_b) > 0$ is achieved, relative to the MI arrangement. From Proposition 16, we know that when $\beta_a > \beta_b$ and $x = 1$, then $Q_a^{SI} < Q_a^{MI}$. By transitivity, $Z_2^{SI}(Q_a^{SI}+N) > Z_2^{SI}(Q_a^{MI}+N) > Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(N)$.

Alternatively, suppose $x = 1$ and $\beta_a < \beta_b$. As before, if $D(\varepsilon, e) = \varepsilon + N \leq Q_a^{SI} + N$, then clearly $Z_2^{SI}(Q_a^{SI}+N) = Z_{S_a}^{MI}(Q_a^{SI}) + Z_{S_b,2}^{MI}(N)$. However, if $D(\varepsilon, e) = \varepsilon + N > Q_a^{SI} + N$, then $Z_2^{SI}(Q_a^{SI}+N) < Z_{S_a}^{MI}(Q_a^{SI}) + Z_{S_b,2}^{MI}(N)$, because for every expected stockout from a patient of S_b under SI, a health benefit of $(\beta_b - \beta_a) > 0$ is not achieved, relative to the MI

arrangement. By definition of optimality, $Z_{S_a}^{MI}(Q_a^{SI}) + Z_{S_b,2}^{MI}(N) < Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(N)$, and by transitivity: $Z_2^{SI}(Q_a^{SI} + N) < Z_{S_a}^{MI}(Q_a^{SI}) + Z_{S_b,2}^{MI}(N) < Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(N)$. □

Proof of Proposition 17:

From Lemma 23a, $Z_2^{SI}(Q_{a,0}^{SI})\big|_{x=0} = \left(Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(Q_b^{MI})\right)\big|_{x=0}$; this scenario realizes with probability $(1 - g(e))$, when effort level e is exerted. From Lemma 23b, when $x = 1$ which occurs with probability $g(e)$, $Z_2^{SI}(Q_{a,1}^{SI} + N) \;|_{x=1} \gtreqless \left(Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(Q_b^{MI})\right)\big|_{x=1}$, depending only on whether $\beta_a \gtreqless \beta_b$. We can therefore rewrite the first order conditions for the optimal effort levels under the multiple and single channel structures, respectively, as follows:

$$\begin{aligned}
&\frac{\partial\left(Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,1}^{MI}(e;\, Q_b^{MI})\right)}{\partial e} = -C'(e) - g'(e)\left(\left(Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(Q_b^{MI})\right)\big|_{x=0}\right) \\
&\qquad +g'(e)\left(\left(Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(Q_b^{MI})\right)\big|_{x=1}\right) = 0 \\
&\frac{\partial Z_1^{SI}(e;\, Q^{SI})}{\partial e} = -C'(e) - g'(e)\left(Z_2^{SI}(Q_{a,0}^{SI})\right)\big|_{x=0} + g'(e)\left(Z_2^{SI}(Q_{a,1}^{SI} + N)\right)\big|_{x=1} = 0
\end{aligned}$$

It is true that $\frac{\partial Z_{S_a}^{MI}(Q_a^{MI})}{\partial e} = 0$, and its inclusion is not required. However, writing it in such way, only the last term is different between the two first order conditions. It is easy to see that the second order condition for concavity is satisfied:

$$\begin{aligned}
&\frac{\partial^2\left(Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,1}^{MI}(e;\, Q_b^{MI})\right)}{\partial e^2} = -C''(e) \\
&+g''(e)\left[\left(\left(Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(Q_b^{MI})\right)\big|_{x=1}\right) - \left(\left(Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(Q_b^{MI})\right)\big|_{x=0}\right)\right] < 0
\end{aligned}$$

because $C''(e) \geq 0$, $g''(e) \leq 0$, and $\left(Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(Q_b^{MI})\right)\big|_{x=1} > \left(Z_{S_a}^{MI}(Q_a^{MI}) + Z_{S_b,2}^{MI}(Q_b^{MI})\right)\big|_{x=0}$.

Therefore at optimality,

$$C'(e^{MI})+g'(e^{MI})\left(\left(Z^{MI}_{S_a}(Q^{MI}_a)+Z^{MI}_{S_b,2}(Q^{MI}_b)\right)\Big|_{x=0}\right)=g'(e^{MI})\left(\left(Z^{MI}_{S_a}(Q^{MI}_a)+Z^{MI}_{S_b,2}(Q^{MI}_b)\right)\Big|_{x=1}\right).$$

If we plug e^{MI} into the first order condition under a single channel, we obtain (*see Lemma 23 for details*):

$$\begin{aligned}
&C'(e^{MI})+g'(e^{MI})\left(Z^{SI}_2(Q^{SI}_{a,0})\right)\Big|_{x=0}=g'(e^{MI})\left(Z^{SI}_2(Q^{SI}_{a,1}+N)\right)\Big|_{x=1}, \text{ when } \beta_a=\beta_b\\
&C'(e^{MI})+g'(e^{MI})\left(Z^{SI}_2(Q^{SI}_{a,0})\right)\Big|_{x=0}<g'(e^{MI})\left(Z^{SI}_2(Q^{SI}_{a,1}+N)\right)\Big|_{x=1}, \text{ when } \beta_a>\beta_b\\
&C'(e^{MI})+g'(e^{MI})\left(Z^{SI}_2(Q^{SI}_{a,0})\right)\Big|_{x=0}>g'(e^{MI})\left(Z^{SI}_2(Q^{SI}_{a,1}+N)\right)\Big|_{x=1}, \text{ when } \beta_a<\beta_b
\end{aligned}$$

Due to concavity, if the right hand-side is larger than (respectively lower than) the left-hand side, then the optimal effort e^{SI} must be to the right (respectively, to the left) of e^{MI}. □

Proof of Theorem 5:

Follows directly from Propositions 16 and 17, and Lemma 23. □

Proof of Lemma 24:

The proof is equivalent to that of Proposition 16, where only c is replaced with the selling price w. □

Proof of Proposition 18:

To show the relationship between e^{MX} and e^{SX}, we simply compare the first order con-

ditions for the optimal effort under multiple and single channels.

$$C'(e^{MX}) = (w-c)\left(g'(e^{MX})\right)N$$
$$C'(e^{SX}) = (w-c)\left(g'(e^{SX})(N + Q_{a,1}^{SX} - Q_{a,0}^{SX})\right)$$

From Lemma 24, $Q_{a,1}^{SX} > Q_{a,0}^{SX} = Q_a^{MX}$, when $\beta_b > \beta_a$; similarly $Q_{a,1}^{SX} < Q_{a,0}^{SX}$, when $\beta_b < \beta_a$, and $Q_{a,1}^{SX} = Q_{a,0}^{SX}$, when $\beta_b = \beta_a$. Due to concavity of Pharma's problem, if the right hand-side is larger than (respectively, lower than) the left-hand side, then the optimal effort e^{SX} must be to the right (respectively, to the left) of e^{MX}. □

Proof of Theorem 6:

Follows from Lemma 24 and Proposition 18. □

Proof of Proposition 19:

Part a) is satisfied by definition since $M^{MX}(e^{MX};Q^{MX}) - T^* = M^{SX}(e^{SX};Q^{SX})$, and $M^{MX}(e^{MX};Q^{MX}) = R^* + M^{SX}(e^{SX};Q^{SX})$.

For part a)$M^{MX}(e^{MX};Q^{MX}) - T_m^* = M^{SX}(e^{SX};Q^{SX}) < M^{MX}(e^{MX};Q^{MX}) = R_s^* + M^{SX}(e^{SX};Q^{SX})$. Parts b) and c) follow almost directly.

For part b), when $T = T_m^*$, then Pharma's profits if $x = 1$ under a multiple channel design are the same as when there was no transfer in a single channel design. Therefore the effort level chosen will be equal to e^{SX}.

For part c), when $R = R_s^*$, then Pharma's profits if $x = 1$ under a single product, are the same as when there was no transfer in a multiple channel design. Therefore the effort level

chosen will be equal to e^{MX}. □

Proof of Proposition 20:

Equivalent to Proposition 19. □

Proof of Proposition 21:

For part a), when $\beta_a > \beta_b$, the efficient design choice is a single channel (Theorem 5). If $C_{i,M}(x) > C_{i,S}(x)$, the incentive for pulling further increases.

For part b), when $\beta_b > \beta_a$, the efficient design choice is multiple channels to prevent the pulling effect (Theorem 5). If $C_{i,M}(x) < C_{i,S}(x)$, the incentive for pulling further decreases.

For part c), when $C_{i,M}(x) = C_{i,S}(x) + K$, the first order condition remains unchanged. The reason why the efficient design may change is given in Propositions 19 and 20. □

Proof of equation (4.7.13):

By use of Leibniz Rule:

$$\begin{aligned}
\left.\frac{\partial Z_2^{SS}}{\partial Q}\right|_{x=0} &= -c + \beta_a \left(Q\phi(Q \mid 0) - Q\phi(Q \mid 0) + \int_Q^\infty \phi(\theta \mid 0)\, d\theta \right) \\
&= -c + \beta_a (1 - \Phi(Q \mid 0)) \\
\Rightarrow\ Q_{a,0}^{SS} &= \Phi^{-1}\left(1 - \frac{c}{\beta_a}\middle| 0\right)
\end{aligned}$$

□

Proof of equation (4.7.14):

Let $\alpha = (\beta_a \mathbb{E}[\varepsilon_a \mid \Theta = \theta,\, x = 1] + \beta_b \left(\mathbb{E}[\varepsilon_b \mid \Theta = \theta,\, x = 1]\right))$.

Then, by use of Leibniz Rule:

$$\begin{aligned}
\left.\frac{\partial Z_2^{SS}}{\partial Q}\right|_{x=1} &= -c + \alpha\phi(Q \mid 1) - \frac{Q\alpha\phi(Q \mid 1)}{Q} + \int_{\Theta=Q}^{\infty} \frac{\alpha}{\theta}\phi(\theta \mid 1)\, d\theta \\
&= -c + \int_{\Theta=Q}^{\infty} \frac{\alpha}{\theta}\phi(\theta \mid 1)\, d\theta \\
&= -c + \int_{\Theta=Q_{a,b}^{SS}}^{\infty} \left(\frac{\beta_a \mathbb{E}[\varepsilon_a \mid \Theta = \theta,\, x = 1] + \beta_b \left(\mathbb{E}[\varepsilon_b \mid \Theta = \theta,\, x = 1]\right)}{\theta}\right) \phi(\theta \mid 1)\, d\theta
\end{aligned}$$

□

Proof of Lemma 25:

Recall that $\phi(y|0) = \psi_a(y)$ for any $y > 0$. The proof then follows by direct comparison of (4.7.15) and (4.7.2).

$Q_{a,0}^{SS} = \Phi^{-1}\left(1 - \frac{c}{\beta_a}\middle| 0\right) = \Psi_a^{-1}\left(1 - \frac{c}{\beta_a}\right) = Q_a^{MS}$ □

Proof of Lemma 26:

The result is equivalent to Proposition 16. Let $\beta_b > \beta_a$. By plugging $Q = \Phi^{-1}\left(1 - \frac{c}{\beta_a}\middle| 1\right)$ into (4.7.15), the left hand side of the equation becomes larger than c, $\Rightarrow Q_{a,b}^{SS} > \Phi^{-1}\left(1 - \frac{c}{\beta_a}\middle| 1\right)$. By plugging $Q = \Phi^{-1}\left(1 - \frac{c}{\beta_b}\middle| 1\right)$ into (4.7.16), the left hand side of the equation becomes less than c, $\Rightarrow Q_{a,b}^{SS} > \Phi^{-1}\left(1 - \frac{c}{\beta_b}\middle| 1\right)$. The opposite is true when $\beta_b < \beta_a$. □