



ESSAYS IN HEALTH ECONOMICS: The Design of Primary Care Incentives

**Thesis submitted in accordance with the requirements of the University of
Liverpool for the degree of Doctor of Philosophy
by**

Yunchou Wu

March 2020

**Management School,
University of Liverpool**

Abstract

The continuous upward trend in health spending and the perpetual rise in the demand for healthcare underlines the importance of designing efficient resource allocation mechanisms. In the healthcare sector, physicians play a crucial role in directing resources. This thesis analyses the design of primary care incentive contracts in presence of two complicating issues. From society's perspective, the physician may provide an inefficient amount of treatment ("moral hazard") or be offered a sub-optimal contract ("adverse selection"), as neither the physician's knowledge about patients' health nor his personal characteristics be perfectly observed. This thesis is structured into three essays.

The first essay discusses theory and evidence related to physician altruism. It also comprehensively surveys optimal incentive contracting when physicians have private information about their characteristics (e.g. costs, abilities), their patients' health, or their actions (efforts, service qualities). The impact of using different types of restrictions for controlling the cost of physician services are also reviewed. Finally, research gaps in the literature are identified and the context of subsequent analysis is established.

The second essay studies the optimal design of publicly funded health insurance schemes when physicians have private knowledge about their preferences and patients' health as well as being paid on a price per service basis ("fee-for-service"). This is formalised in a math model, where the regulator chooses a policy vector of price, insurance premium and cost-control policies, while physicians subsequently choose medical service quantities. In this setup, it is shown that the welfare-maximizing price system fails to align different providers' objectives and major trends such as population ageing further exacerbates this distortion. Widely applied physician cost-control restrictions, such as maximal amount of service, fixed budgets or a payment per patient ("capitation"), can be imposed on the price system to improve efficiency. In particular, capitation performs better than other restrictions.

The last essay extends the analysis by letting primary care physicians select their practice sizes. The optimal payment must ensure that the entire population has access to healthcare. It is shown that the welfare-maximizing price system induces medical service overproduction and distorts physicians' incentives. Nevertheless, physician cost-control restrictions are still efficiency enhancing while the highest welfare level is generated by using a capitation. Overall, the thesis provides a framework which not only captures key features of the health market but also can be used to analyse the impact of using different physician cost-control policies.

Table of Contents

List of Table and Figures	v
List of Abbreviations	vi
Acknowledgements	vii
Chapter 1. Introduction.....	1
1.1 THE RISING GOVERNMENT SHARE OF TOTAL HEALTH EXPENDITURE.....	2
1.2 SOURCES OF HEALTHCARE MARKET FAILURES	3
1.3 SUPPLY-SIDE INCENTIVES FOR PRIMARY HEALTHCARE	7
1.4 TYPOLOGIES OF PCP PAYMENTS	8
1.5 FEE-FOR-SERVICE PAYMENT SYSTEMS.....	11
1.6 MAJOR COST CONTROL INSTRUMENTS.....	12
1.7 THESIS OUTLINE.....	15
Chapter 2. Literature Review	18
2.1. INTRODUCTION	18
2.2 THEORY	23
2.3 EVIDENCE	50
2.4 CONCLUSION.....	65
Chapter 3. The Optimal Design of Fee-for-Service Contract 1	69
3.1 INTRODUCTION	69
3.2 RELATED LITERATURE.....	73
3.3 MODEL	75
3.4 THE PHYSICIAN'S OPTIMIZATION	79
3.5 WELFARE ANALYSIS.....	86
3.6 COST CONTROL POLICIES AND EFFICIENCY ENHANCEMENT	104
3.7 CONCLUSION.....	120
3.8 LIMITATIONS AND EXTENSIONS	121
3.9 APPENDIX.....	123
Chapter 4. The optimal design of fee-for-service contract 2.....	132
4.1 INTRODUCTION	132
4.2 RELATED LITERATURE	134
4.3 SETUP	136
4.4 THE PHYSICIAN'S OPTIMIZATION	138
4.5 WELFARE ANALYSIS.....	149
4.6 COST-CONTROL POLICIES AND EFFICIENCY ENHANCEMENTS	158
4.7 SUMMARY	171
4.8 LIMITATIONS AND EXTENSIONS	172

4.9 APPENDIX.....	176
Chapter 5. Conclusion and Further Research	183
5.1 CONCLUSIONS.....	183
5.2 FUTURE STUDIES	187
References.....	189

List of Table and Figures

Table 1. Healthcare expenditure as a percentage of GDP.....	3
Figure 2. Characteristics and Incentives in Provider Payment Systems	10
Figure 3. L-doctor's supply function for a B-patient	89
Figure 4. L- and H- doctors' supply curves for a T-type patient	90
Figure 5. Welfare and the Optimal Price	91
Figure 6. Effects of varying the initial health of healthier patients	93
Figure 7. Effects of varying the proportion of less healthy patients	95
Figure 8. Effects of varying the proportion of more altruistic physicians	97
Figure 9. Effects of varying the number of patients allocated per physician	99
Figure 10. Effects of varying the technology factor	102
Figure 11. Quantity Restriction.....	108
Figure 12. Revenue Restriction.....	111
Figure 13. Capitation	114
Figure 14. The L-doctor's response functions as price varies	154
Figure 15. L- and H-doctors' response curves for a type-t patient	155
Figure 16. Welfare and Total Number of Patients Treated.....	157
Figure 17. Welfare and total service supplied when quantity restriction varies	161
Figure 18. The optimal price and L- and H-doctors' decisions	161
Figure 19. Welfare and total service supplied when revenue restriction varies	164
Figure 20. The optimal price and L- and H-doctors' decisions	164
Figure 21. Welfare and total quantity of service supplied when capitation varies	167
Figure 22. The optimal price and L- and H-doctors' decisions	168

List of Abbreviations

CAP – capitation

DRG – diagnostic-related group

FFS – fee-for-service

GP – general practitioner

HI – health insurance

OECD – Organisation for Economic Co-operation and Development

PCP – primary care provider

PPS – prospective payment system

SNE – symmetric Nash equilibrium

SSNE – Strong symmetric Nash equilibrium

UHC – universal health care

WHO – World Health Organization

Acknowledgements

The completion of this PhD journey would not have been possible without the help and support of many people in my life over the past four years. I am extremely grateful to each and every one of them mentioned here. Firstly, I wish to thank my supervisor, Prof. Dominique Demougin, who has supported my development throughout my academic education. He inspired me to pursue postgraduate studies and supported me obtaining scholarships from ESRC and University of Liverpool Management School. Moreover, his input in helping me to strengthen the analytical framework of this thesis, in reading multiple drafts meticulously from the very beginning till the very end of the entire writing process, is greatly appreciated. I would also like to thank Christian Bach and Olga Gorelkina, whom supervised me in my PhD final year. Their support and guidance as well as their feedback has been greatly appreciated.

Secondly, I am thankful to the members of the Microeconomics Research Group, Health Economics Research Group at the University of Liverpool Management School and the wider Economics Department. In particular, Alan Haycox provided valuable comments on the third chapter of this thesis. In addition, I wish to thank Rob Edwards for helpful feedback on the third and fourth chapters. Valued comments have also come from audiences at the Internal Seminar of Economics Department of the University of Liverpool 2020, the NWDTC Economics Conference in Liverpool 2018 and in Lancaster 2019, the Health Economics Workshop in the University of Liverpool 2018. I also gratefully acknowledge the financial support of the Economic & Social Research Council and the University of Liverpool Management School.

Thirdly, Harvey Upton has read many parts of this thesis and provided much valuable advice. Indeed, Harvey Upton, Sana Laksa, Sirui Wu, Yusong Miao have been a constant source of support and guidance over the four years of my research, for which I am especially grateful. I am also thankful to Kaori Narita, Xueqin Wang, Yawen, Zheng, Yigit Yahya and Zining Huang who studied with me over the last year of my PhD. Finally, I would like to thank my family and other friends for their great support over the last few years, without which this thesis would not have been possible.

Chapter 1. Introduction

The continuous upward trend in healthcare expenditure around the world has emphasised the importance of using health resources efficiently. Focusing on a policy problem that originates from healthcare market failure, this dissertation examines the design and implementation of a primary care provider remuneration mechanism that could be used as a regulatory policy tool to improve cost-efficiency and quality improvement outcomes. Combining cross-disciplinary theories drawn from health economics, contract design and public administration, this thesis begins with a comprehensive review of the optimal design of primary care incentives, including both theory and evidence. The review is documented in the form of a literature survey in the first essay (Chapter 2). The next two essays (Chapters 3 and 4) are theoretical works which extend the analysis from the literature review and proposes both qualitative and quantitative recommendations for the future reform of health insurance schemes in developed countries.

This chapter provides a broad and contextualised background to the three ensuing essays that form the substance of this thesis. It begins with the discussion of healthcare expenditure trends and the major challenges faced by health authorities in developed countries in Section 1.1. Major factors that lead to health market failures and subsequent resource allocation inefficiencies are addressed in Section 1.2. This is followed by Section 1.3, which discusses the essential roles that primary care providers perform and the importance of their regulation. Section 1.4 reviews the main theories and the empirical evidence related to primary care provider payment methods in developed countries. In particular, Section 1.5 discusses one of the predominant provider payment systems, namely fee-for-service. Section 1.6 introduces the major cost control instruments that have been widely applied in these economies. Finally, Section 1.7 summarises the major results from the following three essays (Chapters 2-4) and highlights their respective contributions.

1.1 The Rising Government Share of Total Health Expenditure

Total health expenditure – usually measured in terms of total health spending as a percentage of GDP – has always been higher in developed than in developing countries (Dieleman et al., 2016; French and Kelly, 2016), with the United States (US) acting as an outlier, spending significantly more than its counterparts in the same income group (i.e. the UK, Canada, Germany, France, etc.) (Anderson et al., 2012; Chen and Goldman, 2016). Moreover, with rapid economic development, increases in national incomes, rising longevity and medical innovations, a substantial increase in health spending has been witnessed over the past decades. For instance, total health expenditure as a fraction of GDP across OECD countries has increased from 5% in 1970 to 9% in 2017, while doubling in high income economies such as Japan, the United Kingdom, France, and Germany (see Table 1). Total health spending as a percentage of GDP in the US has almost trebled over the past 50 years, rising from 6.2% in 1970 to 17.9% in 2017 (Sawyer, 2018). The real annual growth rate of health expenditure per capita across OECD countries was about 3.2% over 1990-2000, which is about 1.1% higher than per capita GDP growth (OECD, 2003 p. 69). During 2000-2009, average annual growth in health spending in real terms increased to 4.1% compared to GDP growth of only 1.5% (OECD, 2013 p. 156). However, from 2010 onwards, growth in health spending has been extremely gradual and often in line with overall economic growth. Annual health spending growth across the OECD between 2009 and 2016 was 1.4% compared to GDP growth of about 1.38% (OECD, 2017, p. 132).

In the major developing countries, the main driver of health spending growth has been the expansion of healthcare coverage as part of the move towards universal health care (UHC) over the last twenty years (OECD, 2015). In China, health expenditure as a percentage of GDP has increased from 4.6% in 2000 to 5.6% in 2014, while real health expenditure per capita increased 11% per annum during 2009-2012 compared to per capita GDP growth of around 9% (OECD, 2015 p.7). Meanwhile, Brazil, spending a similar proportion of its GDP on health as OECD countries, experienced a continued growth in healthcare spending from 2009-2011, averaging 6% per year (OECD, 2015). South Africa also saw an increase in health spending, ranging from 2% to 6% from 2009 to 2013 (OECD, 2015).

Rising government expenditure on health has led to solvency issues for governments as they struggle with inadequate resources to finance healthcare. These predicaments are compounded by a larger share of informal sector employees in the workforce (especially in developing countries), thereby making direct contributions to healthcare become extremely difficult (Durairaj and Evans, 2010). With the population's increasing demand for health services becoming inevitable in both developed and developing countries, ensuring the financial sustainability of health systems is an essential task for health authorities (Tan, 2018).

Country	% of GDP					
	1970	1980	1990	2000	2009	2016
France	5.2%	6.7%	8.0%	9.5%	10.7%	11%
Germany	5.7	8.1	8.0	9.8	11	11.3
Netherlands	N/A	6.6	7.1	7.1	10.4	10.5
UK	4.0	5.1	5.1	6.0	8.5	9.7
US	6.2	8.2	11.3	12.5	16.4	17.2
Japan	4.4	6.3	5.7	7.2	9.2	10.9
OECD	5.3	7.0	7.6	8.4	8.6	9

Table 1. Healthcare expenditure as a percentage of GDP¹

1.2 Sources of Healthcare Market Failures

The first fundamental theorem of welfare economics states that the equilibrium of an ideal market system (i.e. perfect information, no transaction costs, and perfect competition) yields a Pareto-efficient allocation of resources (Greenwald and Stiglitz, 1986). Market failures, accordingly, refer to a situation where a market violates the

¹ See Huber and Orosz (2003) and OECD (2003, 2017).

conditions necessary for a perfect market and does not organise production or goods allocation efficiently (Cunningham, 2011). Perfect market conditions are rarely met with in reality, common types of market failure include returns to scale, asymmetric information, public goods and externalities.

This thesis focuses on understanding one of the above issues that leads to healthcare market failure—information asymmetries. In particular, I discuss two specific types of problems that arise from information asymmetries, namely moral hazard and adverse selection. Since the aim of this thesis is to understand the design of a physician payment system that induces the efficient allocation of health resources,² I review related theories and evidence in the environment with the aforementioned two problems.³

Information asymmetries and different incentives across major parties in a health insurance scheme, including the payer (benevolent regulator), providers and users, constitute multiple principal-agent problems. In this triad, the regulator wants to maximize social well-being and control financial risks, thereby ensuring financial sustainability. Providers, on the other hand, have a propensity to conceal their real costs in order to maximise their claims from users who have health insurance. Consequently, providers might induce excessive healthcare supply, resulting in an overpayment of medical expenses. Meanwhile, health service users are intent on making the most use of their health plans or coverage. Since both patients and healthcare providers have an incentive to maximize their own benefits, it has been argued that they might collude to drive up service provision, potentially resulting in causing unnecessary health service

² To avoid confusion, efficiency/optimality is used to represent the “constrained” efficiency/optimality hereafter. This refers to an allocation or a payment that maximizes social benefits net of costs in the presence of market failures. In contrast, the allocation or a payment that would be chosen under complete information, is denoted afterwards by first best or Pareto-efficiency.

³ Information asymmetries also create difficulties in writing complete contracts, which results in the well-known hold-up problem. This problem refers to “a situation when one party makes a sunk, relationship-specific investment and then engages in bargaining with an economic trading partner (Hermalin and Katz, 2009). That partner may be able to appropriate some of the gains from the sunk investment, thus distorting investment incentives, either towards too little investment or toward investments that are less subject to appropriation” (Hermalin and Katz, 2009, p. 405). For instance, the UK NHS (a dominant area-based purchaser) must negotiate terms with individual health suppliers. If the NHS commits to not recouping any of the return on suppliers’ investment, then the supplier will invest efficiently. However, setting the contractual terms in a way in which the NHS never recoups any of the return on suppliers’ investment is complex, particularly if: (1) the precise details of the investment cannot be specified in advance; or (2) in the time between the investment and provision of the service being made, unverifiable random events occur that affect service delivery. In such situations, the individual supplier may need to bargain with the NHS and reduce the efficient investment.

consumption and escalated health expenditure (Culyer, 1989; Cutler and Zeckhauser, 2000).

Information asymmetry among multiple players in the private health insurance is a classic example of market failure in healthcare (Bloom et al., 2008). More recently, it has been identified that information asymmetry between providers and payers in provider financing (in both private and public health insurance schemes) has also led to market failures. For instance, Makris and Siciliani (2013) and Barham and Milliken (2015) show that physicians do not select the first best practice sizes or service quality if their production efficiencies or concerns for patients cannot be observed by payers (i.e. insurance companies or the regulator). Kantarevic and Kralj (2016) and Jack (2005), on the other hand, show healthcare providers would not provide the first best level of service quality if the purchaser (i.e. the health authority) could neither observe their characteristics (i.e. concern for patients, marginal production costs), nor monitor their efforts. In the following section, I define alternative types of information asymmetries, provide examples, and explain their links to my thesis.

1.2.1 Adverse Selection

‘Adverse selection’ – also known as ‘hidden information’ – describes the situation where some partners in a contract have information about a given characteristic that is already known to them prior to contracting with other partners (Milgrom, 1987). In the private health insurance field, for instance, a higher proportion of people with known health risks will decide to enrol in health insurance while people with relatively better health voluntarily opt out from the same health insurance (Morris et al., 2012). To avoid incurring losses, insurance companies have an incentive to raise premiums which will further discourage additional individuals from purchasing health insurance, with only the extremely ill buying insurance. Insurance companies, as a result, suffer a financial loss, eventually leading to the demise of the health insurance market (Mwachofi and Al-Assaf, 2011).

Adverse selection also makes the payer (i.e. insurance firms in private health insurance or the benevolent regulator in public health insurance) unable to induce the first-best level of patient enrolment and medical service provision. Since a given the payer does not know their provider’s specific characteristics (i.e. ability, background and

experience, concerns for patients), the single payment contract fails to align different preferences across providers while the contract menu leaves an information rent to make providers reveal their types (Baron and Myerson, 1982). In this thesis, I review the design of provider payment mechanism in the presence of different types of adverse selection problems. My analysis captures this adverse selection as unknown physicians' concerns for patients' health benefits, and shows how it prevents the regulator using a single pricing system to induce welfare maximizing patient enrolment or health service provision.

1.2.2 Moral Hazard

'Moral hazard' or a hidden action generally refers to a situation in which one or both parties in a contract are required to undertake non-verifiable action after contracting (Bowles, 2009). In healthcare, moral hazard occurs in many forms. For instance, between an insurance firm and a patient with low-deductible or co-payment, the latter will demand any pharmaceuticals and other treatments promising any benefit at all, net of the risks and side effects of the treatment and without regard to cost. Consequently, patients will over-consume treatments but the insurer will bear the extra cost incurred.

Another example of moral hazard is when neither the regulator nor the patient can fully evaluate the action of the healthcare provider. As a result, the provider may not act in the interests of the patient. For instance, the provider may degrade quality, offer unnecessary treatments or skim off low risk patients. In this thesis, I review the design of a physician payment system in the presence of different types of moral hazard. To capture the moral hazard problem, I introduce a stochastic component of receiving treatment.⁴ The introduction of this stochastic element prevents the regulator from perfectly inferring patient initial health by evaluating the patient's health after treatment and the quantity of service provided. In other words, the regulator is unable to observe the provider's "action" (i.e. health benefits produced per patient) because of the existence of the stochastic factor. In presence of moral hazard, I show that the

⁴ This stochastic component can be understood as the exogenous factors which would affect patient health benefits after receiving treatment. These factors include people that the patient touches, the food he eats, etc.

conventional fee-for-service system fails to induce the Pareto-efficient provision of health services.

In short, the unique features and information issues associated with healthcare markets makes it impossible to rely on market mechanisms alone to enhance the efficiency of allocating healthcare resources, hence, a diversity of cost control policies in the literature have been employed to address aforementioned sources of market failure (Bali and Ramesh, 2017; Leonard et al., 2013). For instance, Neudeck (1991) discusses using quantity rationing to reduce the oversupply of doctors and health service whereas Fan et al. (1998) and Benstetter and Wambach (2006) compare the effect of using a quota or a budget constraint to contain health expenditures and improve social well-being. Similarly, this thesis discusses the impact of introducing three cost-control instruments, namely quantity rationing, revenue cap and a capitation on physicians' behaviour. Furthermore, the thesis evaluates these cost control policies by comparing their total health benefits, total costs, and the social well-being generated.

1.3 Supply-side Incentives for Primary Healthcare

In the health sector, physicians as health service suppliers play a major role in directing health resources (i.e. ordering tests, prescribing therapies and procedures, and deciding hospital admissions). Moreover, they negotiate with insurers and regulators for their reimbursement (Ghali, 2016). The predominant role of healthcare providers suggests that provider-targeted regulation could be effective in containing healthcare spending. In practice, physicians have a lot of control over the type and quantity of medical services consumed, especially in an emergency setting. Doctors are also better at assessing the risks and benefits of various medical procedures. Furthermore, they are better able to bear financial risks, as they are often the dominant decision maker for certain types of care (i.e. dental or optometry services) and can pool risks across individuals. Finally, by shifting responsibilities and incentives to doctors, containing healthcare expenditure need not necessarily degrade patients' insurance coverage (Léger, 2008).

Given the increasingly important roles that provider targeted policies play, understanding how healthcare providers respond to these policies is essential, especially in the primary care sector. Starfield (1991) in her landmark book *Primary care*:

Balancing health needs, services and technology, estimated that around 75-85% of the general population require primary care services in a particular year. In both developed and developing countries, primary healthcare has also been demonstrated to be associated with enhanced access to healthcare services, better health outcomes, a decrease in hospitalisation and the use of emergency department visits (Shi, 2012) alongside a reduction in the costs of care provision.

The main character in primary healthcare is its provider; primary care providers (PCPs) make up a significant proportion of total medical staff (e.g. on average 30% across OECD countries) while acting as the first contact and principal point of continuing care for patients within a given healthcare system (MedlinePlus, 2017). Moreover, they address a large majority of personal healthcare needs, developing a sustained partnership with patients while coordinating other specialist cares that a patient may need (Shi, 2012).

Therefore, the reliance of healthcare performance on PCPs makes it essential to analyse the design of primary care incentives. This thesis examines the design of PCP incentives from the perspective of remuneration schemes. More specifically, I look at how traditional payment mechanisms structure PCPs' incentives and how different cost control policies can be applied in order to improve social well-being. This is achieved by characterising physician decisions in terms of the number of patients enrolled and the intensity of service provided as a function of a particular payment and a cost control policy in a model that captures the main features of healthcare market. More detail about this approach is provided in Section 1.7.

1.4 Typologies of PCP Payments

The majority of OECD countries have a partial or fully publicly-funded health insurance scheme that collects insurance premiums from individuals and uses them to reimburse healthcare providers (Colombo and Tapay, 2004). In general, PCP payment systems can be categorised as either fixed or variable and as either prospective or retrospective. As illustrated in Figure 2, a system is fixed or variable depending on the relationship between activities and payment. A payment system is considered as 'fixed' when the reimbursed amount does not vary as activities (per treatment) increase or decrease. In contrast, a system is regarded as 'variable' if variations in activities induce changes in payment (Jegers et al., 2002).

A PCP payment can also be classified as prospective or retrospective. This feature of a system concerns the relation between the provider's income and his costs for providing the service. In a retrospective payment system, the provider's costs incurred are fully (or partially in certain systems) reimbursed *ex post* (i.e. after the delivery of the treatment and the verification of costs incurred). The commonly used retrospective payment reimburses PCP practices on the basis of per diem or historical budgets (Gosden et al., 2000). As the retrospective payment pays the provider according to the volume of service provided, it promotes a tendency to overprescribe treatment, therefore resulting in cost escalation and welfare loss.

On the contrary in a prospective payment system, the reimbursement does not link to the individual costs of the provider (Jegers et al., 2002). Accordingly, the provider's payment rates or budgets are determined *ex ante* (i.e. before the delivery of the treatment). The frequently used prospective payment mechanisms include capitation and case-based (or diagnostic-related group) payment. Prospective payment, in general, is believed to be better at reducing costs and improving the efficiency of healthcare delivery. However, research shows that providers under this system have incentives to stint on quality (Porter and Kaplan, 2016) and avoid treating excessively costly patients (Chalkley and Malcomson, 1998; Ma, 1994).

In short, the fixed/variable dimension describes the presence/absence of a link between the payment for the provider and his activities. On the other hand, the retrospective/prospective dimension refers to the presence/absence of a link between the reimbursement for the provider and his costs. While activities and costs are related (the amount and the type of activities determines the provider's cost), they are not the same.

In retrospective systems, providers' real costs and extra production are fully (or partially) reimbursed. Since expenditures cannot be forecasted *ex ante*, the sponsor bears the financial risks. Accordingly, a fully retrospective payment system is always variable. However, the reverse does not always hold, i.e. not every variable system is retrospective. The provider under a variable system earns revenues for extra production, but there is no guarantee that these financial flows are sufficient to cover the provider's real costs. A variable system can also be prospective; the well-known examples include Medicare Diagnostic Related Group (DRG) payments and fee-for-service. On the other hand, fully fixed systems are always prospective but prospective systems can be both fixed (i.e.

global budget, salary, capitation) and variable (case-based payment). Figure 2 summarizes characteristics and incentives in payment systems according to the retrospective/prospective and variable/fixed dimensions.

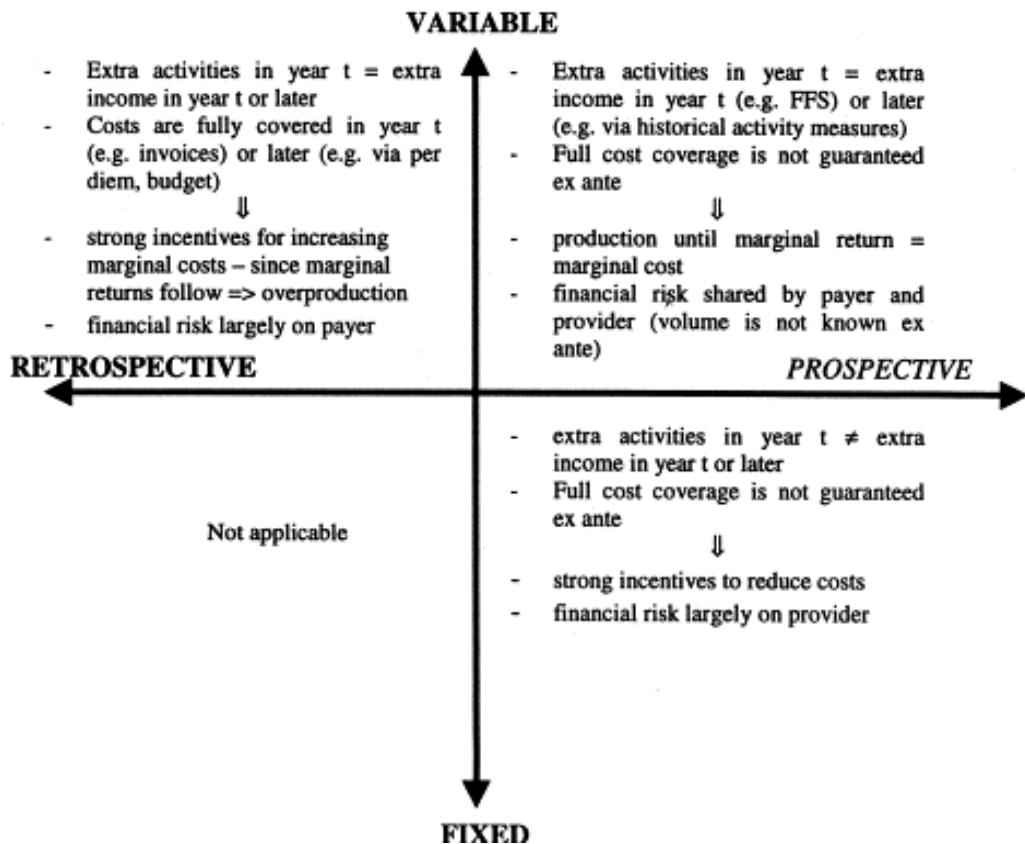


Figure 2. Characteristics and Incentives in Provider Payment Systems⁵

To conclude, the design of a provider payment system has important implications in terms of shaping the healthcare suppliers' incentives and behaviours. Besides understanding incentives created by different payment system, performance measurement and activity monitoring becomes increasingly important (Tan, 2018). The extent to which market and institutional management are mediating the effects of provider payment systems, is an important question to be answered in the governance of health insurance schemes.

⁵ See Jegers et al. (2002), p.263.

1.5 Fee-for-Service Payment Systems

This thesis focuses on analysing incentives embedded in a fee-for-service payment system under a publicly financed HI scheme. Under this remuneration, the HI scheme pays the provider a predetermined rate for each service they render such as an office visit, test, or procedure. The most recent review of the impact of different payment mechanisms shows that FFS remains the predominant PCP payment mechanism in the US⁶ and many other OECD countries (Porter and Kaplan, 2014). According to OECD (2016), 25 out of 34 OECD countries use FFS as the main method of primary care physician payment. In the US, Denmark and France, in particular, over two thirds of GPs' income stems from FFS in 2012.

FFS payment system has been used widely as it directly matches the interests of both providers and patients (Ginsburg, 2015); patients under such a system are likely to receive the level of treatment that maximizes their health benefits and are rarely excluded on the basis of complex treatment requests. Moreover, patients have rapid access to non-emergency appointments and can see any specialist whenever they like (OECD, 2016, p. 42). On the other hand, physicians' incomes increase with their clinical activity and decrease with the time spent on patients. As a result, FFS embeds an incentive for physicians to treat patients effectively and provide them with the required level of treatment.

Despite these advantages, the literature on provider payment has argued that the FFS system induces treatment overprovision and drives up healthcare expenditure (Léger, 2008). The increased costs will finally be passed on to patients through increased premiums and lower social welfare. “The phenomenon that doctors use their ‘discretionary power’ to engage in demand-shifting or inducement activities such that their recommended care differs from that which an informed patient would deem appropriate” is known as supplier-induced-demand (Bickerdyke et al., 2002).

⁶ Rama (2017) has reported that an average of 69 percent of physician practice's revenue in the US coming from fee-for-service in 2012 and an average of 71.6 percent in 2014. In 2015 this share reduced to 62 percent (Nehk, 2018).

1.6 Major Cost Control Instruments

In order to address the multiple causes that have led to aforementioned health market failures as well as to curb the ensuing excessive increase in health expenditure, a diversity of cost-containing instruments have been introduced by many developed countries over the last decades. Of all these different methods, this section focuses on three most commonly used instruments; quantity rationing, budget caps and capitation.

1.6.1 Quantity Rationing

Healthcare regulators attempt to contain costs or to reduce patients' reliance on and overuse of certain types of medical resources. To do these, regulators have introduced limitations on the volume of service supplied by healthcare providers. The restrictions can be direct, in the form of rules or regulations, or indirect, such as practice guidelines that are not mandatory (Stabile et al., 2013). In 2011, family doctors in the UK were asked to ration the number of patients they sent for life-saving cancer scans; they were required to slash the number of patients referred to hospital for tests including ultrasounds, MRIs and CT scans (Borland, 2011). Moreover, GPs' prescribing of certain drugs, including over the counter medications, has been restricted by local healthcare commissioners in order to save on costs (Iacobucci, 2017). For instance, patients' usual 56 days' supply of controlled drugs (i.e. heroin, morphine, remifentanil, pethidine, secobarbital, glutethimide, amphetamine, and cocaine) has been reduced to 28 with the maximum quantity be limited to 30 days from 2005 onwards (Knott, 2019). In 2017, NHS England also consulted nationally on plans to restrict GPs' prescribing of 3,200 products that are available without prescription, such as paracetamol, antifungal treatments, and eczema creams (Iacobucci, 2017, p.1).

In the US, the health authorities have used quantity rationing to control the length of initial prescriptions of opioid pain medication – typically to fewer than seven days – which has been passed in states such as New York and Massachusetts in an attempt to reduce opioid abuse (Scully et al., 2018). Germany introduced restrictions on the volume of physicians' prescription practices. For instance, a limit on the maximum number of antibiotic drugs per patient that can be prescribed was introduced in order to constraint overuse (Zweigner et al., 2018). Moreover, France has encouraged the early retirement

of self-employed physicians, and tightened the total volume of pharmaceutical prescriptions (Stabile et al., 2013).

1.6.2 Revenue Restriction

Revenue restrictions aim to create an upper limit on third-party player spending, typically at the level of the health sector as a whole or for specific service areas. One of the commonly used revenue restrictions is the expenditure cap, which refers to a fixed amount of money containing services delivered during the year (Glaser, 1993). If the organization/sector or individual has reached the limit toward the end of the year, it must either deny some additional service or reduce price for every service (Glaser, 1993). Caps imply strong control and are usually used by government authorities (e.g. a single centralized payer or a few highly coordinated payers). In this subsection, we look at the expenditure caps which have been imposed on an individual level.

In the UK, the National Health Service (NHS) has imposed a strict annual expenditure cap on physicians' spending. For instance, GPs are paid on a fixed budget per patient per annum (£156). As a result, total payment to GPs is predictable and stays within an annual budget except for their drug prescriptions. Individual GP practices have also been provided with a fixed prescribing budget from the Clinical Commissioning Group (British Medical Association, 2018).⁷

Germany have also experimented with individual budgets, such as personal spending accounts, for long-term care. These budgets enable people to choose or purchase services to meet their care needs (Stabile et al., 2013). France introduced a national cap on statutory health insurance expenditure in 1997, which capped healthcare spending for six sectors (Stabile et al., 2013). In the US, a physician compensation cap has been introduced. Typically, the cap is set at the 90th percentile of national pay based on surveys, such as those from the Medical Group Management Association (MGMA), the American Medical Group Association, national consulting firms which specialize in healthcare compensation issues, and others (Medical Economics, 2015).

⁷ CCGs are assured by NHS England, which retains responsibility for commissioning primary care services such as GP and dental services, as well as some specialised hospital services. Many GP services are now co-commissioned with CCGs (National Health Service, 2016).

1.6.3 Capitation

Capitation refers to a payment system which reimburses providers with a per patient lump sum payment under their supervision during a certain period (Jegers et al., 2002).⁸ This mechanism has been widely used for the reimbursement of primary care providers if patients are to be enrolled. Capitation takes two major forms: list capitation and geographic capitation. The former payment ties the aggregate income of the provider to the number of patients enrolled under his/her care while the latter ties reimbursement to the population living within a particular geographical area (Andoh-Adjei et al., 2016).

Capitation incentivises profit-driven providers to reduce costs as the latter would not be rewarded additionally when a patient seeks care several times during a given period. Moreover, capitation motivates providers to pay attention to health literacy to keep their enrollees healthy in order to save costs (Andoh-Adjei et al., 2016). Some evidence suggests that blending capitation and fee-for-service payment is effective in eliminating the supplier-induced demand associated with fee-for-service (Rosen, 1989; Rudmik et al., 2014). By using capitation, the payers of health care services also have the benefit of knowing their budgets in advance. However, the capitation rate has to be risk-adjusted, at least by age and sex, as providers have an incentive to provide less care for perceived risk groups on their list (Park et al., 2007).

Capitation has been widely used in developed countries but with variations. In the UK, GP practices are paid by a mixed system with the weighted capitation the largest element (52%), a significant pay-for-performance (14%) and some fee-for-service (Marshall et al., 2014). The capitation rate in this mixed payment is calculated by applying a full risk-adjustment formula (Andoh-Adjei et al., 2016). In Spain, GPs receive a fixed salary and a capitation payment. The capitation rate depends either on the age of patients enrol or the nature of the population in the service area (e.g. the share of population over 65s).

In countries that mainly pay primary or ambulatory care providers on a fee-for-service basis (the US, Germany, Canada, France), there have also been small shifts toward capitation payments (Stabile et al., 2013). In some other European countries such as Finland, Norway and Italy, primary care providers who were initially paid on the basis

⁸ In contrast to the expenditure cap imposed on an individual level, capitation does not restrict total spending on that given patient. These two remunerations are the same only if the physician is paid by capitation alone.

of fee-for-service, have introduced a capitation in order to contain health expenditure and to reduce hospitals referrals (Park et al., 2007). In the US, capitation payments have been frequently used in both outpatient and inpatient care within the framework of Health Maintenance Organizations (HMOs) or managed care plans (Carrin, 2012). In Canada Ontario, the ‘Primary Care Reform’ launched in late 1990s introduced a menu of primary care physician payments that has blended FFS, capitation and salary elements (Laberge et al., 2016).

To summarize, all three cost-containing instruments have been widely used to align physicians’ incentives and control fast growing health spending. However, given the heterogeneous and complex nature of the healthcare market, comparing these instruments is difficult. In order to evaluate these methods, this thesis takes a representative form of publicly funded HI scheme from high-income countries and assumes that all physicians are paid on a FFS basis. Based on this setup, I compare the total welfare generated by introducing a simplified version of these instruments respectively. Specifically, I compare introducing per patient quantity rationing, per patient expenditure cap, and capitation.

1.7 Thesis Outline

This thesis investigates the design of an optimal primary care payment mechanism in the presence of moral hazard, adverse selection and altruistic providers. Specifically, I assume that physicians not only care about their financial returns, but also have concerns regarding their patients’ health benefits. This assumption has been used by Ellis and McGuire (1986), Ma (1998), Jack (2005), Choné and Ma (2011), Makris and Siciliani (2013), and Barham and Milliken (2015). I begin the analysis by providing a comprehensive review of the relevant literature. Next, in the presence of moral hazard and adverse selection, I investigate the design of an efficient health insurance scheme (HI) when physicians can select the quantity of medical service given one of the aforementioned cost control policies. I then extend the analysis to the design of efficient HI when physicians can also select patient numbers. Overall, this thesis enhances our understanding of how different types of physicians respond to changes in their financial remuneration, the exogenous factors (i.e. population ageing, technology improvement) and cost control policies. Moreover, it outlines the best adjustment plan of a welfare-

maximizing health insurance scheme recalling current trends such as population ageing and technology advances. Finally, it provides a framework that can evaluate the performance of different cost control policies.

Chapter 2 presents a comprehensive review of both theories and evidence relating to the design of optimal PCP payment schemes. The chapter aims to achieve four objectives. First, it provides an in-depth analysis of the similarities and differences between different moral hazard/adverse selection setups and their implications in relation to the design of PCP remuneration. I show that elements of an efficient PCP contract vary according to the models used, but the form of a contract (e.g. a menu or a single payment) remains unchanged. Second, since my formalisation of moral hazard and adverse selection shares similarities with these models, the literature review also establishes the context in which the analysis in the subsequent analysis is conducted. Third, it provides some empirical support for the theoretical results derived from my research. Finally, this literature survey points out some promising avenues for future research.

Chapter 2 begins by reviewing the physician altruism model while highlighting the setup that will be used in the subsequent chapters. Next, an extensive review of the design of physician payment in the presence of moral hazard and/or adverse selection and key results is presented. I show that in an environment with moral hazard, the optimal payment is a single mixed system, and in an environment with adverse selection the optimal payment is often a contract menu. However, as the provision of a contract menu across physicians has been rarely supported in evidence, potential reasons that lead to this result are presented. This survey then considers the alternative, a blended single payment system and cost control policies. In particular, I review the impact of imposing quantity rationing, revenue cap or capitation on physicians paid on a fee-for-service system. Finally, empirical evidence relating to the aforementioned topics is provided.

In Chapter 3, I present a stylised model in which altruistic physicians choose the quantity of medical services while a benevolent HI scheme select the fee-for-service price and insurance premium to maximize social well-being. This model attempts to capture the main difficulties in the design of an optimal payment scheme arising from the informational structure which characterises the primary healthcare environment. In this context, I derive the conditions which would characterize the constrained optimal HI scheme. Since the model is too demanding for a purely analytical inquiry (there are too

many variables in a non-linear system), the study proceeds by employing a numerical simulation. This simulation enables solving the constrained optimal price and insurance premium as a function of the system's exogenous parameters (i.e. the fraction of less healthy individuals, the factor of technology improvement.). As variations with respect to these parameters can be interpreted as the evolution of recent trends, I recommend the optimal adjustment in the HI scheme in accordance with these trends. Since the aforementioned trends drive health expenditure and further exacerbate resource allocation, I evaluate three potential cost-control methods: (1) a per patient quantity restriction; (2) a per patient expenditure cap; and (3) a capitation. I show that all of these instruments can be used to enhance efficiency while mixing fee-for-service with a capitation that does not fully cover per patient a fixed cost which dominates other instruments.

In Chapter 4, I adjust the model presented in Chapter 3 by allowing physicians to choose the total number of patients. This feature is particularly prevalent in a healthcare system experiencing shortages of PCPs and difficulties recruiting new medical staff. I follow the approach developed in Chapter 3, that is, I characterize physicians' decision regarding their practice sizes and service quantity as a function of price, capitation and exogenous parameters. Next, I derive the efficient but constrained pricing system given that the entire population has access. After that, I analyse the impact of introducing different cost control policies by employing a numerical simulation. Finally, I derive policy implication for the optimal design of PCP remuneration scheme and discuss intuitions.

Chapter 5 summarizes key results of the thesis and discusses the contributions of theoretical models in the contexts of the wider literature. Moreover, it presents some limitations associated with my setups and outlines some possible alternative modelling choices. Finally, it highlights some of the ways in which this analysis could be extended.

Chapter 2. Literature Review

2.1. Introduction

In the field of healthcare, ethical considerations are omnipresent (Batifourlier and Da Silva, 2014). Every medical-related profession has a system of morality supported by a “code of deontology”. For example, the American Medical Association’s Code of Medical Ethics stipulates the ethical attitude to be followed with a “Council of the Order” to enforce it (Batifourlier and Da Silva, 2014, p.241), while the “Hippocratic Oath” commits physicians to provide ‘appropriate’ amounts of care (Woodward and Warren-Boulton, 1984). Medical ethical values such as beneficence, non-maleficence, justice and respect for autonomy induce special responsibilities for professionals (Gillon, 1994). Moreover, research reflects a strong belief that physicians perform ethically in their daily work (Glannon and Ross 2002).

Given that medical ethics is an essential coordinating factor, it has to be included in standard health economics. To introduce professional ethics within the frontiers demarcated by economic rationality, economists have confined the professional ethics of doctors to economic calculus as this allows one to explain ethics using tools of mainstream economic theory. The ethical attitude of a doctor is thus formalised in terms of “medical altruism”. This concept is designed to consider the values of doctors by assuming that the paradigm of *homo economicus* is a sufficient starting point (although it may have to be modified) (Batifourlier and Da Silva, 2014). These values are incorporated through the concept of utility. The individual is endowed with classic individual preferences to which are added social preferences. Drawing on the traditional definition of altruism in mainstream economics which is advocated by Harsanyi (1955), medical altruism is defined as the inclusion of the patient’s welfare in the doctor’s utility function. This representation of altruism consists of an internalisation of the patient’s utility function (or a proxy such as health status or income) into that of the doctor. An altruistic doctor remains a utility maximiser, whose utility function includes the well-being of the patient (Batifourlier and Da Silva, 2014, p. 242).

The aforementioned framework of physician altruism has been frequently applied in understanding physician behaviour and decision making (e.g. Chalkley and Malcomson,

1998b; Choné and Ma, 2011; Ellis and McGuire, 1986; Farley, 1986; Jack, 2005; Woodward and Warren-Boulton, 1984), which has a central and noble tradition in health economics (Galizzi et al., 2015). One of the significant topics here is the design of remuneration for primary care providers. In the healthcare market, physician payment systems are often designed by health authorities to influence doctors' incentives for allocating health resources. Accordingly, it is important to investigate the impact of each payment on the behaviour of physicians and the optimal combination of physician payment mechanisms. The purpose of this chapter is to offer a survey of a particular subset of this literature. Specifically, I review optimal incentive contracting for altruistic healthcare providers in an environment with moral hazard or adverse selection. These information asymmetries are popular in healthcare markets and drive different types of health resource allocation inefficiencies (McGuire, 2000).

The first type of information difficulty relates to moral hazard. In the contract theory literature, there is a wealth of theories that economic agents take non-verifiable actions after contracting (Laffont and Martimort, 2002b; Nyman, 1999; Rowell and Connally, 2012; Smelser and Baltes, 2001). In the current context (i.e. a publicly funded health system), the health care provider is the agent while the principal is some body of the state (e.g. the Centres for Medicare and Medicaid Service in United States, or Ministries of Health in Canada and the UK). Moral hazard, therefore, refers to a situation whereby the actions of physicians are difficult to write down in a complete contract. For instance, Ma (1994) has pointed out that neither the efforts of a physician to keep costs down nor the medical service quality they provide can be easily monitored. Moreover, patients' ultimate health can be affected by many random factors such as hygiene conditions, food consumed after their treatments etc. As a result, a physician's performance measuring by total health benefits generated for his patients is also non-contractible. In the presence of moral hazard, standard contract theory suggests that the regulator faces many different types of trade-offs that prevents allocating resources efficiently (Demougin and Fluet, 2001). These trade-offs include rent vs. efficiency (Demougin and Fluet, 1998, 2001), risk premium vs. efficiency (Holmström, 1979), and multi-tasking (Baker et al., 1994; Baker, 2002). For the sake of parsimony, I will only discuss rent vs. efficiency trade-off in the following chapters.

The second type of information asymmetry is adverse selection, referring to a situation where some partners in a contract have information about a given characteristic that is

already known to them prior to contracting with other partners (Milgrom, 1987). In the current context, it describes a situation where some specific characteristics (i.e. ability, background and experience) of a healthcare provider are unknown. Chalkley and Malcomson (2002) claim that the regulator could not observe a patient's severity of illness and the associated costs that might be incurred. Choné and Ma (2011), on the other hand, assume the physician has private information about his degree of concern for his patients. Allard et al. (2014) also state that physicians have unknown and heterogeneous abilities in terms of curing patients. In presence of adverse selection (hidden information), conventional studies suggest it is natural to propose a menu of physician remuneration schemes. However, using a menu of physician payment systems may distort some physicians' quantities of service provided while leaving other physicians informational rents (Hart and Holmstrm, 1986; Laffont and Martimort, 2002a; Laffont and Tirole, 1986, 1993). Furthermore, the utilization of a menu of remuneration schemes may incentivise the health care authority to renegotiate the contract with respective types of physicians, as the latter would reveal their private information (i.e. their types) when selecting their contracts before providing health services (Demougin, 1989; Maskin and Moore, 1999).

Finally, I review information asymmetry which is comprised of both moral hazard and adverse selection. Since this information asymmetry is comprised of the previous two, the regulator may encounter all problems discussed above in the design of physician payment. The number of studies analysed incentive contracting in the presence of this information asymmetry has grown significantly over the past 40 years (Maréchal and Thomas, 2018). In relation to health economics, while the number of publications investigating this issue is small, it is still increasing (Barham and Milliken, 2015). For instance, Jack (2005) assumes that both physician altruism and their cost-reduction efforts are unknown whereas Kantarevic and Kralj (2016) regard physicians' quality-enhancing efforts and marginal costs of production as their private information. Moreover, Wu et al. (2018) assume physicians have unknown and heterogeneous abilities of curing patients while exerting non-contractible effort per patient (i.e. the proxy can be time spent or costly mental and manual work).

It has been widely recognised that information asymmetries among regulators, providers and patients contribute to healthcare market failure and healthcare provision inefficiency (Tan, 2018). Furthermore, altruistic preferences of the physician, the information

advantages of the provider and the incentive offered by traditional payment systems have resulted in supplier-induced demand (Léger, 2008). Hence, this survey further reviews how cost-control instruments affect physicians' behaviours and how the design of provider payment should be adjusted.

In short, this chapter surveys how the design of primary care providers' (i.e. physicians or abstract GP practices) incentive contracting has been affected by altruistic preferences, information asymmetries, and cost control policies. Besides discussing the theoretical findings, a wide variety of evidences show physicians care about their patients' benefits (e.g. Hennig-Schmidt et al., 2011; Hennig-Schmidt and Wiesen, 2014). For instance, empirical studies, including surveys, field studies and laboratory experiments have shown that physicians' trade off patients' health benefits and their financial income in the provision of medical service (e.g. Godager and Wiesen, 2011; Galizzi et al., 2015). These studies provide a basis for introducing altruistic preference in the analytical framework. Moreover, empirical observations of physicians' response to different payment system and cost control policies are presented. The main purpose of reviewing this empirical literature is to highlight areas of similarities and departures among real world observations, assumptions and theoretical predictions.

The chapter is organised as two distinctive section: Theory (Section 2.2) and Evidence (Section 2.3). The theoretical literature presented in Section 2.2 comprise three main streams; firstly, physician altruism is discussed in Section 2.2.1. Specifically, Section 2.2.1.1 defines physician altruism and discusses the similarities and differences between altruism another related concept intrinsic motivation. Section 2.2.1.2 then presents the model of altruistic preferences and intrinsic motivation.

Secondly, the optimal design of incentive contracting for physicians with altruistic preferences is discussed in Section 2.2.2. I begin with reviewing the physician payment design under complete information and the implications of altruism variations in Section 2.2.2.1. Next, I extend the review to an environment with moral hazard (Section 2.2.2.2) and adverse selection (Section 2.2.2.3) respectively. In particular, I outline Barham and Milliken's (2015) model and highlight some potential extensions. In Section 2.2.2.4, I discussed studies investigating physician incentive contracting when moral hazard and adverse selection co-exist.

Thirdly, Section 2.2.3 examines the impact of imposing different types of cost-control instruments on a fee-for-service system. To do this, I present different types of cost control policies and report their respective impacts on improving patients' well-being as well as reducing total health expenditure. Moreover, I discuss the impact of introducing altruistic preferences on both physician behaviour and the implementation of cost control policies. I conclude this section by relating different setups used in this chapter and highlighting their major similarities and differences. Finally, Section 2.2.4 summarize the main results, highlight similarities and differences between setups and addresses gaps.

Section 2.3 reviews evidence relating to the design of the optimal physician payment. This strand of the literature is markedly less developed and has, in some cases, failed to maintain pace with the increasing theoretical interest in this research area. For instance, while the optimal forms of physician remuneration have been derived in many theoretical analyses, these results are rarely supported by real life observations. In Section 2.3.1, I provide different types of empirical analysis supporting the existence of medical altruism, including surveys (2.3.1.1), prescription records (2.3.1.2), field works (2.3.1.3) and laboratorial experiments (2.3.1.4). The next section, 2.3.2, compares observations of physicians' decisions on medical service supply under fee-for-service and capitation. This is then followed by Section 2.3.3, which examines the evidence relating to the impact of imposing different cost control policies on a fee-for-service system. These policies include quantity rationing (Section 2.3.3.1) and revenue restriction (Section 2.3.3.2)

In Section 2.4, I complete this review by linking the respective topics reviewed together, discussing the key results, highlighting major inconsistencies and offering some thoughts on future directions. First, this survey addresses differences in the design of the optimal agent incentive contracting between standard contract theory and health economics. Second, while the literature comparing physician behaviour under commonly used payment systems (e.g. a lump-sum transfer, fee-for-service, capitation) is well-established (McGuire, 2000; Léger, 2008), the commonly used single payment system fails to induce efficient healthcare provision in the presence of heterogeneous types of physicians and patients. This literature considers the main characteristics of healthcare markets, while the impact of using alternative approaches to modelling moral hazard or adverse selection have not been discussed. Moreover, the possibility of

imposing cost-control instruments and their impact on the design of an optimal form of payment under information asymmetry has been ignored. Third, there is a relatively small literature exploring the impact of introducing cost instruments on physicians' behaviour under the fee-for-service payment system. However, studies on this topic often ignore some of the main features observed in the healthcare market; regulators, physicians and patients do not share information. In addition, healthcare providers are partially motivated by their patients' well-being while restricted by non-negative profits. The aforementioned inconsistencies underline the need to construct a more sophisticated analytical framework to understand the impact of using respective physician cost control policies and therefore the optimal design of physician remuneration. This will be a central theme of this thesis and the motivation for the theoretical models outlined in Chapters 3 and 4. Overall, this literature review examines studies which discuss the design of optimal payment for primary care providers in the presence of moral hazard, adverse selection and medical altruism.

While the majority of the existing literature is covered, the survey is not intended to be exhaustive as some papers that partially cover the aforementioned topics have been excluded. First, there are a large number of studies which analyse conceptualizations and empirical findings related to the notion of provider altruism. However, my review only covers theories and evidence in the context of health economics while research on other contexts is generally excluded. Second, while there is a rich and sophisticated literature discussing incentive contracting in the presence of different types of information asymmetry, this survey only includes healthcare studies which investigate payment design in a simplified environment where moral hazard, adverse selection and altruistic providers are present. Finally, there is a well-established literature studying the effect of introducing global budgets on physician behaviour in a fee-for-service system. In this survey I focus on reviewing the literature which uses setups to compare quantity restriction and budget. In particular, I only include evidence focussed on examining the effects of imposing quantity or revenue restriction at an individual level.

2.2 Theory

2.2.1 Altruism in Health Economics

2.2.1.1 Physician Altruism and Intrinsic Motivation

As discussed in the introduction of this chapter, altruism has been believed to be an integral factor in medical professionalism (e.g. moral obligation) and has been applied operationally to the specific case of healthcare providers (Harris, 2018). Since Arrow (1963) which stressed that uncertainty and altruistic behaviour are two important aspects in the organization of healthcare (Martinsson and Persson, 2019), it has become popular to include patients' welfare as an argument in a doctor's utility function on the basis that the physician has an altruistic concern for the patient's health and well-being (Farley, 1986; Woodward and Warren-Boulton, 1984). In a more recent review of the conceptualizations and empirical findings related to altruistic preferences, physician altruism is defined as "the weight in the doctor's utility function attached to the benefit of the patient's health, besides the self-interested monetary considerations" (Galizzi et al., 2015, p. 3). This definition has been widely applied in the literature studying physician behaviour and physician incentive contracting while the degree of physician altruism has been measured in recent empirical studies (Brosig-Koch et al., 2016; Godager and Wiesen, 2013; Hennig-Schmidt et al., 2011). In addition, the important roles that altruism plays make it an essential element in my subsequent analytical framework.

In the parallel stream of the health economics literature, physicians are assumed to be intrinsically motivated, in which case they perform an activity for no purpose other than the activity itself (Galizzi et al., 2015). To illustrate this, a physician may be interested in providing a high-quality health service due to professional ambition and not because of her concern for patient welfare. Hence, Galizzi et al. (2015) concluded in their review that such intrinsic motivation is conceptually distinct but commonly associated with altruism (p. 17). Studies in this strand of the literature focus mainly on measuring the extent of intrinsic motivation or examining whether the introduction of financial incentives leads to the crowding out of intrinsic motivation (Galizzi et al., 2015; Sicsic et al., 2012). While discussing this crowding out effect is not the focus of my literature survey, physicians' intrinsic motivation and altruism respectively has been used interchangeably in much of the literature (e.g. Ellis and McGuire, 1986; Ma, 1997; Chalkey and Malcomson, 1998; Jack, 2005). In the following section, I highlight the similarities and differences of modelling these concepts while selecting the modelling approach to follow in the Chapters 3 and 4.

2.2.1.2 Models of physician altruism and intrinsic motivation

Ellis and McGuire (1986) were among the first to formalize the idea of physician altruism. They analysed the service supply decision of a doctor acting as an agent for both a patient and a hospital. The physician internalizes the external effect of his decision based on the hospital's profit and the patient's health. He cares about the profit of the hospital as his remuneration is positively related to the hospital's income. Formally, Ellis and McGuire use $U(B(q), \pi(q))$ as the utility function of the physician, where $\pi(q)$ represents the profit of the hospital and $B(q)$ his patient's benefit. The variable q measures the service provided to the patient during a hospital episode. In other studies, this variable also refers to the quality of health service provided (see, Ma, 1998; Kaarboe and Siciliani, 2011; Barham and Milliken, 2015). The authors also assume patients are fully insured, indicating that they passively accept the quantity of medical service provided by the doctor. Physician altruism is then defined as the doctor's marginal rate of substitution of a patient's health benefit for the hospital's profit:

$$MRS_{\pi,B} = -\frac{\partial U / \partial B}{\partial U / \partial \pi} \quad (1)$$

As it measures the amount of profit the physician is willing to give up for a marginal increase in the consumer's benefit, keeping the same level of utility.⁹

Later studies studying the impact of altruism in healthcare have used this model but also deploy a linear and separable form of the physician's utility function (e.g. Barham and Milliken, 2015; Godager and Wiesen, 2013; Makris and Siciliani, 2013).

$$U = \pi(q) + \alpha B(q) \quad (2)$$

where $(q) = R(q) - C(q)$. R and C refer to the provider's revenue and costs while the parameter $\alpha \geq 0$ measures the degree of physician altruism towards his patient's well-being. This function is a particular form of utility function used by Ellis and McGuire (1986). Alternatively, Chalkley and Malcomson (1998a) as well as Jack (2005) apply the physician's utility function:

$$u = \pi + \varphi(q) \quad (3)$$

⁹ Clearly, assuming the hospital's profit and the health benefits provided to the patient are goods for the physician is natural (Ellis and McGuire, 1986).

where q is either health service quality or quantity and $\varphi(q)$ the patient's benefit as well as other elements. For example, it can be also interpreted as the level of a physician's intrinsic motivation as the term $\varphi(q)$ can be interpreted as the physician's private value regarding the high-quality production of healthcare or any other utility obtained by providing medical services such as warm glow benefits derived from treating patients. As a result, the utility function (3) is more general than the linear utility function (2) with the altruistic component.

Many observers have also noted that physicians have an interest improving their patients' health (Godager and Wiesen, 2013; Galizzi et al., 2015). Hence, this thesis has followed the aforementioned literature, assuming part of physicians' well-being is derived from improving their patients' health benefits. Accordingly, the equation (2) is used as the physician's utility function in my research while the physician altruism α measures the marginal rate of substitution between the physician's profit and his patients' welfare.

2.2.2 The Optimal Payment for Altruistic Physicians

2.2.2.1 Optimal Physician Payment with Complete Information

The benchmark physician incentive contracting model (Ellis and McGuire, 1986) features a doctor who is hired as an agent for a hospital to decide the quantity of service q to be provided in return for his financial income R and the altruistic benefits $\alpha B(q)$ from serving patients. The doctor's preferences in terms of his payment and altruistic benefits can therefore be expressed as a separable utility function $U(q) = \alpha B(q) + \pi(q)$ and $\pi(q) = R(q) - cq$, where α and c represents the doctor's degree of altruism and the marginal costs of providing medical service respectively. For the patient, he is fully insured, passive and his total benefits from health treatment $B(q)$ are assumed to be inversely U-shaped.¹⁰ The social optimal quantity q^* is defined implicitly by the equation:

$$B'(q) - c = 0 \quad (4)$$

¹⁰ Patients' total health benefits would fall after being provided with a certain level of service, both because of "the time-price of receiving treatment and because of the risk of infection and other iatrogenic illness associated with a continued hospital stay" (Ellis and McGuire, 1986 p.132).

This means that the marginal benefit of providing medical service in the economy should be equal to the marginal cost. Since the quantity of medical service provided by the doctor q_m is implicitly defined by the equation:

$$\alpha B'(q) + R'(q) - c = 0 \quad (5)$$

the cost-based payment such as full cost reimbursement (i.e. $R(q) = cq$) or fee-for-service payment (i.e. $R(q) = pq$) with $p > c$ indicates the service provided by the doctor satisfies:

$$B'(q_m) \leq 0 < c \quad (6)$$

The concavity of the health benefit function $B(q)$ therefore implies $q_m > q^*$. That is, the physician under the full cost reimbursement or fee-for-service system (with $p > c$) overproduces.

Under the fully prospective payment (i.e. $R(q) = T, R'(q) = 0$), the quantity of health service provided by the doctor q_m is implicitly defined by:

$$\alpha B'(q) - c = 0 \quad (7)$$

This equation implies that the fully prospective payment would be optimal if the physician trades off his net financial returns and the patient's benefit in the same way the social planner does (i.e. the physician becomes a "perfect agent" with the degree of altruism $\alpha = 1$). Since the physician typically places a smaller weight ($\alpha < 1$) on the health of patients than the social planner, it is natural to observe the physician under-produce ($q_m < q^*$) as $B'(q_m) = \frac{c}{\alpha} > c$. As this study assumes a one-to-one relationship between costs and medical service quantity/quality (i.e. q_m is a function of c), the physician's under-provision can therefore be fixed by paying a fraction r of incurred costs. In this case, the quantity of service provided by the physician q_M is implicitly defined by:

$$\alpha B'(q) - (1 - r)c = 0 \quad (8)$$

To get $q_M = q^*$, we require $B'(q) = c$ which implies the cost reimbursement rate should be set at $r^* = 1 - \alpha$. Clearly, the optimal cost reimbursement rate reduces with the level of physician altruism. Intuitively, the physician increases both his supply of medical service with his degree of altruism and the cost reimbursed by the regulator

respectively. Hence, the degree of altruism and α the cost-sharing rate r are perfect substitutes in terms of inducing the physician's production. Accordingly, the more altruistic or intrinsically motivated the physician is, the less financial or extrinsic reward is needed. Further, the authors extend the model and show that blending prospective and cost reimbursement payment also helps to reduce low-value admissions, patients' reclassification and providers' financial risk.

While the authors assert that the mixed reimbursement system is optimal, this conclusion relies heavily on some of the assumptions imposed above. For instance, efficiency in their model is defined as a medical service quantity that equalizes marginal social benefits and marginal costs. In other words, Ellis and McGuire implicitly assumes patients' total healthcare expenditure equals physicians' total income. However, as Laffont and Tirole (1986) pointed out, collecting premiums is costly. As a result, the current mixed system may become inefficient when the shadow costs of collecting patients' payment are introduced.

Other important assumptions are that the marginal cost is constant and total costs are contractible. Hence, there exists a one-to-one relationship between cost reimbursement and quantity. If the physician's costs are non-contractible, then the optimal form of payment has to be based on the physician's actions that can be monitored. For instance, the regulator may consider blending prospective payment with fee-for-service rather than cost-sharing when only the quantity of medical service provided can be observed. Moreover, if physician heterogeneity is also considered, offering physicians a menu of contracts or introducing additional cost-control instruments might be a better option. Finally, the inversely U-shaped health benefit function $B(q)$ implies that the optimal level of healthcare provided to a patient is given exogenous. Endogenizing the level of the patient's treatment may affect the social optimal level of healthcare required as well as the optimal contract offered to physicians. For instance, Ellis and McGuire (1990) find that the optimal payment for the physician depends on: (1) the bargaining power of the patient; (2) the patient risk preference; and (3) the degree of moral hazard. Lee (1995) also shows that the rates of a physician's and a patient's cost sharing have to be properly interrelated to induce the efficient provision of medical service. Later studies extend the model of Ellis and McGuire (1986) by relaxing some of the aforementioned assumptions in an environment where physicians' characteristics or their actions are non-contractible.

Physicians make final decisions in relation to diagnosing illness, prescribing medicines, performing surgeries and so on. As they are at least partially self-interested and therefore have an incentive to improve their monetary payoff, compensation rules play a key role in inducing physicians to provide appropriate treatments. The design of optimal compensation depends on how much the regulator knows about the providers. As discussed in Section 2.1, the main information issues between the health authority and providers include: (1) moral hazard; (2) adverse selection; and (3) moral hazard and adverse selection. In the succeeding review, I divide all studies according to these three types of information asymmetries while examining their models. In addition, I address the assumptions that may change their results. Specifically, I explain the model and summarize the main results of each individual study. This is followed by an explanation of how these studies developed from each other. Finally, I discuss some gaps that these studies have not paid attention to.

2.2.2.2 Optimal Physician Payment with Moral Hazard

In this subsection, I review incentive contracting for an altruistic provider in the presence of moral hazard only. Rogerson (1994) has explored the optimal price regulation imposed on an altruistic healthcare provider (i.e. a non-profit hospital) in an environment where the provider's service intensity cannot be observed by the regulator but can be seen by patients. In his model, the payer (a government or the private insurer) selects a price paid per patient treated in order to maximize social health benefits net of his payments. Next, the healthcare provider chooses a treatment intensity/quality (i.e. the level of resources a provider employs to treat a specific illness) to maximize patients' health benefits subject to the constraint of non-negative profit. Finally, patients select a level of demand for the provider's product,¹¹ which increases with the service quality provided by the healthcare provider. Except for non-contractible medical service quality, the payer has both complete and symmetric information for all aspects.

Rogerson shows that the regulator can use fixed price contracts¹² to achieve the efficient provision of quality/intensity as long as patients can observe quality and use this to

¹¹ Patients are assumed to be fully insured so that price does not affect demand.

¹² Fixed price contracts refer to the reimbursement that the regulator pays the healthcare provider, namely a fixed amount of money per treated individual.

choose whether or where they are to be treated. The intuition is that the regulator can observe patients' demand and there is a positive demand response from care quality. Hence, by manipulation of the price paid per patient, the regulator can induce the efficient provision of quality.

Ma (1998) compares alternative types of efforts induced by fixed-price and retrospective cost-based contracts. Specifically, he has adopted a multi-task agency approach¹³ where the regulator selects a payment policy to maximize social welfare measured by the sum of patients' benefits net of total cost of production. The altruistic provider subsequently allocates his non-contractible efforts between (1) quality enhancement and (2) cost reduction. Similar to Rogerson (1994), quality is assumed to be a single dimension, cannot be contracted upon, and drives patient demand. Moreover, the healthcare provider has to meet all patient demands and his profit is assumed to be non-negative. The only difference, however, is that total costs of treatment are assumed to be verifiable ex-post.

Based on this model, Ma (1998) shows that first best cost reduction and quality enhancing efforts can be induced by using a fixed price contract. Intuitively, the provider fully internalizes all production costs under the fixed payment contract, therefore the price can be selected to fully internalize the benefits of improving quality (Ma, 1994). In contrast, cost reimbursement provides no cost-reduction incentives but induces a constrained optimal level of quality efforts. Ma (1998) also shows that, when the provider's profit is positive, the optimal price, the margin (i.e. the payment in addition to reimbursement of the provider's costs) and the provider's profit decreases with the degree of the provider's altruism. When the provider's profit is binding at zero, the price and the margin does not change in line with the level of physician altruism. Finally, when the provider is allowed to refuse patients, the optimal payment system should blend both prospective payment and cost reimbursement.

Chalkley and Malcomson (1998b) find that the results from the previous two studies are satisfied only when quality has only a single dimension and it is efficient to treat all patients who demand treatment. The main departure from the previous model is that the healthcare provider can additionally select the number of patients to be treated, which differs from the number of patients demanding health service and the provider's

¹³ To be clear, 'multi-task' means that the physician has to enhance quality and to reduce costs.

maximum capacity. When it is inefficient or impossible to treat all patients demanding medical service,¹⁴ they show that a price-quantity schedule¹⁵ induces a provider either to attract too many patients or to provide too low quality. This is because the fixed price contract which induces efficient quality would make the number of patients demanding treatment greater than either the efficient number of patients that can be treated or the provider's maximum capacity. However, this inefficiency can be improved by introducing a payment based on a measure of patient demand, whenever such a measure is available. Further, they show that this result extends to any number of dimensions of quality when patients' relative valuations of the different dimensions of quality are the same as the regulator's (Chalkley and Malcomson, 1998a).

Chalkley and Malcomson (1998a) further explore the case in which patient demand does not reflect quality.¹⁶ They show that a prospective payment will induce an optimal cost reducing effort while resulting in a sub-optimal level of quality if the provider places less than full weight on patients' well-being. Hence, they find that adding a cost sharing component to a prospective payment system will improve efficiency. While introducing cost sharing has the positive impact of enhancing quality and the negative impact of raising costs, the former impact dominates the latter. In addition, they find that the optimal cost-sharing rate decreases in line with the degree of the provider's benevolence, as the provider's extrinsic motivation is substituted by his intrinsic concern for patients. Finally, as the provider's total production costs in Chalkley and Malcomson (1998a) depends on the number of patients treated, the service quality provided and the cost reduction effort exerted, the one-to-one relationship between costs and quality does not hold. Hence, subsidizing costs is no longer a perfect substitute for subsidizing quality, thus, the corrective subsidy inherent in Ellis and McGuire's (1986) cost sharing rule would not result in first-best efficiency.

The previous three studies show that when the provider's action(s) are unverifiable, cost reimbursement would induce the provider exert the lowest cost-reducing effort but provide a constrained optimal level of quality whereas prospective payment would induce the provider to offer the first-best cost reducing effort. Moreover, the provider

¹⁴ Hospitals are constrained by limited capacity for certain treatments or patients often do not pay full cost of treatment.

¹⁵ Quantity in their paper refers to the number of patients who received medical treatment.

¹⁶ Patients may not be aware of all aspects of the quality of service or may not know about the medical implications for their treatment (Arrow, 1963).

would improve service quality as they receive more reimbursement (i.e. get higher price per treatment or take lower cost sharing rate). The optimal contract for a partially benevolent provider is a prospective or mixed payment; the former induces first-best level of efforts when it is efficient to treat all patients demanding healthcare and patient demand reflects quality (e.g. Ma, 1998). The latter, however, becomes optimal when the quality is multi-dimensional; it is not able to treat all patients demanding healthcare (Chalkley and Malcomson, 1998b); patient demands does not reflect quality (Chalkley and Malcomson, 1998a) etc. The size of the cost reimbursement employed in the blended payment depends on the extent to which the provider values consumer benefits. In contrast to the standard contract theory literature, these studies analyse the effect of physician concerns for patients in the design of physician payment while comparing the physician's efforts under prospective payment and cost-reimbursement.

In reviewing the above studies, it is worth paying attention to the following aspects. First, the moral hazard problem of the physician could be modelled in different ways. In addition to non-contractible physician efforts, the health benefits that the physician generates for his patients are difficult to measure as a patient's health after treatment can be affected by many factors such as his hygiene conditions or food consumed. These factors could be interpreted as the random shocks associated with the patient's health, as shown in Kantarevic and Kralj (2016). The moral hazard problem can also be motivated by patients' unknown severities of illness. Since neither the regulator nor patients have sufficient knowledge about their health, whether physicians produce appropriately could not be determined, even if the quantity of service provided is verifiable.

Second, the above studies assume that the physician's costs incurred can be verified, therefore cost-sharing is one of the recommended remuneration instruments. In reality, cost-sharing contracts have been used rarely and may lead to the problem of cost padding (i.e. charging non-allowable expenditures or costs for some patients). To avoid cost padding, the regulator has to monitor the physician's costs sufficiently closely, which might be very expensive (Chalkey and Macomson, 2002). As implementing cost-sharing is problematic, the regulator often considers substituting cost-sharing by other remuneration instruments such as fee-for-service (Allard et al., 2014). Finally, and as will be discussed in the following review, the above studies do not discuss the optimal design of physicians' contract in the presence of unknown heterogeneities.

2.2.2.3 Optimal Physician Payment with Adverse Selection

More recent studies have considered adverse selection frameworks that focus on unobserved heterogeneity among physicians (i.e. ability, altruism, efficiency) or patients (i.e. severity of illness) (e.g. Chalkley and Malcomson, 2002; Choné and Ma, 2011; Liu and Ma, 2013). These studies can be divided into two categories: the first category examines the adverse selection between patient and regulator, in particular for the case that the former's characteristics are unknown to the latter. The second category, which my research is closely related to, analyses adverse selection between provider and regulator mainly for the case that the provider's characteristics could not be observed by the regulator.

Patient type is unknown

Chalkley and Malcomson (2002) analyse the optimal design of a remuneration scheme for each healthcare provider who performs non-contractible cost-reduction efforts and treats a patient with unknown health. The cost incurred by the patient is verifiable ex-post and decreases with the provider's effort as well as the patient's initial health. Based on this model, Chalkley and Malcomson show that the regulator can reduce total expected payment to healthcare providers by introducing in the fixed price payment system a proportion of the realized costs when there are substantial variations in the cost of patient treatment unknown to the regulator. This contract reduces total expenditure since low cost patients would receive less reimbursement so that the supplier has lower informational rent. However, in reality cost-sharing requires a close monitoring of the provider's production while the associated costs may dominate savings derived from the reduction of informational rent. As a result, cost sharing contracts are not frequently used in reality as the theory predicted. While Chalkley and Malcomson consider unknown heterogeneity among patients, it can be also regarded as modelling unknown heterogeneity among providers. To illustrate, the provider who is allocated a patient with high (or low) costs can be considered as the “high (low) costs” provider while the optimal contract is a menu with the cost reimbursement rate positively related to the actual costs incurred as shown in Laffont and Tirole (1993).

Choné and Ma (2011) study the effect of unknown and heterogeneous physician altruism as well as patient health benefit on the design of payment and healthcare quantity. In

their model, the physician's provision of healthcare to each individual patient is constrained by a minimum profit restriction. They show that the optimal payment depends only on the weighting of physician agency and exhibits extensive pooling, with the quantity and payment provided being insensitive to physician altruism or consumer benefit. The choice of pooling is an outcome of a tension caused by incentive compatibility between quantities and non-negative profit.¹⁷ This tension, according to the authors, is caused by the patients' unobserved heterogeneity. The optimal choice of pooling system has also been discussed by Liu and Ma (2013), Allard et al. (2014), and Burani and Palestini (2016) etc. In addition to the reason highlighted in Choné and Ma (2011), other reasons support offering a single payment system for all physicians, and will be discussed in this subsection's summary (p.21).

Liu and Ma (2013) also analyse the process of the decision-making delegation to altruistic physicians. Physicians are paid to match patients with different illness severities to different treatment protocols, specifying recovery probabilities and costs. As in Choné and Ma (2011), providers possess private information about their degree of altruism and the severity of patients' illnesses while having a non-negative restriction on their expected profit. However, they further model insurance, consumer risk aversion, treatment plans (i.e. sequences of treatments) and plans commitment. If the physician can commit to treatment plans before learning about the severity of the patient's illness, the authors show that the first best can be implemented via a single contract (optimal for the least altruistic providers) and applying to all types of altruistic physicians. Intuitively, physicians tend to select the more generous treatment plans, however, they would not choose these plans as their profit would otherwise become negative. If physicians cannot commit to treatment plans, all healthcare providers (except the least altruistic one) earn positive profit and treatment decisions are deviated from the first-best.

The above two studies analyse how a payer who operates in a competitive market should design the optimal contract for a physician who needs to treat different types of patients. Therefore, it will be interesting to see how their results would change if different assumptions are imposed. For instance, they assume that the physician requires to earn a non-negative profit from treating each individual. However, the physician in reality

¹⁷ In an incentive compatible mechanism, quantities must be non-decreasing while profits are non-increasing in the physician's concern for patients. Setting quantities for a range of degree of physician altruism implies the corresponding profits are zero.

often needs to treat many patients and is more likely to be constrained by a non-negative expected profit from treating all of his patients. Next, while these studies analyse the physician's incentive to contract in the presence of adverse selection, they ignore the issue of the physicians' unknown actions. The payer in their study could also be a centralised but benevolent health authority (e.g. the healthcare financing system of the UK, Germany), which cares about both patients' total benefits and healthcare providers' financial returns. Finally, in the situation that an individual physician knows the degree of his altruism and patients' health status before accepting a contract while not performing a sequence of non-contractible actions, he accepts or rejects the contract provided. As a result, commitment in Liu and Ma (2013) becomes irrelevant.

The provider's type is unknown

De Fraja (2000) studies incentive contracting for healthcare providers with unknown and heterogeneous costs. Each contract specifies the number of cases that a provider is required to treat, the payment per case and a lump-sum reimbursement. The healthcare provider's costs depend on two parameters: (1) the provider's intrinsic efficiency and (2) the treated patients' life expectancy. In her model, it is assumed that the above two parameters are distributed independently. Moreover, the provider's costs decrease with his intrinsic efficiency and patients' life expectancy. After observing the contract being provided, the provider selects the type and number of patients to be treated.

De Fraja (2000) shows that the optimal remuneration is a menu of contracts which reduces the payment per case but raises lump-sum reimbursement in line with the provider's costs. As the provider with relatively lower costs would dump less but treat more expensive patients. While De Fraja does not consider physician altruism, decisions (i.e. service quantity provided or the number of cases to be treated) made by the less or more costly provider is similar to the decision made by the more or less altruistic provider.

Makris (2009) further extends the study of incentive contracting for health providers with unknown and heterogeneous costs by introducing physician altruism, which is assumed to be a common knowledge.¹⁸ In contrast to De Fraja (2000) in which the

¹⁸ The preferences of civil servants are often regarded in the literature as common knowledge (see Wilson, 1989; Delfgaauw and Dur, 2007).

physician selects the number and the type of patients to be treated, the physician in this framework decides the level of a verifiable performance (i.e. the quantity of service provided to a patient with a known disease). Furthermore, the physician is wealth constrained so that his incurred costs have to be borne by the payer.

The introduction of altruism reduces the low-cost provider's incentive to over-report his type whereas the administrative constraint limits the high-cost provider's incentive to under-report. As a result, the author shows that altruism weakly reduces the power of optimal incentives while determining the type of equilibrium emerges; when the level of physician altruism is low, the health service quantity provided by the efficient (low-cost) and inefficient (high-cost) provider is distorted upwards and downwards respectively. In this case, the inefficient provider makes a zero profit. When the level of altruism is intermediate, there is no distortion in quantities and both types of provider have a zero profit (i.e. first best). Makris and Siciliani (2013) further shows that when the degree of physician altruism is sufficiently high, the quantity of inefficient type is distorted upwards while the quantity of efficient types is distorted either downwards or upwards. In this case, the inefficient type makes a zero profit.

Notice that Makris (2009) and Makris and Siciliani (2013) could also assume providers have unknown and heterogeneous degrees of altruism while having homogenous productivities. This may be the case when the healthcare provider is an individual primary healthcare physician rather than a civil service organization (i.e. a hospital or a group of physicians). Accordingly, it may be very difficult to measure each individual's degree of altruism. It can be also argued that primary care physicians often provide standardized healthcare services and are allocated a fixed proportion of different type of patients. As a result, assuming they have the same cost function could be reasonable.

Allard et al. (2014) analyse the optimal incentive contracting of GPs, when they are allowed self-selecting payment forms. In particular, this paper examines the role of multi-dimensional heterogeneity (abilities and concerns for patients' health) on GPs' choice of alternative payment forms (capitation and fee-for-service) and the subsequent implications for treatment and referral decisions to specialty care. They show that the optimal form of remuneration depends on the regulator's objective. When the regulator's main concern is to save the high costs of specialized care, it is optimal to pay GPs a fee-for-service. Alternatively, if the main concern is to limit potential errors that can be made

by GPs, the optimal form of remuneration depends on the distribution of GPs' profiles. If high ability GPs constitute a relatively large proportion, it is optimal to allow physicians to select their favourite payment systems. On the contrary, if low ability physicians constitute a relatively large proportion, then capitation is optimal. Capitation is also optimal when failures to identify severe conditions are extremely costly in terms of health losses.

Barham and Milliken (2015) take a population-based approach to compare costs and treatment incentives embedded in different payment systems, including the choice of more than one. Similar to Choné and Ma (2011) and Liu and Ma (2013), a patient's initial health and physician's degree of altruism is assumed to be the latter's private information. However, Barham and Milliken (2015) is the first study in the literature assuming physicians can choose both the number and type of patients as well as the intensity of treatment provided. Moreover, the optimal payment system trades off social benefits and costs while ensuring the entire population's access to healthcare.

To capture above features, Barham and Milliken (2015) assume a physician can be either altruistic ($j = 1$) or non-altruistic ($j = 2$).¹⁹ Patients are also of alternative types, namely frail ($i = F$) or healthy ($i = H$). The total number of frail and healthy patients is N_F and N_H whereas the number of patients selected by an individual physician's practice is n_F and n_H . The target level of service that the regulator wants to see is \bar{q}_F and \bar{q}_H where $\bar{q}_F > \bar{q}_H$.²⁰ Altruistic providers derive positive non-pecuniary benefits from the level of service provided to frail patients, however, they do not derive any non-pecuniary benefits from treating healthy patients. The physician's utility is expressed as:

$$V^j = I^j - C(n_F^j q_F^j + n_H^j q_H^j) + n_F^j A^j(q_F^j) \quad (9)$$

where I^j is the income of the provider, q_F^j and q_H^j the intensity of service provided to a frail and a healthy patient respectively, with n_F^j and n_H^j the number of the type F - and H -patients enrolled respectively. The cost function $C(\cdot)$ is increasing, convex and depends only on the total caseload. In other words, costs associated with enrolling

¹⁹ Hereafter I use superscript and subscript to denote the physician and the patient respectively.

²⁰ The authors argue that the given targets reflect the consensus of medical opinion, namely that primary care doctors often refer to clinical practice guidelines which indicate the care that should be provided to patients with given sets of symptoms. Alternatively, they suggest the targets can be thought of as the levels of service that would be selected as the solution to the optimal contracting problem in a full information world (p.898).

additional patients are ignored. The altruism function is bell-shaped, that is, $A^{1'}|_{q_F^1 < \bar{q}_F} > 0$, $A^{1'}|_{q_F^1 = \bar{q}_F} = 0$, $A^{1'}|_{q_F^1 > \bar{q}_F} < 0$, $A''(0) > 0$ and $\exists \hat{q} < \bar{q}_F$ such that $A''(q) > 0$ when $q < \hat{q}$ and $A''(q) < 0$ when $\hat{q} < q < \bar{q}_F$. For the non-altruistic physicians, $A^{2'} = 0$ for all q . The physician's optimization problem under capitation or fee-for-service is:

$$\begin{aligned} \max_{n_F^j, q_F^j, n_H^j, q_H^j} V^j &= k(n_F^j + n_H^j) + p(n_F^j q_F^j + n_H^j q_H^j) - C(n_F^j q_F^j + n_H^j q_H^j) \\ &\quad + n_F^j A^j(q_F^j) \end{aligned} \quad (10)$$

$$\text{Subject to } \bar{q}_H \leq q_H^j \leq \bar{q}_F, \bar{q}_H \leq q_F^j \leq \bar{q}_F$$

Where the constraints reflect the contractual requirement to provide the service intensity in the certain range, k is the per patient capitation fee, $n_F^j + n_H^j$ the total number of patients selected, and p the price per unit of service provided. The Lagrange of the system (10) is:

$$\begin{aligned} \mathcal{L} &= k(n_F^j + n_H^j) + p(n_F^j q_F^j + n_H^j q_H^j) - C(n_F^j q_F^j + n_H^j q_H^j) + n_F^j A^j(q_F^j) \\ &\quad + \mu_F^j(q_F^j - \bar{q}_H) + \eta_F^j(\bar{q}_F - q_F^j) + \mu_H^j(q_H^j - \bar{q}_H) \\ &\quad + \eta_H^j(\bar{q}_F - q_H^j) \end{aligned} \quad (11)$$

The solution to the system (11) has to satisfy the following first order conditions:

$$\left\{ \begin{array}{l} k + q_F^j(p - C') + A^j = 0 \\ k + q_H^j(p - C') = 0 \\ n_F^j(p - C') + n_F^j A^{j'} + \mu_F^j - \eta_F^j = 0 \\ n_H^j(p - C') + \mu_H^j - \eta_H^j = 0 \\ \mu_F^j(q_F^j - \bar{q}_H) = 0 \\ \eta_F^j(\bar{q}_F - q_F^j) = 0 \\ \mu_H^j(q_H^j - \bar{q}_H) = 0 \\ \eta_H^j(\bar{q}_F - q_H^j) = 0 \\ \mu_F^j, \eta_F^j, \mu_H^j, \eta_H^j \geq 0 \end{array} \right. \quad (12)$$

where the first four lines are first order conditions with respect to $n_F^j, q_F^j, n_H^j, q_H^j$ respectively. The next four lines show the corresponding complementary slackness conditions associated with the inequalities in the optimization problem. The last line shows restrictions on the respective multipliers of the non-negative requirements for services.

From the system (12), the non-altruistic physician ($j = 2$) would provide the quality of service $q_F^{2*} = q_H^{2*} = \bar{q}_H$ and select the number of patients $n_F^{2*} = n_H^{2*}$ which is implicitly defined by:

$$k + \bar{q}_H(p - C'(n_F^{2*}\bar{q}_H)) = 0 \quad (13)$$

Accordingly, the non-altruistic provider increases his practice size and total caseload with price or capitation while providing the minimum quality of service for both types of patients.

The altruistic physician ($j = 1$) would prefer to serve only frail patients and to provide the service $q_F^{1*} \geq q_H^{1*} = \bar{q}_H$ (see the first and the third equation of the system (12). When $q_F^{1*} > \bar{q}_H$, n_F^{1*} is implicitly determined by:

$$\begin{cases} k + q_F^{1*}(p - C') + A(q_F^{1*}) = 0 \\ p - C' + A'(q_F^{1*}) = 0 \end{cases} \quad (14)$$

When $q_F^{1*} = \bar{q}_H$, n_F^{1*} and q_F^{1*} are implicitly determined by:

$$k + \bar{q}_H(p - C'(n_F^{1*}\bar{q}_H)) + A(\bar{q}_H) = 0 \quad (15)$$

In the former case $q_F^{1*} > \bar{q}_H$, by combining both equations in system (14), we have:

$$k + A(q_F^{1*}) - qA'(q_F^{1*}) = 0 \quad (16)$$

Since $A(q_F^{1*}) - qA'(q_F^{1*}) < 0$, the curvature assumption of A indicates $A'' < 0$.

Calculating the first order condition with respect to k and p , we have:

$$\frac{\partial q_F^{1*}}{\partial k} = \frac{1}{qA''} > 0 \quad (17)$$

$$\frac{\partial q_F^{1*}}{\partial p} = 0 \quad (18)$$

Next, calculating the first order condition with respect to k and p on the first equation of (14) and substitute $\frac{dq_F^{1*}}{dk}$ and $\frac{dq_F^{1*}}{dp}$ derived from (17) and (18), we have:

$$\frac{\partial n_F^{1*}}{\partial k} = \frac{1 - \frac{nC''}{A''}}{(q_F^{1*})^2 C''} \quad (19)$$

$$\frac{\partial n_F^{1*}}{\partial p} = \frac{1}{q_F^{1*} C''} > 0 \quad (20)$$

Finally, the total caseload $Q_F^{1*} = n_F^{1*} q_F^{1*}$ satisfies:

$$\frac{\partial Q_F^{1*}}{\partial k} = n_F^{1*} \frac{\partial q_F^{1*}}{\partial k} + q_F^{1*} \frac{\partial n_F^{1*}}{\partial k} = \frac{1}{q_F^{1*} C''} > 0 \quad (21)$$

$$\frac{\partial Q_F^{1*}}{\partial p} = n_F^{1*} \frac{\partial q_F^{1*}}{\partial p} + q_F^{1*} \frac{\partial n_F^{1*}}{\partial p} = \frac{1}{C''} > 0 \quad (22)$$

Equation (16) indicates that the frail patient will be provided a higher level of service quality under capitation ($k > 0, p = 0$) than under fee-for-service ($k = 0, p > 0$). Under both systems, however, the service quality provided is suboptimal (i.e. $q_F^{1*} < \bar{q}_F$). The equations (17)-(22) characterize the altruistic physician's decisions regarding the number of patients selected and the service quality as a function of price and capitation rate. These functions show that the altruistic physician increases service quality and total caseload as the capitation rate goes up. However, the altruistic provider raises practice size and total caseload while the quality of service remains unchanged as price increases.

Given this model, Barham and Milliken (2015) then compare FFS, capitation, mixed payment (capitation plus a partial reimbursement of costs incurred), a menu comprised of FFS and the mixed mechanism by the respective social welfare are provided with the whole population having access to health service. This welfare is measured as benefits of providing service net of social costs. Further, the authors show that physicians' total caseload provided under a single payment can also be induced under a menu of menu of contracts, but at a relatively lower cost. Accordingly, they conclude that the optimal payment is a menu of contracts only when social benefits associated with providing a higher quality of care to frail patients (e.g. the improvement in frail patients' health and

the non-pecuniary benefits obtained by altruistic physicians) are greater than the additional costs of providing this care.

By reviewing the above analysis, a few adjustments or extensions should be considered. First, the authors assume that the optimal service quality to be provided to each type of patient is given exogenous. It would be interesting to see whether their results still hold if the optimal level of care provided to each type of patient depends on remuneration instruments and other exogenous parameters (endogenization). Second, the altruism function $A(q)$ in Barham and Milliken (2015) represents the physician's concern for his patients' welfare. In reality, the physician may care only about his patients' health benefits rather than their welfare. For instance, while smoking maximizes a patient's welfare (utility), the physician would not allow him to do so because smoking has a negative impact on the patient's health. Third, the altruism function $A(q)$ is assumed to be "bell shaped" to ensure that the altruistic physician obtains an equilibrium between enrolling patients and producing quality.²¹ Given that there is neither theoretical nor empirical evidence to support this assumption, the alternative approach of modelling the physician's choice of practice size and service quality should be considered. Further, the research does not consider an important element in that there is a cost associated with getting to know additional patients. Finally, their analysis can also be extended to the case where physicians are constrained by a non-negative profit constraint.

Studies outside health economics that have investigated the selection of workers with intrinsic motivation are also relevant. For instance, Heyes (2005) and Delfgaauw and Dur (2007) have studied optimal wage schemes when workers are privately informed about their motivations. Burani and Palestini (2016) have analysed the screening problem of a firm hiring workers with unknown ability but an observed intrinsic motivation. Handy and Katz (1998), Delfgaauw and Dur (2007a, 2010), Barigozzi and Turati (2012), and Burani and Palestini (2016) have investigated optimal incentive schemes that sort motivated workers into different sectors. Common to these analyses is that the principal cannot observe his workers' abilities or motivations. Moreover, under the optimal remuneration scheme physicians have access to self-selecting a menu of lump-sum reimbursements.

²¹ If the altruism function is strictly concave, there is an incentive to severely undertreat patients (but treat many patients); if it is strictly convex, there is an incentive to severely over treat patients (but treat a few patients) (Barham and Milliken, 2015, p.898).

To summarize, the heterogeneity of the physician population makes it natural to propose a menu of physician reimbursement schemes. This menu often increases the power of financial incentives with the physician's productivity/altruism and therefore induces the more altruistic/productive physician to select the contract with higher payment and service provision. However, the exact form of the menu depends on the available payment mechanisms, the types of heterogeneity, and financial constraints imposed. The optimal payment system can also be a single payment system which comes from tensions caused by incentive capabilities between quantity and financial constraints (e.g. non-negative profit) in the presence of multi-dimensional adverse selection.

Before moving to the next part of this survey, there are a few points that need to be addressed. First, it may be impossible or very costly to offer physicians a menu of contracts. In addition to the reason discussed in Choné and Ma (2011), a primary care physician is often allocated many patients with different severities of illness. When physicians differ only in their degrees of concern for their patients, offering them a menu of contracts requires that the physicians' altruism functions satisfy the single crossing property. As patients differ in their initial health and the marginal benefits of consuming medical service, there is no guarantee that the physicians' altruism function only crosses once.

Furthermore, providing physicians with a menu of contracts requires a strong commitment capability on the part of the payer (Hart and Moore, 1988). After a physician selects a contract, she reports her type (i.e. ability, costs etc.) and her informational advantages are removed. At this stage, it would usually be to the advantage of both parties to renegotiate their initial agreement (Demougin, 1989). The possibility of renegotiation would therefore destroy the properties of the contract scheme. Allowing physicians to select their reimbursement schemes also makes the cream-skimming of healthy patients feasible and financially lucrative, thus difficult to resist (Matsafanis and Glennerster, 1994). The above difficulties associated with using a menu of contracts suggest applying a single physician remuneration scheme might be a better option. In fact, it is possible to dispense a menu of contracts as the direct revealing mechanism can be indirectly implemented by a single linear/non-linear schedule which depends only on the outcome of the production process (i.e. the delegation principle) (Hart and Moore, 1988; Demougin, 1989).

Second, physicians are often constrained by a non-negative profit condition for treating his patients. Third, the studies discussed in the above subsection does not examine the problem associated with physicians' non-contractible actions. Finally, the principal's objective function might differ; if it concerns an insurance firm operating in a competitive market, the function sums up patients' benefits only. If it is a centralised health authority, the objective function sums up both patients' net benefits and providers' net profits.

2.2.2.4 Optimal physician payment with both moral hazard and adverse selection

Studies analysing the optimal design of physician remuneration in an environment with both moral hazard and adverse selection are small but growing continuously.

Jack (2005) was the first to study incentives for quality and cost choices by physicians with unknown altruism. He has assumed the healthcare provider's utility increases with quality of health service, however, this utility is unknown to the regulator. Moreover, the regulator is not able to monitor the physician's quality and cost-reducing efforts. Jack adopts a reservation utility constraint, allowing physicians to suffer financial losses from treating patients. Finally, he assumes collecting public funds is costly while the provider's total costs incurred can be observed ex-post.

Later studies analysing the design of physician remuneration capture adverse selection and moral hazard problems in different ways. For instance, Mougeot and Naegelen (2009) as well as Kantarevic and Kralj (2016) assume the healthcare provider has private information about his costs incurred per treatment and exerts a non-contractible effort. In the former paper, the unobservable effort reduces the cost of treatment while in the latter it enhances medical service intensity. The former paper also considers the effects of the provider's degree of benevolence, which is assumed to be common knowledge. Moreover, the provider can select a cost threshold and refuse to treat patients whose cost is beyond (the outlier). The provider receives a fixed payment per case, however, an additional cost-sharing will be paid for when treating the outlier.

Jelovac and Kembou Nzale (2017) as well as Wu et al. (2018) also capture moral hazard by assuming an unobservable effort on the part of the healthcare provider. While this effort contributes to patient health benefits increase in both papers, it improves the

benefits by providing input jointly with the regulator's in the first paper and by raising the patient's recover probability in the second paper. With respect to adverse selection, the former and the latter study assume the healthcare provider has private information about his altruistic preference and ability to cure the patient respectively. The healthcare provider only chooses the level of effort in the former study. However, he decides the level of effort, the quantity of health service and the size of practice simultaneously in the latter. In contrast to previous studies in this literature, the provider is constrained by a non-negative profit constraint. Finally, in addition to deciding the fixed payment, the regulator also determines the level of his input in the former while deciding the fee-for-service and pay-for-performance price in the latter.

The main conclusion from this literature is that the optimal physician remuneration scheme is a menu of contracts rather than a single payment system (Jack, 2005; Kantarevic and Kralj, 2016; Wu et al., 2018). This contract can be approximated by a menu of the linear contracts comprised of a fixed salary component and a variable payment.²² In particular, the rates of fixed payment and variable reimbursement are negatively related across contracts. Under this remuneration scheme, the more productive/altruistic physician would choose the contract with a higher fixed payment and lower cost/performance reimbursement.

However, the single payment system can also be optimal in some specific contexts. For instance, Mougeot and Naegelen (2009) show that the optimal physician payment is a fixed price contract when the provider's benevolence is high whereas a mixed system with fixed price per case and a cost sharing for the outlier (i.e. the patient whose cost is greater than the predetermined cost threshold), when the provider's benevolence is low. On the other hand, Jelovac and Kembou Nzale (2017) argue that it is too costly to use a menu of contracts to extract providers' private information if both the regulator and provider contribute to a non-contractible outcome.

As shown in the discussions in the discussions in Sections 2.2.2.2 and 2.2.2.3, moral hazard can be captured not only by the physician's unknown efforts but also the non-contractible patients' health as well their health benefits generated. Adverse selection

²² The exact form of the variable payment depends on the contractible outputs provided by the physician. In Jack (2005) and Kantarevic and Kralj (2016), it is regarded as the rate of cost-reimbursement. In Wu et al. (2018), it is the price paid per unit of service provided or the reimbursement rate per unit of an approximation of the physician's improved 'performance' and per patient treated.

problem can also be modelled in different ways; it can be captured by the unknown and heterogeneous physicians' ability, altruism, productivities or patients' initial health, abilities to be cured from illness etc. I show that the degree of motivation/altruism tends to be a common knowledge when healthcare provider is a civil service organization while a private information when healthcare provider is an individual physician.

Moreover, the objective function of the principle depends on its role and where it operates. Finally, the heterogeneity of physicians' population or their patients' severities of illness makes it natural to propose a menu of physician remuneration schemes. However, the tension caused by incentive capabilities between quantity and financial constraints (e.g. non-negative profit), the unachievable single crossing property, the incentives to renegotiate the initial contract and cream-skimming low-cost patients, make implementing contract menus impossible or inefficient. As a result, physicians are rarely offered a menu of contract in reality.

2.2.3 The impact of cost control mechanisms on physician behaviour

There are also studies that examine and compare incentives provided by different policies to control physicians' costs. In this subsection I review how the introduction of these instruments alter physician incentives when they are paid on a fee-for-service basis.

Neudeck (1991) considers an environment where the social insurance fund (SIF) chooses a price to maximize health output (i.e. efficiency generated per unit of service), assuming that there are enough doctors willing to accept the contracts offered. The selected level of price then determines the productivity and the quantity of service provided by each doctor. Similar to the result derived from efficiency wage model in the labour market, the author shows that the SIF should pay each physician a price above the market clearing level to induce the efficient provision of medical service. However, paying the price at such a level would create additional demands for physician contracts and result in medical service overprovision. As a consequence, rationing the number of contracts and service quantity is a rational strategy for the SIF. Finally, this paper further shows that the above result can be generalized in an environment with target caseload/income or adverse selection.

Notice that the analytical framework presented by Neudeck (1991) is one of the simplest representations of healthcare system which does not capture many essential health market characteristics. For instance, it does not consider the role of a patient in determining the level of medical service provided. It also does not consider physicians' agency and their limited liabilities. Different issues regarding information asymmetries between the payer, providers and patients are also ignored. As shown in Neudeck (1991), SIF trades-off efficiency and the quantity of service to be provided. When the aforementioned health market characteristics are also taken into consideration, the SIF may face additional difficulties in terms of inducing the efficient provision of healthcare (e.g. moral hazard, externalities etc.).

Fan et al. (1998) have presented a model that compares two alternative methods of controlling the costs of physician service under a fee-for-service payment, namely, the expenditure target, which involves quantity control and the expenditure cap, which is a retrospective price-setting mechanism. To do this, they compare the total quantity of service that can be provided by different types of physicians under these alternative physician cost control policies, given the same level of health spending. A payment system is considered to be “better” if physicians can provide more services. Physicians in their model differ in their productivity (measured by marginal costs of providing healthcare) and decide the quantity of a single service q to be provided at a given price p . Under expenditure target, each physician is given a quota \bar{q} . If he produces over \bar{q} , he receives only αp ($\alpha < 1$). Under the expenditure cap B , the price is determined as the budget divided by total quantity provided by all physician (i.e. $p = \frac{B}{\sum q_i}$), where q_i is health service provided to each patient.

The authors show that the expenditure cap induces physicians to produce a larger quantity of medical service than the expenditure target given the same budget under the symmetric Nash equilibrium. It follows that, by introducing an expenditure cap, the retrospective price setting mechanism introduces competition into the environment, which raises efficiency. This theoretical result is also supported by laboratory-controlled experiments, where subject behaviour is well captured by individual profit maximization and symmetric Nash equilibrium predictions.

Benstetter and Wambach (2006) compare the physician supply of medical service under a price system (i.e. FFS system) and under a point system (the retrospective payment of

a fixed budget). Their key assumption is that the individual physician has to borrow a certain amount of money and would bear bankruptcy costs if he fails to pay back. Accordingly, they show that the physician may even raise the quantity of service provided as price decreases in order to avoid the high costs of bankruptcy. They find that a shift from a price to a point system may lead to a ‘treadmill effect’ where physicians increase total health service provision, receive less payment per service and even choose to exit the market. Similar to Fan et al. (1998), this effect is derived from the fact that physicians under FFS with a budget cap are induced to compete in order to increase their income. In order to moderate the treadmill effect, the authors further suggest introducing either a price floor or a maximum number of treatments can be provided per physician.

Notice that the simple representations of alternative cost control instruments in this literature may not represent many widely used quantity or revenue restrictions in reality. For instance, in contrast to Fan et al. (1998) the service provided beyond the quota may not be reimbursed; during the German healthcare financing revolution in 1997, each individual physician was imposed a maximum amount of treatment to be claimed (Benstetter and Wambach, 2006). Moreover, expenditure caps have not only been imposed on the global level but also on the regional or individual level. For instance, Germany has introduced per surgery spending caps for pharmaceutical expenditure (Busse et al., 2005) whereas the Netherlands has imposed a ceiling on physicians' income (Kroneman et al., 2009). Future studies may also evaluate the impact of introducing these cost control restrictions while capturing essential features of healthcare market (e.g. information asymmetry, limited liability etc.) in developing models.

2.2.4 Summary

The previous sections reviewed theories studying physicians' responses to incentives embedded in different payment systems in the presence of respective types of information asymmetry. Given the physicians' medical service supply decisions, the optimal design of physician incentive contracting mechanisms are analysed. After that, the effects of introducing alternative cost control policies to physicians paid on a fee-for-service basis are discussed.

In the presence of moral hazard only, the cost-reimbursement payment often induces the provision of constrained optimal quality and zero cost-reduction efforts. Moreover, physicians increase their level of efforts and quantity of service as the rate of prospective or cost reimbursement increases. Prospective payment, however, may induce the first best levels of quality and cost reduction efforts if: (1) there is a single product; (2) quality reflects patient demand; and (3) physicians are unable to dump patients. When the above conditions do not hold, first best quality and cost reduction efforts can only be induced by mixing prospective payment and cost-reimbursement.

I show that studies analysing the optimal design of physician payment often capture the moral hazard problem by physicians' unknown efforts although an alternative modelling approach has rarely been applied. While the optimal physician remuneration in the presence of moral hazard often has an element of cost sharing, implementing this remuneration requires a sufficiently close monitoring. As closely monitoring physicians' costs may be very expensive and inefficient, later studies propose a fee-for-service system as one substitute for the cost-sharing mechanism (McGuire, 2000).

In the presence of physician or patient heterogeneity, the optimal physician remuneration is often proposed as a menu with negatively related prospective and retrospective payments across contracts. Under this payment, the more productive/altruistic physician selects the contract with a higher fixed payment while the less productive/altruistic one selects the contract with a higher cost-reimbursement. The more productive/altruistic physician would have an informational rent whereas the less altruistic/productive doctor would underproduce. This survey also shows that a single physician remuneration system may perform better than a menu of remuneration schemes as the allocation under the latter system can be indirectly implemented by the former, which does not require type-reporting and truth-telling. Moreover, the tension caused by the incentive compatibility between quantities and non-negative profit, the difficulties to satisfy the single crossing property, and the incentives to renegotiate the current contract and cream-skim low costs patients, make single physician payment system a better option.

Cost control instruments such as quantity restrictions (e.g. quotas) and revenue caps (e.g. individual, sectorial, regional and global budgets) are often imposed on a fee-for-service system to control physicians' healthcare spending. The existing literature provides a

theoretical support in that quantity restriction could be effective in terms of improving efficiency while revenue restrictions may lead to a treadmill effect. Given the same level of expenditure, fee-for-service physicians under a global budget will provide more services but receive a lower payment per service than under a quota as the budget induces competition between physicians. This literature contains a limited number of studies, which focus on investigating FFS physicians' supply of health service when a global budget cap is imposed. Only a few studies compare physicians' supply decisions under a quota and an expenditure cap. Moreover, the quantity and the revenue restriction fail to represent their applications in the real world. To make studies in this literature provide more valuable predictions and insights, it is also necessary for them to capture key characteristics of the healthcare market in their models.

By reviewing the literature and addressing potential gaps, my thesis focuses on an area where less attention has been paid to the design of more sophisticated physician incentive contracting (compared to a simple transfer) in the presence of moral hazard and adverse selection. Specifically, I investigate the optimal implementation of a fee-for-service payment system and the most common cost control instruments (quantity rationing, revenue cap, capitation). Moreover, I characterize the physician's choice of practice size or medical service intensities as a function of price, the respective cost control instrument, and other exogenous parameters. The moral hazard problem is captured by the fact that the regulator cannot determine whether the physician provide the optimal level of service for each patient, as the regulator does not know patients' health. Since the healthcare provider in my analysis is a primary care provider, I capture adverse selection in terms of physicians' unknown degrees of altruism rather than as productivities or abilities. Each physician is allocated many patients with different health statuses and is expected to earn non-negative profit. Facing the both theoretical and practical difficulties of implementing a menu of contracts as evident in the literature, my analysis will focus on designing a single payment system for all physicians. My research will also be compare the impact of introducing different types of cost control instruments on FFS physicians. In contrast to the existing literature, I will examine the cost control methods that have been imposed on an individual doctor or patient.

2.3 Evidence

The foregoing section introduces physician altruism and discusses physicians' incentives embedded in cost-reimbursement, prospective payment and different cost control policies. Moreover, it discusses the optimal design of physician incentive contracting. While the literature investigating physician altruism and physicians' response to different payment systems is extensive (e.g. Galizzi et al., 2015; Ellis and McGuire, 1986; Chalkley and Malcomson, 1998; Ma, 1998; Chalkley and Malcomson, 2002; Jack, 2005; Choné and Ma, 2011; Makris and Siciliani, 2013; Kantarevic and Kralj, 2016; Wu et al., 2018), the evidence regarding the optimal design of physician payment in practice is limited. One explanation is that the real-life design of a physician payment system depends on institution- and country-specific contexts (e.g., US physician group practice, the UK's fund-holding system). As a result, it is difficult to achieve a consensus regarding which payment is optimal in general. In the ensuing chapter, I offer empirical observations of physician altruism and the impact of using different payment system and cost control policies. This range of evidence make it reasonable to assume altruistic physicians' preferences and are shown to be consistent with the results I derived in Chapters 3 and 4. Section 2.3.1 surveys the existing empirical literature which discusses provider altruism. Specifically, it offers an overview of the existing evidence on healthcare provider altruism in different branches of the empirical literature. The evidence relating to physicians' behaviour and the design of their payment is discussed in Section 2.3.2. Finally, Section 2.3.3 provides evidence regarding the impact of introducing physician cost control policies.

2.3.1 Provider Altruism

Galizzi et al. (2015) has grouped evidence of provider altruism into four main categories based on degrees of control applied in empirical strategies including: 1) survey and interview data; 2) prescription records; 3) field experiments; and 4) laboratory experiments.

2.3.1.1 Survey and Interview

Some systematic questionnaires have been designed to measure altruistic motivation in surveys. For example, the Prosocial Personality Battery developed by Penner et al. (1995) includes self-reporting of other-oriented empathy and participation in helpful behaviours or activities. The Penn State College of Medicine Professionalism Questionnaire by Blackall et al. (2007) considers altruism as one of seven elements of professionalism. Pawlikowski et al. (2012) assesses Scale of Attitudes towards the patient in his test which is based on respect for autonomy, altruism, empathy, and a holistic approach to the patient. Gutiérrez et al. (2006) developed the Nursing Motives for Helping scale, where altruism is identified by response to items such as the effort to identify patients' needs, the level of support offered, and the level of attention paid to patients.

Altruistic motivation among healthcare providers has also been identified by ad-hoc surveys and interviews. For instance, Allaby (2003) has analysed local doctors' motivation to serve in charitable clinics in urban Nepal. He found the main reason that doctors decide to work in a charitable clinic is "a desire to serve the poor and improve society" (p.84). Siddiqi et al. (2011) found that the "opportunity to serve people" (p.1632) is important in both public and private setups in Pakistan after surveying 300 medical doctors. De Costa et al. (2008) has shown that altruism was one of two motivators for collaborating with a government scheme aimed at improving patient welfare. Meanwhile, Desquins et al. (2007) has predicted around 20 percent of doctors adopt altruistic behaviours.

2.3.1.2 Prescription Records

Empirical evidence of physician altruism can be found in studies investigating prescription choices in primary healthcare (e.g. Hellerstein, 1998; Lundin, 2000; Crea et al., 2015). In their models, physicians care about patients' welfare and insurance expenditures. Given the model, the authors analyse the physician's trade-off between marginal utility from patient welfare improvement and marginal disutility from insurance spending. For instance, Hellerstein (1998) used data from the 1989 National Ambulatory Medical Care Survey (NAMCS) to study physicians' choice of drug prescription. He found that the majority of physicians in the survey prescribed both generic and trade-name drugs to their patients, but some physicians were more likely to

prescribe the less expensive generic drugs while others were more likely to prescribe trade-name drugs. In addition, Lundin (2000) has utilized data collected from two pharmacies in Sweden in 1992 and 1993 to analyse the role of health insurance in the drug prescription market. His probit estimates show that physicians weigh patients' benefits from health insurance more than insurance expenditure.

Crea et al. (2015) has analysed Finnish pharmaceutical prescriptions records from the Social Insurance Institution (2001-2011) to estimate the probability that doctors prescribe generic versus branded versions of statins to their patients, as a function of the shares of the difference in prices that patients have to pay out of their own pocket compared to what is covered by insurance. Results from their panel logit models provide strong support for the existence of physician altruism and moral hazard hypotheses in Finland. In these studies, the degree of physician altruism is estimated by comparing the variation of patients' out-of-pocket payments and health benefits obtained from the consumption of healthcare (Galizzi et al., 2015, p.15).

2.3.1.3 Field experiment

Results from artefactual field experiments also provide evidence for physician altruism. For instance, Jacobsen et al. (2011) asked two samples of 88 nursing and 73 real-estate broker students in Norway to play a Dictator Game (DG) with Amnesty International as the recipient. They found nursing students donated about 75% of their endowments, compared to 61% of real-estate broker students, which implies that nursing students were more generous than real-estate broker students.

Smith et al. (2012) asked 1,064 final year nursing students in Kenya, South Africa and Thailand to play a DG. In their experiment, nursing students decided the proportion of their endowment to be allocated to themselves, a fellow student, a patient or a poor person respectively. They found nursing students in three countries donated one third of their endowment and showed greater generosity to patients and the poor than to fellow students. Similar results are also supported by Kolstad and Lindkvist (2012), who asked two samples of medical and nursing students in Tanzania to play a DG game with medical or nursing students as recipients. They found that students who prefer to work in the public health sector have stronger pro-social preferences than those who prefer to work in the private for-profit sector. Meanwhile, Serra et al. (2011) asked 219 nursing

students and 90 medical students in Ethiopia to play a Generalized Trust Game (GTG). They found some correlation between their measure of generalized trustworthiness in GTG and the self-reported intention to work in the non-profit health sector.

2.3.1.4 Laboratory experiments

In comparison to the observational studies discussed above, laboratory experiments allow researchers to analyse trade-offs between patients' health benefits and physician profit under controlled conditions using incentivized behavioural data (Galizzi, 2015). Another major advantage is that choice situations can be implemented with trade-offs that relate closely to theoretical models of physician behaviour.

Hennig-Schmidt et al. (2011) first used a controlled laboratory experiment to analyse the effect of variation in the payment system regarding physicians' quantity choice. In their experiment, medical students (as representatives of physicians) chose the quantities of medical services to determine both their financial incomes and patients' health benefit (measured in monetary terms) outside the laboratory. Subjects therefore face a trade-off as they were unable to maximize their own profit and patients' welfare simultaneously. While this study does not discuss the impact physician altruism on physician behaviour, it provides useful data for later studies testing or measuring physician altruism.

Godager and Wiesen (2013) used data provided by Hennig-Schmidt et al. (2011) to estimate the degree of physician altruism. Specifically, the level of physician altruism is measured as the marginal rate of substitution between patient benefit and a physician's profit. Their multinomial logit and mixed logit regressions results indicate that each medical student puts a positive weight on his patients' health benefits though the weight differs. Substantial variations in physicians' altruism were also observed in Godager et al. (2016), which investigated how physicians' quantity choices were affected by disclosing outcome information.

Brosig-Koch et al. (2017) have applied a laboratory experiment designed in the fashion of Hennig-Schmidt et al. (2011) to identify physician altruism. They inferred physician altruism according to subjects' decisions concerning service quantity, which determines profit and patient benefits. They found that patient benefits play an important role in subjects' decision making. As in the previous study, they also report that physicians

differ significantly in their levels of altruism. Hennig-Schmidt and Wiesen (2014) have reported a considerable difference between medical students and students with other majors in terms of their altruistic motivation towards a patient, based on the set up developed by Hennig-Schmidt et al. (2011).

Finally, results from laboratory and artefactual field experiments by Brosig-Koch et al. (2016) show that physicians, medical students and students from other majors respond in a similar way to incentives embedded in fee-for-service and capitation. However, physicians were found to have the highest altruistic motivation in their medical service provision. Kesternich et al. (2015) show that when professional values are made more salient or when social incentives benefit actual patients (rather than students), medical students behave more altruistically.

To summarize, studies based on surveys, interviews, prescription records, field and laboratory experiments all provide evidence of physician altruism. Accordingly, the assumption that physicians are concerned about patients' benefits is well supported by both theory and empirical evidence. Moreover, recent studies have applied laboratory experimentation in order to analyse physicians' choice of medical service quantity to maximize both profit and patient health benefits. These studies have not only measured physicians' altruistic motivation but also show a substantial heterogeneity in the degree of physician altruism (Hennig-Schmidt et al., 2011; Godager and Wiesen, 2013; Hennig-Schmidt and Wiesen, 2014; Brosig-Koch et al., 2017). These observations support the assumptions made in Chapters 3 and 4, where physicians differ in their degree of altruism and they impose a positive weight on their patients' health benefits.

2.3.2 Physician behaviour and payment design

In this subsection, I review the empirical studies analysing the number of primary care physicians' office visits, the quantity and quality of service provided as well as referral rates under fee-for-service and capitation systems. Moreover, I discuss the impact of fee-for-service price or capitation variations on the above decisions made by patients' or doctors. The main purpose, however, is to highlight those observations which relate to my results derived in Chapters 3 and 4.

Physician incentives embedded in standard remuneration rules have been well described in the literature (e.g. McGuire, 2000; Léger, 2008). However, testing the effect of doctors' remuneration has been challenging as the data obtained is often biased due to the problem of self-selection or the confounding effects based on other contextual factors (Gosden et al., 1999; Scott et al., 2011). Furthermore, a provider's performance is hard to measure (Lagarde and Blaauw, 2017). As a result, evidence of physicians' decisions on the basis of the same remuneration rule are mixed (Hennig-Schmidt et al., 2011).

First, there is much evidence supporting physicians respond to different monetary incentives (Hemenway et al., 1990; Hillman et al., 1989). For instance, Gaynor and Gertler (1995) have showed that physicians reduce the number of weekly office visits when a fee-for-service payment system is substituted by capitation. A similar behavioural pattern has also been observed for US office-based primary care physicians who participated in a randomized controlled trial conducted by Davidson et al. (1992). They showed that physicians in an FFS group have a higher frequency of visits than in a CAP group.

In the UK, Croxson et al. (2001) have found evidence that physicians respond to financial incentives embedded in the fund-holding system (a prospective payment system). Before getting involved, physicians increase their number of activities in order to be allocated a larger budget for the duration of the scheme. However, once they become fund-holders, they reduce activities to raise their fund surplus. Madden et al. (2005) have used a difference-in-difference methodology to analyse the effect of a change in financial incentives on Irish GPs' behaviour. More specifically, all GPs were paid on a fee-for-service up to 1989, since when the state has started remunerating GPs on a capitation basis for low income patients (medical card holders). This study shows that a low-income patient has a higher rate of GP visits than other patients. However, the average visit rates of the medical card patient have fallen since GPs were paid by capitation. The rest of the population's visit rates have fallen even more as they may decide to use other healthcare services.

Krasnik et al. (1990) have reported that, based on their before-and-after study, GPs in Denmark raised diagnostic and curative services while decreasing referrals to secondary care when pure lump-sum payments were replaced by a CAP supplemented by a FFS

component. Iversen and Lurås (2000) have obtained similar results, finding that Norwegian GPs have more referrals under a CAP system with a low FFS component than under a capitation system complemented by a FFS payment. However, the increase in referral rates may not only be due to CAP payments but also the result of a low FFS payment. Meanwhile, Sørensen and Grytten (2003) have found that salaried physicians have fewer consultations and patient contacts but higher referral rates than FFS physicians.

Dumont et al. (2008) have compared data on primary care services in a Canadian province in Quebec before and after a change from a FFS system to a mixed system with a fixed salary and a reduced FFS reimbursement. They found that physicians reduced quantity of service but increased the time spent per service and per non-clinical service under the mixed payment framework. Devlin and Sarma (2008) also found that FFS payments incentivised physicians to see many more patients per week than alternative payment systems such as CAP, after disentangling selection and incentive effects. However, they showed that physicians who do not select a FFS system appear to have characteristics that would result in them engaging in more patient visits per week than those who choose the FFS scheme.

The aforementioned literature tends to use more traditional designs including surveys, controlled trials and field studies. However, these designs are usually expensive, time consuming and cannot avoid the self-selection bias of doctors in relation to payment systems. Moreover, it is difficult to ensure that only one component of the payment system varies while patient characteristics remain comparable for the samples under study (Hennig-Schmidt et al., 2011). Consequently, more recent studies (Lagarde and Blaauw, 2017; Brosig-Koch et al., 2017, 2016) have applied laboratory experiments. Such experiments allow researchers to: (1) compare physicians' behaviour under *ceteris paribus* conditions (only the payment system varies); (2) conduct thorough robustness checks of the findings (e.g. it can be repeated by different scientists under the same conditions); and (3) offer a test-bed for large-scale studies or institutional revolutions (Hennig-Schmidt et al., 2011).

Hennig-Schmidt et al. (2011) were among the first to conduct a lab experiment to compare physicians' supply of medical services under FFS and capitation. They found that patients are overserved under FFS but underserved under CAP. Moreover, when

physicians are paid on a capitation basis at a uniform rate, more costly patients receive a diminished quantity of service compared to less costly patients. While their experiment design has been adopted by many later studies (Godager and Wiesen, 2013; Hennig-Schmidt and Wiesen, 2014; Brosig-Koch et al., 2016; Brosig-Koch et al., 2017), their experiment failed to examine the effect of capitation on the number of patients treated and the impact of the multi-tasking environment faced by providers (Lagrade and Blaauw, 2017).

The aforementioned gaps have been partially addressed by Green (2014), who first used a real effort experiment where the subjects are paid for reading and correcting spelling mistakes on the basis of salary, FFS, CAP, report cards with CAP and report cards with FFS respectively. He found that the highest quantity (the number of edits) is offered under FFS while the same quantity is produced under salary and CAP. Green also found that the quality of service (the number of correct edits) are the same under the three different payment systems. The real effort experiment has been considered as the better experimental approach as it better reproduce some aspects of real work that physicians' efforts are not hypothetical and negative but may yield utilities (Lagrade and Blaauw, 2017).

Brosig-Koch et al. (2017) has extended Hennig-Schmidt et al. (2011) by studying the effect of introducing blended payments; the results are consistent with Hennig-Schmidt et al. (2011), demonstrating that physicians significantly over-produce or under-produce under FFS and CAP while the quantity of service supplied increases with the severity of patient illness. Moreover, Brosig-Koch et al. have found that blending FFS and CAP significantly reduces deviations from patient-optimal treatments (i.e. the level of treatment which maximizes the patient's welfare). This recall Ellis and McGuire (1986), which predicts that a mixed system can mitigate the incentives embedded in FFS and CAP.

Brosig-Koch et al. (2016) have systematically analysed how different subject pools (non-medical students, medical students, physicians) respond to FFS and CAP. Their results show that subject pools react to monetary incentives consistently; all subjects provide significantly more service under FFS than under CAP. However, the extent to which subjects respond to financial incentives varies. Specifically, physicians' supply of medical services is less affected by the change of financial incentives compared to

medical and non-medical students and the results are robust regarding subjects' gender, age, and personality traits.

Lagarde and Blaauw (2017) have compared physician incentive embedded in different compensation policies (i.e. FFS, CAP, salary) by conducting a real effort experiment where medical students are paid to enter patients' blood test results into computers. Like Green (2014), this study has tested the impact of different payment systems on physicians' performance in a multi-tasking environment, in which the physician decides both the quantity of output (the total number of blood tests entries) and the quality of output (the total number of correct entries). However, their experiment depends less on subjects' prior knowledge and abilities, thereby making direct evaluation of the causal effect of incentives more reliable. Moreover, the experiment is closer to the health setting as it adopts a medical framing, uses medical students as subjects and social incentives are implemented to benefit real patients outside the lab. This study shows that FFS and salary yields the highest and lowest total number of entries respectively. However, the total number of correct entries is lowest under FFS but highest under salary. Finally, when quality benefits patients directly, research subjects improve their output quality without reducing their output quantity.

There is also some evidence which shows that physician supply of medical services does not respond to financial incentives (Gosden et al., 2001; Sørensen and Grytten, 2003). Hurley and Labelle (1995) failed to find evidence of a clear-cut response to different payment incentives in health service provision among Canadian physicians. In Norway, Grytten and Sørensen (2003) have found that the impact of payment systems on physicians' behaviour is not significant after controlling for patient and GP characteristics. These observations reflect that not all theoretical predictions can be supported by empirical evidence.

Physician responses to service price or capitation rate changes have been widely studied. Since price variation has both income and substitution effects, evidence of physicians' responses is often contradictory. For instance, Grytten et al. (2008) have found that a rise in fee-for-service price results in an increase in the total number of consultations and laboratory tests conducted. Kantarevic et al. (2011) have also found that primary care physician significantly increases the total number of services provided and patient visits after fee-for-service price increases. Clemens and Gottlieb (2014) also found that

the Medicare programme in the US increases healthcare supply as health service prices rise. Specifically, they estimated that a two percent increase in reimbursement rates leads to a three percent increase in the provision of healthcare service.

However, there are studies which offer evidence contradicting these results. Some this evidence shows that fee changes do not have a significant effect on physicians' supply of medical services whereas others illustrate that the effect of price changes will be partially countered by changes to medical service volume. For instance, Zuckerman et al. (1998) estimated that a 1% price reduction led to an increase in the provision of management services, tests and procedures of about 25%, 15% and 50% respectively. Yip (1998) also found that a price cut leads to a significant increase in the volume of coronary artery bypass graft (CABG) surgeries in both the Medicare and private markets. Carlsen and Grytten (1998), Grytten and Sørensen (2001), Grytten et al. (2001), and Madden et al. (2005) have found no evidence of supplier-induced demand under FFS regimes in countries with publicly funded healthcare systems.

To summarize, evidence from field studies and surveys shows that the impact of financial incentives on physicians' supply of medical service is mixed. While most of these studies have found that physicians under FFS have more patient visits/consultations, less quality of service supplied/time spent per patient, and lower referral rates than CAP, some evidence suggest that using different payment systems has little impact on physician behaviour. Since surveys or field studies face various methodological difficulties, recent studies have applied different types of laboratory experiments thereby offering more reliable results. The results from these experiments are similar and consistent; first, the total quantity of medical service provided is significantly higher under FFS than under CAP. Second, FFS induces a higher quantity of service provision while CAP induces a higher quality of service provision per patient. Third, blending FFS and CAP mitigates incentives embedded in the respective payment system while patients are more likely to be provided the optimal quantity of medical service. Finally, a physician's supply of health service does not always increase with price. These observations will be used to echo some of the theoretical results derived in the ensuing chapters.

2.3.3 The impact of imposing cost-control instruments

In this final subsection, I review the evidence examining how cost-containment mechanisms affect physicians' behaviour. Specifically, the first part of this subsection reviews the impact of imposing expenditure caps while the second part surveys the influence of using quantity restrictions such as quotas. Finally, some evidence comparing the impact of using both revenue and quantity restrictions are provided. As in the previous subsection, the main purpose of this subsection is to highlight some evidence that will echo the theoretical results derived in Chapters 3 and 4.

2.3.3.1 Revenue Restriction

Revenue restriction generally refers to the maximum amount of money that can be spent on a particular industry ex-ante for a specific time period (Vogler et al., 2009). Revenue restrictions differ in the level at which they have been implemented and in their specific design. Broadly speaking, the restriction can be imposed at a national level, at a regional level or at the level of an individual (Fischer et al., 2018). For instance, in Italy, France and Ireland the public system as a whole negotiated the level of healthcare budget with the pharmaceutical industries. In Sweden, the health authority negotiates or sets the level of health expenditure for all physicians in a given region. However, in most countries (e.g. Germany, the Czech Republic, the UK, etc.), individual-specific prescribing targets are imposed. In the ensuing discussion, I review the impact of introducing different types of revenue restrictions on total health spending, the total quantity of service provided and the quality of service provided.

Poterba (1994) found that Germany, after adopting a global budget system in the mid-1980s, experienced a reduction in the growth of real health spending per capita, from 4.5% during the 1970s to 1.5% during the 1980s. This observation provides some support for the assertion that budget caps can hold costs. Poterba (1994), however, argues that the aforementioned observation only reflects that budget caps can be effective for one-time cost reduction but cannot guarantee a permanent reduction in the growth rate of health outlays. As a result, he concluded that global budgets are unlikely to reduce the share of national resources devoted to healthcare.

Benstetter and Wambach (2006) have analysed the impact of imposing administered caps on the total budget for German outpatient care. They found that the introduction of a budget cap system in 1992 resulted in a 25% medical service price reduction and a 16% decrease of real average physician income between 1993 and 1997 while a 8.9% increase of costs was incurred during 1993-1995. These stylized facts are consistent with the theoretical predictions highlighted in Section 2.3.3, which shows the introduction of an expenditure cap on a FFS system may lead to a severe coordination problem, namely a treadmill effect.

Sood et al. (2009) have examined how revolutions in regulatory policies have influenced pharmacy has shown that physician budgets and global budgets together reduced total pharmaceutical spending by 6%. Furthermore, when disaggregating budget effects, they found that physician-specific budgets are more effective in controlling pharmaceutical spending than global budgets. This is because budgets are borne directly by prescribing doctors who are individually accountable, thereby creating stronger incentives for physicians to reduce their spending.

Rashidian et al. (2015) have systematically reviewed the impact of pharmaceutical interventions using financial incentives on prescribers' behaviour (i.e. drug use, healthcare utilisation and expenditure). After evaluating eighteen studies of pharmaceutical policies from six high-income economies, they found that the utilization of budget caps may lead to a modest reduction in overall drug use (item per patient or prescription) and an increase in the use of generic drugs. However, these effects are uncertain because of the limitations recognized in the studies included. The authors also found some evidence that using pharmaceutical budget caps resulted in a reduction in drug cost per item or per patient/prescription (e.g. UK fund-holding studies) and therefore a decrease in total drug expenditure.

Fischer et al. (2018) have investigated the impact of physician-level drug budgets (i.e. a predetermined maximum level of spending on pharmaceuticals for a specific period) in relation to the cost and quality of prescriptions based on panel data from 440 German outpatient physicians over 2005-2011. They estimated that the mean utilization of drug budgets was 92.3 percent across all physicians and years. The results from their regression model provide evidence that drug budgets' utilization affects the cost and quality of prescribing. By analysing indicators of prescribing costs, the authors found

that a rise in imposing drug budgets in the previous year leads to a significant increase in generic share and concentration among generic brands. However, the number of prescriptions per visit and the number of branded prescriptions remain unaffected. By examining the indicators of prescribing quality, they found an increase in drug budgets raises the share of prescriptions deemed inappropriate for the elderly while reducing the concentration among therapeutic substances.

The above evidence shows that the utilization of revenue restrictions often leads to reductions in the volume of service prescribed and the total pharmaceutical industry spend. Physicians under such restrictions also tend to prescribe less expensive pharmaceuticals. Furthermore, the restrictions implemented at the individual level are more effective in controlling costs than those imposed at a global or regional level. However, some of the above evidence also shows that global budgets only reduce total healthcare expenditure in the short term and may lead to an increase in service provision as the revenue restriction drives competition between physicians. In Chapters 3 and 4 I show that my results are consistent with observations that imposing the physician specific budget restriction is effective in (1) limiting the growth of quantity of service provided per patient and total health spending and (2) improving the efficiency of healthcare system.

2.3.3.2 Quantity Restriction

Quantity restriction in this literature review refers to the maximum quantity of health service that a physician can prescribe to a given patient. The purpose of introducing a quantity restriction is to maintain patients' health (e.g. to prevent addiction to or reliance on certain types of drugs) and to control costs. This type of restriction has been widely used in terms of limiting physician drug prescriptions (Stabile et al., 2013). In the following section I discuss some examples of implementing drug prescription limits and their impact. In Chapters 3 and 4, I show that my research findings reflect some of these results.

Recently, the misuse of and addiction to opioids (e.g. pain relievers, heroin, and fentanyl) has become a national crisis impacting on public health and social welfare in developed countries such as the US and the UK (National Institute on Drug Abuse, 2019). In the US, opioid-related deaths have increased nearly fourfold from 1999 to 2014 (Jones et

al., 2018). In 2018, opioids overdose increased a further 30% compared to 2017 while the country had already provided 30 times more opioid pain relief medication than necessary (BBC, 2018). In the UK, the NHS has been accused of fuelling a rise in opioid addiction; from 2007 to 2017, opioid prescriptions increased by 10 million and over 50% of drug overdose deaths were related to opioids (Rhodes, 2018). To counter the aforementioned adverse effects, states such as New York and Massachusetts limited the length of initial prescriptions of opioid pain medication – typically to fewer than seven days (Scully et al., 2018). Moreover, prescription drug monitoring programmes have been introduced on an increasingly common basis (Scully et al., 2018).

There is some evidence that the utilization of opioid prescription limits has had a positive impact. For instance, Express Scripts launched a one-year pilot programme limiting new opioid users to seven-day prescriptions (Salter, 2017). The analysis of 106,000 patients in the pilot programme showed a 38% reduction in hospitalizations and a 40% reduction in emergency room visits compared to a control group (Chua, et al., 2019). Another research studying the impact of a similar programme by CVS Caremark also found that physicians have reduced opioid prescriptions by 17% a few years after the programme was initiated and directed patients to other forms of pain management, including physical and cognitive behavioural therapy (Chua et al., 2019).

However, there is also some evidence showing that the utilization of opioid prescription limits has failed to achieve its intended effects (i.e. preventing excessive opioid prescribing and maintaining adequate pain control). For example, a recent report by the U.S. Substance Abuse and Mental Health Services Administration in 2014 found that about half of those who misused prescribed pain killers in the previous year obtained the pain relievers from a friend or relative for free. Moreover, around 10% frequent users of pain relievers said they have bought pain relievers from either friends or relatives. A total of 22% of patients said they obtained their drugs from one doctor. Finally, around 4% of painkiller users indicated that they obtained their most recently misused pain relievers from sources other than their physicians (Lipari and Hughes, 2017).

Chua et al. (2019) have also argued that imposing limits on opioid prescribing is unlikely to achieve its intended effects (i.e. preventing excessive opioid prescribing and maintaining adequate pain control) because a uniform restriction may not satisfy patients' heterogeneous combinations of opioid use and pain needs. In the presence of this

heterogeneity, the restriction may either be set too high to reduce excessive prescribing or set too low to avoid the potential for inadequate pain control.

When the regulator imposes a certain opioid prescribing limit (e.g. a 5-day supply restriction), Kao-Ping Chua et al. (2019) suggest that all patients can be roughly divided into three groups: (1) patients who would initially be provided an amount equal to or less than the given limit; (2) patients who would receive an amount that is more than the given limit but consume the given limited amount or less; (3) patients who would receive and consume an amount more than the given limit. The authors suggest that the utilization of a prescription limit would not affect the volume of opioid prescription and pain control in the first group but could prevent some excessive prescribing in the second group. In the last group, the restriction may prevent excessive prescribing but increase the potential for inadequate pain control. Overall, they conclude that imposing limits on opioid prescription will likely have some positive effects, particularly among physicians who would not change their practice in the absence of a mandate. However, this desirable effect tends to be small and may even be offset by undesirable effects in relation to pain control.

Fan et al. (1998) have compared the quantity of service supplied by fee-for-service physicians under quotas and a global budget cap. They found in both symmetric (SNE) and strong symmetric Nash equilibrium (SSNE), physicians would provide a greater quantity of health service under a global expenditure cap than under a quantity restriction when physicians are reimbursed at the same budget level. Their experimental results show that physicians' profit maximization prediction is well captured by subjects' behaviour under the quantity restriction. However, under the expenditure cap only SNE predictions are supported.

To conclude, this subsection provides some evidence regarding the impact of utilizing prescription limits (quantity restriction) to restrict excessive medical service provision. This evidence suggests that the implementation of a prescription limit does indeed incentivise physicians to reduce prescriptions. This decreases patients' reliance on certain types of drugs and reduces their total costs incurred. However, given the heterogeneity of patient demands, a uniform limit may not always control excessive drug prescribing while lead to inadequate pain control for some patients. Moreover, given the same level of spending, laboratory experiments show physicians under an expenditure

cap would provide more service than under quantity restriction (Fan et al., 1998; Benstetter and Wambach, 2006).

My numerical results are consistent with some of the aforementioned evidence. For instance, my numerical findings are in line with the findings that imposing quantity rationing (1) does not always help controlling excessive prescribing, (2) may control medical service provision and result in a social welfare improvement, and (3) may not be as effective as imposing revenue restrictions.

2.4 Conclusion

In this survey of the literature, I have provided a comprehensive review of the theory and evidence relating to altruism, physician behaviour, the design of physician payment mechanisms and the various implementations of physician cost control restrictions. Several core themes emerged: First, scarcity plays a key role in allocating health resources. Consumers have unlimited demands but healthcare providers and social planners are restricted in their resources or inclination to engage with patients; Second, physicians are not solely extrinsically motivated but care about their patients' welfare. This additional motivation has important implications in terms of the design of social objectives and physician remuneration. I show physician altruism has often been used in models analysing physician behaviour. This assumption has also been supported by different types of evidence; Third, information asymmetries lead to inefficiencies and these costs are borne by society. Information asymmetries can take varies forms and are associated with problems such as moral hazard and adverse selection. Fourth, commonly used physician remuneration scheme such as fee-for-service, capitation or cost-reimbursement may fail to provide physicians with the correct incentives (McGuire, 2000; Léger, 2008). This motivates a thorough understanding of physicians' incentives under each individual payment system (McGuire, 2000). The literature shows that physicians increase quantities of service provided in line with price and may overproduce under the fee-for-service system (Léger, 2008). Under full cost-reimbursement, physicians often provide the constrained optimal level of quality while exerting zero cost-reduction efforts (Ma, 1998; Chalkley and Malcomson, 1998a). In contrast, under the prospective/capitation system, physicians often under-produce (McGuire, 2000; Léger, 2008). Moreover, they provide the first best cost reduction

efforts and their quality of service provided increases with capitation rate when physicians do not select their practice sizes.

Fifth, this survey discusses the optimal design of physician payment in an environment with respective types of information asymmetry. While a menu of contracts has often been proposed as the optimal physician remuneration scheme in the presence of adverse selection (Chalkley and Malcolmson, 2002; Makris, 2009; Makris and Siciliani, 2013), the same allocation under this remuneration can be indirectly implemented by a single payment based on the outcome of physicians' production. Moreover, the tension caused by incentive compatibility between quantities and non-negative profit (Choné and Ma, 2011), the unachievable single crossing property, the incentive to renegotiate and select low costly patients indicate a single physician payment might be a better option. There is also limited evidence supporting the implementation of a menu of contracts.

Since a single remuneration scheme may not be able to align all physicians' incentives while a menu of contracts is difficult to implement, it is reasonable to consider an alternative method- introducing a single physician payment system some cost-control restrictions. In particular, I discussed a small literature analysing the impact of introducing alternative physician cost control restrictions on the most commonly used fee-for-service system (Neudeck, 1991; Fan, et al., 1998; Benstetter and Wambach, 2006). This literature shows that introducing FFS system a quantity rationing maybe effective in improving patients' health benefits. Moreover, imposing an expenditure cap on FFS system induce physicians to produce more service than imposing a quota, if physicians are provided with the same level of budget. These results are shown to be consistent with some empirical observations. However, this literature neither considers main features of the healthcare market nor does it recommend the optimal form of physician payment. As a result, the main objective of the ensuing chapters is to address this gap, analysing physicians' behaviour under a mixture of standard remuneration and cost-containing instruments, while recommending the optimal payment system without assuming away key features of the healthcare market.

Despite the aforementioned contributions, this literature survey has some limitations. First, while altruistic preference plays an essential role in modelling physician behaviour and the design of their payment, this review only highlights the main approach of modelling and provides some evidence regarding the existence of physician altruism.

Research on the modelling of altruism in non-health related fields and its impact on economic agents' behaviour are largely ignored. Second, this literature review does not consider the emerging literature analysing physician behaviour under either pay-for-performance or cost-sharing. The former system was proposed in recent years and has been shown to lack an empirical foundation and evaluation (Mannion and Davies, 2008). Developed countries have been reluctant to use the cost sharing system (e.g. the UK and US) as it may lead to cost padding and needs to be closely monitored (Chalkley and Malcomson, 2002). Finally, this chapter also does not discuss the effect of introducing cost-control mechanisms in a significant detail. The reason for delaying a systematic review of this literature is threefold: (1) There is little consensus on the definition of the cost-control instruments such as quantity rationing (Keliddar et al., 2017). For instance, quantity restriction not only means limiting the amount of services, equipment, and time provided to patients, but also reflects constraints regarding the number of contracts offered to doctors (Neudeck, 1991); (2) My study focuses exclusively on the quantity rationing or revenue cap imposed at per patient or physician level. As a result, studies analysing the effect of introducing cost containing instruments at a regional or global level are largely excluded; (3) Cost control instruments have been imposed worldwide and under different health insurance schemes. In this survey, I mainly consider those instruments that have been imposed on fee-for-service systems.

In addition to the above limitations, there are several interesting questions which have received relatively little attention, or have yet to be studied. These include whether the theoretical evaluation of cost control policies can be more inclusive? Could models in the literature extend to the environment with multi-dimensional adverse selection and moral hazard? How would results be changed if patients can negotiate with their physician about the quantity of service to be supplied or if either providers or the regulator are risk-averse?

Most studies only analyse the FFS physician behaviour under one cost control policy (e.g. Neudeck, 1991; Mousgeot and Naegelen, 2005; Benstetter and Wambach, 2006). Fan et al. (1998) have compared the physician's supply of medical service under a quota and an expenditure cap. The most recent study, Wu et al. (2018), has investigated the optimal design of a menu of contracts when the regulator can use fee-for-service, capitation and pay-for-performance. Hence, it would be interesting to see a model which

is able to compare the impact of introducing the FFS system capitation, salary, pay-for-performance, cost-sharing, quantity and revenue rationing simultaneously.

Ma and McGuire (1997) have derived the optimal insurance for patients and the optimal remuneration method for physicians when neither the patient selected medical service quantity nor the physician decided effort are contractible. Lien et al. (2004) then generalized this analysis by introducing the possibility of persuasion, the physician-set quantity ceiling and unknown patients' preferences. While these studies consider patient decisions regarding the quantity of service supplied and the design of their payment, their design of physician payment only includes fee-for-service and a lump sum payment. The impact of introducing different types of cost control policies are ignored. Moreover, their models do not consider essential features of the healthcare market such as physician altruism and unknown heterogeneity. Hence, it will be interesting to see how the optimal design of patient insurance and physician payment should be adjusted if the above characteristics are considered.

The contract theory of economics shows that risk preferences influence a variety of behaviours under the condition of uncertainty (Holmstrom, 1979). Given the prevalence of such conditions regarding questions of health care provision, there is a rich literature on risk attitudes in the health sector, focused on measuring patient preferences and propensity towards risk (Arrieta et al., 2017; Liu and Ma, 2013). While physicians' risk attitudes also play an essential role in determining health service provision, this issue has received comparatively little attention (Arrieta et al., 2017; Galizzi et al., 2013). Eeckhoudt et al. (1985) first developed a simple model analysing how a physician's decision concerning the provision of treatment intensity is affected by his level of risk aversion and technological innovations. However, their setup did not consider many essential features of the healthcare market such as asymmetric information, physician altruism, physician and patient heterogeneities, as well as cost-control policies. Recent studies such as Arrieta et al. (2017), Bories et al. (2018) and Lawton et al. (2019) have discussed how physicians' risk attitudes affect their diagnostic and therapeutic patient decisions. However, these studies and the literature cited in these papers are mainly empirical or experimental. Accordingly, it would be interesting to develop analytical frameworks that can analyse how physicians' decisions regarding their practice size, effort, quantity or the quality of service provided is affected by asymmetric information as well as altruistic and risk-averse preferences.

Chapter 3. The Optimal Design of Fee-for-Service Contract 1

3.1 Introduction

In the first chapter of the thesis I emphasise that the importance of improving health resource usage efficiency as health expenditure has already constituted a significant proportion of GDP and the trend is continuously upward. It is shown that primary care physicians (PCPs) play a vital role in directing health resource allocation, therefore, it is important to investigate the design of physicians' incentive contracting. In particular, I focus on analysing incentive contracting on the basis of fee-for-service, which is one of the most commonly used PCP payment systems in the US, Canada, Belgium and many other countries (Emery et al., 1999; Kringos et al., 2015; Porter and Kaplan, 2014).

The second chapter discusses the literature, which analyses physician altruism and the optimal design of physician incentives. This literature provides a strong support for the existence of provider altruism;²³ as a result, I follow the well-established approach by Ellis and McGuire (1986) and introduce physician altruism in the setup in this chapter. This literature survey also shows that it is natural to propose a menu of contract across physicians if they have private knowledge about their degree of altruism and patients' illness severity. However, as discussed in the survey, the allocation of a menu of contracts can be indirectly implemented by a single payment system depending on an output of the production process and there are many difficulties associated with implementing a menu of contracts directly. Hence, this chapter focuses on the design of an optimal single health insurance scheme.

While providing heterogeneous physicians with a single payment is popular among developed economies, results from the survey suggest that it fails to align different providers' incentives and drives up health expenditure. As a result, cost control policies²⁴ should be introduced to enhance efficiency. In the literature review, I find that analytical frameworks used in these studies do not consider some of the key features of health insurance systems. In this chapter, however, I analyse the optimal design of a

²³ 'Provider altruism' refers to physicians not only caring about financial returns but also having concerns about their patients' health benefits (Ellis and McGuire, 1986, p.131).

²⁴ I also refer to these cost-control policies as "cost-controlling" or "cost-containing" instruments hereafter.

single compensation scheme given that physicians are paid on a FFS basis while analysing the impact of information asymmetry and the utilization of physician cost control restrictions. Intuitively, in the presence of unknown and heterogeneous types of PCPs, not all physicians' incentives will be affected by the introduction of cost-containing instruments. As a result, social planners can align incentives of different types of physicians by using cost-controlling methods and price respectively.

This chapter makes three distinctive contributions: First, my setup captures the key features of health insurance schemes in developed countries. For instance, I have modelled patient insurance, physician altruism and issues relating to information asymmetry; Second, comparative static analysis based on a numerical exercise allows us to recommend the optimal adjustment of the HI scheme to current trends such as ageing populations, declining medical altruism and technological advances; Finally, my analysis provides a framework for comparing different healthcare expenditure control instruments.

In line with the results obtained from my literature survey, this chapter shows that an optimal pricing system is not able to align physicians' incentives and therefore fails to achieve Pareto-efficiency. This study also shows that current trends such as ageing populations, the increase in the proportion of the less altruistic doctors and medical innovations have further distorted physicians' incentives and driven up total healthcare expenditures. In addition, I find in my numerical analysis that the commonly used cost control policies such as quantity rationing, expenditure caps and capitation can be imposed on FFS payment systems to improve social well-being as measured by the sum of patients' health benefits and the costs associated with running a healthcare system. In particular, it is shown that introducing a negative capitation²⁵ to pricing system yields the highest social well-being.

The main implication of this result is that the regulator can use the "bonding contract" or the "franchise agreement" to contain healthcare spending and to align different types of doctor incentives. Specifically, the negative reimbursement per patient is used to limit high intrinsically motivated physicians' overproduction and to extract informational rent

²⁵ I obtain this result since I do not consider the administrative/fixed costs of treating patients. Once these costs are introduced, it can be shown that the optimal payment system requires a positive capitation that does not fully cover the fixed/administrative costs associated with enrolling an additional patient.

from extrinsically motivated doctors. On the other hand, the price per unit of service is raised to incentivise medical service production by low intrinsically motivated doctors.

To capture these notions, I developed a model in which the regulator (the principal) chooses a per patient insurance premium and reimbursement price per unit of service provided to maximize social welfare measured by summing up all individuals' utilities. Physicians subsequently decide the quantity of medical service provided per patient to maximize their weighted sum of treated patients' health benefits and their financial returns under the constraint that their profit is non-negative. A degree of the physician's altruism is represented by the weight in the doctor's utility function attached to his patients' health benefits (Galizzi et al., 2015). This approach is applied in the study not only because it has been well-established in theory, but also because fieldwork and laboratory experiments have provided evidence for physician altruism. For instance, Jacobsen et al. (2011), Smith et al. (2012) and Kolstad and Lindkvist (2012) found that medical students tend to donate significantly more of their endowments than students from other subjects. Brosig - Koch et al. (2017), Godager and Wiesen (2013), and Godager et al. (2016) show in their laboratory experiments that physicians' decisions on medical service quantity is affected significantly by information regarding patients' health benefits. Finally, based on physicians' response to price variations, the regulator selects an optimal price (insurance premium) and decides whether cost-control instruments should be used to further enhance efficiency. While the cost containing instruments analysed in this chapter are relatively crude, they represent some important applications in reality. For instance, the quantity rationing and expenditure restriction can be respectively interpreted as the current drug/pain killer prescription limit imposed per patient, as well as the expenditure ceiling introduced in quality adjusted life years (QALY) or diagnosis-related groups (DRG).

In the first part of this chapter, I solve the physician's optimization problem and characterize his supply of medical service per type of patient as a function of price. Specifically, these supply curves are solved when the number of patients allocated is fixed,²⁶ while the fraction of alternative types of patients and the degree of physician altruism are assumed exogenously given. I find that the physician increases his medical

²⁶ To make the analysis tractable, I first ignore the fact that physicians in reality may be able to select their roster sizes. I also ignore competition between physicians for patients. As a result, the number of patients allocated per physician does not depend on the intensity of medical service provided by a given physician.

service provided per patient in line with price but at a different rate depending on whether his profit is binding. Less healthy individuals are always provided with a higher intensity of medical service. Finally, the medical service provided per patient is increasing whereas the profit is non-increasing in relation to the degree of physician altruism.

Next, given the physician's supply functions to alternative types of patient, I derive the welfare-maximizing price and insurance premium in the presence of moral hazard and adverse selection. The former issue is captured by randomized patient health benefits whereas the latter is represented by an unknown degree of physician altruism. For the sake of simplicity, only two types of physicians are introduced – namely, more or less altruistic – while their fractions are also assumed to be fixed. Because of moral hazard and adverse selection, the optimal pricing scheme is incapable of bridging the difference in intrinsic and extrinsic motivation across medical staff, resulting in medical service over- and underproduction by strongly (H -) and weakly (L -) motivated doctors respectively. The over (and under) production is defined by comparing the current production of $H(L)$ - type physicians with what would be produced if there are only $H(L)$ - type doctors in the economy.

In the third part of this study I present a detailed comparative static analysis of an optimal HI scheme by employing a numerical simulation (due to the large system and non-differentiable physician supply functions), which enables a solution of the optimal price and insurance premium as a function of the system's exogenous parameters (i.e. patients' initial health level, the proportion of the less altruistic physicians etc.). As variations with respect to some of these parameters can be interpreted as the evolution of recent trends (e.g. ageing or scientific progress etc.), the study recommends the best response of the HI scheme to these trends. My numerical analysis suggests that the optimal HI scheme would raise the price and insurance premium per person as population ages, as the proportion of less altruistic doctors increases, and as technology advances²⁷. However, the optimal HI scheme would reduce the price and insurance premium as healthy life-styles such as eating vegetables and walking 10,000 steps every day are

²⁷ In the model these are formally captured by an increase in the share of the less healthy patients, the proportion of less altruistic providers, and a technology parameter.

promoted. These predictions are found to match related trends in reality (Bauchner and Fontanarosa, 2018; Papanicolas et al., 2018).

In the last part of this chapter, I evaluate three simple cost-control instruments – per patient quantity rationing, per patient expenditure caps, and capitation that might be imposed on the pricing system – to further improve efficiency. My result shows that all three instruments can be imposed on the optimal pricing system to enhance efficiency, but to different extents; I found that blending price system with “negative capitation” as a bonding contract dominates the other instruments, as the payment by doctors helps to align physicians’ incentives and extract their informational rent. Moreover, imposing a per patient revenue cap is superior to per patient quantity restriction, as the latter is not able to react to providers’ local knowledge.

The remainder of this chapter is structured as follows. Section 3.2 provides a short summary of the most relevant studies while Section 3.3 introduces the model. Section 3.4 solves the physician’s constrained optimization problem as a function of a price set by the regulator. Section 3.5 sets up the optimization of the HI scheme and proves the existence of the solution while providing a comparative static analysis of the optimal HI scheme as a function of exogenous parameters based on numerical experiments. Finally, Section 3.6 evaluates different cost control instruments and derives the optimal form of a HI scheme. Sections 3.7 and 3.8 conclude the chapter and outlining some potential ways of extending this analysis. Section 3.9 provides all proofs.

3.2 Related Literature

This study is closely related to the literature which analyses the impact of introducing cost-controlling methods in a fee-for-service system (Benstetter and Wambach, 2006; Fan et al., 1998). As discussed in the previous survey, major issues discussed in this literature are (1) how physicians’ incentives are affected, and (2) how is the impact of using cost-contain instruments measured? Most papers from this literature have investigated the effect of cost-control methods on a self-interested physician’s decision on medical service quantity and the overall impact on society, which are measured respectively as: the total quantity of medical service produced under the same budget level (Fan et al. 1998; Benstetter and Wambach, 2006); patients’ total health benefits (Neudeck, 1991); and the sum of both patients and providers’ well-being (Mougeot and

Naegelen, 2005). However, these studies have ignored some main features of the primary healthcare market such as the application of health insurance schemes, information asymmetries, as well as physician altruism.

Altogether, this chapter contributes to the literature by introducing in the existing setups moral hazard and adverse selection problems between the regulator and PCPs, patient heterogeneity, physician altruism and the publicly financed health insurance system. This is important because the above extensions help us to understand FFS-based PCPs' decisions concerning the supply of medical service and their incentive contracting in a more realistic manner. My contributions also enable us to evaluate different cost-control mechanisms by comparing their respective generation of social well-being. This allows us to make both qualitative and quantitative recommendations for future PCP remuneration reforms, including which type of cost-control instruments should be used and the policies reflect the maximum quantity or expenditure that PCPs are allowed to produce or spend respectively.

Since inefficient health resource allocation mainly arises from asymmetric information between providers and patients, this chapter also relates to the literature which has investigated physician incentive contracting in the presence of moral hazard and adverse selection (Choné and Ma, 2011; Jack, 2005; Kantarevic and Kralj, 2016). The main conclusion is that physicians should be offered a menu comprising negative-related retrospective and prospective components across contracts. However, this result is typically based on the setup where the physician only treats a single patient. When the physician has multiple patients with different severities of illness, their aggregated health benefits may violate a single crossing property. In reality, a menu of contracts have rarely been used to compensate physicians as it may lead to a two-tier healthcare system (Kay, 2002) and leave high informational rent for some physicians (Croxson et al., 2001). Furthermore, under this payment system the regulator and physicians have an incentive to renegotiate their contracts when physicians' private information is revealed by their selection of contracts. Finally, the literature on "the delegation principle" shows that the allocation of a contract menu can be implemented by a single payment system which depends on an outcome of the production process.

3.3 Model

Consider an economy consisting of a benevolent regulator and two types of risk-neutral individuals, namely patients and doctors. Patients are indexed by the subscript $i = 1, 2, \dots, N$ and doctors by the superscript $j = 1, 2, \dots, M$ where the parameters N, M and $n = N, M$ are sufficiently large to apply the law of large numbers whenever needed. The benevolent regulator determines the health insurance scheme (hereafter HI), including the insurance premium and the price of medical services.

From patients' perspectives, all doctors appear alike. As a result, I assume that patients are allocated randomly to doctors. Patients have identical preferences over their health level and their health insurance costs. These preferences are represented by the linear utility function:

$$u(h, \tau) = \hat{h} - \tau \quad (23)$$

where \hat{h} denotes the indicator of patient's health level and τ the health insurance premium.²⁸ The patient's health indicator is itself a function of three variables, $\hat{h}(z_i, q_i, \xi_i)$, where z_i denotes the initial health level of patient i , q_i the quantity of medical service provided to him, and ξ_i denotes the realization of random shock. The initial health status of an individual is itself a random variable $z_i \in \{0, z\}$ with $0 < z < 1$ and $Pr[z_i = 0] = \beta$. For any individual, the realization of z_i is assumed to be observable only by the doctor. In contrast, the quantity of health service is assumed to be verifiable. Finally, the random shock ξ_i are assumed to be *i.i.d* log-normally distributed across patients with $\xi_i \sim \ln\mathcal{N}\left(-\frac{1}{2}, 1\right)$,²⁹ hence the distribution of ξ_i implies $\mathbb{E}[\xi_i] = 1$. For parsimony of analysis, I impose a specific functional form:

$$\hat{h}(z, q, \xi) = h(z, q)\xi \quad (24)$$

Where $h(z, q)$ denotes the real health of a patient and it takes the form $h(z, q) = z + (1 - z)f(q)$. $f(q)$ is a strictly increasing concave function with $f'(0) = +\infty$, $f(0) = 0$ and $\lim_{q \rightarrow +\infty} f(q) = 1$.³⁰ Accordingly, an individual's health increases with his initial

²⁸ For similar approach, see Ellis and McGuire (1990) and Lee (1995).

²⁹ A random variable X is $\ln\mathcal{N}(\mu, \sigma)$ if $\ln(X)$ is $\mathcal{N}(\mu, \sigma)$. It is well known that $X > 0$, $\mathbb{E}[X] = \exp\left[\mu + \frac{1}{2}\sigma^2\right]$.

³⁰ While $h \in [0, 1]$ and $\hat{h} \in [0, +\infty)$, we can find a function $g(\cdot)$ such that $g(\hat{h}) \rightarrow h$.

health and the quantity of health service being provided. The marginal patient benefits of consuming health service decrease with the patient initial health and quantity of service provided respectively. Moreover, individual's health has both lower and upper bounds. The introduction of the random shock ξ_i prevents that patient or the health insurance scheme from deducting the patient initial health status by looking at his quantity of service provided.

Each doctor provides medical services to n individuals. The set of patients allocated to doctor j is indexed by $I(j) = \{i = (j - 1)n + l \mid l = 1, \dots, n\}$.³¹ Moreover, from the foregoing description of patients, doctor j 's clientele is fully described by the vector of initial health status associated with each individual denoted hereafter by the vector $\mathbf{z}^j = (z_{(j-1)n+1}, \dots, z_{nj})$. In analogy, $\mathbf{q}^j = (q_{(j-1)n+1}, \dots, q_{nj})$ denotes the vector of medical services provided by doctor j to his patients. This production generates the costs:

$$\tilde{C}(\mathbf{q}^j) = C\left(\sum_{i \in I(j)} q_i\right) \quad (25)$$

where $C(\cdot)$ is a strictly increasing and convex function which satisfies $C(0) = C'(0) = 0$ and $C'''(\cdot) > 0$. For a given per unit price of medical service- p , doctor j who produces the vector of services \mathbf{q}^j obtains the profit:

$$\Pi(\mathbf{q}^j, p) = p \sum_{i \in I(j)} q_i - C\left(\sum_{i \in I(j)} q_i\right) \quad (26)$$

Doctors have preferences that are others regarding³², meaning that physicians are not only motivated by their financial returns but the health benefits of their patients. Since the physician can observe the initial health of each of his patient, the health outcome of his patients is measured by:

$$H(\mathbf{z}^j, \mathbf{q}^j) = \sum_{i \in I(j)} h(z_i, q_i) \quad (27)$$

³¹ For simplicity, I denote $(j - 1) \times n + 1$ and $n \times j$ by $(j - 1)n + 1$ and nj respectively hereafter.

³² "Preferences over another individual's payoffs, in addition to one's own". The other regarding preferences is considered to be altruistic, if one's utility is increasing with another person's payoff. For more details, see Kaplow (2010) and the literature cited therein.

Using this notation, I assume that a doctor's preferences can be represented by the utility function:³³

$$V(H, \Pi, \alpha^j) = \alpha^j H + \Pi \quad (28)$$

where $\alpha^j > 0$ measures that doctor's concern for his patient³⁴ which is his private information. For the sake of simplicity, I assume that the altruism level α^j takes only two values - $\alpha^j \in \{\alpha^L, \alpha^H\}$ with $0 < \alpha^L < \alpha^H \leq 1$ and denote by Λ the proportion of α^L -type doctors. Finally, a non-negativity restriction is imposed on the profit of doctors. This restriction can be interpreted as a legal constraint in accordance with the idea that work must be paid.

The benevolent regulator determines the health insurance scheme in order to maximize expected social welfare under a self-financing constraint. I follow the standard approach of regulation and procurement (Laffont and Tirole, 1993), which assumes the welfare sums up the utilities of all the participants in the industry, including the provider.³⁵ Altogether, let $\mathbf{u} = (u_1, u_2, \dots, u_N)$ denote the vector of a patient's utilities and $\mathbf{\Pi} = (\Pi(\mathbf{q}^1, p), \Pi(\mathbf{q}^2, p), \dots, \Pi(\mathbf{q}^M, p))$ the vector of a doctor's profit. Social welfare is:

$$W(\mathbf{u}, \mathbf{\Pi}) = \sum_{i=1}^N u_i + \sum_{j=1}^M \Pi_j \quad (29)$$

The self-financing constraint on the HI scheme requires:

$$N\tau = (1 + \lambda) \sum_{i=1}^N pq_i \quad (30)$$

where the parameter $\lambda > 0$ denotes the shadow price associated with the HI scheme. Intuitively, it captures the administrative costs associated with collecting the insurance premium, verifying the medical services provided by doctors and running healthcare system. This equation shows that the money collected from all patients has to cover

³³ For more examples, see Woodward and Warren-Boulton (1984), Ellis and McGuire (1986), Lee (1995), Hennig-Schmidt et al. (2011), and Galizzi et al. (2015).

³⁴ From the point view of an individual doctor, the number of patients n is exogenously given. Accordingly, doctor j 's preferences can be equivalently described by $V(H, \Pi, \alpha^j) = \alpha^j \bar{H} + \bar{\Pi}$ where \bar{H} and $\bar{\Pi}$ denote the respective average health and profit per patient.

³⁵ For more examples, see Buchanan (1988), Chalkley and Malcomson (1998), Jack (2005), and Jelovac and Kembou Nzale (2017).

payment to doctors and the shadow costs generated by the process of collecting insurance premium.

In order to gain further insight regarding society's objectives, observe that $W(\mathbf{u}, \boldsymbol{\Pi})$ can be written differently. From the equation (23) and (26), $\sum_{i=1}^N u_i = \sum_{i=1}^N \hat{h}_i - N\tau$ and $\sum_{j=1}^M \Pi_j = \sum_{j=1}^M [p \sum_{i \in I(j)} q_i - C(\sum_{i \in I(j)} q_i)]$. Hence using the equation (30), substituting the definition of the respective doctor's profit (26) and cancelling common terms, we obtain

$$W(\mathbf{u}, \boldsymbol{\Pi}) = \sum_{i=1}^N \hat{h}_i - \left[\lambda \sum_{i=1}^N pq_i + \sum_{j=1}^M C\left(\sum_{i \in I(j)} q_i\right) \right] \quad (31)$$

where $\sum_{i=1}^N \hat{h}_i$ measures total expected health benefits across patients. In this formulation of welfare function, the term $\lambda \sum_{i=1}^N pq_i + \sum_{j=1}^M C(\sum_{i \in I(j)} q_i)$ represents the total health expenditure of the medical system. These expenditures sum up the costs of collecting the insurance premium, the costs of verifying the medical services through the HI scheme and the doctor's actual costs of providing medical services. Altogether, the regulator problem is to find a HI scheme which maximizes (31) considering the doctor's choices with respect to the provision of medical services.

To conclude this section, a few remarks are in order. First, while I used a specific form of patient health function $h(z, q) = z + (1 - z)f(q)$, results derived from the ensuing sections does not rely on this specification. My results will be hold as long as health function $h(z, q)$ satisfies $h_z, h_q > 0, h_{zq}, h_{qq} < 0, h_q(z, 0) = +\infty$ and $\lim_{q \rightarrow +\infty} h(z, q) = 1$. However, using this specification helps to provide simple intuitions for my results.

Second, the multiplication of the random shock element $\xi \sim \ln \mathcal{N}\left(-\frac{1}{2}, 1\right)$ to the health function (24) ensures that the value of patient health remains positive. In contrast to Ellis and McGuire (1986) and Barham and Milliken (2015), who respectively assume an inversely U-shaped and a Bell-shaped health function, I assume that health function $h(z, q)$ is increasing and concave in q . This assumption reflects the idea that health service is provided in order to improve the patient's health. Moreover, when physicians are allocated the same number of patients, the more altruistic physician provides more service to a patient than his less altruistic colleague (i.e. $\frac{\partial q_t^*}{\partial \alpha} > 0$).

Third, the equation (29) is chosen as the welfare function as the payer is assumed to be a benevolent regulator who (e.g. the NHS in the UK) cares about all participants' well-being. However, there are many other studies which only maximizes patients' utilities. These studies assume that the payer is a risk-neutral insurance company or a regulator that operates in a competitive market (e.g. Wu et al., 2018). I do not use this approach as healthcare markets in most developed countries are not competitive. Finally, I exclude the altruistic benefits from social welfare function (29) to avoid the issue of double counting (Chalkley and Malcomson, 1998; Hammond, 1987; Jack, 2005). As Chalkley and Malcomson (1998) pointed out, "there is strong case to exclude the benevolent component from social welfare on the grounds that benevolence represents a desire to do what is in the social interest and, as such, should have no role in determining what the social interest is" (p. 6).

3.4 The Physician's Optimization

In this section, I take the HI policy (p, τ) as given and solve the decision problem of a physician characterised by the generic parameter α . From the previous section, the doctor has only two types of patients; n_G of them are characterised by an initial health level $z_G = z$ while the remaining n_B patients have the initial health level $z_B = 0$, denoting the set of types by $T = \{B, G\}$. Slightly abusing the notation, I write $q_t \geq 0$ for $t \in T$ and obtain the doctor's constrained optimization problem:

$$\begin{aligned} U &= \max_{q_t} \sum_{t \in T} n_t [\alpha(z_t + (1 - z_t)f(q_t)) + pq_t] - C \left(\sum_{t \in T} n_t q_t \right) \\ &\quad p \sum_{t \in T} n_t q_t - C \left(\sum_{t \in T} n_t q_t \right) \geq 0 \end{aligned} \tag{32}$$

$$q_t \geq 0$$

where the first line represents that the physician selects medical service (q_B, q_G) to maximize his utility defined in the equation (28). The second and third lines show that the physician's profit defined in the equation (26) and the medical service quantity

supplied to B - and G -type patients are non-negative. Writing the Lagrangian associated with the doctor's optimization problem (32) yields:

$$\begin{aligned}\mathcal{L} = & \alpha \sum_{t \in T} n_t [z_t + (1 - z_t)f(q_t)] + \sum_{t \in T} \mu_t q_t \\ & + (1 + \eta) \left[p \sum_{t \in T} n_t q_t - C \left(\sum_{t \in T} n_t q_t \right) \right]\end{aligned}\tag{33}$$

where η is the Lagrange multiplier for the profit constraint and the μ_t are the respective multipliers of the non-negative service restrictions. Keeping in mind that $z_B = 0$ and $z_G = z$, the first order conditions of (33) become:

$$\left\{ \begin{array}{l} \alpha n_B f'(q_B) + (1 + \eta)n_B[p - C'(Q)] + \mu_B = 0 \\ \alpha(1 - z)n_G f'(q_G) + (1 + \eta)n_G[p - C'(Q)] + \mu_G = 0 \\ \eta[pQ - C(Q)] = 0 \\ \mu_B q_B = 0 \\ \mu_G q_G = 0 \\ \eta, \mu_B, \mu_G \geq 0 \end{array} \right.\tag{34}$$

where $Q = n_B q_B + n_G q_G$ denotes the total production of medical services by the doctor. The first two equations in (34) are the first-order conditions with respect to q_B and q_G . The next three lines in (34) are the corresponding complementary slackness conditions associated with the inequalities in the optimization problem. Finally, the last line provides restrictions on the respective multipliers of non-negative requirements for profit and services.

In order to simplify notation, I use f'_G as $f'(q_G)$, f'_B as $f'(q_B)$, and C, C', C'' as $C(Q), C'(Q)$ and $C''(Q)$ respectively whenever this is possible without confusion. Moreover, I denote the solution to the equation system (34) by the superscript “*”. Keeping in mind that the assumptions $f'(0) = +\infty$ and $\alpha > 0$, as a result the physician will always provide $q_B^*, q_G^* > 0$ whenever his service provision generates income (i.e. $p > 0$). Accordingly, the above system (34) can be simplified as:

$$\left\{ \begin{array}{l} \alpha f'_B + (1 + \eta)[p - C'(Q)] = 0 \\ \alpha(1 - z)f'_G + (1 + \eta)[p - C'(Q)] = 0 \\ \eta[pQ - C(Q)] = 0 \\ \eta \geq 0 \end{array} \right. \quad (35)$$

Lemma 3.1 For all $p > 0$, $q_B^* > q_G^* > 0$.

Proof. From the first two equations in (35), we obtain $f'_B = (1 - z)f'_G$. As $f'' < 0$ and $0 < z < 1$, we have $q_B^* > q_G^* > 0$. ■

Intuitively, at the point where the doctor provides the same service to both types of patients (i.e. $q_B^* = q_G^*$), price and marginal costs are the same but the marginal return from the altruistic component is larger for less healthy patients. Accordingly, holding the total amount of medical services constant and substituting services for the healthier patients while increasing the provision of the less healthy raises the doctor's utility.

Depending on whether the physician's profit (i.e. $\Pi^*(p)$) is binding, there are two possible cases: $\Pi^*(p) > 0$ or $\Pi^*(p) = 0$. I start with the former case where the profit is strictly positive. Specifically, I define the set $\mathbf{P} = \{p > 0 | \Pi^*(p) = pQ^* - C(Q^*) > 0\}$ and consider a price $p \in \mathbf{P}$. According to the third line of the system (35), we know that the multiplier $\eta^* = 0$. As a result, the system (35) simplifies to:

$$\left\{ \begin{array}{l} \alpha f'_B + p - C'(Q) = 0 \\ \alpha(1 - z)f'_G + p - C'(Q) = 0 \end{array} \right. \quad (36)$$

Lemma 3.2 For all $p \in \mathbf{P}$, we have $\frac{\partial q_B^*}{\partial p} \geq \frac{\partial q_G^*}{\partial p} > 0$.

Proof. See Appendix A1. ■

This lemma shows the physician increases the medical service quantity per patient in line with the price and the increase in the service provided to the B -patient is higher than that provided to the G -patient. Intuitively, the price increase raises the marginal benefit of providing medical service to both types of patient so that the medical service provided per patient is adjusted upwards. However, as the production per patient goes up, the

marginal benefit of providing each unit of service is reduced. Since this reduction as a result of providing additional service to the B -patient is lower than that of the G -patient, the marginal increase in the service provided to the B -patient is therefore higher.

Lemma 3.3 For all $p \in P$, we have $\frac{\partial q_B^*}{\partial \alpha} \geq \frac{\partial q_G^*}{\partial \alpha} > 0$.

Proof. See Appendix A2. ■

This lemma shows that the physician increases the intensity of medical service provided per patient as his altruism level increases. Intuitively, the rise in physician altruism ceteris paribus increases the marginal benefit of providing medical service. As a result, the physician increases the medical service intensity. This lemma also shows that the physician raises intensity of medical service provided to the less healthy individual more because the subsequent reduction in marginal benefit as a result of providing additional service is relatively lower.

Lemma 3.4 For all $p \in P$, we have $\frac{\partial q_B^*}{\partial \beta} \leq \frac{\partial q_G^*}{\partial \beta} < 0$.

Proof. See Appendix A3. ■

In contrast to Lemma 3.3, this lemma shows that the physician reduces the intensity of service provided per patient as the proportion of less healthy individuals increase. Intuitively, the less healthy individual demands a higher amount of treatment than the healthier individual. As a result, the increase in the proportion of the less healthy individual ceteris paribus increases the marginal cost of providing medical service. To counter the effect, the physician reduces health service provided per patient. In particular, the reduction of health service provided to the less healthy individual is higher, as the subsequent raise in the marginal health benefit as a result of reducing additional service is relatively larger.

Substituting $q_B^*(p)$ and $q_G^*(p)$ ³⁶ into the total quantity of service provided yields $Q^*(p) = n_B q_B^*(p) + n_G q_G^*(p)$ which I use to define the doctor's profit function:

³⁶ In the ensuing analysis, I assume α, β constant except in the Lemma 3.7 and 3.8.

$$\Pi^*(p) = pQ^*(p) - C(Q^*(p)) \quad (37)$$

Taking the first-order condition of (37) with respect to p , we obtain

$$\frac{\partial \Pi^*(p)}{\partial p} = Q^*(p) + [p - C'(Q^*(p))] \frac{\partial Q^*(p)}{\partial p} \quad (38)$$

I now verify the following result.

Lemma 3.5 *Let p_0 denote the smallest p such that $\eta^*(p) = 0$, we have $\mathbf{P} = (p_0, +\infty)$.*

Proof. In order to verify the claim, I show that for all $p > p_0$, $\Pi^*(p) > 0$. Clearly, when $p \rightarrow 0$, the non-negative profit constraint will become binding, i.e. $\eta^*(p) > 0$. Intuitively, as p becomes arbitrarily close to zero, the doctor's revenue converges to zero but his marginal utility remains large due to the altruistic component. In contrast, for $p \rightarrow +\infty$ the profit constraint is not binding, i.e. $\eta^*(p) = 0$.³⁷ Since $\eta^*(p)$ is continuous in p , while there exists at least a price p_0 such that the doctor's profit is just binding, i.e. $\Pi^*(p_0) = 0$.

Next, I verify that given $C''' > 0$,³⁸ $\frac{d\Pi^*}{dp}(p_0^+) > 0$.³⁹ From (38) and using the definition of Q^* , we have:

$$\frac{d\Pi^*}{dp}(p_0^+) = Q^* - (p_0 - C') \frac{n_G f''_B + n_B(1-z)f''_G}{\alpha(1-z)f''_B f''_G - C''[n_G f''_B + n_B(1-z)f''_G]} \quad (39)$$

Hence, we want to show:

$$\begin{aligned} \alpha(1-z)Q^*f''_B f''_G - [Q^*C'' + p_0 - C'(Q^*)][n_G f''_B + n_B(1-z)f''_G] \\ > 0 \end{aligned} \quad (40)$$

Keeping in mind that $\alpha(1-z)Q^*f''_B f''_G \geq 0$ and $n_G f''_B + n_B(1-z)f''_G < 0$, the condition $C''' > 0$ ensures $Q^*C'' - C'(Q^*) + p_0 > 0$. As a result, the inequality (40) is

³⁷ *Ceteris paribus*, the price increase directly raises the physician's profit.

³⁸ This assumption is sufficient but clearly not necessary. For instance, lemma 3.5 will still hold if we assume $Q^2 C' - Q C'(Q^*) + C > 0$. However, this assumption does not have non-technical meaning in reality. Instead, $C''' > 0$ is chosen as it can be naturally interpreted as an increasing and convex marginal cost.

³⁹ The function $\Pi^*(p)$ is not differentiable at $p = p_0$. I denote by p_0^+ and p_0^- the respective right- and left-hand derivatives w.r.t. p at $p = p_0$.

satisfied.⁴⁰ To conclude the proof, suppose that in contradiction to the claim there exists a price $p > p_0$ with $\Pi^*(p) = 0$. Then, by continuity, there exists a $\hat{p} > p_0$ with $\Pi^*(\hat{p}) = 0$, and $\frac{\partial \Pi^*}{\partial p}(\hat{p}^-) < 0$. However, this is not possible because following the same logical steps (38)-(40) would yield $\frac{\partial \Pi^*}{\partial p}(\hat{p}^-) > 0$, which is a contradiction, thereby verifying the claim. ■

Lemma 3.5 tells us that the physician's profit can be partitioned by a unique critical price p_0 . Intuitively, as the physician obtains a positive profit for a sufficiently high price but a zero-profit if the payment is zero, there must be some price that the physician's profit is just binding. Furthermore, as the physician's profit is non-decreasing at the critical price, there exists only one critical price since otherwise the physician's profit at critical price is negative.

For the other case where the physician's profit is binding, i.e., $p \notin \mathbf{P}$ the system (35) simplifies to the following equation because we know that $\Pi^*(p) = 0$. Hence, the quantity of medical services which maximize the doctor's utility are implicitly defined by:

$$\begin{cases} f'(q_B) - (1-z)f'(q_G) = 0 \\ p(n_B q_B + n_G n_G) - C(n_B q_B + n_G n_G) = 0 \end{cases} \quad (41)$$

We can now use this system to complete the derivation of q_B^* and q_G^* . Altogether, we obtain the following result that there is only one way to obtain zero profit (i.e. set $MR = MC$ and $\Pi = 0$). For all $p < p_0$, we can again apply the implicit function theorem to system (41) and obtain the following Lemma.

Lemma 3.6 *For all $p \in \mathcal{R}^+ / \{p_0\} > 0$, the service supply function $q_t^*(p)$ for $t \in \{B, G\}$ are differentiable and increasing in p . Moreover, at a given price, we have $\frac{dq_B^*}{dp} > \frac{dq_G^*}{dp} > 0$. At p_0 , we have $\frac{dq_t^*}{dp}(p_0^-) > \frac{dq_t^*}{dp}(p_0^+)$.*

Proof. See Appendix A4. ■

⁴⁰ While the condition (18) is sufficient, it is clearly not necessary.

The result with respect to the left- and right-hand slope of the physician's supply function $q_t^*(p)$ at the point p_0 follows directly from the assumption $C''' > 0$. The intuition is straightforward for the case where both types of patient have the same health status, i.e. technically for $z_B = z_G = z = 0$, so that $q_G = q_B$ is denoted hereafter by q . Using the above notation, if the doctor's profit is binding, i.e. $p \notin P$, the solution q^* is implicitly defined by $npq - C(nq) = 0$. Hence, calculating the first order condition with respect to p and applying the implicit function theorem yields:

$$\frac{dq^*(p)}{dp} = \frac{q^*}{C'(nq^*) - p} \quad (42)$$

Alternatively, if $p \in P$, the doctor profit is not binding, and the solution is defined by $\alpha f'(q) + p - C'(q) = 0$. Following the same steps of deriving equation (42), we have:

$$\frac{dq^*(p)}{dp} = \frac{q^*}{C''(nq^*) - \alpha f''(q^*)} \quad (43)$$

In Lemma 3.5 we show the intuitive result that $P = (p_0, +\infty)$ which simply states that there is a critical price level above which doctors make a positive profit. Note, however, that at the point p_0 , while profit and quantity are continuous in price, they are not differentiable. For $z = 0$, it is now immediate from $C''' > 0$ that the left-hand derivative in (42) at p_0 becomes steeper than the right-hand derivative (43) at the same point.

Lemma 3.7. *For all $p \in \mathcal{R}^+ / \{p_0\} > 0$, we have $\frac{dq_B^*}{dp} \leq \frac{dq_G^*}{dp} < 0$.*

Proof. See Appendix A5. ■

Intuitively, as the less healthy individual requires a higher amount of treatment, the physician's total costs ceteris paribus increase. Constrained by the non-negative profit condition, the physician has to reduce the quantity of medical service provided per patient.

Lemma 3.8. *The critical price p_0 is increasing in α .*

Proof. See Appendix A6. ■

To provide an intuition for the result, consider a doctor characterised by the altruism level α . From the previous results, we know that there exists a price $p_0(\alpha)$ such that the doctor's profit is just binding. Suppose there exists a doctor with a slightly larger α (i.e. a marginally more altruistic individual). Holding price constant that doctor would want to provide a bit more service. Note, however, that at $p_0(\alpha)$ we have $p_0(\alpha) < C'$ so that profit would become negative and the constraint becomes strictly binding.

Keeping in mind that a physician's concern for patients is a random variable with two possibilities, $\alpha^j \in \{\alpha^L, \alpha^H\}$, and that there are two kinds of patients, we denote hereafter a low and a high altruistic doctor's supply functions by $q_t^L(p)$ and $q_t^H(p)$ and the respective critical price as $p_0^L(p)$ and $p_0^H(p)$ given $\alpha^L, \alpha^H, \beta$ and z are holding constant.

3.5 Welfare Analysis

3.5.1 Welfare Maximization

Taking the self-financing constraint of the health system into account, the results from the foregoing section are used to rewrite welfare equation (31) as a function of price. First, observe that expected health sums up individual patients' expected health. Moreover, individual health depends on the doctor's type and behaviour. Altogether, we obtain the total health function of p :

$$\sum_{i=1}^N h_i = \Lambda M \left\{ n_B f(q_B^L(p)) + n_G [z + (1-z)f(q_G^L(p))] \right. \\ \left. + (1 - \Lambda) M \{ n_B f(q_B^H(p)) + n_G [z + (1-z)f(q_G^H(p))] \} \right\} \quad (44)$$

The price p also determines the overall medical costs generated by doctors over the entire health system. Specifically, let the total service provided by a L - and a H -doctor be denoted by $Q^L(p)$ and $Q^H(p)$. Using this notation, the overall medical costs can be written as:

$$\sum_{j=1}^M C \left(\sum_{i \in I(j)} q_i \right) = M [\Lambda C(Q^L(p)) + (1 - \Lambda) C(Q^H(p))] \quad (45)$$

Finally, the non-medical costs associated with running the health system are given by the shadow price λ multiplied by the sum of payments to doctors, i.e.

$$\lambda \sum_{i=1}^N pq_i = \lambda p M[\Lambda Q^L(p) + (1 - \Lambda)Q^H(p)] \quad (46)$$

Substituting (44)-(46) in (31) yields welfare as a function p . Slightly abusing notation,⁴¹ we write $W(p)$ and note that it is a continuous function.

Lemma 3.9. *There exists a price $p^* > 0$ such that $W(p^*) \geq W(p)$ for all $p > 0$.*

Proof. See Appendix A7. ■

Lemma 3.9 verifies that the initial welfare optimization with respect to the health insurance scheme is well defined. Intuitively, the proof proceeds as follows: first, I show that there exists a price $p_N > 0$ such that for all $p > p_N$ welfare is negative. Next, I apply the Extreme Value Theorem over the closed interval $[0, p_N]$. Once solved for p^* , the equation (30) can be used to obtain the associated insurance premium required to run the medical system. Specifically, multiplying the LHS of (30) by $\frac{1+\lambda}{\lambda N}$ yields:

$$\tau^* = \frac{1 + \lambda}{N} \sum_{i=1}^N pq_i^*(p^*) \quad (47)$$

Further investigating the optimal solution would require proceeding to a comparative statics analysis using some of the endogenous parameters of the system. An analytical approach presents two difficulties. First, it may not be possible to characterise the optimal solution by using the first-order condition of the welfare optimization problem because of the potential non-differentiability of the optimal production of medical services at the prices p_0^L and p_0^H (see Lemma 3.5). Second, even if the solution were differentiable everywhere, the system of first-order equations may be too large for a meaningful analytical solution. As an alternative, a numerical simulation will be used in the ensuing analysis.

3.5.2 Specifications of Numerical Experiments

In order to perform a simulation, different parameters affecting welfare are specified.

⁴¹ Formally, in order to distinguish this welfare with the welfare represented by equation (9) a different representation $W^*(p)$ should be applied. However, we drop the superscript “**” to make the notation tractable.

- With respect to doctors, we initially impose the following parameters and functional forms. Doctors' altruism parameters are given by $\alpha^L = 0.15$, $\alpha^H = 1$. I used these parameters as they represent a relatively large difference between more and less altruistic doctors. Moreover, I set the initial proportion of low altruistic doctors at $\Lambda = 0.7$ to capture the fact that the majority of primary care physicians attach a relatively higher weight to their own financial interest rather than patients' health benefits (e.g. Godager and Wiesen 2013, 2014).
- I assumed that the health production function $f(q)$ takes the form $\sqrt{1 - \frac{1}{q+1}}$ whereas a doctor's cost for medical services is represented by the quadratic form $C(Q) = \frac{1}{200}Q^2$.⁴² These functions satisfy all the slope and curvature requirements introduced in Section 3.3.
- With respect to patients, I start with the following parameters; $N = 100000$ (total number of patients), $n = 100$ (average number of patients allocated per physician), $z = 0.4$ (initial health of G -type patients) and $\beta = 0.3$ (the proportion of B -type patients). The parameter of N , z is selected to satisfy law of large numbers and to contrast initial health between alternative types of patients respectively. The parameter of β is selected to capture the average proportion of less healthy patients (i.e. population over 65s) in developed countries.
- I assumed that the shadow price of taxation is $\lambda = 0.1$ as in Laffont and Tirole (1986).

3.5.3 Numerical Results

I apply the foregoing specification and use Mathematica to solve the respective doctor's problem and obtain all the supply functions. In the next subsection, these supply functions will be represented. Subsequently, I aggregate the supply decisions and calculate the welfare function which will be used to find the optimal price for medical services p^* . This initial setup will then be used to perform some numerical experiments.

⁴² It can be shown that this cost assumption satisfies $Q^2C'' - QC' + C \geq 0$, which is sufficient to prove Lemma 3.5.

3.5.3.1 Supply Curves

Figure 3 represents the individual supply curves for L -doctors. It maps the quantity supplied to a B - or a G -patient as a function of the regulatory price p , i.e. $q_B^L(p)$ and $q_G^L(p)$:⁴³

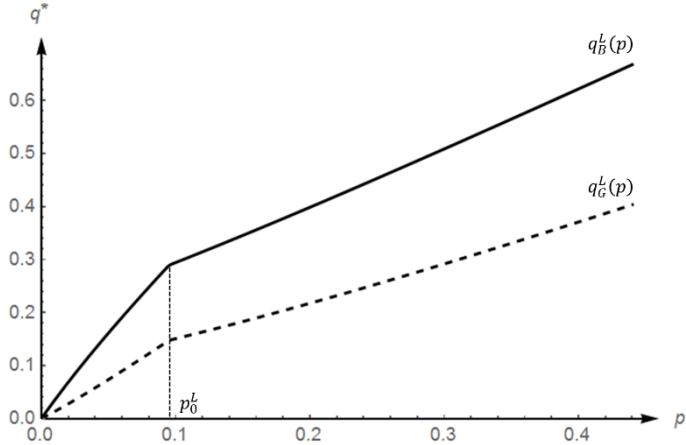


Figure 3. L -doctor's supply function for a B -patient

Figure 4 plots the individual supply functions of L - and H -doctors with respect to a type t patient. These graphics clearly depict the results summarised by the Lemmas 3.1, 3.2, 3.3 and 3.6; for any give price $p > 0$, the figure shows the physician provides more service to a less healthy individual than to a healthier one. Next, the supply curves have positive slopes, i.e. the physician increases service supply with price for B - and G -type patients. For the same patient, he tends to be provided more service from the physician with a higher level of altruism at a given price. Moreover, at the same price, the slope of healthcare supply function of B -patients is larger or equal to that of G -patients. Finally, the supply curves are not differentiable at the critical price p_0^L and the slope of $q_t^L(p)$ satisfies $\frac{dq_t^L}{dp}(p_0^-) > \frac{dq_t^L}{dp}(p_0^+)$, where p_0^L is defined by the kink.

⁴³ In order to avoid an overly cumbersome notation, for the L -type doctor we write $q_B^L(p)$ for $q_B^{L*}(p)$.

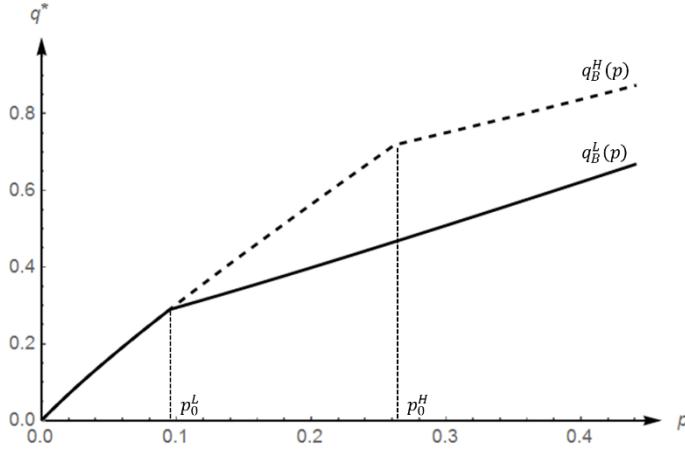


Figure 4. L- and H-doctors' supply curves for a T-type patient

Just as before, the critical prices, p_0^L and p_0^H are defined by the respective kink of $q_t^L(p)$ and $q_t^H(p)$. The graphic verifies for the numerical example the finding from Lemma 3.9 that the critical price of high altruistic provider exceeds that of the low altruistic provider (i.e. $p_0^H > p_0^L$). The definitions of p_0^L and p_0^H indicate that, for $p \in [0, p_0^L]$, both types of doctors obtain zero profit. Hence, for these prices, altruism does not affect the quantities of medical service provided by either doctor. Intuitively, the total quantity is determined solely by the zero-profit condition. As a result, doctors have the same marginal cost and, therefore, supply the same quantity of service independent of their differing level of altruism.

Second, for all $p \in [p_0^L, p_0^H]$ we have $q_t^H(p) > q_t^L(p)$. In order to provide intuition for this result, consider that the *H*-doctor provides the quantity of medical service that is less or equal to that of the *L*- doctor (i.e. $q_t^H(p) \leq q_t^L(p)$). The curvature assumption of the cost function and his higher level of altruism immediately imply that his marginal return of producing $q_t^H(p)$ would exceed its marginal cost, which is indeed a contradiction. Finally, for $p > p_0^H$, $q_t^H(p) \geq q_t^L(p)$, applying the preceding logic yields $q_t^H(p) > q_t^L(p)$. However, doctors are increasing their service supply in price and their marginal return from altruistic component is zero given a large quantity provision of medical service. Hence, for very high payment rates, the quantity of service supplied will no longer depend on the level of altruism.

3.5.3.2 Welfare Function and the Optimal Health Insurance Scheme

I now substitute the supply decisions of the H - and L -doctors with respect to their G - and B -type patients into the welfare function in order to obtain $W(p)$. The next figure provides the geometrical representation.⁴⁴

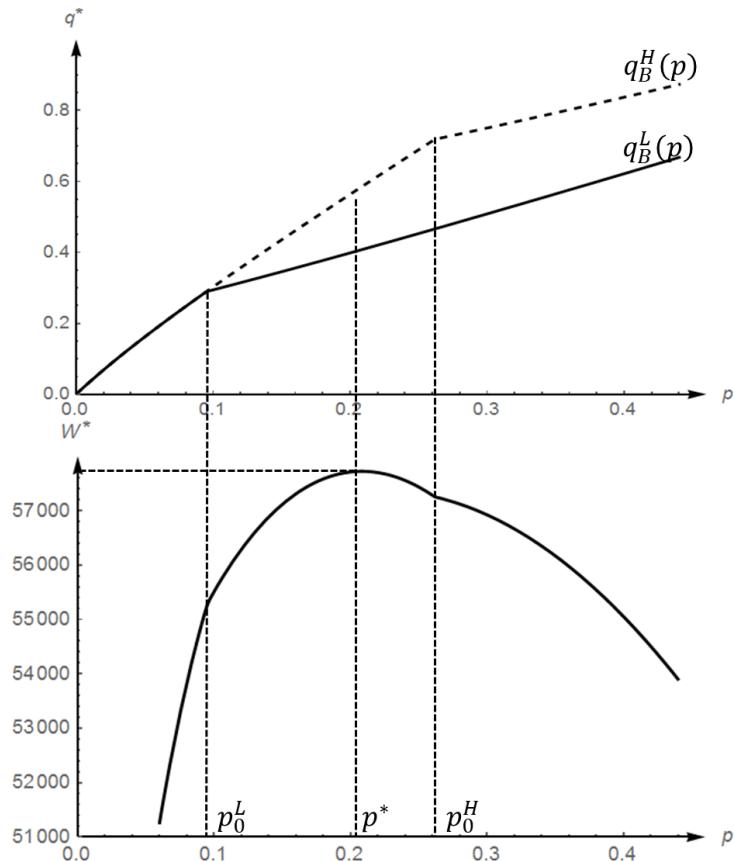


Figure 5. Welfare and the Optimal Price

Figure 5 provides a visual representation of lemma 3.9. The upper graph depicts the supply function of L -and H -supply curves to the type t -patients (as shown in Figure 4), while the lower graph represents welfare as a function of price. This example shows the welfare optimal price is $p^* = 0.2084$, which lies between the critical price level of L - and H -type physicians (i.e. $p_0^L = 0.095$ and $p_0^H = 0.26$). As a result, the less altruistic doctors (the L -type) obtain a strictly positive profit, while their more altruistic counterparts have zero profit at the welfare-maximizing price. Intuitively, the health insurance scheme does not have the required information to distinguish between H - and L -doctors as well as between G - and B -patients. As a result, the scheme sets an “intermediate” price which trades off the over-production from H -doctors against under-

⁴⁴ In order to see the welfare function clearly the starting point is set at $p = 0.05$.

provision by L -doctors. To understand this “intermediate price”, suppose the economy is only comprised of H -type physicians, then this example shows that the welfare-maximizing price would become $p_H^* = 0.189$.⁴⁵ This price is lower than p^* if both types of doctors exist as the highly altruistic doctors require a low price in order to be motivated. On the contrary, if the economy is only comprised of L -type physicians, the welfare-maximizing price would become $p_L^* = 0.248$. This price is higher than p^* as low altruistic providers require high financial rewards to be motivated.

3.5.4 Comparative Static Experiments

In this subsection, I perform a number of numerical exercises. The purpose of doing these exercises is to evaluate the optimal response of the health insurance scheme to variations in the exogenous variables. To do this, I vary one exogenous parameter at a time and solve the welfare maximizing health insurance scheme for each parameter constellation. Next, I use the solution (p^*, τ^*) to find the respective doctor’s quantity of service provided (q_t^{L*}, q_t^{H*}) , profits (Π^{L*}, Π^{H*}) , utilities (U^{L*}, U^{H*}) , expected patient health h_t^* , and welfare to society W^* .⁴⁶ Finally, based on the above information, I present a real life interpretation of each analysis.

These experiments are structured as follows: Experiments 1 and 2 analyse the impact of varying the initial health of healthier individuals (z) and the proportion of less healthy individuals (β), which represent health consciousness promotion and the ageing population respectively. Experiments 3 and 4 examine the impact of varying the number of patients allocated per doctor (n) and the proportion of less altruistic providers (Λ) respectively. These variations reflect recent trends, namely that hiring new medical staff becomes more difficult and that the proportion of more altruistic doctors is declining. Finally, Experiment 5 predicts the impact of technological innovation, which is represented by an increase in the technology factor (T).

⁴⁵ See the graph “the effect of varying Λ on the constrained optimal level of price” in the attached Mathematica file- Comparative static_lambda.

⁴⁶ The superscript “*” represents the value of the respective function at the solution (p^*, τ^*) .

3.5.4.1 Variations in Patients' Initial Health

In the first experiment, I vary the initial health of G -type individuals (z). To understand the aforementioned results, imagine that society were divided between the health conscious and other individuals. Moreover, suppose that raising the awareness of the health benefits associated with fruit and vegetables in the diet only affects G -type patients so that their average health increases.

In the numerical experiment, I varied z between 0 and 1 (i.e. initially B - and G -individuals are the same). The main results from this exercise are summarised in Figure 6.

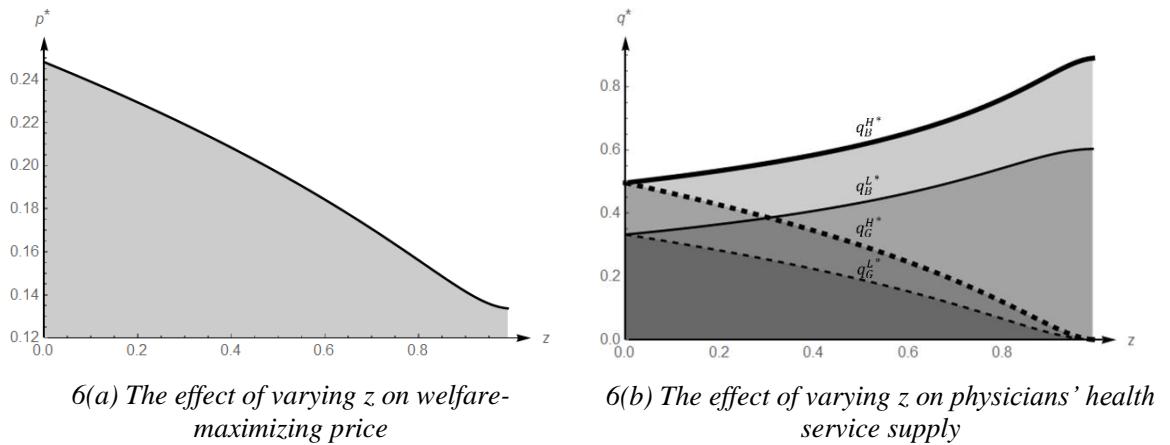


Figure 6. Effects of varying the initial health of healthier patients

Result 1: Promoting health consciousness will reduce price, insurance premiums, and result in a better allocation of health services between patients.

In order to derive an intuition for the underlying effects, consider a situation where the HI scheme equalizes the marginal cost associated with the health premium in terms of the marginal benefit thereof (in relation to health). Now, imagine a *ceteris paribus* (c.p. hereafter) increase in the initial health level of G -individuals due to a rise in z . This implies that a doctor's marginal return $\beta(1-z)f'_G + p$ in relation to a G -type individual decreases. Hence, doctors will reduce the level of health service which they provide for this type of patient so that Q goes down. As a result, doctors' marginal costs $C'(Q)$, are also lowered. This produces a countervailing effect for B -type patients who receive more services. Altogether, all patients see their expected health levels increase. From the point of view of the HI, it means that the marginal benefit of health

services has gone down while the marginal costs associated with health premiums stay constant. Accordingly, the initial equality is no longer satisfied, thereby inducing HI to reduce p^* as illustrated in the Figure 6a.

The change in p^* has a feedback effect in the production of health services for all patient-doctor pairs which are summarised in Figure 6b. As can be seen, the feedback effect does not reverse the initial directions in the variation of services; G -individuals still receive less services and B -patients continue to receive more services. The total effect on health is positive for both type of patients; B -individuals receive more medical services while the z -effect on G -individuals dominates the negative impact of the reduction of their services.

Following the same reasoning used in the preceding experiment, the profit of H -doctors remains constrained by the non-negative requirement throughout. For L -doctors, the direct effect of the price reduction dominates the indirect effect of the reduction in output so that their profit declines. Nevertheless, my exercise shows both types of doctors attained higher utilities. For the H -doctors, this result follows directly from the observation that their profit remains unchanged while the expected health level of their patients goes up. For the L -doctors, the finding is specific to the set of parameters chosen. Intuitively, the positive utility effect of the improvement in initial health dominates the negative utility impact of their reduction in profit.

Finally, from the point of view of HI, the insurance premium is decreasing in the initial health of healthier individuals (z). To illustrate this, first start with the above c.p. argument; τ^* would necessarily go down as Q decreases in z . The additional feedback effect due to the reduction in p^* and the further decrease in Q only reinforce that initial impact.

Overall, this experiment simulated the impact of an increase in the level of health consciousness by a fraction of the population, which can be interpreted as a fraction of the population is willing to adjust behaviour in order to improve their respective health (for instance by eliminating smoking, reducing alcohol consumption, improving eating habits, etc.). In that context, I imagined the impact of an advertising campaign promoting healthier behaviour (e.g. 10,000 steps a day, the “five fruit and vegetables” campaign, etc.). By assumption, only the fraction of the population that is health conscious would alter their behaviour and obtain a higher level of initial health as a result.

In the experiment G -patients were assumed to be the health-conscious individuals and a change in their awareness results in an increase in their initial health levels. This experiment therefore suggests that the HI scheme should reduce the insurance premium and the price for medical services. Despite the price reduction, altruistic physicians raised the service provided to B -patients but reduced that offered to G -individuals. Altogether, the evolution and the response of the HI scheme induced an improvement for all participants.

3.5.4.2 Variation in the Proportion of Patients

In the second experiment, I vary the proportion β which measures the fraction of B -type patients in the economy. A real-life example of what we have in mind is the current dynamic in most countries where the rise in the average age of the population also means that the proportion of less healthy individuals within a given economy is increasing. Figures 7a and 7b summarises some of the results from the numerical exercise.

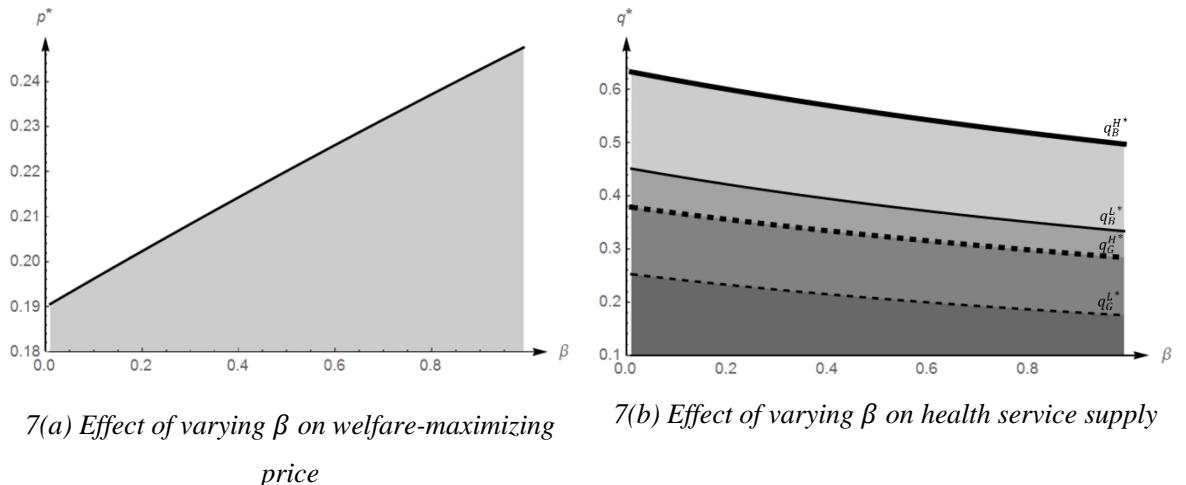


Figure 7. Effects of varying the proportion of less healthy patients

Result 2: Population ageing will lead to a price and insurance premium increase and a reduction of medical service supplied.

Figure 7a represents the optimal adjustment of the price for medical services as a response to the change in β . Intuitively, the increase in the proportion of less healthy patients requires that the overall production of medical services (i.e. Q^j) increases. In

order to incentivize doctors to increase production, the health insurance scheme adjusts the price upward. Since the total production of medical services increases and p^* rises, the health insurance premium τ^* also increases. Figure 7b maps the resulting quantity of health services from each possible pairing of patient and doctor. Not surprisingly, the quantity of services provided turned out to decrease in line with the fraction of less healthy individuals (B-type patients). As a result, expected health for both types of individual goes down as β increases.

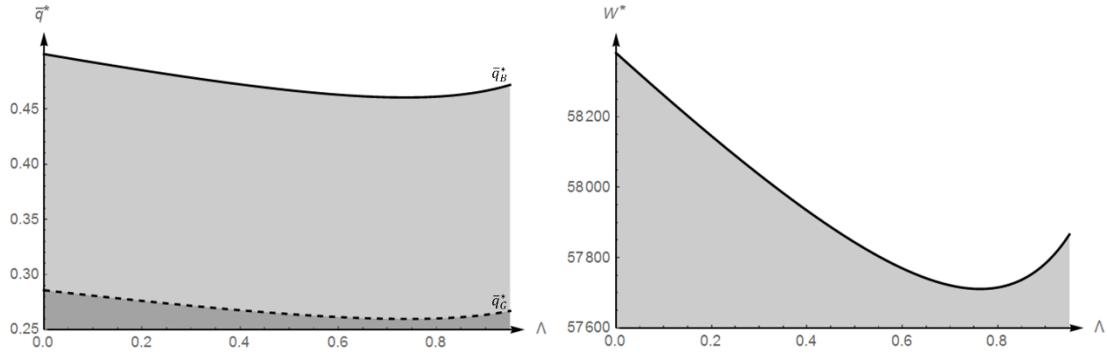
In the numerical exercise, the decision of the more altruistic doctors was found to be constrained throughout by the zero-profit requirement. In contrast, the profit for the less altruistic doctors increased throughout. This reflects that the regulator selects a price to balance overproduction by H -doctors with underproduction on the part of L -doctors.

It is noteworthy that in the experiment the well-being of both types of doctors was found to decrease in β . For the more altruistic doctors, this follows directly from the observation that their profit remains constant while the average health of their patients goes down. For the less altruistic doctors, the finding is an artefact of the specific numerical experiment where the negative impact of the health reduction outweighs the positive utility impact of the increased profit. Overall, this experiment reproduces the key ingredients associated with an ageing population. Assuming that older individuals are on average less healthy than younger people, the idea of an ageing population is captured through an increase in the proportion of B -patients in society. The main finding of this experiment is that the optimal scheme would induce a reduction in the service provided to all types of patients, increasing the insurance premium while leading to deteriorating levels of satisfaction for medical staff. In other words, even under ideal conditions ageing would lead to patients, insurers and doctors' increased dissatisfaction.

3.5.4.3 Variation in the Proportion of Doctors

In the third experiment, I study an increase in the parameter Λ which measures the fraction of L -type doctors. Intuitively, the experiment captures the observation that altruism in medicine has been declining (see Jones, 2002).⁴⁷

⁴⁷ For instance, many observers suggest that the average physician pays less attention to patients (e.g. GP home visit rates have been falling while many doctors are reluctant to provide 24-hour services to patients).



8(a) Effect of increasing Λ on expected supply of service provided per doctor 8(b) Effect of increasing Λ on social welfare

Figure 8. Effects of varying the proportion of more altruistic physicians

Result 3: Price and insurance premium will rise as the proportion of less altruistic doctors increases and the presence of multiple types of physicians will deteriorate social well-being.

To illustrate, I let Λ vary from zero (all the doctors are characterised by a high level of altruism) to one. Not surprisingly, the optimal price and insurance premium are found to increase in Λ . Intuitively, with a lower proportion of H -type doctors, the regulator is forced to use more extrinsic motivation to align incentives, thereby requiring a larger premium. However, the presence of both types of doctors creates a tension, increasing the extrinsic motivation of L -doctors induces H -doctors to produce too much relative to the efficient level.⁴⁸ Obviously, this tension disappears at either extreme of the support ($\Lambda = 0$ or $\Lambda = 1$), which explains why the expected quantity of services initially decreases but then increases (Figure 8a). The effects in the expected health of patients follows a similar pattern, directly reflecting the level of expected medical services. Figure 8b captures the effect of changes in Λ on welfare. Clearly, having only H -doctors is best because motivating them is easy so that costs are low. However, similar to the expected level of services with mix of both types of doctors, this implies that the price system works poorly; p is too low to motivate L -doctors but too high for H -doctors who produce excessively relative to the efficient level.

In healthcare market, it has often been claimed that the extensive provision of financial incentives has produced a crowding out of intrinsic motivation in the medical professions (Berdud et al., 2016). As a result, the current experiment examines the

⁴⁸ The service quantity maximizes welfare when only L -or H -type doctors exist.

impact of a reduction in the level of altruism in relation to the optimal HI scheme by increasing the fraction of doctors with low altruism. My experiment predicts that both price and insurance premium would increase in line with the decline of more altruistic doctors. As a result, both types of physicians raised their production of services to their patients, thereby improving the associated level of satisfaction. From the point of view of the total medical service production, there is a countervailing force due to the reduction in the proportion of H -type doctors. In the experiment, this countervailing effect was found to dominate when the fraction of low altruism doctors was small. For the selected parameters, a decrease in the well-being of patients was found because the countervailing effect was always sufficiently large to overturn the indirect impact of price increases. However, welfare was not monotonic; as the fraction of low altruism doctors became large, the increase in doctors' profits became large enough to compensate for the small reduction in patients' utility. The welfare function is inversely U-shaped (Figure 8b), indicating that a large variance in the level of altruism among doctors made the price scheme not very effective.

3.5.4.4 Variation in the Allocated Number of Patients

In the fourth experiment, I vary the parameter representing the number of patients allocated per doctor. The example which comes to my mind is that the regulator alters the number of medical staff. With this in mind, observe that (29) can also be written as

$$W = N\Lambda[\beta u_B^L + (1 - \beta)u_G^L] + N(1 - \Lambda)[\beta u_B^H + (1 - \beta)u_G^H] + M[\Lambda\Pi^L + (1 - \Lambda)\Pi^H] \quad (48)$$

where u_i^j denotes the expected utility of an i -patient allocated to a j -doctor and Π^j is the profit of the j -doctor. Accordingly, a c.p. increase in the number of medical staff (M) automatically raises welfare in our initial model. Obviously, this result ignores that adding medical staff is costly because it requires training them. While omitting training costs did not matter for the former experiments, it is clearly essential in the current case. Hence, in this subsection I include in the welfare function the linear educational costs

c^{49} that are associated with the training of medical staff. The welfare function in this experiment therefore becomes:

$$W = N\Lambda[\beta u_B^L + (1 - \beta)u_G^L] + N(1 - \Lambda)[\beta u_B^H + (1 - \beta)u_G^H] \\ + \frac{N}{n}[\Lambda\Pi^L + (1 - \Lambda)\Pi^H] - c \frac{N}{n} \quad (49)$$

Moreover, I use a specific cost function $C(Q) = \frac{1}{100}Q^{1.4}$ which is increasing and convex as well as satisfying $Q^2C'' - QC' + C \geq 0$. Finally, in the experiment I use the definition $n = \frac{N}{M}$ to substitute $M = \frac{N}{n}$ while varying the average number of patients n . For the numerical experiment, I set the parameter related to per doctor training at 35, which ensures that the ratio of training to operational costs remained within a reasonable range.⁵⁰

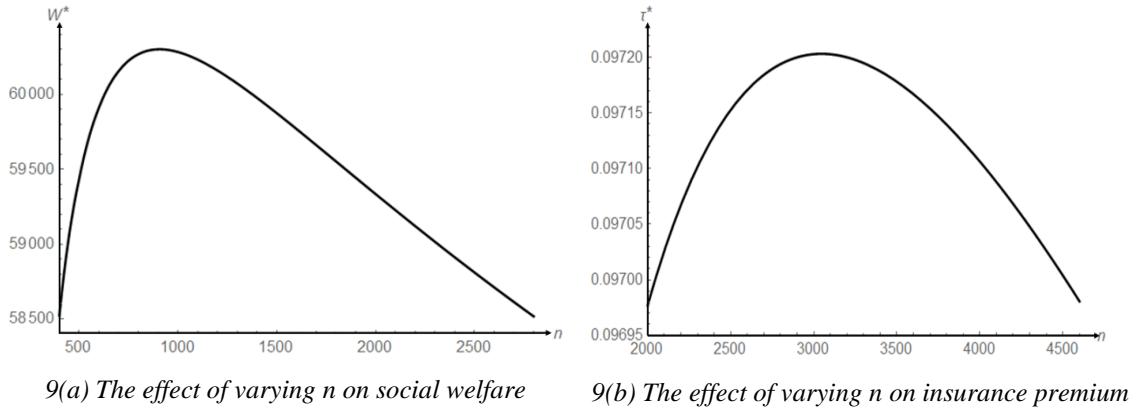


Figure 9. Effects of varying the number of patients allocated per physician

Result 4: There exists a welfare and insurance premium that maximizes the number of patients allocated per doctor.

To illustrate, we first show that p^* is increasing in terms of the average number of patients per doctor (n) throughout. The intuition is as follows: suppose doctors were to keep the per patient amount of medical service q_i^j constant. Holding p^* constant would

⁴⁹ In the UK, for instance, research shows it costs around £220,000 to train a primary care doctor over their five-year degree (Brown, 2016). To analyse parsimony without losing generality, in this model I assume a linear training costs.

⁵⁰ My example captures the fact that the average training costs of a primary care physician's is significantly less than his average income (per year). As shown in Mathematica file Comparative static_n.pdf- the calculated ratio, selecting $c = 35$ and the specific cost function makes the simulated ratio equals around 0.36.

imply that initial marginal benefits are unaffected. However, the increase in n would result in an increase in the total amount of medical services provided by doctors, thus raising their marginal costs. Restoring equality between doctors' marginal benefits and costs would require a reduction in services. To reduce the negative impact on health, the HI wants to motivate doctors to increase output, therefore it finds it advantageous to raise p^* . The price increase generates a countervailing incentive inducing doctors to reduce services (and therefore expected health levels) less severely.

The experiment can also be reinterpreted in terms of finding the optimal number of patients per doctor. In Figure 9a, we varied the number of patients per doctor between 50 and 500. As can be seen from this figure, the welfare maximizing n (denoted by n^* hereafter) turned out to be around 1000 individuals per doctor ($n^* = 985$ to be precise). At that point, the marginal educational costs are just equal to the marginal benefit of adding one more doctor to the system. This marginal benefit includes the increase of patients' expected health, the reduction in physicians' marginal costs, and the resulting decrease in p^* .

Figure 9b depicts the impact of n on the health insurance premium. As can be seen, τ^* is increasing around $n = n^*$, while for large n it is decreasing. Intuitive variations in n have a direct and an indirect impact. In direct terms, raising n reduces the number of doctors and thus c.p. total medical expenditures and total profit paid to L -type doctors. The indirect effect derives from the quantity adjustments and price changes affecting profits. In the numerical exercise, I find that L -doctors' profit was increasing in n while the profit of H -doctors remained constrained by zero. In the example, the indirect effect was dominant for small n while the direct effect prevailed for large n .

The fourth experiment captures the idea of controlling medical cost through restricting the total number of physicians. To do this, the experiment raises the patients per doctor ratio and adjusts the model to account for the cost of educating doctors. Unsurprisingly, the resulting increase in workload induced physicians to reduce the level of services provided to all types of patients. To counter this tendency, the HI scheme raises the price of medical services to counter the negative impact on health. However, the insurance premium was not monotonous as neither the effect of the price increase nor the effect of the quantity reduction dominated throughout.

Patients' utility was reduced throughout because it did not account for the costs associated with medical training. The total impact on welfare was also not monotonic. Specifically, when the ratio of patients to physicians is small, the reduction in the cost of physician training became large enough to compensate for the reduction in patients' well-being. A by-product of this experiment was that it verified that an extension of the model which would take the number of medical staff into account is well behaved. In other words, the policy vector could easily be extended to include the decision of a patients-per-doctor ratio. Including that ratio as a policy choice would also allow one to expand the comparative static analysis and examine questions like the optimal variation in the number of patients-per-doctor as a response to ageing, technical change etc. While these are undoubtedly very important and interesting questions, they go beyond the scope of the current analysis.

3.5.4.5 Variation in Medical Technology

In this last experiment I try to capture one of the major trends that has affected HI schemes over the last half century. Major pharmaceutical and medical innovations have significantly improved health levels for all age categories. These innovations have led to an overall increase in both the quality of life and life expectancy. A by-product of this increased longevity is that the average age of the population is rising. However, given that the demand for medical services increases with age, the rising proportion of older people is placing an upward pressure on the overall demand for healthcare.

In order to model this evolution within the current framework, we introduce a technology factor T which is assumed to increase over time. Specifically, we use the functional form

$$h(z, q, T) = z + (1 - z)f(q, T) \quad (50)$$

where $f(q, T) = \left(1 - \frac{1}{1+q}\right)^{\frac{1}{T}}$ while this equation captures that a rise in T pivots expected health outcomes upward. To capture the population ageing that results from an improvement in medical technology, I assume that in the model increasing T also raises the fraction of the B -type individuals in the economy. Since B - and G -type individuals can be interpreted as reflecting the population over and under the age of 65 in the economy respectively, this assumption therefore reflects that an increase in technology leads to an increase (or decrease) in the proportion of elderly (or young) individuals. As

a result, the average age of an individual in the economy increases, which reflects an ageing population. Specifically, I assume $\beta(T) = aT^b$ ($a, b > 0$). For the numerical experiment, I let T vary from [2,3] and use the parameter values $a = 0.01$, and $b = 3$. These particular parameters are selected so that when T increases from 2 to 3, β goes from approximately 0.08 to 0.27. Since β in this model measures the proportion of less healthy individuals, the aforementioned increase in β matches the changes in the fraction of the over-65 population in developed countries such as Japan, Germany, and France from approximately 10% in 1960 to 25% in 2020 (Hawe, 2008).

In contrast to these experiments, the dissimilarity between G - and B -patients is increased by raising z from 0.4 to 0.8. This change is introduced to match the difference in the per person expenditure across the population over- and under-65. For instance, U.S. data shows that spending for the over-65 group is approximately four times what it is for the 34-44 group (similar findings exists for the UK and the EU overall, though the ratio is slightly lower than 4, see Robineau, 2016). Figure 10 summarises the key findings of the experiment.

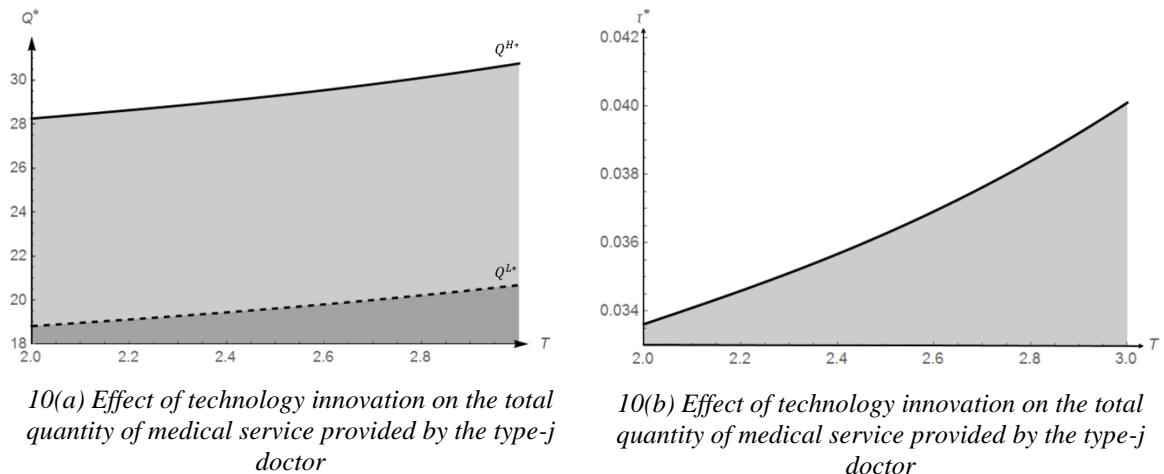


Figure 10. Effects of varying the technology factor

Result 5: *Medical innovation increases total health service provision and patient expected health but drives up health expenditure per person.*

The evolution of T implies that the marginal effect of q on health changes over time. From the functional form, however, note the effect of T on the health margin of q depends on the initial level of medical services; it is increasing for small but decreasing

for large q . In all cases, the evolution of T feedback into the motivation of doctors via the altruistic component of their preferences. Given the initial parameter set, I find that B -patients *ceteris paribus* received less services whereas G -patients received more services for low T but less services for high T . However, in all cases the technological improvement remained beneficial.

The fraction of B -types present in the economy β also increased over time in line with technological advances. This created an additional negative impact on the provision of services, as the increase in β drives up total medical service demand and therefore the marginal cost of providing medical services. As a result, per patient services provided to both types of patients are reduced. In order to partially countervail the reduction of services, the HI benefited from raising price. The combination of more low health individuals with the increase in p^* led to a rise in the supply of services by both types of doctors, $Q^j, j = H, L$, as represented in Figure 10a. Finally, there was an expansion of medical service together with a rise in the price required to increase the insurance premium as shown in Figure 10b.

The total effect on expected health remained positive. With respect to profit levels, L -doctors was found to benefit from the price evolution while H -doctors remained constrained by the non-negativity condition. Regarding the level of doctors' well-being, I found that it increased for both L - and H -doctors. Intuitively, the L -type physicians have higher profits and their patients' health goes up. For the H -type physicians, their profit remains zero but their patients derived higher health benefits. In this numerical exercise, it turned out that patients' health improvement and/or physicians' rise in profits outweigh the growth in the B -individuals' fraction.

To conclude, this experiment reflects the argument that technology has been the key driver for burgeoning expenditure as pointed out by (Sorenson et al., 2013). I represented the idea as a continuously increasing "technology factor" making the health function more productive over time and raising the proportion of B -patients. The latter requirement encompassed the idea that better medicine increases life expectancy, but in doing so also raises the fraction of individuals with high medical needs. Major findings are that the HI scheme should raise both price and insurance premium while continuously reducing the quantity services provided to all types of patient. With respect to total medical service production, there is an additional countervailing effect from the

change in the fraction of high need patients. Since the countervailing effect dominated throughout, the total production of medical services increased. Altogether, the well-being of all types of patients are improved because the positive impact on patients' health outweighs the negative impact on their premium.

3.6 Cost Control Policies and Efficiency Enhancement

In most developed countries, authorities responsible for health systems have been challenged by the evolution of fundamental trends and the implications for their respective health insurance schemes. Required adjustments to different systems have been marred by a plethora of criticisms from medical professionals, insurance companies, pharmaceutical firms and the public at large. A fundamental difficulty in evaluating the validity of these criticisms is the lack of an objective perspective. For instance, doctors may denounce a policy because their utility goes down while pharmaceutical firms may prefer increasing the supply of medication, and fully insured patients always favour an increase in medical services.

Modelling the health insurance scheme from the perspective of a benevolent welfare maximizing planner is an attempt to provide a more objective viewpoint. With this in mind, the initial model introduced in Section 3.3 was characterised by a constrained optimal health insurance scheme, as discussed in Section 3.5. Furthermore, this section provides experiments aimed at evaluating the best response of the constrained efficient HI scheme to reflect trends that capture some of the salient features in recent evolutions.

Section 3.5 emphasises that the presence of informational asymmetry using the pricing scheme to align incentives of doctors displays a major flaw; in of itself it is incapable of bridging differences in intrinsic/extrinsic motivation across medical staff. As a result, maximizing welfare induces doctors with a high intrinsic motivation (in the model referred to as 'high altruism') to produce more than those who require a high extrinsic motivation. The subsequent numerical experiments further show that any evolution of the underlying parameters which require an increase in medical services exacerbates this difference. Intuitively, inducing an increase in medical services requires raising the price which worsens the misallocation of resources. Accordingly, the next step of this analysis is to extend the set of policy tools used to align medical incentives.

In order to alleviate the tension presented in the last two sections resulting from the presence of private information associated with doctors' preferences, many HI schemes resort to quantity rationing, costs restrictions or mixed payment in the production of medical services. For instance, with respect to the quantity rationing, the NHS system in the UK restricted the time spent on GP consultations and applied numerous examples of therapy and drug rationing. Well known examples include restrictions for cancer treatments, painkillers and arthritis medications (Press Association, 2016).

With respect to the second type of restriction, the German HI scheme introduced different levels of expenditure cap in order to ensure that the health system does not spend more than the income generated by the aggregate insurance premium (Ehrbeck et al., 2010). For instance, the German health authority established an overall budget for the system at the regional level. This forces hospitals and physicians into negotiation with insurers in order to arrive at collective annual budgets (Ehrbeck et al., 2010). At the individual level, pharmaceutical expenditure caps per surgery have been implemented over the last decades (Busse et al., 2005).

Finally, the mixed payment system is comprised of both fee-for-service and fixed payment per patient has been widely applied in the OECD economies (OECD, 2016). For instance, in the UK, Finland, Norway and Italy, primary care physicians are paid based on a combination of capitation and FFS for the purpose of containing health expenditure (CAP) and reducing referrals to hospitals (Park et al., 2007). In Ontario, Canada, the "Primary Care Reform" launched in the late 1990s introduced a menu of PCP payment that blends FFS, capitation and salary elements (Laberge et al., 2016).

By construction, quantity and costs restrictions are very crude policy instruments since the medical information relating to specific patients' health remains the private information of doctors. The purpose of this section is to explore whether quantity and revenue restrictions as well as capitation can nevertheless help to improve welfare. In order to keep the analysis tractable, each restriction will be examined separately.

3.6.1 Quantity Restriction

In this subsection, I introduce an additional policy variable denoted by \bar{q} which represents the maximum quantity of per patient treatment that a doctor can perform.

Accordingly, a HI policy becomes a triplet (p, \bar{q}, τ) . Just as in Section 3.4, doctors take the policy vector as given and maximize their utility. Accordingly, the optimization problem (32) needs to be slightly adjusted to integrate the requirement $\bar{q} \geq q_t$ for $t \in \{B, G\}$, when it becomes

$$\begin{aligned}\bar{U} = \max_{q_t} \sum_{t \in T} n_t [\alpha(z_t + (1 - z_t)f(q_t)) + pq_t] - C \left(\sum_{t \in T} n_t q_t \right) \\ p \sum_{t \in T} n_t q_t - C \left(\sum_{t \in T} n_t q_t \right) \geq 0 \\ \bar{q} \geq q_t \geq 0\end{aligned}\tag{51}$$

where α stands for degree of altruism of a generic doctor. In Appendix A8, I prove that the system (51) has a unique solution for any given policy vector (p, \bar{q}, τ) . Slightly abusing the notation, I denote that, for this subsection, $q_t^H(p, \bar{q})$ and $q_t^L(p, \bar{q})$ is the respective solution for H - and L -doctors. Following the same procedures as used in Section 3.4, I substitute $q_t^H(p, \bar{q})$ and $q_t^L(p, \bar{q})$ into the equation (31) and obtain the welfare function $\bar{W}(p, \bar{q})$. Observe that for any restriction level \bar{q} , the same process as outlined in Sections 3.4 and 3.5 can be followed, i.e. I solved for the welfare maximizing price, denoted hereafter by $p_Q^*(\bar{q})$, then substituted the solution into the welfare function to obtain $\bar{W}(p_Q^*(\bar{q}), \bar{q})$. Clearly, for \bar{q} to be sufficiently large, the restriction never binds and $\bar{W}(p_Q^*(\bar{q}), \bar{q}) = W(p^*)$. Accordingly, we know

$$W(p^*) \leq \bar{W}(p_Q^*(\bar{q}^*), \bar{q}^*)\tag{52}$$

where \bar{q}^* denotes the optimal level of quantity rationing imposed to maximize $\bar{W}(p_Q^*(\bar{q}), \bar{q})$. As a result, the inequality (52) shows that the optimal welfare under the quantity rationing is larger equal to the optimal welfare without imposing the restriction.

From the foregoing section, one would think that a small reduction of \bar{q} at the point $\bar{q} = q_B^H(p^*)$ should lead to an improvement in welfare. The logical argument is as follows: at the price p^* which maximizes (31), we know that, relative to the first-best solution, H -doctors over-produce while L -doctors under-produce. Intuitively, at p^* the HI trade-offs the benefit of an increase in the production of L -doctors against the reduction in welfare from an increase in production by H -doctors. Hence, one would think that

introducing a restriction which forces H -doctors to marginally reduce production for the high need patients would positively impact this trade-off.

Unfortunately, while this intuition is appealing, it is nevertheless false in many cases. To see why, consider the examples discussed above. In almost all cases, I found that H -doctors were constrained by the zero-profit requirement. In such a situation, it is easily verified that a marginal reduction of \bar{q} at $\bar{q} = q_B^H(p^*)$ would not lead H -doctors to reduce overall production. Instead, they would marginally reduce q_B^H but simultaneously increase q_G^H while keeping total output and profit constant. Taken together, one would expect this to worsen welfare since it would take away medical services from the individuals who need them more to give to those who need them less. This observation implies that introducing a quantity restriction generates a countervailing incentive via the doctors' zero-profit constraint. At *a priori*, there are no natural restrictions to guarantee which effects dominate. Accordingly, the standard comparative static technique will in general fail. In the remainder of this subsection, the foregoing numerical examples will be adapted to show that there are cases where a quantity restriction does improve welfare.

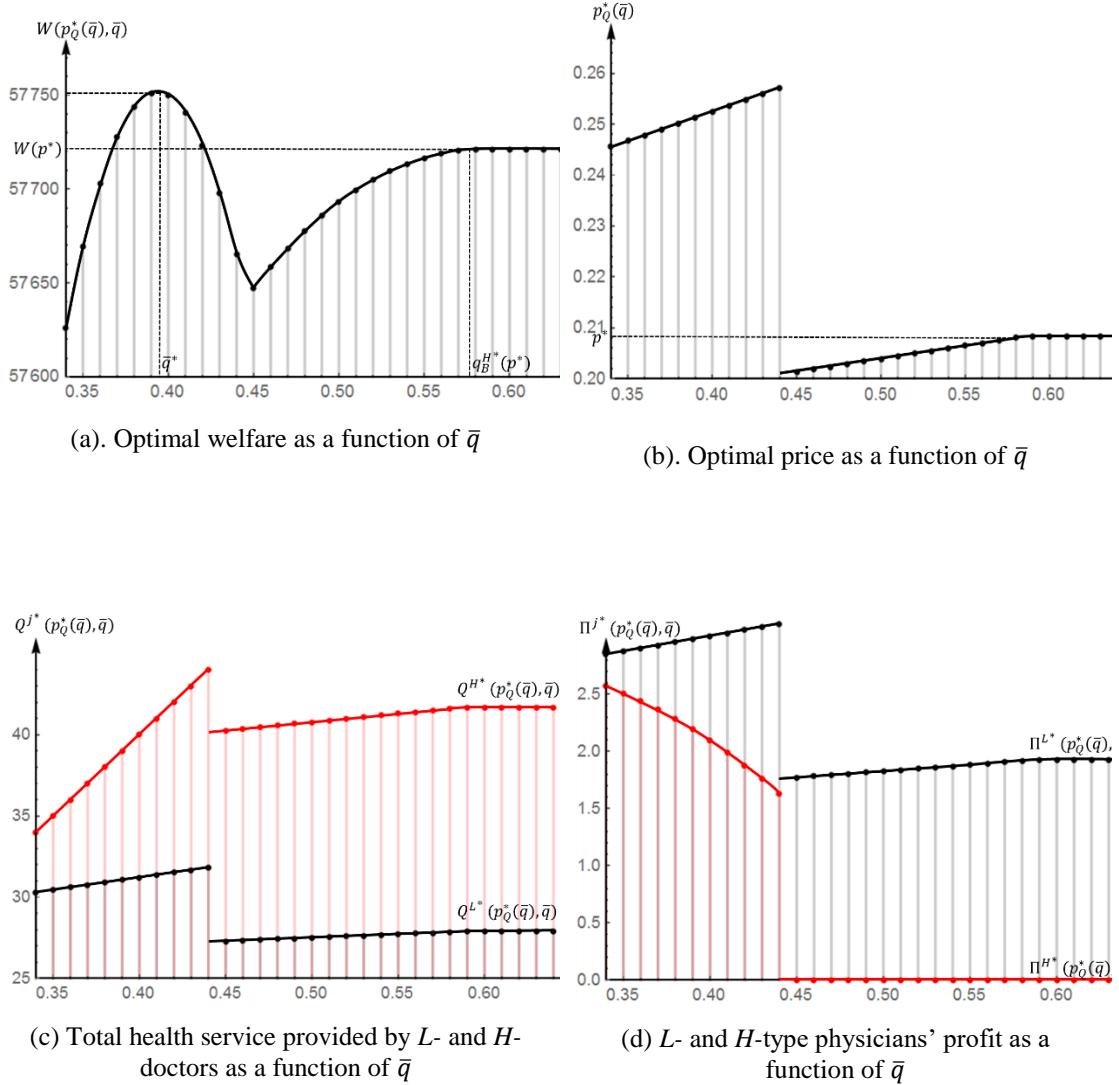


Figure 11. *Quantity Restriction*

Specifically, I use the example described in Section 3.5.3 and follow the procedure outlined just above in order to obtain the function $\bar{W}(p_Q^*(\bar{q}), \bar{q})$, which has been plotted in Figure 11a. The graph reflects the general intuition discussed above. Initially, reducing the level of rationing \bar{q} below $\bar{q} = 0.58$ worsens the welfare $\bar{W}(p_Q^*(\bar{q}), \bar{q})$ below $W(p^*)$. The numerical example is used to gain some further insights. Figure 11b shows that the initial reduction in \bar{q} induced the HI scheme to reduce price. As explained above, H -type doctors react to a small variation in \bar{q} by holding their total output constant (see Figure 11c), while shifting some of the production between B and G patients. This worsens the allocation of medical resources since it shifts services away from needy patients. In order to reduce this negative impact, the HI scheme finds that it is advantageous to lower the price, thereby curtailing q_G^H and Q^H . This process continues

until the reduction in \bar{q} is sufficiently large to constrain the level of services so that the HI can sharply raise the price (see Figure 11c) to obtain $q_B^H(p_Q^*(\bar{q}), \bar{q}) = q_G^H(p_Q^*(\bar{q}), \bar{q}) = \bar{q}$ without reducing welfare. As a result, Q^L and Q^H goes up and the L -type doctors increased medical service production more than H -type doctors (see Figure 11c). At this point, any further restriction on \bar{q} would, *ceteris paribus*, reduce the total medical service provided by H -doctors (see Figure 11c) and the HI scheme would adjust price downwards to counter this effect (see Figure 11a). Overall, the positive effect of the reduction in Q^H dominates the negative effect of price adjustment and welfare increases.

As anticipated in the previous paragraph, the discontinuity of the response function $p_Q^*(\bar{q})$ (and the feedback effect on all the other functions) implies that a marginal analysis cannot lead to a conclusive result for the general case. In particular, there is no guarantee that there is always a restriction \bar{q} such that $\bar{W}(p_Q^*(\bar{q}), \bar{q}) > W(p^*)$. Overall, I find that even though imposing a per patient quantity restriction on a FFS system could improve welfare (as in the numerical example), it does involve numerous drawbacks. First, imposing the restriction may generate a countervailing incentive via physicians' zero-profit constraint and there is no guarantee that the positive impact would dominate. Second, while rationing might improve welfare, it cannot induce efficiency. For instance, in the example, at the welfare optimum H -doctors ended up providing the same level of medical services to B - and G -patients despite their significant difference in medical needs. Note that such a policy is likely to face many criticisms; it would frustrate doctors (as they obtain lower utility), in particular the intrinsically motivated as well as health needy patients as their services may be reallocated to their less needy counterparts.

3.6.2 Revenue Cap

In this second subsection, I investigate an alternative regulation where the HI scheme restricts the average spending per doctor across patients. This subsection proceeds in the same way as the last section. First, a new policy variable is introduced denoted by \bar{r} which measures the maximal level of expenditure that a doctor can spend across all his patients. Accordingly, a HI policy is a triplet (p, \bar{r}, τ) . Second, the doctor's behaviour for a given policy vector (p, \bar{r}, τ) is derived, i.e. we solve:

$$\begin{aligned}
\tilde{U} &= \max_{q_t} \sum_{t \in T} n_t [\alpha(z_t + (1 - z_t)f(q_t)) + pq_t] - C \left(\sum_{t \in T} n_t q_t \right) \\
&\quad p \sum_{t \in T} n_t q_t - C \left(\sum_{t \in T} n_t q_t \right) \geq 0 \\
&\quad \bar{r} - p \sum_{t \in T} n_t q_t \geq 0 \\
&\quad q_t \geq 0
\end{aligned} \tag{53}$$

The first line of the system (53) represents the physician's objective function and the second, the third, and the last line shows the non-negative profit, the expenditure limit, and the non-negative service production respectively.

Third, I show in Appendix A9 that (53) has a unique solution. The respective medical service provided by L - and H -type doctors are denoted by $\tilde{q}_t^H(p, \bar{r})$ and $\tilde{q}_t^L(p, \bar{r})$. These solutions are then substituted into the equation (31) to generate the welfare $\tilde{W}(p, \bar{r})$. Next, I denote by $p_R^*(\bar{r})$ the price of medical services which maximizes $\tilde{W}(p, \bar{r})$ for a given expenditure cap \bar{r} . Lastly, I substitute the solution into the welfare function, yielding the optimal welfare for a given \tilde{r} which is denoted by $\tilde{W}(p_R^*(\bar{r}), \bar{r})$.

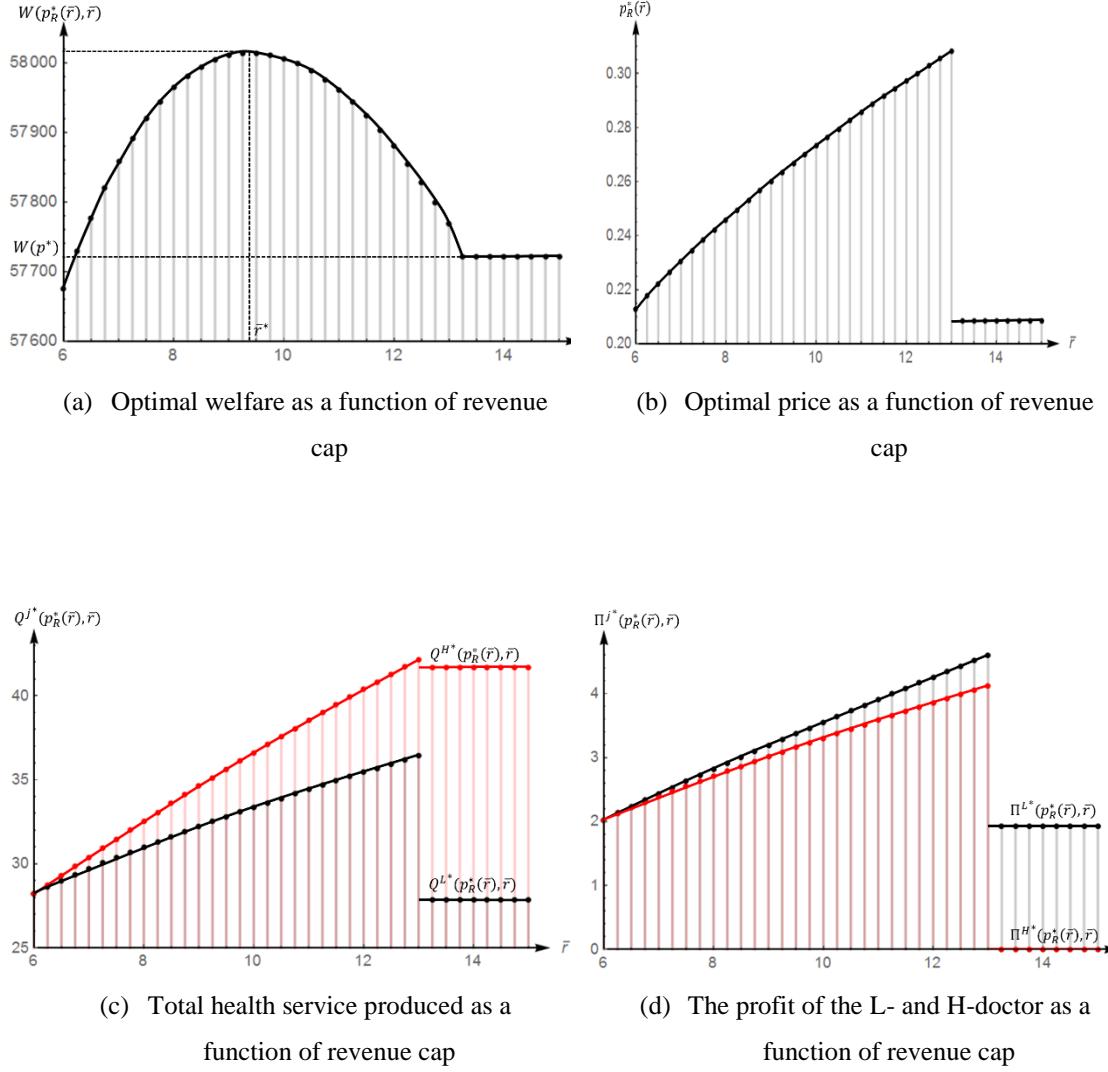


Figure 12. Revenue Restriction

In order to gain some insight, the numerical exercise in Section 3.6.1 is adjusted to evaluate the impact of a variation in the revenue cap. Figure 12 plots the result of the exercise for the parameter selected in Section 3.5.2. Observe that $\bar{r} \rightarrow +\infty$, the revenue cap never binds so that $\tilde{W}(p_R^*(\bar{r}), \bar{r}) = W(p^*)$. Initial reductions in the revenue cap have no impact until reaching the point $\bar{r} \approx 13$. First, consider Figure 12c for value $\bar{r} > 13$. From the foregoing section, we know that at p^* the HI trades off the benefit of an increase in price which would increase Q^L against the loss of an increase in Q^H . Next, suppose we are at the point $\bar{r} = 13$. From Figures 10c and 10d, we find that a small reduction in \bar{r} induces the HI to increase price such that

$$\bar{r} = p_R^*(\bar{r}) \tilde{Q}^H(p_R^*(\bar{r}), \bar{r}) \quad (54)$$

In other words, the revenue cap is such that the regulator is indifferent between imposing and relaxing the cap. To illustrate this further, at $\bar{r} = 13$ the regulator has raised the price to a level such that $Q^H(p_R^*(\bar{r})) > \tilde{Q}^H(p_R^*(\bar{r}), \bar{r}) > Q^H(p^*)$ and $\tilde{Q}^L(p_R^*(\bar{r}), \bar{r}) > Q^L(p_R^*(\bar{r})) > Q^L(p^*)$. As a result, price and total quantity of medical service go up and the regulator increases the health insurance premium. Altogether, the marginal reduction in \bar{r} at $\bar{r} = 13$ does not affect the welfare level as the positive impact of Q^L increases cancelled out the negative effect of the \tilde{Q}^H and τ increase (see Figure 12a). Finally, more reduction in \tilde{r} makes H -type doctors *ceteris paribus* reduce the \tilde{Q}^H and allows the HI scheme to reduce the price (see Figures 12b and 12c). The L -type doctors would therefore reduce \tilde{Q}^L and the insurance premium goes down (see Figure 12c). Altogether, the positive effect on the reduction in \tilde{Q}^H and insurance premium outweighs the negative effect on the reduction in Q^L , therefore welfare goes up (see Figure 12a). This process will continue until all the margins are realigned and the welfare maximum is attained at \bar{r}^* .

To conclude this subsection, a few remarks are in order. First, from Figure 12c note that, compared with the optimal policy without revenue restrictions, H -doctors are induced to produce less while L -doctors produce more. In other words, the optimal revenue cap helps the HI scheme to overcome one of the weaknesses associated with the standard FFS system which is linked to the informational asymmetry between HI and doctors. Second, compared with the quantity restriction, the revenue cap provides doctors with more flexibility which allows them to better adjust healthcare resources to the specific needs of individual patients. One would expect that this additional flexibility would make the revenue cap policy more palatable to doctors than the foregoing quantity restriction. Moreover, in this numerical example patients with initially low health were also better off with the revenue cap.

3.6.3 Capitation

In this last subsection, I explore another popular form of remuneration which pays the physician based on the number of patients treated and the total quantity of medical service produced. In line with the steps outlined in the above subsections, a fixed payment rate per patient denoted by $k \in (-\infty, +\infty)$ is introduced. A HI policy is

therefore a triplet (p, k, τ) . Next, the physicians' best responses to a given policy vector (p, k, τ) are derived, i.e. a general physician solves:

$$\begin{aligned} \widehat{U} &= \max_{q_t} \sum_{t \in T} n_t [\alpha(z_t + (1 - z_t)f(q_t)) + pq_t] + k \sum_{t \in T} n_t - C \left(\sum_{t \in T} n_t q_t \right) \\ p \sum_{t \in T} n_t q_t + k \sum_{t \in T} n_t - C \left(\sum_{t \in T} n_t q_t \right) &\geq 0 \\ q_t &\geq 0 \end{aligned} \tag{55}$$

where the first line of the system (55) is the physician's utility function with an additional term k representing the reimbursement (capitation) per patient. The second and third lines represent the profit and production constraint respectively.

The optimization problem (55) has a unique solution, as can be shown in the Appendix A10. I denote by $\hat{q}_t^L(p, k)$ and $\hat{q}_t^H(p, k)$ the respective quantity of medical service provided by L -and H -type doctors. By substituting these solutions into the initial welfare representation, one obtains the welfare function $\widehat{W}(p, k)$, finally, solving the welfare optimizing price $p_K^*(k)$ and substituting this price into $\widehat{W}(p, k)$ which yields the function $\widehat{W}(p_K^*(k), k)$.

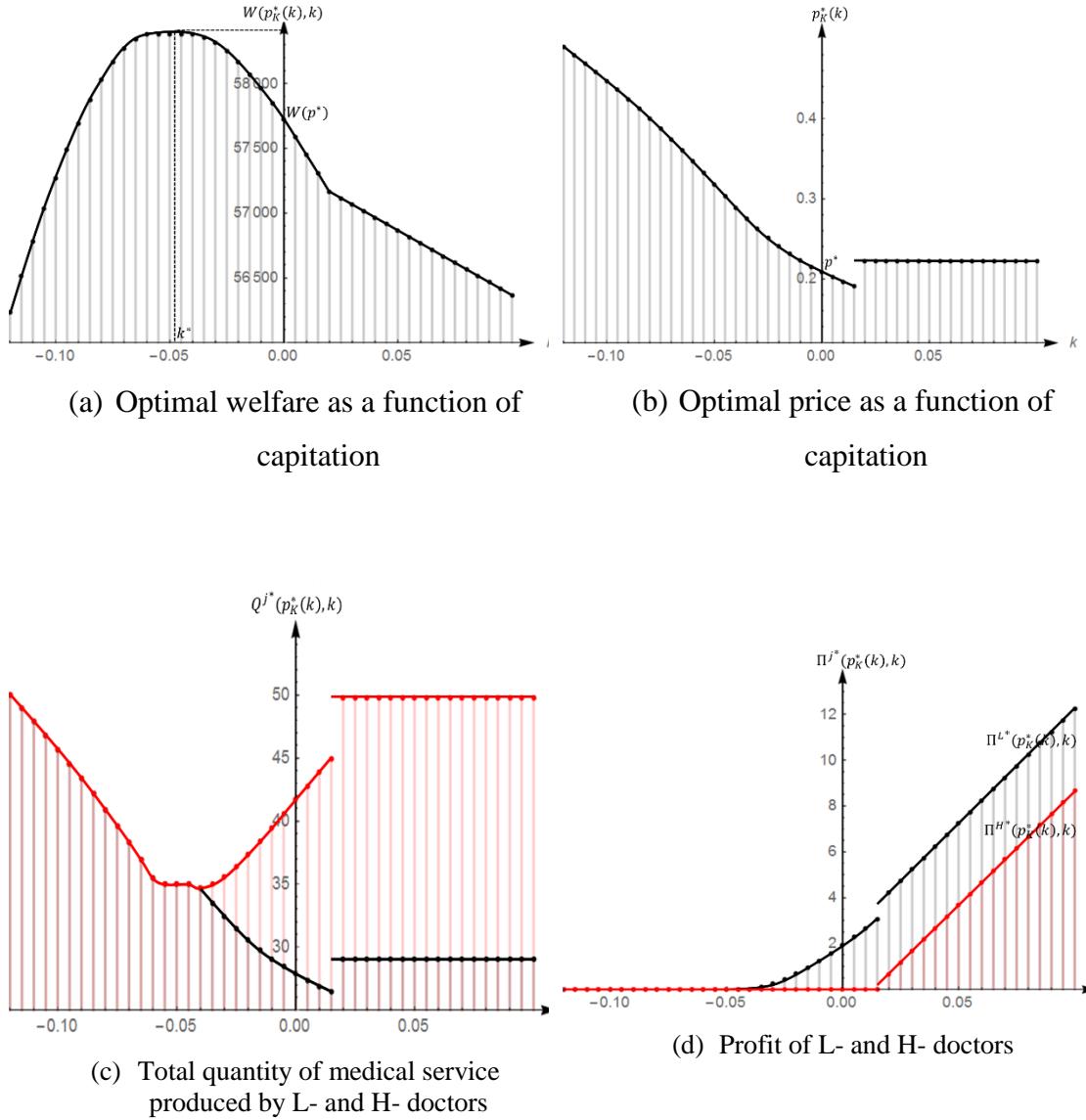


Figure 13. Capitation

To proceed, the numerical experiment conducted in Sections 3.5.1 is adjusted to examine the effect of changing the capitation rate. In Figure 13, I plot the results of the exercise for the parameter set introduced in Section 3.5.2. At the capitation rate $k = 0$, $p_K^*(0) = p^*$ the HI scheme trades-off the benefit of an increase in the production of L -doctors against the deterioration in welfare from an increase in production by H -doctors. Moreover, at p^* the H -type provider's profit is just binding. Notice that when $k > 0$, Figure 13b shows that the $p_K^*(k)$ initially goes down, then jumps upwards and become constant. Intuitively, marginal increase in k relaxes H -type physicians' binding profit constraint so that they raise their production of medical service. As a result, the average marginal cost of production increases and the HI scheme found it is advantageous to

adjust price downward. As a result, \hat{Q}^L decreases, the insurance premium goes up, and welfare goes down (see Figure 13a). When k is sufficiently high so that H -doctors are no longer constrained by the non-negative profit condition, the HI scheme will raise the price sharply to the level that maximizes unconstrained welfare optimization. Further increases in k would only increase physicians' profits and therefore the total costs of the HI scheme, leading to a deterioration of social well-being (see Figure 13a). Hence, when the regulator imposes the capitation $k > 0$, the optimal welfare $\hat{W}(p_K^*(k), k)$ is lower than the optimal welfare without imposing capitation $W(p^*)$.

However, when k becomes negative, the H -type doctor will reduce \hat{Q}^H to avoid negative profit (see Figures 13c, d). To counter the effect of the reduction in \hat{Q}^H , the HI scheme adjusts the price upward and L -type physicians increase their production of medical services (see Figures 13b, c). In terms of total expenditure, the effect of the reduction in \hat{Q}^H and the decrease in k dominates the increase in price and \hat{Q}^L , thereby lowering the insurance premium. Overall, introducing a negative capitation aligns different physicians' incentives, reduces the insurance premium, thus social welfare goes up (see Figure 13a). This process continues until all the margins are aligned at the rate $k^* \approx -0.05$, where $\hat{Q}^H(p_K^*(k), k) = \hat{Q}^L(p_K^*(k), k)$ and $\hat{\Pi}^H(p_K^*(k), k) = \hat{\Pi}^L(p_K^*(k), k) = 0$.

In short, this subsection shows that the optimal physician payment system should blend the negative capitation and positive FFS. This result differs in an important way from what has become the conventional wisdom (see Kantarevic and Kralj, 2016), as mixing positive capitation and retrospective reimbursement would leave physicians with informational rent while reducing the production incentives of low altruistic providers. By contrast, blending the negative capitation with high retrospective payments would reduce H -type doctors' over-production while motivating L -type doctors to raise their production.

As a result, the HI scheme no longer trade off the benefit of an increase in the production of L -doctors against the reduction in welfare from an increase in production by H -doctors. In reality, the negative capitation can be interpreted as the spending physicians paid to buy rights to serve patients while the payment can be compensated by the rise in health service price. The capitation is not only a tool that aligns physicians' incentives but also an instrument that extracts physicians' informational rent. The experiment also

shows that the negative capitation outperforms quantity rationing and revenue restriction, as these cost-containment instruments cannot completely align physicians' incentives and is not able to extract physicians' informational rent.

3.6.4 Discussion

Standard analysis of the FFS scheme focuses on the use of fees to align doctors' incentives. However, asymmetric information between physicians, patients and health insurance systems has been found in the literature, resulting in medical service overproduction, health spending escalation and inefficient allocation of health resource (Léger, 2008; McGuire, 2000). Specifically, the foregoing sections have underlined that ignoring extrinsic motivation and the difference thereof across doctors necessarily leads to an inefficiency in the allocation of medical resources. The main reason for this inefficiency is that, for the same fee structure, high intrinsically motivated staff automatically supply more medical services than low motivated individuals. A natural response to this inefficiency is to regulate medical service price. However, as all doctors increase their service supply with price, raising or reducing price would further exacerbate highly motivated doctors' overproduction or low motivated doctors' underproduction. Hence, imposing a conventional price regulation does not guarantee an improvement in the efficiency of health resource allocation.

A logical reaction to the weakness of price regulation is to adjust the scheme by introducing additional restrictions aimed at capping the supply of medical services by high intrinsically motivated medical staff. Doing so helps to curb highly motivated doctors' excessive supply while the HI scheme can simultaneously increase the price for services with the aim to reduce the medical service under-provision of less motivated medical staff.

In that respect, although individual quantity rationing is shown to be useful, it can only be used in a relatively restrictive environment where the quantity of medical service supplied to individual patient can be closely and accurately monitored. Without such monitoring, quantity rationing may induce highly intrinsically motivated doctors to shift service supply away from patients with more health needs to their less needy counterparts, holding the total service supply constant and therefore resulting in inefficient resource allocation. Since the costs associated with monitoring is high

(Chalkley and Malcomson, 2002) and the regulator does not know patients' health, it may be difficult to implement individual quantity rationing.

My experiments pointed to an additional difficulty with a pure quantity restriction of services, namely, welfare may not react smoothly (see Figure 11). Theoretically, this is not a problem and welfare would increase if the quantity rationing is sufficiently restrictive. However, from a practical point of view, this may be much more complicated, as it is very difficult for the regulator to decide the exact level of the maximum service quantity per patient that can be provided. Further, potential deviations from the quantity rationing target may result in a deterioration of current social well-being.

Finally, even when the HI scheme pays monitoring costs and successfully implements efficiency enhancing quantity rationing, the policy may receive many criticisms from both patients and doctors. For instance, physicians might observe that they cannot use their local knowledge while some of them – in particular those who are highly motivated – would even complain that their well-being is deteriorated. Patients with high levels of health needs are also likely to complain as they receive the same amount of service as their low health needs counterparts but have to pay a higher premium.

Some of the previous unintended effects associated with imposing individual quantity rationing can be also found in reality. For instance, Chua et al. (2019) has shown that imposing limits on opioid may result in the excessive prescriptions and inadequate pain control in the US, as a uniform restriction does not satisfy patients' heterogeneous combinations of opioid use and pain needs. Moreover, their evidence shows that the quantity restriction is either set too high to reduce excessive prescribing or too low to avoid the potential for inadequate pain control. While physicians' prescription of opioids has been restricted, patients can nevertheless obtain opioids from other resources. A recent report by the US Substance Abuse and Mental Health Services Administration (SAMHSA) found that about half of people who misused prescription opioids in the past year obtained them from a friend or relative for free (Lipari and Hughes, 2017).

Overall, this study shows while individual quantity rationing could be used to enhance resource allocation efficiency, the unintended effects, high costs associated with monitoring, as well as potential criticisms from both patients and doctors has restricted its further application. Given the aforementioned limitations of quantity rationing, one would consider controlling total medical expenditure within a primary care practice to

be a better alternative as the revenue cap requires less monitoring and allows physicians to use their local knowledge, thereby reducing costs and improving patients' health benefits.

Specifically, this policy reduces high intrinsically motivated physicians' over-production by directly limiting their total expenditure while the HI scheme simultaneously raises the fee-for-service price to incentivise service provision by low intrinsically motivated physicians. However, compared to quantity rationing, this policy does not provide physicians with a countervailing incentive which induces highly motivated physicians to shift service from the more health needy to less needy individuals. Furthermore, I find this policy is always efficiency enhancing, as long as the maximum expenditure per physician can be set below the remuneration level that the most motivated physician would receive while exceeding the reward that the least motivated doctors would gain under a pure FFS system. This specific range of effective revenue restriction will be a lot easier and less costly for the regulator to determine in comparison to quantity rationing.

The policy also does not require a detailed monitoring of physicians' medical service supply, thereby saving all monitoring costs. Finally, as the expenditure cap allows physicians to use their local knowledge, patients with high severe illnesses are provided with more services, therefore average patient health is improved.

While an expenditure cap overcomes most weaknesses associated with using quantity rationing, my study reflects some of its disadvantages. First, while the initial reduction of the expenditure cap per physician limits more altruistic physicians' overproduction, the subsequent sharp increase in price provides all physicians with additional informational rents. Next, to further reduce the more motivated physicians' overproduction, the health authority keeps reducing the maximum health spending per physician. However, the HI scheme would adjust the price downward to counter this reduction, which weakens the low motivated physicians' incentives to increase their service production. These two observations show that introducing an expenditure cap in a FFS system could not extract physicians' informational rents (as a result of informational asymmetry) and align different physicians' incentives (thereby enhancing efficiency) simultaneously.

Indeed, the real-world applications of expenditure cap demonstrate some of the insights derived above. First, many more applications of revenue restriction compared to quantity rationing can be found in the real-world applications. For instance, the UK has applied quality adjusted life years, on the basis that, for each extra year of life delivered, if it costs more than £30,000 then, adjusted for quality, it should not be publicly funded (Cookson, 2013). Germany has also extensively used sectoral and regional budgets for hospitals and ambulatory care (Stabile et al., 2013). Moreover, France introduced a national cap on statutory health insurance expenditure in 1997, capping healthcare spending for six sectors (Sandier et al., 2004).

Second, comparing revenue cap and quantity rationing, both theory and experiments suggest the former policy induces more service provision than the latter given the same budget (Fan et al., 1998). Mougeot and Naegele (2005) further demonstrate that a revenue cap cannot solve the conflicts between allocation efficiency and rent extraction, therefore it can never act as a means of achieving first-best social welfare. Poterba (1994), Benester and Wambach (2006) and Fischer et al. (2018) have also found that using revenue caps may drive health spending in the long-term, exacerbating medical service overproduction while inducing inappropriate drug prescription for elderly patients. As a result, one should further consider a policy which provides physicians incentives that align with the one embedded in FFS, while extracting physicians' informational rents. The former requirement keeps patients' total health benefits unchanged while reducing the overall costs of medical operations. The latter reduces the regulator's total health spending on physicians and the costs generated from collecting insurance premium.

My numerical analysis shows that negative capitation is the best cost-control instrument; the negative capitation curbs the high intrinsically motivated doctors' excessive service provision while extracting informational rents from low motivated physicians without affecting their incentives in supplying medical service. The HI scheme, therefore, found that it is advantageous to adjust price upwards, which incentivises low motivated physicians to provide more services. This process continues until both types of physicians provide the same amount of service for respective types of patients while having a zero profit. In short, negative capitation is used to align all physicians' incentives while an FFS payment reimburses the costs that physicians incur. Notice my model does not consider fixed costs associated with enrolling a patient, therefore

negative capitation is an artefact. In reality, the HI scheme should blend price with a positive capitation less than the fixed costs associated with treating each individual.

Mixed payment systems comprise of fee-for-service payments and capitation has been widely used. For instance, primary care physicians in the UK, Finland, Norway and Italy are paid based on a combination of capitation and FFS for the purpose of containing health expenditure and reducing hospital referrals (Park et al., 2007). In the US, capitation payments have also been used in both outpatient and inpatient care within the framework of Health Maintenance Organizations (HMOs) or managed care plans (Carrin, 2012). Despite these applications, there is limited evidence that the capitation has been set at less than the administrative cost per patient in a blended payment system.

However, outside of the healthcare sector, there are many applications of negative capitation. For instance, a bonding contract comprised of a positive bonus and a negative fixed payment is frequently used in the construction sector to guarantee that the contractor's performance in accordance with the conditions of their contract as well as protecting against the contractor's default (Awad and Fayek, 2012). Moreover, franchise agreement also requires the franchisee pays a franchisee (i.e. a firm set up to market a product or service in a specific location) pays a franchisor (i.e. the parent company which developed some products or services) a certain sum money for the right to sell the franchisor's services or products (Rubin, 1978). For instance, McDonalds require at least £110,000 unencumbered funds from its franchisee to invest in the UK, with £40,000-£80,000 to buy a restaurant and a one-off franchise fee of £30,000.

3.7 Conclusion

In this chapter, I derived the optimal physician payment mechanisms based on numerical simulations. My results show that FFS system is inefficient as it fails to align incentives of heterogeneously motivated physicians. The study then reproduces the impact of the recurrent trends; these findings suggest that price and insurance premiums increase with ageing and science innovations, which are consistent with some stylized facts presented for healthcare systems in most developed countries. Moreover, revolutions in current trends also exacerbate the misallocation of health resources.

The numerical analysis then provides both qualitative and quantitative recommendations for policy changes. I found that introducing FFS system a ceiling on the maximum number of medical services provided per patient, an expenditure cap per physician, or a negative capitation could enhance efficiency in terms of the use of healthcare resources. In particular, social welfare will be improved if the foregoing restrictions are selected at a level that could cap the health services provided by strongly motivated physicians. Finally, I find that the negative capitation is superior to other cost-containing methods as it perfectly aligns L- and H-type doctors' incentives while extracting all informational rents.

3.8 Limitations and Extensions

The theoretical framework can be extended in two important ways. First, primary care providers often face a plurality of problems and have multiple sets of activities. They not only need to decide the service quantity/quality provided per patient but also the type or the number of patients to be treated. In addition, they determine the level of cost reducing efforts and whether or not to refer patients to specialized treatment. Accordingly, a natural extension would attempt to allow physicians select the number of patients to be treated, which will be discussed in Chapter 4. Furthermore, it is worth to check whether my current findings hold when physicians can also select patient types. Physicians' decisions on the level of their quality enhancing efforts and speciality care referral have been discussed by Allard et al. (2011) and Allard et al. (2014). However, these studies do not consider how these decisions are influenced by cost control policies.

Second, my work has assumed that patients differ only in their initial health while physicians differ only in their degree of altruism. Furthermore, my model considers the simplest situation where there are only two types of physicians and patients. It is worthwhile to examine whether my current findings can be generalized when there are more than two types of patients or physicians. Meanwhile, it will also be interesting to determine how the introduction of additional heterogeneities such as individual benefits from treatment or marginal costs of service provision will change my results. These extensions have been discussed respectively by Chalkley and Malcomson (2002), Makris and Siciliani (2013) and Kantarevic and Kralj (2016), although these studies

focused on designing a menu of contracts. However, they have not considered the implementation of cost control policies.

3.9 Appendix

A1

Applying the implicit function theorem to the system (36), we have:

$$\begin{pmatrix} \frac{\partial q_B^*}{\partial p} \\ \frac{\partial q_G^*}{\partial p} \end{pmatrix} = - \begin{pmatrix} \alpha f_B'' - n_B C'' & -n_G C'' \\ -n_B C'' & \alpha(1-z)f_G'' - n_G C'' \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad (56)$$

Inverting the matrix yields:

$$\begin{pmatrix} \frac{\partial q_B^*}{\partial p} \\ \frac{\partial q_G^*}{\partial p} \end{pmatrix} = - \frac{1}{\det} \begin{pmatrix} \alpha f_B'' - n_B C'' & -n_G C'' \\ -n_B C'' & \alpha(1-z)f_G'' - n_G C''' \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad (57)$$

where \det denotes the determinant of the matrix. Substitute and simplify yields:

$$\frac{\partial q_B^*}{\partial p} = \frac{-(1-z)f_G''}{\alpha(1-z)f_B''f_G'' - C''[n_G f_B'' + n_B(1-z)f_G'']} > 0 \quad (58)$$

$$\frac{\partial q_G^*}{\partial p} = \frac{-f_B''}{\alpha(1-z)f_B''f_G'' - C''[n_G f_B'' + n_B(1-z)f_G'']} > 0 \quad (59)$$

where the respective signs follow from the curvature assumptions $f'' < 0, C'' > 0$. Moreover, since $\lim_{q \rightarrow 0} f'(q) = +\infty, \lim_{q \rightarrow +\infty} f'(q) = 0$, $f' > 0, f'' < 0$, we must have $f''' > 0$. For parsimony, I assume $0 > f_B'' \geq (1-z)f_G''$. Hence, we have $-(1-z)f_G'' \geq -f_B'' > 0$ and $\frac{\partial q_B^*}{\partial p} \geq \frac{\partial q_G^*}{\partial p}$, thereby verifying the claim.

A2

Following the same proof procedures as in the appendix A1, we have:

$$\begin{pmatrix} \frac{\partial q_B^*}{\partial \alpha} \\ \frac{\partial q_G^*}{\partial \alpha} \end{pmatrix} = - \frac{1}{\det} \begin{pmatrix} \alpha f_B'' - n_B C'' & -n_G C'' \\ -n_B C'' & \alpha(1-z)f_G'' - n_G C'' \end{pmatrix} \begin{pmatrix} f'_B \\ (1-z)f'_G \end{pmatrix} \quad (60)$$

Hence,

$$\frac{\partial q_B^*}{\partial \alpha} = \frac{-(1-z)f_G''f'_B}{\alpha(1-z)f_B''f_G'' - C''[n_Gf_B'' + n_B(1-z)f_G'']} > 0 \quad (61)$$

$$\frac{\partial q_G^*}{\partial \alpha} = \frac{-f_B''(1-z)f'_G}{\alpha(1-z)f_B''f_G'' - C''[n_Gf_B'' + n_B(1-z)f_G'']} > 0 \quad (62)$$

As shown in the appendix A1, $-(1-z)f_G'' \geq -f_B'' > 0$. Moreover, as $f'_B = (1-z)f'_G$, we have $\frac{\partial q_B^*}{\partial \alpha} \geq \frac{\partial q_G^*}{\partial \alpha} > 0$ verifying the claim.

A3

Denote $q_B - q_G$ as Δq , following appendix A1 and A2, we also have:

$$\begin{pmatrix} \frac{\partial q_B^*}{\partial \beta} \\ \frac{\partial q_G^*}{\partial \beta} \end{pmatrix} = \frac{1}{\det} \begin{pmatrix} \alpha f_B'' - n_B C'' & -n_G C'' \\ -n_B C'' & \alpha(1-z)f_G'' - n_G C'' \end{pmatrix} \begin{pmatrix} \Delta q C'' \\ \Delta q C'' \end{pmatrix} \quad (63)$$

accordingly,

$$\frac{\partial q_B^*}{\partial \beta} = \frac{(1-z)f_G''}{\alpha(1-z)f_B''f_G'' - C''[n_Gf_B'' + n_B(1-z)f_G'']} < 0 \quad (64)$$

$$\frac{\partial q_G^*}{\partial \beta} = \frac{f_B''}{\alpha(1-z)f_B''f_G'' - C''[n_Gf_B'' + n_B(1-z)f_G'']} < 0 \quad (65)$$

As shown in appendix A1, $0 > f_B'' \geq (1-z)f_G''$. Hence, we have $\frac{\partial q_B^*}{\partial \beta} \leq \frac{\partial q_G^*}{\partial \beta} < 0$, verifying the claim.

A4

From the system (41), we have:

$$\begin{pmatrix} \frac{\partial q_B^*}{\partial p} \\ \frac{\partial q_G^*}{\partial p} \end{pmatrix} = - \begin{pmatrix} f_B'' & -(1-z)f_G'' \\ n_B(p - C') & n_G(p - C') \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ n_B q_B + n_G q_G \end{pmatrix} \quad (66)$$

Inverting the matrix, solving and applying the curvature assumptions yields:

$$\frac{\partial q_B^*}{\partial p} = \frac{-Q^*(1-z)f_G''}{(p-C')[n_Gf_B'' + n_B(1-z)f_G'']} > 0 \quad (67)$$

$$\frac{\partial q_G^*}{\partial p} = \frac{-Q^*f_B''}{(p-C')[n_Gf_B'' + n_B(1-z)f_G'']} > 0 \quad (68)$$

This verifies the first part of the claim. Next, as shown in Appendix A, we have $0 > (1-z)f_G'' > f_B''$. As a result, the above slopes satisfy $\frac{\partial q_B^*}{\partial p} > \frac{\partial q_G^*}{\partial p}$. Finally, we verify $\frac{\partial q_B^*}{\partial p}(p_0^-) > \frac{\partial q_B^*}{\partial p}(p_0^+)$. The other possibility where $t = G$ is perfectly symmetrical and is left to the reader. Using the equations (58) and (67), we need to show:

$$\begin{aligned} & \frac{-(1-z)f_G''}{\alpha(1-z)f_B''f_G'' - C''[n_Gf_B'' + n_B(1-z)f_G'']} \\ & < \frac{Q^*(1-z)f_G''}{(C'-p)[n_Gf_B'' + n_B(1-z)f_G'']} \end{aligned} \quad (69)$$

Rearranging terms and taking the respective signs of f'' and C'' into account, we obtain:

$$[n_Gf_B'' + n_B(1-z)f_G''](Q^*C'' - C' + p_0) < Q^*\alpha(1-z)f_B''f_G'' \quad (70)$$

This is true since the RHS is positive while the LHS is negative. The observation follows because the first square bracket is negative whereas the second is positive by $C''' > 0$ and $p_0Q^* - C(Q^*) = 0$.

A5

As proved in A4, we have:

$$\begin{pmatrix} \frac{\partial q_B^*}{\partial \beta} \\ \frac{\partial q_G^*}{\partial \beta} \end{pmatrix} = - \begin{pmatrix} f_B'' & -(1-z)f_G'' \\ n_B(p-C') & n_G(p-C') \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ -n\Delta q(p-C') \end{pmatrix} \quad (71)$$

Following the same procedures as in the A6 and denote $\Delta q^* = q_B^* - q_G^*$, we obtain

$$\frac{\partial q_B^*}{\partial \beta} = \frac{-n\Delta q^*(1-z)f_G''}{n_G f_B'' + n_B(1-z)f_G''} < 0 \quad (72)$$

$$\frac{\partial q_G^*}{\partial \beta} = \frac{-n\Delta q^*f_B''}{n_G f_B'' + n_B(1-z)f_G''} < 0 \quad (73)$$

The assumption $0 > f_B'' \geq (1-z)f_G''$ indicates that $\frac{\partial q_B^*}{\partial \beta} < \frac{\partial q_G^*}{\partial \beta} < 0$, therefore, verifying the claim.

A6

Consider the production vector for services which solves (35). We use this solution to define the ensuing total production of services denoted by $\hat{Q}(p, \alpha)$.⁵¹ The critical price p_0 is implicitly a function of α which is determined by solving:

$$p_0 \hat{Q}(p_0, \alpha) - C(\hat{Q}(p_0, \alpha)) = 0 \quad (74)$$

Calculate the first order condition of (74) with respect to α , we have:

$$\left[\hat{Q} + (p_0 - C') \frac{\partial \hat{Q}}{\partial p} \right] \frac{\partial p_0}{\partial \alpha} + (p_0 - C') \frac{\partial \hat{Q}}{\partial \alpha} = 0 \quad (75)$$

Following the same steps as outlined in Appendix A, it is immediately evident that $\frac{\partial \hat{Q}}{\partial p}(p_0, \alpha) > 0$ and $\frac{\partial \hat{Q}}{\partial \alpha}(p_0, \alpha) > 0$. Moreover, the system (35) implies that $p_0 - C' < 0$. Following the same procedures from (38) to (40) it can be proved that $\hat{Q} + (p_0 - C') \frac{\partial \hat{Q}}{\partial p} > 0$. As a result, we must have $\frac{\partial p_0}{\partial \alpha} > 0$, verifying the claim.

A7

Proof. First, observe at $p = 0$, patients' total health is positive ($z > 0$), keeping in mind that welfare sums up individuals' health and the system's overall medical cost $W(0) >$

⁵¹ Slightly abusing the notation, observe that $\hat{Q}(p, \alpha) = Q^*(p, \alpha)$ iff $p \in P$.

0. Moreover, for $p \in [0, p_0^L)$, L - and H - doctors' profits are strictly binding so that $q_t^L(p) = q_t^H(p) = \tilde{q}(p)$ and the slope of the welfare as a function of p is:

$$\frac{\partial W}{\partial p}(p) = [f'\tilde{q}_B(p) - \lambda p - C(\tilde{Q}(p))] \frac{\partial \tilde{Q}}{\partial p}(p) - \lambda \tilde{Q}(p) \quad (76)$$

For $p \rightarrow 0^+$, $\tilde{q}_B(0) = \tilde{Q}(0) = C(\tilde{Q}(0)) = 0$. However, with $f'(\tilde{q}_B(0)) = +\infty$ and $\frac{\partial \tilde{Q}}{\partial p}(0^+) > 0$, we have $\frac{\partial W}{\partial p}(0^+) > 0$.

Second, q_t^L, q_t^H are increasing in p , so the slope and curvature assumptions ensures that $\sum_{i=1}^N h_i(p)$ is increasing concave and bounded (i.e. $\sum_{i=1}^N h_i(p) \leq N$). However, the total cost $\lambda \sum_{i=1}^N p q_i(p) + \sum_{j=1}^M C\left(\sum_{i=(j-1)n+1}^{nj} q_i(p)\right)$ is increasing and unbounded.

Hence, as p becomes sufficiently large, we have $\lambda \sum_{i=1}^N p q_i(p) + \sum_{j=1}^M C\left(\sum_{i=(j-1)n+1}^{nj} q_i(p)\right) > N$ and $W(p) < 0$. Moreover, notice that $\lambda \sum_{i=1}^N p q_i(p) + \sum_{j=1}^M C\left(\sum_{i=(j-1)n+1}^{nj} q_i(p)\right)$ is a continuous function of p and $\lambda \sum_{i=1}^N p q_i(0) + \sum_{j=1}^M C\left(\sum_{i=(j-1)n+1}^{nj} q_i(0)\right) = 0$, with the intermediate value theorem showing that there exists a price p_N such that $\lambda \sum_{i=1}^N p q_i(p_N) + \sum_{j=1}^M C\left(\sum_{i=(j-1)n+1}^{nj} q_i(p_N)\right) = N$, hence, for $p = p_N$, $W'(p_N) < 0$. Altogether, the extreme value theorem implies that there exists at least a price $p^* \in [0, p_N]$ that maximizes $W(p)$. Finally, given $W(p) < 0$ for all $p > p_N$, $\max_{p \in [0, p_N]} W(p) = \max_{p \in [0, +\infty]} W(p)$, this verifies the claim.

A8

The Lagrange of (51) is:

$$\begin{aligned} \mathcal{L} = & \alpha \sum_{t \in T} n_t [z_t + (1 - z_t)f(q_t)] + \sum_{t \in T} \mu_t (\bar{q} - q_t) \\ & + (1 + \eta) \left[p \sum_{t \in T} n_t q_t - C\left(\sum_{t \in T} n_t q_t\right) \right] \end{aligned} \quad (77)$$

while the necessary condition that (77) has a solution for q_B and q_G is:

$$\left\{ \begin{array}{l} \alpha n_B f'(q_B) + (1 + \eta) n_B [p - C'(Q)] + \mu_B = 0 \\ \alpha(1 - z) n_G f'(q_G) + (1 + \eta) n_G [p - C'(Q)] + \mu_G = 0 \\ \eta [pQ - C(Q)] = 0 \\ \mu_B (\bar{q} - q_B) = 0 \\ \mu_G (\bar{q} - q_G) = 0 \\ \eta, \mu_B, \mu_G = 0 \end{array} \right. \quad (78)$$

The first two equations in (78) are the first-order conditions with respect to q_B and q_G .⁵² The next three lines in (78) are the corresponding complementary slackness conditions associated with the inequalities in the optimization problem. Finally, the last line gives the restriction on the respective multipliers of non-negative requirements for profit and services. The first order condition with respect to q_B and q_G in (78) can be set equal to zero, therefore we have the necessary condition to solve utility maximization q_B . In addition, the functions f' and C' are monotonous while the respective solution of q_B and q_G is unique for a given policy vector.

Finally, we derive the second order condition of (78) with respect to q_t and get the Hessian:

$$\begin{bmatrix} \mathcal{L}_{BB} & \mathcal{L}_{BG} \\ \mathcal{L}_{GB} & \mathcal{L}_{GG} \end{bmatrix} = \begin{bmatrix} \alpha(1 - z)f''_B - n_B^2(1 + \eta)C'' & -n_B n_G (1 + \eta)C'' \\ -n_B n_G (1 + \eta)C'' & \alpha n_G (1 - z)f''_G - n_G^2(1 + \eta)C'' \end{bmatrix} \quad (79)$$

Since $\alpha(1 - z)f''_B - n_B^2(1 + \eta)C'' < 0$, and $[\alpha(1 - z)f''_B - n_B^2(1 + \eta)C''][\alpha n_G (1 - z)f''_G - n_G^2(1 + \eta)C''] - (n_B n_G (1 + \eta)C'')^2 > 0$, the sufficient and necessary condition all support the presence of q_B and q_G to maximize (51) therefore our problem is well-defined.

⁵² Slightly abusing the notation, I denote the solution of the system as q_t, μ_t and η as in the previous section.

A9

Writing the Lagrange of the problem (53), we have:

$$\begin{aligned}\mathcal{L} = & \alpha \sum_{t \in T} n_t [z_t + (1 - z_t)f(q_t)] + \sum_{t \in T} \mu_t q_t \\ & + (1 + \eta - \omega)p \sum_{t \in T} n_t q_t - (1 + \eta)C \left(\sum_{t \in T} n_t q_t \right) + \omega \tilde{r}\end{aligned}\quad (80)$$

Following the same procedure as in the last subsection, we derive the necessary conditions such that (80) has a unique solution:

$$\left\{ \begin{array}{l} \alpha n_B f'(q_B) + (1 + \eta)n_B[p - C'(Q)] - \omega n_B p + \mu_B = 0 \\ \alpha(1 - z)n_G f'(q_G) + (1 + \eta)n_G[p - C'(Q)] - \omega n_G p + \mu_G = 0 \\ \omega(\tilde{r} - pQ) = 0 \\ \eta[pQ - C(Q)] = 0 \\ \mu_B q_B = 0 \\ \mu_G q_G = 0 \\ \eta, \omega, \mu_B, \mu_G \end{array} \right. \quad (81)$$

The first two equations in (81) are the first-order conditions with respect to q_B and q_G . The next four lines in (81) are the corresponding complementary slackness conditions associated with the inequalities in the optimization problem. Finally, the last line gives the restriction on the respective multipliers of non-negative requirements for profit, services and the cap. The first order condition with respect to q_B and q_G in (81) can be set equal to zero, then we have the necessary condition to solve utility maximization q_t . In addition, the functions f' and C' are monotonous, while the respective solution of $q_B(p, \tilde{r})$ and $q_G(p, \tilde{r})$ is unique.

Finally, the second order condition of (81) is:

$$\begin{bmatrix} \mathcal{L}_{BB} & \mathcal{L}_{BG} \\ \mathcal{L}_{GB} & \mathcal{L}_{GG} \end{bmatrix} = \begin{bmatrix} \alpha(1-z)f_B'' - n_B^2(1+\eta)C'' & -n_B n_G(1+\eta)C'' \\ -n_B n_G(1+\eta)C'' & \alpha n_G(1-z)f_G'' - n_G^2(1+\eta)C'' \end{bmatrix} \quad (82)$$

As $\alpha(1-z)f_B'' - n_B^2(1+\eta)C'' < 0$, and $[\alpha(1-z)f_B'' - n_B^2(1+\eta)C''][\alpha n_G(1-z)f_G'' - n_G^2(1+\eta)C''] - (n_B n_G(1+\eta)C'')^2 > 0$. The second order condition of the system (81) is negative, thereby verifying the claim.

A10

As in the last section, the Lagrange of the system (55) is:

$$\begin{aligned} \mathcal{L} = & \alpha \sum_{t \in T} n_t [z_t + (1-z_t)f(q_t)] + \sum_{t \in T} \mu_t q_t \\ & + (1+\eta) \left[p \sum_{t \in T} n_t q_t + k \sum_{t \in T} n_t - C \left(\sum_{t \in T} n_t q_t \right) \right] \end{aligned} \quad (83)$$

Following the same procedures as in the last subsections, the necessary condition that (83) has a unique solution is:

$$\left\{ \begin{array}{l} \alpha n_B f'(q_B) + (1+\eta)n_B[p - C'(Q)] + \mu_B = 0 \\ \alpha(1-z)n_G f'(q_G) + (1+\eta)n_G[p - C'(Q)] + \mu_G = 0 \\ \eta[pQ + nk - C(Q)] = 0 \\ \mu_B q_B = 0 \\ \mu_G q_G = 0 \\ \eta, \mu_B, \mu_G \end{array} \right. \quad (84)$$

The first two lines of the equation are the first order conditions with respect to q_B and q_G and the third to fifth lines are the corresponding complementary slackness conditions associated with the optimization inequalities. The last line has the restriction on the respective multipliers of non-negative requirements. Given f' and C' are continuous, the respective solution $q_B(p, k)$ and $q_G(p, k)$ are unique.

The second order conditions of (84) can be written as:

$$\begin{bmatrix} \mathcal{L}_{BB} & \mathcal{L}_{BG} \\ \mathcal{L}_{GB} & \mathcal{L}_{GG} \end{bmatrix} = \begin{bmatrix} \alpha(1-z)f_B'' - n_B^2(1+\eta)C'' & -n_B n_G(1+\eta)C'' \\ -n_B n_G(1+\eta)C'' & \alpha n_G(1-z)f_G'' - n_G^2(1+\eta)C'' \end{bmatrix} \quad (85)$$

Given $\alpha(1-z)f_B'' - n_B^2(1+\eta)C'' < 0$, and $[\alpha(1-z)f_B'' - n_B^2(1+\eta)C''][\alpha n_G(1-z)f_G'' - n_G^2(1+\eta)C''] - (n_B n_G(1+\eta)C'')^2 > 0$. The second order condition of the system (84) is negative, thereby verifying the claim.

Chapter 4. The optimal design of fee-for-service contract 2

4.1 Introduction

The previous chapter analysed the design of the optimal compensation scheme for altruistic providers in an environment where physicians decide the quantity of medical service (e.g. the number of prescriptions or the length of hospital stay) and have private information about their degree of altruism as well as their patients' initial health. The main finding is that the optimal remuneration blends both positive payment per service and a negative reimbursement per patient. The negative payment has been applied widely in practice, well-known examples include franchise agreements (Rubin, 1978) and bonding contracts (Awad and Fayek, 2012).

The aforementioned result is obtained in a parsimonious setup in which the physician faces no fixed costs in terms of providing medical service and is allocated a given number of patients. Accordingly, the purpose of this chapter is to generalize the previous analysis by removing the above two restrictions. Specifically, I follow Wu et al. (2018) as well as Barham and Milliken (2015) who allow the physician to choose the size of his practice.⁵³ Moreover, the physician faces fixed costs associated with enrolling patients. The generalizations, however, make it essential to design a health insurance scheme that ensures the entire population's access to healthcare. In short, this chapter improves our understanding regarding how price variations and different cost control instruments affect physicians' choices regarding their practice size and service quality. Moreover, it provides qualitative recommendations in relation to optimal incentive contracting for primary healthcare providers.

The chapter follows the same procedures adopted in Chapter 3 to solve the optimal health insurance scheme. First, I characterize the physician's decision on medical service quantity and patient numbers as a function of price and capitation.⁵⁴ The main departure from the previous analysis is that the price rises may lead to a decrease in the level of medical service intensities/qualities provided. This is because, compared to improving

⁵³ According to Barham and Milliken (2015), allowing physicians to select their practice sizes is particularly relevant in countries facing medical staff shortages.

⁵⁴ I denote capitation as fixed payment per patient, which has also been defined in the Chapter 1.

quality, enrolling an additional patient provides the physician with extra marginal altruism benefits. As a result, when constrained by a zero-profit condition, the physician increases his practice size to maximize utility while reducing service intensities to maintain zero profit.

In the second part of the chapter, I derive the constrained optimal fee-for-service system that ensures the entire population's access to healthcare. As in the previous chapter, the price system fails to align more altruistic doctors' incentive to overproduce⁵⁵ and less altruistic doctors' incentive to underproduce as a result of moral hazard and adverse selection. However, in the current extension, the regulator faces additional trade-offs between productive efficiency and full patients' access.

In the final part of this chapter, I examine whether three cost-control mechanisms⁵⁶ introduced in the previous chapter – i.e. per patient expenditure control, per patient quantity rationing, and a capitation less than fixed costs – can be used in the current context to enhance resource allocation efficiency and ensure all patients have access to healthcare. I find that all three methods can be used to achieve these objectives. This is because cost control policies directly limit the total quantity of medical service produced while combining price with cost-control instruments aligns physicians' heterogeneous incentives. Specifically, the cost-control instruments force low altruistic providers to reduce service quality per patient but increase the number of patients treated. Meanwhile, by adjusting price downward, high altruistic providers are incentivised to reduce their practice size and to improve medical service quality.

Overall, I find that the revenue cap and capitation perform better than quantity rationing, as the former two instruments exploit the private knowledge of doctors. While imposing a fee-for-service system a capitation or a revenue cap improves social welfare at the same level, the former payment is considered as the best remuneration mechanism as it is more tolerant of potential mistakes that could be made by the regulator compared to the revenue cap. The result that capitation is superior to other cost control instruments

⁵⁵ As discussed in the Chapter 3, the more/less altruistic physician would provide a higher/lower total caseload than they would do in an economy with the more/less altruistic physicians only.

⁵⁶ I refer to these cost-control instruments as policy tools used by health authority to contain health expenditure growth.

is therefore consistent with the one obtained in Chapter 3, reflecting that my previous findings can be further generalized in a less restrictive context.

The rest of the chapter is organized as follows. Section 4.2 briefly discusses the related literature while Section 4.3 introduces the model. Section 4.4 describes the physician's utility optimization problem, derives his decisions in terms of the number of patients treated as well as medical service quality, and provides a comparative static analysis of these decisions. Section 4.5 describes the regulator's optimization problem and introduces parameters for the succeeding numerical analysis. Section 4.6 evaluates different cost control instruments and derives the optimal form of HI scheme. Finally, Section 4.7 concludes, section 4.8 discusses potential extensions, while Appendix 4.9 provides all proofs.

4.2 Related literature

Broadly speaking, this study relates to three types of literature. First, it relates to the literature that compares physician responses to the incentives embedded in traditional payment systems (Ellis and McGuire, 1986; Léger, 2008). This literature often uses a principal-agent model whereby the physician is viewed as a representative making treatment decisions for a single patient (e.g. Choné and Ma, 2011; Ellis and McGuire, 1986; Jack, 2005), or the physician has a utility function per patient (Allard et al., 2014; Eggleston, 2000; Ellis, 1998). Many studies use this approach for investigating the ways in which differences in information structure, physician preferences or patient characteristics affect the quality or the intensity of treatment provided. In contrast, this chapter allows the physician to treat multiple types of patients and addresses the extent to which the physician reacts to a different remuneration system by adjusting his roster size compared to the quality/quantity of care provided.

Studies in this literature examining incentives embedded in different payment systems to accept low-cost patients (cream-skimming) while limiting access to excessively expensive patients (dumping) are also related (Barros, 2003; Ellis, 1998; Ma, 1994; Ma and McGuire, 1997). Their main focus, however, is to characterize the consequences of physician selection of patients and to derive the payment that induces an efficient level of quality or cost efforts (Ma, 1994,1998). Different to all of these studies, this chapter analyses the design of physician payment, not only to ensure the efficient provision of

quality but also to ensure the entire population's access to healthcare (Barham and Milliken, 2015).

Recent studies in this literature also discuss the design of physician payment for heterogeneous patients and providers (Jack, 2005; Kantarevic and Kralj, 2016; Wu et al., 2018). In contrast to all of these studies proposing a menu of physician compensation systems, my study derives an optimal single payment which blends both FFS and capitation. This remuneration scheme can be interpreted as an indirect implementation of the direct revelation mechanism via a single and linear schedule. The single payment system has been widely applied in the healthcare and other sectors, as it does not require reporting information (i.e. physicians' types) and therefore is not subject to the problem of renegotiation (Demougin, 1989; Maskin and Moore, 1999). Moreover, the linear payment schedule does not embed incentives that induce physicians to pick up low-cost patients.

Second, this chapter relates to the literature on physicians' competition for patients (Allard et al., 2011; Ma and McGuire, 1997; McGuire, 2000), in which the physician's practice size is determined endogenously. This literature considers an environment where patients are relatively scarce and argues that competition may provide physicians with an incentive to improve their quality of care or to keep their patients who have an option to switch to an alternative provider. The main departure from this literature is that I do not model physicians' competition. Accordingly, my model is more suited to countries facing primary care physician (PCP) shortages while theirs are more appealing to countries with a surfeit of PCPs.

Finally, this chapter relates to the literature which compares the impact of imposing different cost-control instruments on the supply of health service by self-interested primary care physicians (PCPs) who are paid on a FFS basis (Benstetter and Wambach, 2006; Fan et al., 1998; Mougeot and Naegelen, 2009; Neudeck, 1991). As discussed in Chapter 2, this study contributes by introducing important features of the healthcare market in the stylised model, including moral hazard and adverse selection problems as well as the altruistic preferences.

4.3 Setup

I use the same model as in the previous chapter except the following extensions. First, the physician j can select both the number of patients n^j and the quantity vector \mathbf{q}^j for patients with alternative health status. Since there are only two types of patients and their distributions are given exogenous,⁵⁷ the respective number of B - and G -type patients is denoted by $n_B^j = \beta n^j$ and $n_G^j = (1 - \beta)n^j$. The vector of the number of patients treated can be defined as $\mathbf{n}^j = (n_B^j, n_G^j)$. Denoting the medical service provided by doctor j to a type B - and type- G patient by q_B^j and q_G^j , the quantity of service provided is therefore denoted by a vector $\mathbf{q}^j = (q_B^j, q_G^j)$. The total caseload (total quantity of service provided to n^j patients) is again denoted by $Q^j = n_B^j q_B^j + n_G^j q_G^j$.

Second, I use a more general form of the total cost function $\tilde{C}(n^j, Q^j, F)$. For simplicity, I assume the cost function is additively separable. This cost function form takes a specific form:⁵⁸

$$\tilde{C}(n, Q, F) = nF + C(Q) \quad (86)$$

where F is exogenously given and the first term nF can be interpreted as the total fixed costs whereas the second term $C(Q)$ represents the treatment costs. This cost function $\tilde{C}(n, Q, F)$ therefore captures the notion that the physician cares about both the number of patients treated and his total caseload. Moreover, the treatment cost function $C(Q)$ is increasing and convex, which satisfies $C''' > 0$ and $C(0) = C'(0) = 0$.

Third, the physician is not only reimbursed by fee-for-service but also a capitation denoted by k . As a result, the doctor j who produced $\mathbf{q} = (q_B, q_G)$ obtains the profit:

$$\tilde{\Pi}(n, \mathbf{q}, p, k, F) = pQ(\mathbf{q}) + nk - \tilde{C}(n, Q, F) \quad (87)$$

The welfare again sums up patients' expected utility and doctors' expected profit over the entire economy under a self-balancing restriction. With the introduction of the capitation rate k , this restriction can be written as:

⁵⁷ From Chapter 3, $z_i \in \{0, z\}$ with $0 < z < 1$. Define $z_B = 0, z_G = z$, we have $Pr[z_i = z_B] = \beta$ and $Pr[z_i = z_G] = 1 - \beta$.

⁵⁸ To simplify the cumbersome notation, I drop the physician's identification superscription j in the following profit function.

$$N\tau = (1 + \lambda) \left(\sum_{j=1}^M pQ^j + k \sum_{j=1}^M n^j \right) \quad (88)$$

where the left side of the equation represents total premium collected from patients and the right side the total payment to physicians. Finally, by assuming $-\tau$ the patient's utility if he failed to receive treatment, the simplified social welfare in the current setup is:

$$W(\mathbf{u}, \tilde{\Pi}) = \sum_{i=1}^N h_i - \left[\lambda \sum_{j=1}^M (pQ^j + kn^j) + \sum_{j=1}^M C(Q^j) + F \sum_{j=1}^M n^j \right] \quad (89)$$

where the former term measures total expected health benefits across patients and the latter the net health expenditure of the medical system. In contrast to the previous Chapter 3, the regulator not only has to find a HI scheme that maximizes (89) but also need to ensure that the entire population have access to healthcare (i.e. $\sum_{j=1}^M n^j = N$), given the doctor's choices with respect to the number of enrolled patients and the provision of medical services.

The foregoing setup closely parallels the one used by Barham and Milliken (2015), which is the first study deploying a population approach⁵⁹ to analyse physician incentive contracting in the presence of patient and physician heterogeneities. My model differs from theirs in four ways: First, I do not assume a bell-shaped altruism function. Instead, a concave patient health function and a fixed cost associated with enrolling an individual patient is introduced; second, my model imposes a minimum profit condition on the physician; third, I do not assume the exogenously given target level of medical service quantity provided per patient, this target level now depends on the physician remuneration scheme and other exogenous factors. Finally, I analyse the impact of introducing different cost control instruments on physician behaviour and the design of optimal physician payment when cost policies are considered. Overall, these extensions allow us to derive rather than assume the optimal quantity of service provided per patient while analysing how the physician trades-off his patients' enrolment and health service quality improvement under a binding profit as well as under different cost control policies.

⁵⁹ Physicians determine the size of their practice and the quality of care provided as well as ensuring that all patients have access to healthcare (Barham and Milliken, 2015, p. 896).

4.4 The Physician's Optimization

Following the same procedures as in the last chapter, I solve the utility maximization problem of a physician characterized by the generic parameter α given the HI policy (p, τ, k) . Denote the set of types by $t \in T = \{B, G\}$, β_t the given proportion of the respective type of patients (i.e. $\beta_B = \beta, \beta_G = 1 - \beta$), α the level of the physician's altruism, q_t the quantity of medical service provided per patient, and n the total number of patients treated, the doctor's optimization problem is:

$$\begin{aligned} U = \max_{n, q_t} & \sum_{t \in T} n\beta_t [\alpha(z_t + (1 - z_t)f(q_t)) + pq_t + k - F] - C \left(n \sum_{t \in T} \beta_t q_t \right) \\ & \sum_{t \in T} n\beta_t (pq_t + k - F) - C \left(n \sum_{t \in T} \beta_t q_t \right) \geq 0 \\ & n, q_t \geq 0 \end{aligned} \tag{90}$$

the Lagrangian associated with the doctor's optimization problem (90) is:

$$\begin{aligned} \mathcal{L} = & \alpha \sum_{t \in T} n\beta_t [z_t + (1 - z_t)f(q_t)] + \sum_{t \in T} \mu_t q_t + \psi n \\ & + (1 + \eta) \left[\sum_{t \in T} n\beta_t (pq_t + k - F) - C \left(n \sum_{t \in T} \beta_t q_t \right) \right] \end{aligned} \tag{91}$$

where η is the Lagrange multiplier for the profit constraint and the μ_t, ψ the respective multipliers of the non-negative restrictions. Notice $z_B = 0, z_G = z$ and denote by $\bar{q} = \beta_B q_B + \beta_G q_G$ the average quantity of service provided per patient. The first order condition of (91) is:

$$\left\{ \begin{array}{l} \alpha[\beta f_B + (1-\beta)(z + (1-z)f_G)] + (1+\eta)[\bar{q}(p - C') + k - F] + \psi = 0 \\ \alpha n \beta f'_B + (1+\eta)n\beta(p - C') + \mu_B = 0 \\ \alpha n(1-\beta)(1-z)f'_G + (1+\eta)n(1-\beta)(p - C') + \mu_G = 0 \\ \eta[pQ - C(Q) + n(k - F)] = 0 \\ \psi n = 0 \\ \mu_B q_B = 0 \\ \mu_G q_G = 0 \\ \eta, \psi, \mu_B, \mu_G \geq 0 \end{array} \right. \quad (92)$$

The first three equations in (92) are the first-order conditions with respect to n, q_B and q_G . The next four lines in (92) are the corresponding complementary slackness conditions associated with the inequalities in the optimization problem. Finally, the last line gives the restriction on the respective multipliers of non-negative requirements for profit, services, and the number of patients. In order to simplify notations, I denote $f(q_G), f(q_B)$ by f_G, f_B and $f'(q_G), f'(q_B)$ by f'_G, f'_B . Moreover, $C(Q), C'(Q), C''(Q)$ are denoted as C, C', C'' respectively without causing confusion wherever possible. Finally, I denote the solution to the system (92) by the superscript “*”. In the following analysis, I assume the parameters $\alpha, \beta, F, z \in (0,1)$ are holding constant, except for comparative static analysis with respect to them.

Lemma 4.1 *For all $p > 0$, the solution of system (92) satisfies $q_B^*, q_G^* > 0$.*

Proof. The system (92) has eight potential solutions. First, when $\psi > 0$, there are four types of solutions: 1) $\mu_B, \mu_G > 0$, 2) $\mu_B > \mu_G = 0$, 3) $\mu_G > \mu_B = 0$ and 4) $\mu_B = \mu_G = 0$. Second, when $\psi = 0$ there are also four types of solution: 5) $\mu_B, \mu_G > 0$, 6) $\mu_B > \mu_G = 0$, 7) $\mu_G > \mu_B = 0$ and 8) $\mu_B = \mu_G = 0$. Given the assumptions $f'(0) = +\infty$, $\alpha, \beta, p, k, F > 0$ and $f(\cdot)$ is increasing, concave and bounded, the solutions 1), 2), 3) and 5), 6), 7) do not satisfy either the second or the third equation of the above system, thus verifying the claim. ■

This result is consistent with the Lemma 3.1, which shows that the intensity of service provided per patient is always positive. The main difference, however, is that the physician in the current setup needs to select the number of patients and may choose not to enrol any patients. Since the physician always provide the positive service quantity. The system (92) can be simplified as:

$$\left\{ \begin{array}{l} \alpha[\beta f_B + (1 - \beta)(z + (1 - z)f_G)] + (1 + \eta)[\bar{q}(p - C') + k - F] + \psi = 0 \\ \alpha f'_B + (1 + \eta)(p - C') = 0 \\ \alpha(1 - z)f'_G + (1 + \eta)(p - C') = 0 \\ \eta[pQ - C(Q) + n(k - F)] = 0 \\ \psi n = 0 \\ \eta, \psi \geq 0 \end{array} \right. \quad (93)$$

The following lemma characterizes the necessary condition that the physician will enrol patients.

Lemma 4.2 *For all $p > 0, -\alpha \leq k - F < 0$, the system (93) has an interior solution $n^* > 0$ (i.e. $\psi^* = 0$).*

Proof. Substituting the second and the third equations into the first equation in the system (93) yields:

$$\alpha\{\beta(f_B - q_B f'_B) + (1 - \beta)[z + (1 - z)(f_G - q_G f'_G)]\} + (1 + \eta)(k - F) + \psi = 0 \quad (94)$$

$q_B, q_G > 0, \eta \geq 0, \psi \geq 0$, the concave and bounded f implies that $0 < f_B - q_B f'_B \leq 1$ and $0 < z + (1 - z)(f_G - q_G f'_G) \leq 1$. The equality of (94) requires $k - F < 0$. Next, as the first term of the above equation $\alpha\{\beta(f_B - q_B f'_B) + (1 - \beta)[z + (1 - z)(f_G - q_G f'_G)]\} \leq \alpha$ and $\eta \geq 0$, suppose on the contrary $k - F < -\alpha$, then the equality of (94) requires $\psi^* > 0$, a contradiction. The necessary condition of $\psi^* = 0$ is $k - F \geq -\alpha$. ■

To provide intuition, notice that the concavity of patient health function implies that the physician obtains higher marginal altruism benefits from enrolling an additional patient rather than improving service quality for existing patients. Therefore, the physician has

an incentive to enrol too many patients while severely under-treating each individual. To counter this incentive, the marginal financial return per patient has to be set negative.⁶⁰ However, this negative payment cannot be set lower than the maximum altruism benefits each patient generates, as otherwise the physician would have no incentive to enrol any patients (marginal financial return per patient becomes negative).

Since $\alpha \in (0,1)$, the restriction $-\alpha \leq k - F < 0$ implies that the regulator has very limited choices of capitation rate k if the fixed cost F is sufficiently large. This problem comes from the fact that the patient health function is scaled in the range zero to one, while there is no restriction on the value of F that can be taken. This issue can be solved by providing the health function and the fixed cost F at the same scale. In this paper, I solve this issue by selecting a F within the range between zero and one.

In the following analysis, I assume the price and capitation is set at the respective range $p > 0$ and $-\alpha \leq k - F < 0$. While there is an interior solution ($n^* > 0$ and $q^* > 0$), the boundary solution ($n^* = 0$ and $q^* = +\infty$) may also exist. Since the latter solution does not make sense in reality, I only consider the interior solutions.

Lemma 4.3 *For all $p > 0$ and $-\alpha \leq k - F < 0$, we have $q_B^* > q_G^* > 0$*

Proof. From the second and third equations in (93), we obtain $f'_B = (1-z)f'_G$. Since $f'' < 0$, $0 < z < 1$ and $q_B^*, q_G^* > 0$, we have $q_B^* > q_G^* > 0$. ■

Intuitively, the physician receives the same financial returns from serving B - or G -type patients. However, the extra marginal altruistic benefits obtained from treating the B -type patient would induce the doctor to provide a higher q_B . While this result is similar to what has been found in Barham and Milliken (2015), this Chapter follows a completely different process. In this Chapter, the physician is allocated a fixed fraction of a respective type of patients and select quantities to maximize his utility. In Barham and Milliken (2015), however, the altruistic physician would only treat frail individuals while the non-altruistic physician would treat healthy patients. Treated by the altruistic

⁶⁰ The same problem also encounters in Barham and Milliken (2015). As they do not assume fixed costs, the physician has an incentive enrol too many patients while severely undertreat each individual when altruism function is concave or enrol too few patients while severely overtreat each individual when altruism function is convex. As a result, altruism function has to be bell-shaped in their paper to ensure that the physician obtains an equilibrium between enrolling patient and improve service quality.

physician who receives a relatively higher marginal altruism benefit from providing medical service, the frail patient in Barham and Milliken (2015) is therefore provided with a higher amount of service.

The main purpose of this section is to analyse how the physician trades off the choice of enrolling additional patients and raising medical service intensity (quality). In order to simplify the explanations, I will discuss the intuition of the simplest case where all patients have the same health status $z_B = z_G = 0$ and therefore receive $q_B = q_G = q$. However, for the derivation of the ensuing results, I provide the general proof in the appendix. The system (93) can be therefore rewritten as:

$$\left\{ \begin{array}{l} \alpha f(q) + (1 + \eta)[q(p - C'(Q)) + k - F] = 0 \\ \alpha f'(q) + (1 + \eta)(p - C'(Q)) = 0 \\ \eta[pQ - C(Q) + n(k - F)] = 0 \\ \eta \geq 0 \end{array} \right. \quad (95)$$

The profit of the physician can be either positive or zero, therefore we have alternative solutions. First, we consider the situation where the physician's profit is strictly positive. Given α, β, F and capitation payment k , denote the set $\tilde{\mathbf{P}} = \{p > 0 \mid \tilde{\Pi}^*(p) = pQ^* + n^*(k - F) - C(Q^*) > 0\}$ and consider the price $p \in \tilde{\mathbf{P}}$.⁶¹ By definition, $\eta^* = 0$ and the system (95) can be further simplified to:

$$\left\{ \begin{array}{l} \alpha f(q) + q(p - C'(Q)) + k - F = 0 \\ \alpha f'(q) + p - C'(Q) = 0 \end{array} \right. \quad (96)$$

Notice that all the functions are continuously differentiable, while an infinitesimal variation around p will therefore keep the profit positive. I use this observation to derive the following results:

Lemma 4.4 *For all $p \in \tilde{\mathbf{P}}$ and $-\alpha \leq k - F < 0$, $\frac{\partial n^*}{\partial p} = \frac{\partial Q^*}{\partial p} > 0$ and $\frac{\partial q_B^*}{\partial p} = \frac{\partial q_G^*}{\partial p} = 0$.*

Proof. See Appendix B1. ■

⁶¹ To avoid the cumbersome notation, I denote the $\tilde{\mathbf{P}}$ as $\tilde{\mathbf{P}}(\alpha, \beta, k, F, z)$ hereafter.

As in Barham and Milliken (2015), I show that a rise in service price does not lead to a higher level of medical service provided per patient. To provide the intuition, notice that the price increase raises the physician's marginal benefit derived from providing medical service; as a result, the physician increases his total caseload Q . Since raising one unit of quality or enrolling one more patient yields the same payoff $p - C'$, the physician's decision regarding patient number and quality depends on the level of marginal altruistic benefit from enrolling an additional patient net of administration cost and the level of marginal glow derived from raising quality. In other words, the equilibrium decision of q and n can be reached if and only if:

$$\alpha[f(q) - qf'(q)] + k - F = 0 \quad (97)$$

Where the term $\alpha f(q)$ represents the marginal altruism benefit of enrolling an additional patient, $\alpha qf'(q)$ the marginal altruism benefit of raising one unit of health service intensity, and $k - F$ the net marginal financial return from enrolling one more patient. From equation (97), the physician's decision on quality does not depend on price. Since the physician increases his total caseload while keeping service quality constant as the price goes up, the number of patients treated must increase⁶² according to the definition of total caseload.

Lemma 4.5 For all $p \in \tilde{P}$ and $-\alpha \leq k - F < 0$, we have $\frac{\partial n^*}{\partial k}, \frac{\partial Q^*}{\partial k} > 0 > \frac{\partial q_G^*}{\partial k} \geq \frac{\partial q_B^*}{\partial k}$.

Proof. See Appendix B2. ■

Lemma 4.5 shows that the physician increases his roster size but reduces service quality for existing patients when the capitation rate goes up. Intuitively, as capitation rate increases, the marginal benefit of enrol additional patient becomes higher than the marginal cost. As a result, the physician will increase his roster size. The increase in roster size *ceteris paribus* raises the marginal cost of providing quality, thus requiring the physician to reduce quality. Altogether, while the physician reduces quality, he increases his size of total caseload. Finally, as the increase in the marginal warm glow

⁶² Notice that the number of patients treated n^* may not take integer values. As explained in Barham and Milliken (2015), patients could be 'part-time patients of more than one doctor.'

benefit increase as a result of reducing q_B^* is lower than that of q_G^* , the reduction of q_B^* is therefore higher than q_G^* .

This result is also consistent with Barham and Milliken (2015) but further generalizes their analysis to the case in which the physician treats more than one type of patient. Overall, Lemma 4.4 and 4.5 show that their results do not require a specific physician's preference (i.e. bell-shaped altruism function). Furthermore, since Barham and Milliken (2015) assume altruistic/non-altruistic physicians, frail/healthy patients, and altruistic (non-altruistic) physicians only treat frail (healthy) patients. As a result, they do not investigate the effect of varying a physician's altruism level or the proportion of a respective type of patient on the physician's choices of practice size and quality. The following Lemmas 4.6 and 4.7 therefore step forward and analyse how these changes affect the physician's decision on his practice size as well as quality of service provided.

Lemma 4.6 *For all $p \in \tilde{\mathbf{P}}$ and $-\alpha \leq k - F < 0$, we have $\frac{\partial Q^*}{\partial \alpha}, \frac{\partial n^*}{\partial \alpha} > 0 > \frac{\partial q_G^*}{\partial \alpha} \geq \frac{\partial q_B^*}{\partial \alpha}$.*

Proof. See Appendix B3. ■

In contrast to Chapter 3, Lemma 4.6 shows that the more altruistic doctors would provide less service intensity to each individual. Intuitively, as the physician can select patient number and quality, the equilibrium is only obtained when the marginal net benefit of enrolling an additional patient is equivalent to the loss of warm glow benefits by raising service quality. A *ceteris paribus* increase in the relative degree of altruism increases the marginal warm glow benefit derived from enrolling an additional patient. Since the per patient administration cost does not change, the physician has to reduce medical service quality in order to re-establish the balance.

As the increase in the level of altruism raises the physician's marginal return for service production, the physician would increase his total caseload. By definition, the increase of total caseload and the reduction in service quality implies that the number of patients treated increases with the physician's altruism. Similar to Lemma 4.5, the reduction of q_B raises less marginal warm glow benefits than that of reducing q_G , thereby the physician reduces more q_B than q_G as his level of altruism increases. Finally, as $-\alpha \leq$

$k - F < 0$, $p - C' < 0$ ⁶³ and the physician increases the number of patients treated and total caseload with the degree of his altruism, the more altruistic physician therefore has a lower profit.

Lemma 4.7 For all $p \in \tilde{P}$ and $-\alpha \leq k - F < 0$, $\frac{\partial n^*}{\partial \beta}, \frac{\partial Q^*}{\partial \beta} < 0$, $\frac{\partial q_B^*}{\partial \beta} \geq \frac{\partial q_G^*}{\partial \beta} > 0$.

Proof. See Appendix B4. ■

This Lemma shows that the physician reduces the number of patients and service total caseload as the proportion of less healthy individuals increases. In contrast to Chapter 3, the individual patient is provided a higher intensity of medical service as the proportion of less healthy individuals increases. This result derives directly from the extension that the physician can choose both service intensities and the number of patients to be treated. Intuitively, a ceteris paribus increase in the proportion of less healthy individuals leads to the same reduction in the proportion of healthier individuals, resulting in a reduction in average health per patient treated. Moreover, since the less healthy individual requires a higher quantity of medical service to be provided, total caseload and the marginal cost of production goes up ceteris paribus. As a result, the physician will reduce his practice size. The increase in the proportion of less healthy individuals also reduces net altruism benefits between enrolling patients and raising quality (see equation (94)). To counter this effect, the physician raises quality. Overall, the physician reduces his total caseload. As the payment per service and per patient does not fully cover their costs. Hence, the reduction in the number of patients and total caseload leads to an increase in the physician's profits.

Substituting $n^*(p, k, \alpha)$, $q^*(p, k, \alpha)$ ⁶⁴ into the total caseload function yields $Q^*(p, k, \alpha) = n^*(p, k, \alpha)q^*(p, k, \alpha)$. The doctor's profit function is therefore:

$$\tilde{\Pi}^*(p, k, \alpha) = pQ^*(p, k, \alpha) + (k - F)n^*(p, k, \alpha) - C(Q^*(p, k, \alpha)) \quad (98)$$

⁶³ See the second equation of the system (95).

⁶⁴I assume the parameters β, F, z are given constant and only analyse the effect of varying price, capitation and the degree of physician altruism.

Taking the first-order condition of (98) with respect to p and simplify, we have:

$$\frac{\partial \tilde{\Pi}^*(p, k, \alpha)}{\partial p} = Q^* + [q^*(p - C') + k - F] \frac{\partial n^*}{\partial p} + n^*(p - C') \frac{\partial q^*}{\partial p} \quad (99)$$

Lemma 4.8 Denote $p_0(k, \alpha)$ the smallest p such that $\eta^*(p_0, k, \alpha) = 0$, then $\tilde{P} = (p_0(k, \alpha), +\infty)$.

Proof. To verify the claim, we need to show $\tilde{\Pi}^*(p, k, \alpha) > 0$ for all $p > p_0(k, \alpha)$ at a given capitation rate net of fixed cost $k - F$. First, when $p \rightarrow 0$ the non-negative profit constraint is binding, i.e. $\eta^*(p_0, k, \alpha) > 0$. From the system (95) and the equation (98), as the price converges to zero, therefore the revenue of the physician also converges to zero but his marginal benefit remains large⁶⁵ given the assumption of the altruistic component. In contrast, as $p \rightarrow +\infty$ the physician's profit for a given $-\alpha \leq k - F < 0$ is not binding, i.e. $\eta^* = 0$. By construction, there is a price $p_0(k, \alpha)$ such that the physician's profit is just binding, i.e. $\Pi^*(p_0(k, \alpha), k, \alpha) = 0$. Next, I prove $\frac{\partial \tilde{\Pi}^*(p_0^+, k, \alpha)}{\partial p} > 0$ given $C''' > 0$. From the profit function (98), the definition of Q^* and $\frac{\partial Q^*}{\partial p}$ obtained from Appendix B1 we have:

$$\frac{\partial \tilde{\Pi}^*(p, k, \alpha)}{\partial p} = Q^* + \frac{q^*(p - C') + k - F}{q^* C''} \quad (100)$$

We want to show the equation (100) is positive, which is the same as verifying:

$$Q^* q^* C'' + q^*(p - C') + k - F > 0 \quad (101)$$

Multiply both sides by n^* , notice that $p_0 Q^* = C(Q^*) - n^*(k - F)$ and the price level is p_0 . The objective becomes to prove:

$$(Q^*)^2 C'' - Q^* C' + C > 0 \quad (102)$$

$C''' > 0$ implies $Q^* C'' - C' > 0$ and therefore $(Q^*)^2 C'' - Q^* C' > 0$. Given $C \geq 0$, I verified that (102) is positive. To conclude the proof, suppose that in contradiction to the claim that $p > p_0$ with $\tilde{\Pi}^*(p_0, k, \alpha) = 0$, then by definition of continuity there exists a price $\hat{p} > p_0$ with $\tilde{\Pi}^*(\hat{p}, k, \alpha) = 0$ and $\frac{\partial \tilde{\Pi}^*(\hat{p}, k, \alpha)}{\partial p} < 0$. However, this is not possible

⁶⁵ This condition will be hold as long as we have an interior solution $n^* > 0$.

as following the aforementioned logic would show $\frac{\partial \tilde{\Pi}^*(\hat{p}^-, k, \alpha)}{\partial p} > 0$, which is a contradiction and thus verifying the claim. ■

The result is that the physician's profit is increasing in price when he has a positive profit, which is consistent with the finding in Chapter 3. The intuition is that the positive (direct) impact of the price increase dominates the subsequent and negative impacts (indirect) of the patient number as well as the total caseload increase. Since this work imposes non-negative profit conditions on physicians (Barham and Milliken, 2015), we can move forward to characterize the physician's decision on the number of patients treated and service quality as a function of the price or capitation when his profit is binding.

When the physician's profit is strictly binding (i.e. $p \notin \tilde{\mathbf{P}}$), the profit of the physician satisfies $\tilde{\Pi}^* = 0$. Accordingly, the number of patients selected and the quantity of medical services provided which maximize the doctor's utility are implicitly defined by:

$$\begin{cases} (p - C')(f - qf') - (k - F)f' = 0 \\ pnq - C(nq) + n(k - F) = 0 \end{cases} \quad (103)$$

where the first equation of (103) is derived from dividing the first by the second equation in the system (95).⁶⁶ Intuitively, this equation shows the marginal rate of substitution between quality and the number of patients should be the same in the iso-profit and iso-utility curves. The second equation of (103) represents the binding profit condition. We can now use this system to complete the derivation of the interior solution of $n^*(p, k, \alpha)$ and $q^*(p, k, \alpha)$. First, it is obvious that the equilibrium between the number of patients and medical service does not change with physicians' degrees of altruism. Second, for all $p < p_0$, I apply the implicit function theorem to system (103) and derive the following result.

Lemma 4.9 *For all $p \notin \tilde{\mathbf{P}}$ and $-\alpha < k - F < 0$, the selected number of patients $n^*(p, k, \alpha)$ and the service supply functions $q_t^*(p, k, \alpha)$ for $t \in T = \{B, G\}$ satisfy*

⁶⁶ The purpose of dividing these two equations is to eliminate the Lagrange multiplier η .

$\frac{\partial Q^*}{\partial p}, \frac{\partial n^*}{\partial p} > 0 > \frac{\partial q_G^*}{\partial p} \geq \frac{\partial q_B^*}{\partial p}$. At the critical price p_0 , we have $\frac{\partial n^*}{\partial p}(p_0^-, k, \alpha) > \frac{\partial n^*}{\partial p}(p_0^+, k, \alpha)$ and $\frac{\partial Q^*}{\partial p}(p_0^-, k, \alpha) > \frac{\partial Q^*}{\partial p}(p_0^+, k, \alpha)$.

Proof. See Appendix B5. ■

In contrast to Chapter 3, I find that the physician reduces service quality when price goes up, though he does raise his practice size and total caseload. As illustrated in the Lemma 4.5, a *ceteris paribus* increase in price raises the marginal return of providing total caseload. Accordingly, total caseload goes up. When the physician has a binding profit, he has an incentive to substitute quality for practice size. In this case, the physician's optimization problem is to maximise the altruistic benefits $nf(q)$ while holding his profit $pQ - C(Q) + n(k - F)$ constant at zero. The equilibrium level of (n^*, q^*) is therefore obtained when (1) the marginal rates of substitution between patient number and quality are the same in both the iso-profit line and the in-difference curve as well as (2) the physician's profit is zero (see the first and the second equation of system (94) respectively). Since the marginal altruism benefit of enrolling an additional patient is higher than raising one unit of quality, the physician increases his practice size in order to maximize utility. However, constrained by the non-negative profit condition, the physician has to reduce quality. This Lemma also shows that the decrease in q_G is lower than q_B , as the increase in the marginal warm glow benefit as a result of reducing q_B is lower than that derived from reducing q_G .

Lemma 4.10 For $p \notin \widetilde{P}$ and $-\alpha < k - F < 0$ the selected number of patients $n^*(p, k)$ and the service supply functions $q_t^*(p, k)$ for $t \in T = \{B, G\}$ satisfies $\frac{\partial n^*}{\partial k}, \frac{\partial Q^*}{\partial k} > 0 > \frac{\partial q_G^*}{\partial k} \geq \frac{\partial q_B^*}{\partial k}$.

Proof. See Appendix B6. ■

This Lemma shows that the physician increases his roster sizes but reduces service qualities as the capitation rate goes up when his profit is binding. This result is again different to what has been shown in Chapter 3 that service intensities provided do not decrease with capitation rate. Intuitively, when the capitation rate increases, the physician has more incentive to increase practice size. Similar as the case when price increases, the physician increases the number of patients treated to maximize utility

while reducing quality to maintain a zero profit. As shown in Lemma 4.6, raising capitation increases the physician's marginal return of production. Hence, he raises his practice size and total caseload.

Lemma 4.11 *For a given $-\alpha < k - F < 0$, the critical price p_0 is increasing in α .*

Proof. See Appendix B7. ■

Lemma 4.11 is consistent with Lemma 3.8 in Chapter 3 with a similar intuition. Consider a physician characterized by a parameter α at a given capitation rate net of fixed cost k . From the lemma 4.7 we know that there exists a price $p_0(\alpha, k)$ such that the doctor's profit is just binding. Suppose there is a physician with a slightly larger α (i.e. a marginally more altruistic individual). Holding price constant that doctor would want to treat more patient and provide a higher total caseload (see Lemma 4.6). However, since we have $p_0(\alpha, k) < C'$ and $k - F < 0$ so that the more altruistic physician's profit would become negative. As a result, to prevent the α -type doctor from having a negative profit, $p_0(\alpha, k)$ has to be raised.

Keep in mind that a physician's concern for patients only takes two values $\alpha^j \in \{\alpha^L, \alpha^H\}$ with $0 < \alpha^L < \alpha^H \leq 1$, $Pr[\alpha^j = \alpha^L] = \Lambda$, $Pr[\alpha^j = \alpha^H] = 1 - \Lambda$ and that there are two kinds of patients $z_i \in \{z_B, z_G\}$ with $0 = z_B < z = z_G < 1$, $Pr[z_i = z_B] = \beta$, $Pr[z_i = z_G] = 1 - \beta$. Given that $\alpha^L, \alpha^H, \beta, F, \Lambda, z$ are constant, I denote a low and a high altruistic doctor's response functions by $n^L(p, k), q_t^L(p, k)$, $n^H(p, k), q_t^H(p, k)$ while the respective critical price is $p_0^L(k)$ and $p_0^H(k)$.⁶⁷

4.5 Welfare Analysis

4.5.1 Welfare Maximization

Since physicians can now select the number of patients, the regulator in this chapter not only needs to derive a health insurance scheme that maximizes welfare but also to ensure the entire population have access to healthcare. Taking the self-financing constraint of the health system into consideration, I follow the same procedures as in Chapter 3

⁶⁷ To simplify the expression of welfare function, I drop all superscript “*” in the solutions from physicians' optimization problem for the ensuing analysis.

applying the results from the foregoing section to rewrite welfare function (89) as a function of price and capitation. Next, by assuming the interior solution of welfare maximizing price and capitation, I use the equation (88) to calculate the optimal insurance premium.

Total health benefits function adds up across all treated patients' expected health, which depends on the doctor's type and behaviour. Assuming the expected health of untreated patients is zero, the total health function as a given policy (p, k) can be written as:

$$\begin{aligned} \sum_{i=1}^N h_i = & \Lambda M \left\{ n^L(p, k) \left[\beta f(q_B^L(p, k)) + (1 - \beta) \left(z + (1 - z)f(q_G^L(p, k)) \right) \right] \right\} \\ & + (1 - \Lambda) M \left\{ n^H(p, k) \left[\beta f(q_B^H(p, k)) + (1 - \beta) \left(z + (1 - z)f(q_G^H(p, k)) \right) \right] \right\} \end{aligned} \quad (104)$$

The payment vector (p, k) also determines the overall medical costs generated by doctors over the entire health system. Specifically, let the total service provided by a L - and a H - doctor be denoted by $Q^L(p, k)$ and $Q^H(p, k)$. Using this notation, the overall medical costs can be written as:

$$\sum_{j=1}^M C(Q^j) + F \sum_{j=1}^M n^j = M \{ \Lambda [C(Q^L(p, k)) + Fn^L(p, k)] + (1 - \Lambda) [C(Q^H(p, k)) + Fn^H(p, k)] \} \quad (105)$$

Where the first term and the second term represent the total operational and fixed costs incurred respectively. Finally, the non-medical costs associated with running the health system are given by the shadow price, λ , multiplied by the sum of payments to doctors, i.e.

$$\lambda \sum_{j=1}^M (pQ^j + n^j k) = \lambda M \{ \Lambda [pQ^L(p, k) + kn^L(p, k)] + (1 - \Lambda) [pQ^H(p, k) + kn^H(p, k)] \} \quad (106)$$

Substituting (104)-(106) in (89) yields welfare as a function of (p, k) . Slightly abusing the notation, I write the welfare function as $W(p, k)$. From Section 4.4's analysis, the interior solution of $n^L(p, k)$, $q_B^L(p, k)$ and $n^H(p, k)$, $q_B^H(p, k)$ is continuous in p for all $p > 0$ and $F - \alpha^L \leq k < F$.⁶⁸ Therefore, $W(p, k)$ is also continuous in p if $p > 0$ and $F - \alpha^L \leq k < F$.

⁶⁸ Suppose on the contrary $k - F < -\alpha^L$, the L -type physician would not obtain an interior solution.

Lemma 4.11 For a given $F - \alpha^L < k < F$, there exists a price $p_N(k) > 0$ such that $M[\Lambda n^L(p, k) + (1 - \Lambda)n^H(p, k)] = N$.

Proof. From lemma 4.4 and 4.8, we know that $n^L(p, k)$ and $n^H(p, k)$ are continuous and increasing in price for any given $p > 0$ and $F - \alpha^L \leq k < F$. As a result, there must exist a price $p = p_N(k) > 0$ that makes $M[\Lambda n^L(p, k) + (1 - \Lambda)n^H(p, k)] = N$. ■

Lemma 4.11 shows that for a given capitation rate k , there exists a unique price $p_N(k)$ such that physicians would treat all patients in the economy. Substitute $p_N(k)$ into $W(p, k)$ yields $W(p_N(k), k)$. As $p_N(k)$ is continuous in k , $W(p_N(k), k)$ is also continuous in k .

Assumption 1: The exogenous parameters α, β, F, z and functions $h(\cdot)$ and $C(\cdot)$ are selected in a way that $W(p, k)$ has an interior solution, i.e. there exists a price p_M such that $W(p_M, k) \geq W(p, k)$ for all $p > 0$ and $F - \alpha^L < k < F$.

From the Lemma 4.4 and 4.9, we know that at a given capitation rate $k \in (F - \alpha^L, F)$ and a price $p \rightarrow 0^+$, physicians enrol a relatively a small number of patients while providing large intensities of medical service. Both types of physicians have a zero profit and would raise the number of patients to be treated but reduce medical service intensities to be provided as price increases. The assumptions of health function therefore implies a relatively large marginal health benefit associated with treating an additional patient while a small marginal benefit associated with raising quality. Hence, $W(p, k)$ would have an interior solution if the expected marginal health benefit of enrolling an additional patient dominates the associated net fixed costs.

Assumption 2: Welfare maximization FFS system neither implies the shortage of patient demand nor the full patient access to health care (i.e. $M[\Lambda n^L(p_M, k) + (1 - \Lambda)n^H(p_M, k)] < N$).

My research investigates the design of a primary care physician remuneration scheme when the patient demand of health service is larger than the (optimal) amount that can be supplied by a healthcare system ($M[\Lambda n^L(p_M, k) + (1 - \Lambda)n^H(p_M, k)] < N$). The

cases where a healthcare system can benefit from creating more healthcare demand ($M[\Lambda n^L(p_M, k) + (1 - \Lambda)n^H(p_M, k)] > N$) or all healthcare demand has already been satisfied $M[\Lambda n^L(p_M, k) + (1 - \Lambda)n^H(p_M, k)] = N$ do not reflect challenges faced by HI systems in developed countries. Hence, the design of the optimal PCP under these two situations will be not be discussed.

Denote the constrained optimal price by $p^*(k)$ ⁶⁹, one can use the equation (88) to derive the associated insurance premium required to run the medical system. Specifically, multiplying both side of (88) by $\frac{1+\lambda}{N}$, we have:

$$\tau^* \equiv \frac{1+\lambda}{N} \sum_{j=1}^M (pQ^j(p^*(k), k) + kn^j(p^*(k), k)) \quad (107)$$

Since the first order equations of the aforementioned system are too large for a meaningful analytical solution and the system contains non-differentiability at the price level $p_0^L(k)$ and $p_0^H(k)$, I proceed to the ensuing analysis by employing a numerical analysis.

4.5.2 Specifications of Numerical Experiments

In order to perform the numerical exercise, I use the same specification of parameters as in Chapter 3.5.2 except for the following adjustments:

- The administration cost per patient is set at $F = 0.1$. The parameter of F is selected at this level, which ensures a well-behaved $W(p, k)$.
- From Lemma 4.2, the necessary condition that L and H -type doctors will enrol a positive number of patients is $k \in [F - \alpha^L, F]$ (i.e. $k \in [-0.05, 0.1]$). Since the per patient reimbursement in reality is non-negative, I assume in the following analysis that the capitation rate can be taken in the range $0 \leq k < 0.1$.
- With respect to patients, the specification of the total number of patients (i.e. $N = 100000$) captures that there are more patients demanding health services than the efficient number of patients who can be treated by the HI scheme. This

⁶⁹ In the current remuneration scheme, we have $p^* = p_N > p_M$.

reflects the reality that the health authority does not have abundant resources. As a result, it faces a two-way trade-off between efficiency and full patient access.

4.5.3 Numerical Results

I apply the foregoing specifications and use Mathematica to solve the respective doctors' problems and obtain all the supply functions. Next, I investigate the case where there is only a FFS payment (i.e. $k = 0$) and depict supply functions of different types of physician. Finally, I aggregate the supply decisions and calculate the welfare function which will be used to find the optimal price for medical services p^* which achieve the highest welfare while ensuring that all patients access healthcare.

4.5.3.1 Physicians' Response Curves

Figure 14 depicts L-doctors' responses to price changes. The Figure 14a, 14b, 14c and 14d respectively maps the number of patients selected to be treated, the total caseload provided, the quantity supplied to a B - and a G -patient, and the profit as a function of the regulatory price at $k = 0$, i.e. $n^L(p)$, $Q^L(p)$, $q_t^L(p)$, $t \in T = \{B, G\}$, and $\Pi^L(p)$.⁷⁰

⁷⁰ In order to avoid an overly cumbersome notation, I write $n^L(p)$, $q_t^L(p)$, $Q^L(p)$ and $\Pi^L(p)$ as $n^{L*}(p)$, $q_t^{L*}(p)$, $Q^{L*}(p)$ and $\Pi^{L*}(p)$ ($t \in T = \{B, G\}$) of the L-type doctor respectively.

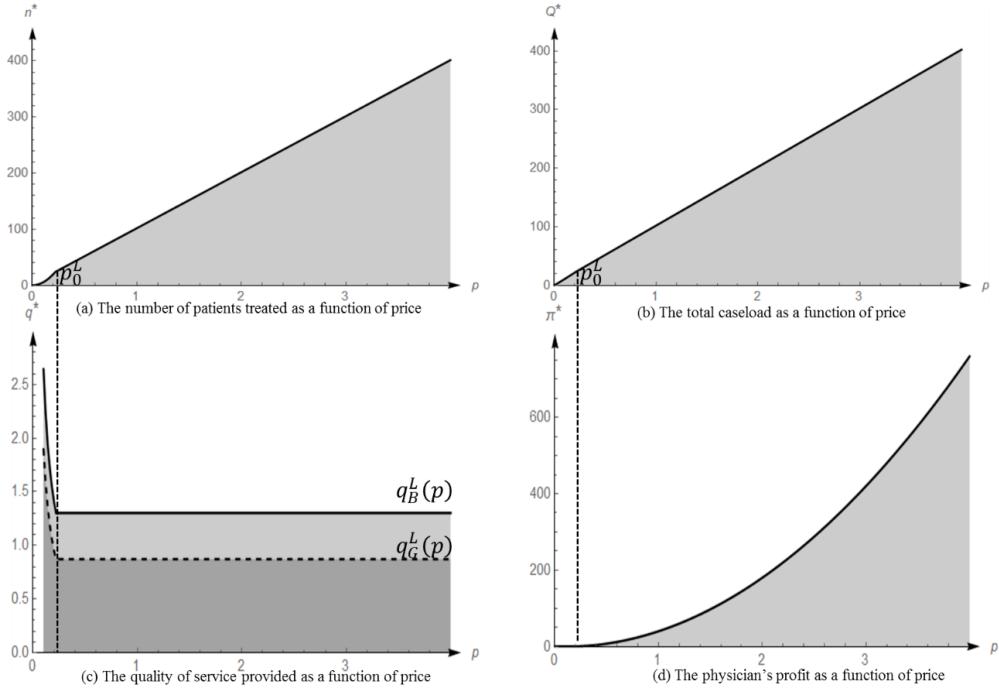


Figure 14. The L -doctor's response functions as price varies

The graphic demonstrates the results summarized by the Lemmas 4.2, 4.3, 4.4, 4.8 and 4.9; first, for all $p > 0$ and at a given $k \in [0,0.1]$, the system has interior solutions for $n^L(p)$ and $q_t^L(p)$, $t \in T = \{B, G\}$. Second, the L -type physician provides a higher medical service intensity to individuals with lower health status ($q_B^L(p) > q_G^L(p) > 0$, figure 14c). Third, in contrast to the result in the Chapter 3, the intensity of medical service provided $q_t^L(p)$ is not increasing in price. It is decreasing in price ($\frac{\partial q_t^L}{\partial p}(p) < 0$) in the range $p \in (0, p_0^L)$ while remaining unchanged ($\frac{\partial q_t^L}{\partial p}(p) = 0$) in the range $p \in (p_0^L, +\infty)$ (Figure 14c). Fourth, the number of patients enrolled is increasing in p but with a different slope at the kink p_0^L (i.e. $\frac{\partial n^L}{\partial p}(p_0^L^-) > \frac{\partial n^L}{\partial p}(p_0^L^+)$, Figure 14a). Fifth, similar to the function of patient numbers, total caseload also increases in p but with a kink p_0^L (Figure 14b). Finally, the L -doctors' profit is binding at zero for the price range $p \in (0, p_0^L)$ but increasing for the range $p \in (p_0^L, +\infty)$ (Figure 14d).

The next figure plots the individual decision functions of L - and H -doctors with respect to a type t patient. This figure mainly reflects result obtained in the Lemma 4.6. As in the previous figure, Figure 15a and 15c depict decisions of alternative types of doctor in relation to roster sizes and the medical service quantity as a function of the price. The figure 15b and 15d plot the total caseload and profit of L - and H -physicians respectively.

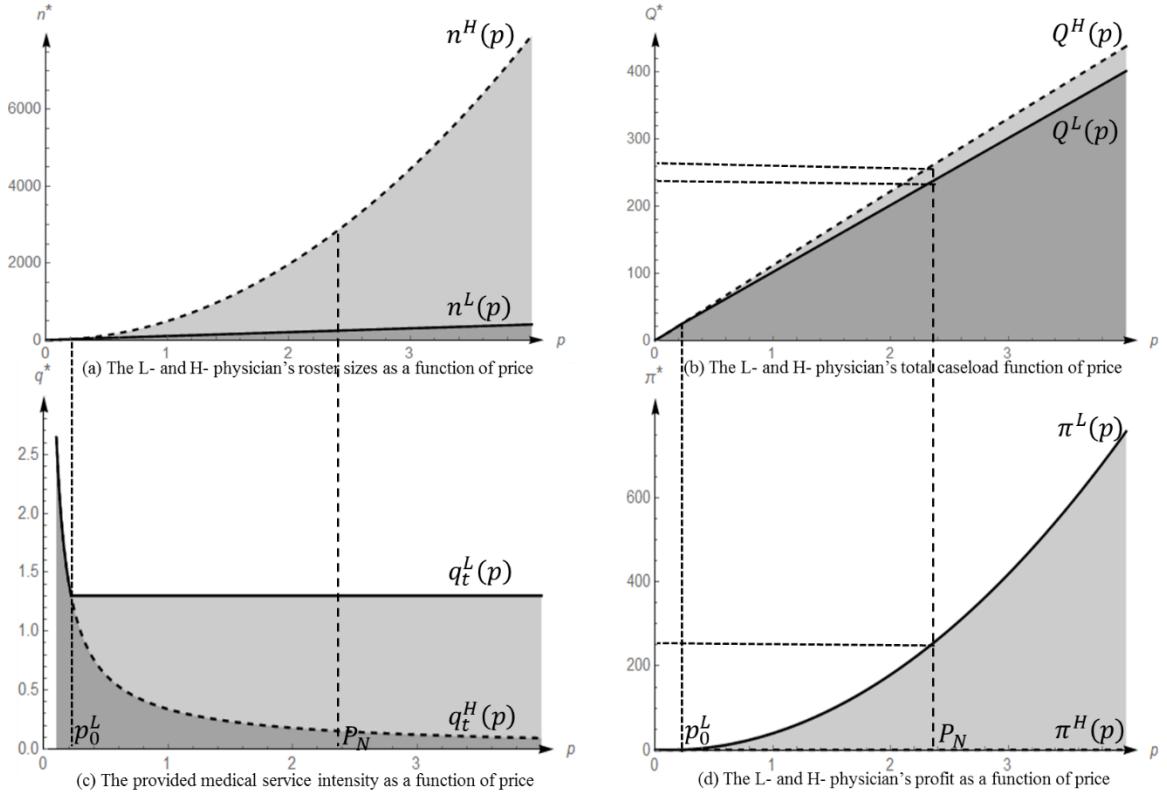


Figure 15. L- and H-doctors' response curves for a type-t patient

The critical price p_0^L is defined by the kink of $n^L(p)$ and $q_t^L(p)$ (Figure 15 a,c). However, for the parameters associated with H -type doctors, we have $a\alpha^H \left[\beta f(q_B^H) + (1 - \beta) \left(z + (1 - z)f((q_G^H)) \right) \right] > F$, meaning the marginal benefit of enrolling additional patient is always higher than the associated administration costs. According to the equation (94), the system has an equilibrium only if the net reimbursement per patient is negative ($k - F < 0$) and the Lagrange multiplier must satisfy $\eta > 0$. As a result, H -type doctors in this example always have a zero profit, therefore we have $p_0^H = +\infty$ (Figure 15d).⁷¹

The definition of p_0^L indicates that, for $p \in (0, p_0^L)$, both types of doctors obtain zero profit. While L - and H -type doctors have different degrees of altruism, to maximize their utilities their willingness to exchange the number of patients to be treated for the medical service intensity to be provided are always the same. Moreover, L - and H -doctors' profits are binding at zero. Restricted by these two conditions, the number of patients

⁷¹ It can be shown that if the value of z selected are sufficiently low, we will also have a critical price p_0^H for H -type doctors, which is defined at the kink of $q_t^H(p)$ or $n^H(p)$. Nevertheless, the current specification of parameters does not change the results.

enrolled and the intensity of medical service provided by both types of doctors are the same (Figure 15-a, c).

Second, for all $p \in (p_0^L, +\infty]$, the H -type physicians treat more patients ($n^H(p) > n^L(p)$, Figure 15a), provide higher total caseload ($Q^H(p) > Q^L(p)$, Figure 15b) while provide lower service intensity ($q_t^H(p) < q_t^L(p)$, Figure 15c) and reduce less service quality ($\frac{\partial q_t^H}{\partial p}(p) < \frac{\partial q_t^L}{\partial p}(p) = 0$, Figure 15c) than the L -type physicians as the price goes up.

Intuitively, the H -type doctors obtain higher altruism benefits from enrolling and serving patients, they have higher roster sizes and total caseload. At equilibrium, a pair of practice size and intensity is selected such that the net marginal altruism benefits from enrolling additional patient are the same as the benefits of raising medical service quality. As physicians face the same fixed cost of enrolling additional patient, the physician has higher altruism therefore provide the lower intensity of medical service. As shown in the previous section, the net marginal health benefits of enrolling additional patients or raising quality are independent of price when the physician's profit is positive (see the equation (94)). When the physician's profit is binding, he increases the number of patients enrolled to maximize utility while reducing quality to ensure a non-negative profit. Accordingly, the service intensity provided by L -type physicians remains the same whereas the quality produced by H -type physicians is reduced as price increases.

Finally, as the capitation net of fixed cost $k - F$ is negative, price is less than the marginal cost of providing total caseload, and the H -type doctors have higher roster sizes and total caseload, their profit is therefore lower (see lemma 4.6). In particular, for the parameters selected in this example, the more altruistic providers always have a zero profit (Figure 15d).

4.5.3.2 Welfare Function and the Number of Treated Patients under FFS

Substituting the supply decisions of the H - and L -doctors with respect to their G - and B -patients into the welfare function given $k = 0$ yields $W(p)$. Figure 16 provides a geometrical representation.

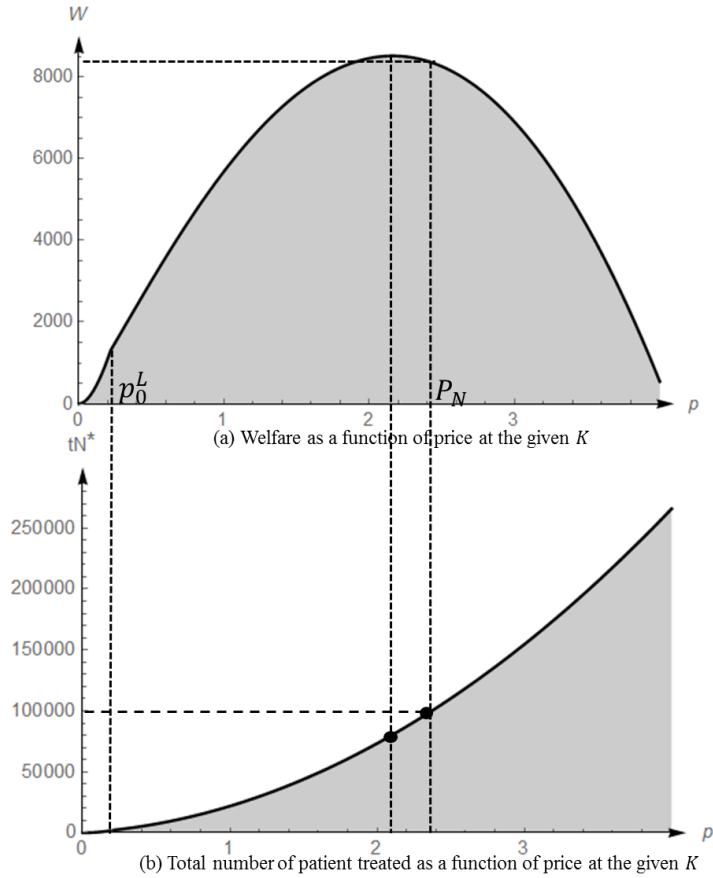


Figure 16. Welfare and Total Number of Patients Treated

In Figure 16a, notice that the welfare is maximized at the price $p_M = 2.1$ which is higher than $p_0^L = 0.22$, indicating that less altruistic doctors (the *L*-type) have a strictly positive profit while the more altruistic physicians are constrained by the zero-profit condition. Intuitively, the health insurance scheme does not have the required information to distinguish between *H*- and *L*-doctors as well as between *G*- and *B*-patients. As a result, the scheme sets an “intermediate” price which trades off the total caseload over-provision from *H*-doctors against under-provision by *L*-doctors (see Figure 15b). Figure 16b represents the result from Section 4.11 that the total number of patients treated increase in price and there is a price at the given capitation rate to ensure the entire population have access to healthcare. However, as pointed out in the Section 4.5.2, the price that ensures all patients’ access does not maximize social well-being (i.e. $p_M < p_N = 2.38$).

As shown in the previous chapter, the pricing scheme fails to align doctors’ incentives in the presence of information asymmetry. In particular, the more intrinsically motivated physicians will treat many patients (but provide too low quality) whereas the less

intrinsically motivated providers will provide very high quality (but enrol too few patients). Overall, welfare-maximizing price induces doctors with a high intrinsic motivation (in the model this is referred to as high altruism) to produce more than those who require a high extrinsic motivation.

In contrast to the last chapter, the current optimal HI scheme has to set the price higher than the welfare maximization level to ensure all patients have access to healthcare, which leads to total caseload over-provision. The next step of the analysis therefore uses a set of policy tools to control medical service over-provision and to align the incentives of different types of healthcare providers.

4.6 Cost-control policies and efficiency enhancements

In order to alleviate the tensions outlined in the previous section resulting from the presence of private information associated with doctors' preferences and patients' initial health, improve patients' health, control costs while guaranteeing that all patients can access healthcare, many healthcare systems have imposed quantity and expenditure limitations or introduced blended prospective and retrospective payments, as discussed in the Chapter 1 and Chapter 3.

By construction, quantity and costs restrictions as well as the blended payment are very crude as patients' severity of illness and providers' preferences is information only retained by doctors. The purpose of this section is to explore whether quantity rationing, revenue restrictions or capitation can be imposed on a pricing system to improve welfare in the current context. In order to keep the analysis tractable, I examine each restriction separately in Sections 4.6.1-3. After that, I explain relations between these restrictions and highlight departures from Chapter 3's results.

4.6.1 Quantity Rationing

In this section, I introduced to the setup in the Section 4.3 a new variable \bar{q} which represents the highest intensity of per patient treatment that a doctor can perform. Given $k = 0$, a HI policy is therefore a triplet (p, \bar{q}, τ) . Similar to Section 4.4, the doctor takes

this policy vector as given and maximize his utility. As a result, the optimization problem (90) has to be adjusted to include $q_t \leq \bar{q}$ for $t \in T = \{B, G\}$, therefore we have:

$$\begin{aligned}
U_Q &= \max_{n, q_t} \sum_{t \in T} n \beta_t [\alpha(z_t + (1 - z_t)f(q_t)) + pq_t + k - F] - C \left(n \sum_{t \in T} \beta_t q_t \right) \\
&\quad \sum_{t \in T} n \beta_t (pq_t + k - F) - C \left(n \sum_{t \in T} \beta_t q_t \right) \geq 0 \\
&\quad \bar{q} \geq q_t \\
&\quad n, q_t \geq 0
\end{aligned} \tag{108}$$

where the first line of the system shows the physician's utility optimization problem and the second, third line represents the non-negative profit restriction and the quantity restriction respectively. Finally, the last line shows the roster size and the intensity of service are non-negative. In this system, α represents a generic doctor's degree of altruism. Similar as the discussion in Section 4.5.2, the system (108) can have both boundary and interior solutions for all $p > 0$, $\bar{q} > 0$ and $0 \leq k < F$, if patient health function, the physician's cost function as well as exogenous parameters are chosen appropriately. In the ensuing analysis, I only consider those interior solutions.⁷²

Slightly abusing notation, I denote in this subsection the respective solution for H - and L -doctors by $n^L(p, \bar{q})$, $q_t^L(p, \bar{q})$ and $n^H(p, \bar{q})$, $q_t^H(p, \bar{q})$. Following the same procedures as in Section 4.5, I substitute $n^L(p, \bar{q})$, $q_t^L(p, \bar{q})$ and $n^H(p, \bar{q})$, $q_t^H(p, \bar{q})$ into the equation (89) and obtain a welfare function $W_Q(p, \bar{q})$. Notice that for any restriction level \bar{q} , the following process can be followed, i.e. solve for the price that maximizes welfare guarantees all patients can access health services, denoted hereafter by $p_Q^*(\bar{q})$,⁷³ then substitute this solution into the welfare function to obtain $W_Q(p_Q^*(\bar{q}), \bar{q})$. Clearly, for a

⁷² Theoretically, the system may have a boundary solution $n^{j*}(p, \bar{q})$ and $q_t^{j*}(p, \bar{q}) = \bar{q}$ ($j \in J = \{L, H\}$) if \bar{q} becomes sufficiently large.

⁷³ Notice that when \bar{q} is sufficiently large, the welfare maximizing price under quantity rationing is $p_Q^*(\bar{q}) = p_N$.

sufficiently large \bar{q} , the restriction does not bind and $W_Q(p_Q^*(\bar{q}), \bar{q}) = W(p_N)$.

Accordingly, we have:

$$W(p_N) \leq \max_{\bar{q}} W_Q(p_Q^*(\bar{q}), \bar{q}) \quad (109)$$

which means the optimal welfare under quantity rationing is at least the same as that under a pure pricing system given the entire population have access to healthcare. From the analysis of foregoing section, one would expect that a marginal reduction of \bar{q} at the point $\bar{q} = q_B^H(p_N)$ should lead to an improvement in welfare. Intuitively, the price p_N which ensures all patients have access to health services is too high compared to the welfare maximizing level p_M . As a result, the total health service provided is larger than the efficient level.

Moreover, at p_N the HI would benefit from reducing total caseload while maintaining the entire population's access to healthcare. As a result, one would expect that introducing a restriction which forces L -doctors to marginally reduce production for high health needy patients would positively impact this trade-off. More specifically, the reduction in \bar{q} would first induce L -type doctors to increase their number of patients treated. This increase allows the HI system to adjust price downwards, thereby decreasing the total service provided and raising the medical service intensity provided by H -type doctors per patient. To see these, I use the example described in Section 4.5.2 and follow the procedures described above to obtain $p_Q^*(\bar{q})$, $n^L(p_Q^*(\bar{q}), \bar{q})$, $n^H(p_Q^*(\bar{q}), \bar{q})$, $q_t^L(p_Q^*(\bar{q}), \bar{q})$, $q_t^H(p_Q^*(\bar{q}), \bar{q})$, $Q^L(p_Q^*(\bar{q}), \bar{q})$, $Q^H(p_Q^*(\bar{q}), \bar{q})$, $\tilde{\Pi}^L(p_Q^*(\bar{q}), \bar{q})$, $\tilde{\Pi}^H(p_Q^*(\bar{q}), \bar{q})$ and $W_Q(p_Q^*(\bar{q}), \bar{q})$ which have been plotted respectively in Figures 17 and 18.

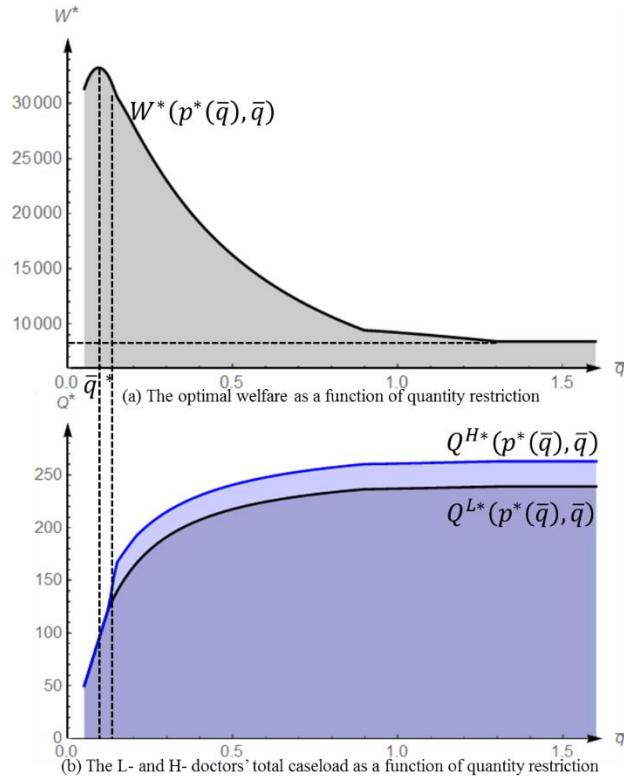


Figure 17. Welfare and total service supplied when quantity restriction varies

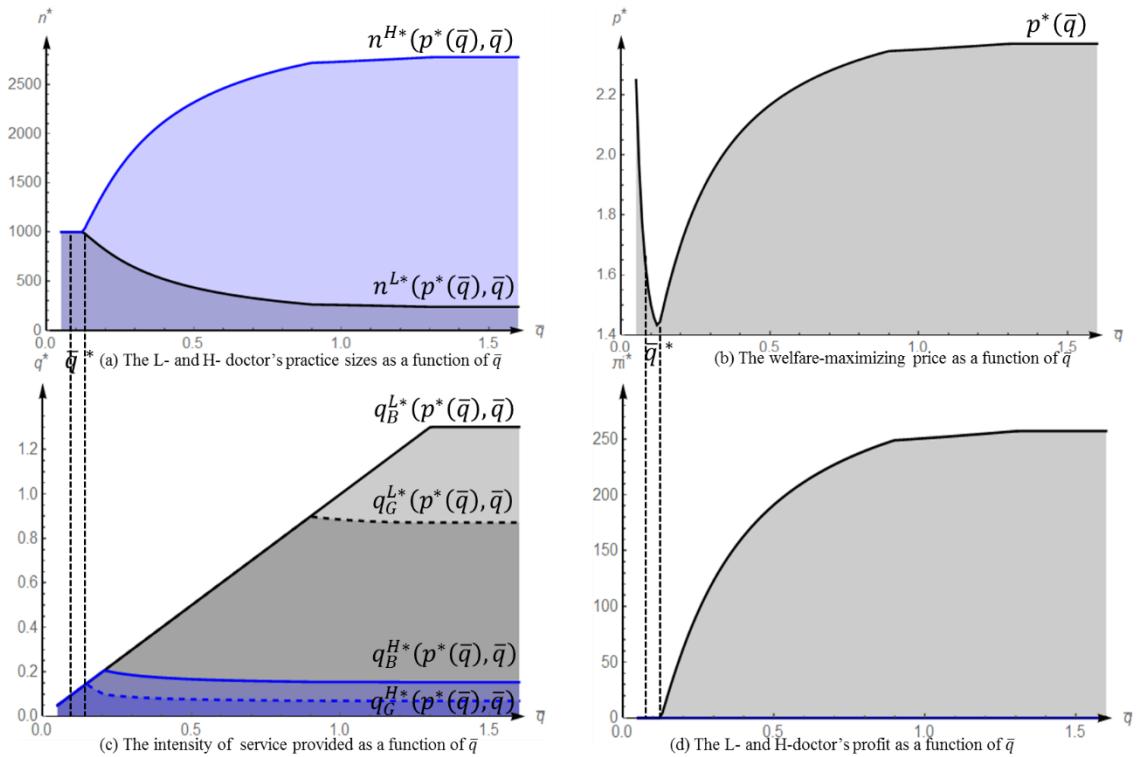


Figure 18. The optimal price and L- and H-doctors' decisions

The Figure 17a shows the intuition discussed above, that is, reducing \bar{q} below $\bar{q} = q_B^L(p_N) = 1.28$ improves welfare above $W(p_N)$. Other graphics provide more insights; the initial reduction in \bar{q} below $q_B^L(p_N)$ incentivizes L -type doctors to increase their roster sizes (Figure 18a), which subsequently induces the HI scheme to reduce the price (Figure 18b). As a result, H -type doctors decrease their number of patients treated and increase their service quality (Figure 18c). Overall, limitation on \bar{q} reduces total caseload provided by both types of doctors (Figure 17b) and the profit of L -type doctors (Figure 18d). However, the reduction in \bar{q} does not affect the profit of H -type doctors (Figure 18d). Since total caseloads are reduced and incentives for both types of doctors are realigned, welfare goes up. This process continues until both types of doctor enrol the same number of patients and provide the same quality of medical service (i.e. $\bar{q} = 0.13$). At this point, any further reduction in \bar{q} will reduce physicians' revenue and the zero-profit condition would induce them to decrease the total number of patients treated. In order to counter this effect, the HI system raises the price. Finally, welfare will reach the maximum level when both marginal social cost and marginal social benefits are equal (i.e. $\bar{q} = 0.1$).

Altogether, while imposing per patient quantity restrictions on a FFS system could improve welfare (as in the numerical example), it cannot induce efficiency. In the example at the welfare optimum H -doctors ended up providing the same level of medical services to B - and G -patients despite their significant difference in medical needs. Note that such a policy is likely to face many criticisms; this would frustrate doctors, in particular those who are intrinsically motivated as well as drawing criticism from more needy patients.

4.6.2 Revenue Cap

In the second subsection, I analyse an alternative cost-containing method where the HI scheme restricts the average expenditure per patient.⁷⁴ This subsection proceeds in the same way as the previous one. First, a new policy variable denoted by \bar{R} which measures the maximum expenditure that a doctor can spend on each individual is introduced.

⁷⁴ While in Chapter 3 the revenue restriction is on individual physician, the number of patients allocated per physician is fixed. As a result, limiting total expenditure per physician is the same as restricting health spending per patient.

Given $k = 0$, a HI policy is again a triplet (p, \bar{R}, τ) . Second, doctors' behaviour for a given policy vector (p, \bar{R}, τ) is derived, i.e. I solve:

$$\begin{aligned} U_R &= \max_{n, q_t} \sum_{t \in T} n \beta_t [\alpha(z_t + (1 - z_t)f(q_t)) + pq_t + k - F] - C \left(n \sum_{t \in T} \beta_t q_t \right) \\ &\quad \sum_{t \in T} n \beta_t (pq_t + k - F) - C \left(n \sum_{t \in T} \beta_t q_t \right) \geq 0 \tag{110} \\ &\quad n \bar{R} \sum_{t \in T} \beta_t - np \sum_{t \in T} \beta_t q_t \geq 0 \\ &\quad n, q_t \geq 0 \end{aligned}$$

where the first line shows the physician's utility maximization problem, the second and the third line represent the non-negative profit constraint and the revenue restriction respectively. The last line is non-negative constraint imposed on the practice size and medical service intensities provided. Third, slightly abusing the notation, the interior solution of the respective number of patients treated and the medical service provided are denoted by $n^L(p, \bar{R})$, $q_t^L(p, \bar{R})$ and $n^H(p, \bar{R})$, $q_t^H(p, \bar{R})$.

These solutions are then substituted into (89) to generate the function $W_R(p, \bar{R})$. Next, I denote by $p_R^*(\bar{R})$ the price maximizing social welfare while ensuring all patients obtain health service for a given expenditure cap \bar{R} . Lastly, substituting the solution into the welfare function yields $W_R(p_R^*(\bar{R}), \bar{R})$. Since as $\bar{R} \rightarrow +\infty$, $p_R^*(\bar{R}) = p_N$, we have:

$$W(p_N) \leq \max_{\bar{R}} W_R(p_R^*(\bar{R}), \bar{R}) \tag{111}$$

Substitute the constrained optimal price into the supply functions yields $n^L(p_R^*(\bar{R}), \bar{R})$, $n^H(p_R^*(\bar{R}), \bar{R})$, $q_t^L(p_R^*(\bar{R}), \bar{R})$, $q_t^H(p_R^*(\bar{R}), \bar{R})$. The total caseload and profit of the respective type of patient are then $Q^L(p_R^*(\bar{R}), \bar{R})$, $Q^H(p_R^*(\bar{R}), \bar{R})$, $\tilde{\Pi}^L(p_R^*(\bar{R}), \bar{R})$ and $\tilde{\Pi}^H(p_R^*(\bar{R}), \bar{R})$. Finally, social welfare is $W_R(p_R^*(\bar{R}), \bar{R})$. In order to gain more insights, the numerical exercise from Section 4.6.1 is adjusted to examine the impact of varying the revenue cap.

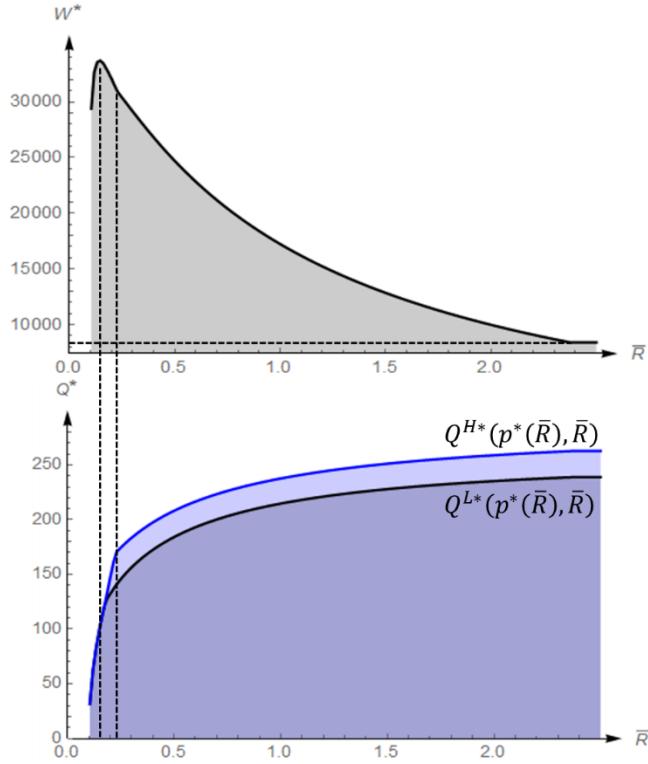


Figure 19. Welfare and total service supplied when revenue restriction varies

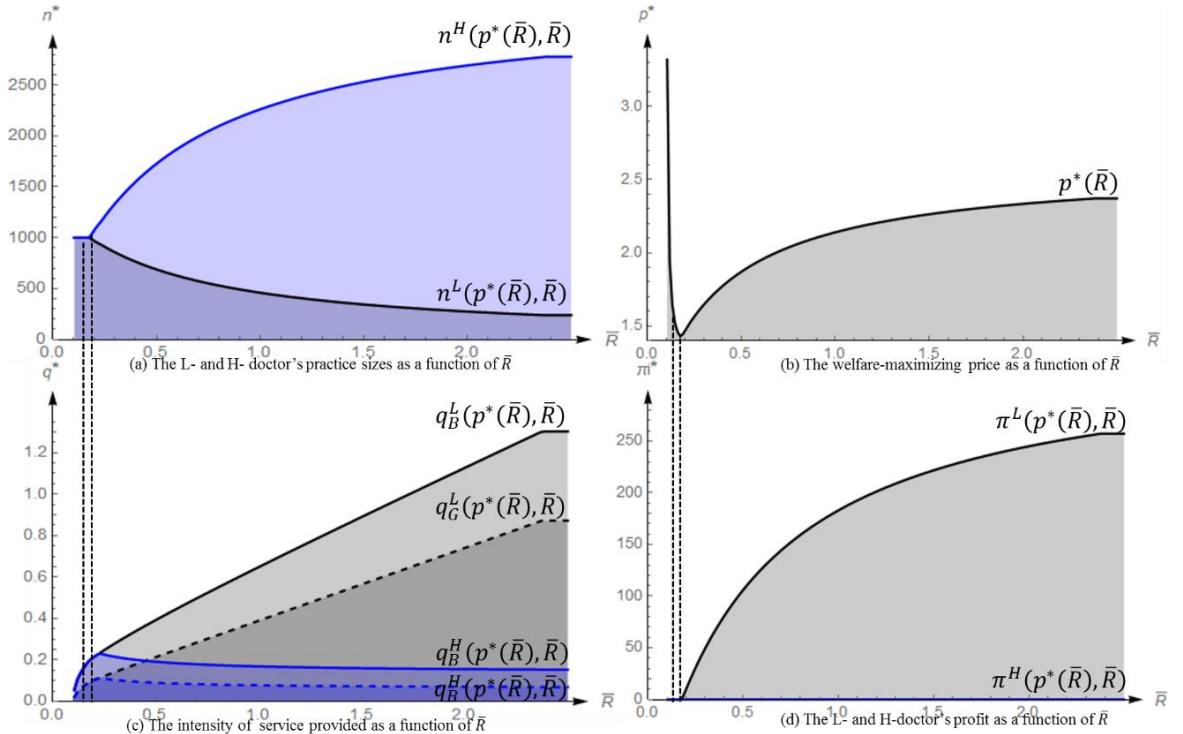


Figure 20. The optimal price and L- and H-doctors' decisions

The Figure 19a plots the result of the exercise for the parameter introduced in Section 4.5.2. Observe that \bar{R} is sufficiently large while the revenue cap never binds so

that $W_R(p_R^*(\bar{R}), \bar{R}) = W(p_N)$. Initial reductions in the revenue cap starts affecting physicians' decisions regarding their quality of care and patient numbers when \bar{R} approaches the point $\bar{R} \approx 2.43$, where L -type providers' spending per patient just binds. At this point, the reduction in \bar{R} would force L -type doctors to reduce quality (see Figure 20c) while increasing the number of patients to treat (Figure 20a). The HI system hence found it is advantages decreasing the price (Figure 20b), resulting in H -type doctors reducing their roster sizes and raising medical service quality. As the price goes down, both types of doctors reduce their total caseload (Figure 19b) and the less altruistic providers have less profits (Figure 20d).

Overall, social welfare improved (Figure 19a) as the total caseload provided in the society is reduced and incentives for different types of doctors are realigned. This process will continue until $\bar{R} \approx 0.22$, where incentives for both types of doctor are completely aligned (i.e. they select the same number of patients and provide the same quality of medical service). At this level of restriction, any further reduction beyond this expenditure level will reduce physicians' revenue and therefore resulting in the reduction of patient enrolment. To counter this effect, the HI scheme has to increase the price. The maximum level of welfare will be attained at $\bar{R} \approx 0.15$, where all the margins are aligned.

A few remarks are provided to conclude this subsection. First, notice from Figures 19 and 20 that the current system reduces the total caseload provided while ensuring the entire population's access to treatment. Second, compared with the optimal pricing policy without introducing expenditure cap, H -doctors are induced to treat fewer patients and provide a higher quality of service while L -doctors are motivated to treat more patients and reduce service quality. In other words, the optimal expenditure control helps HI overcome one of the weaknesses related to the standard FFS system and linked to the informational asymmetry between HI and doctors.

Finally, compared with the quantity restriction, the revenue cap provides doctors with more flexibility which allows them to better adjust healthcare resources to the specific needs of individual patients. One would expect that this additional flexibility would make the revenue cap policy more palatable to doctors than the foregoing quantity restriction. I also find in this numerical example that patients with lower initial health

levels obtain more service and therefore have higher expected health under the expenditure control is higher than under the quantity rationing.

4.6.3 Capitation

In the last subsection, I investigate the effect of blending the pricing system and one of the most frequently used prospective remuneration method, namely fixed payment per patient or capitation. I refer capitation as a “cost-control” instrument as it will be set below the level of per patient fixed cost, thereby reducing physicians’ incentive to enrol patients and increase total caseload. Following the procedures outlined in the previous two sections, I first assume that the capitation rate takes value within the range $0 < k < F$. As a result, a HI policy becomes triplet (p, k, τ) . Second, the physician selects the size of his roster and quality of medical care at the given policy to maximize his utility, i.e.

$$\begin{aligned} U &= \max_{n, q_t} \sum_{t \in T} n\beta_t [\alpha(z_t + (1 - z_t)f(q_t)) + pq_t + k - F] - C \left(n \sum_{t \in T} \beta_t q_t \right) \\ &\quad \sum_{t \in T} n\beta_t (pq_t + k - F) - C \left(n \sum_{t \in T} \beta_t q_t \right) \geq 0 \\ &\quad n, q_t \geq 0 \end{aligned} \tag{112}$$

where the first line of the system is the physician’s utility optimization problem whereas the second line and the third line represent the non-negative constraints imposed on profit, practice size and medical service intensity provided. Then, considering the interior solution only, I denote the respective solution of *L*-and *H*-type doctors based on the foregoing system by $n^L(p, k), q_t^L(p, k)$ and $n^H(p, k), q_t^H(p, k)$.

Substituting these solutions into (89) yields the welfare function $W_K(p, k)$. Next, I solve the price which maximizes social welfare while ensuring all patients access health services and denote the solution by $p_K^*(k)$. Finally, substituting this price back into the physician’s supply functions, total caseloads and profits yields $n^{L*}(p_K^*(k), k), q_t^{L*}(p_K^*(k), k), n^{H*}(p_K^*(k), k), q_t^{H*}(p_K^*(k), k); Q^{L*}(p_K^*(k), k)$,

$Q^{H*}(p_K^*(k), k), \Pi^{L*}(p_K^*(k), k), \Pi^{H*}(p_K^*(k), k)$. The social welfare function is then $W_K(p_K^*(k), k)$. The following two figures show the impact of varying k based on the parameters introduced in Section 4.5.2.

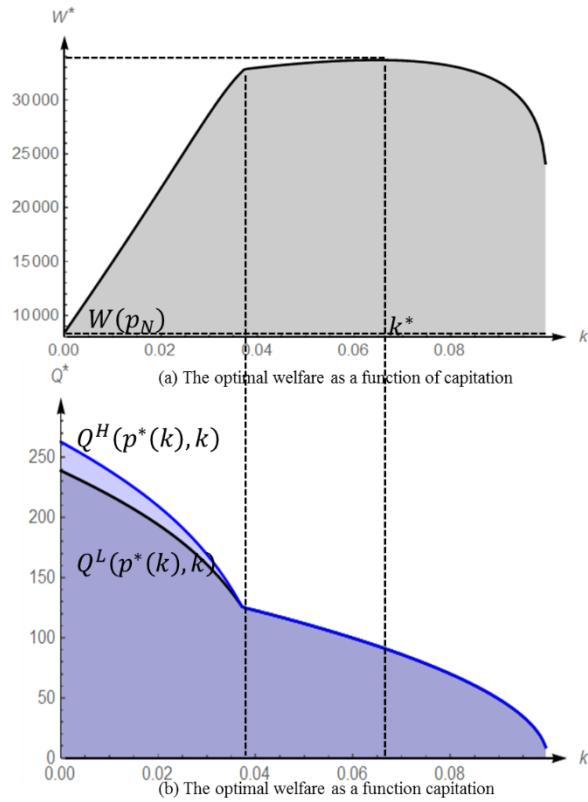


Figure 21. Welfare and total quantity of service supplied when capitation varies

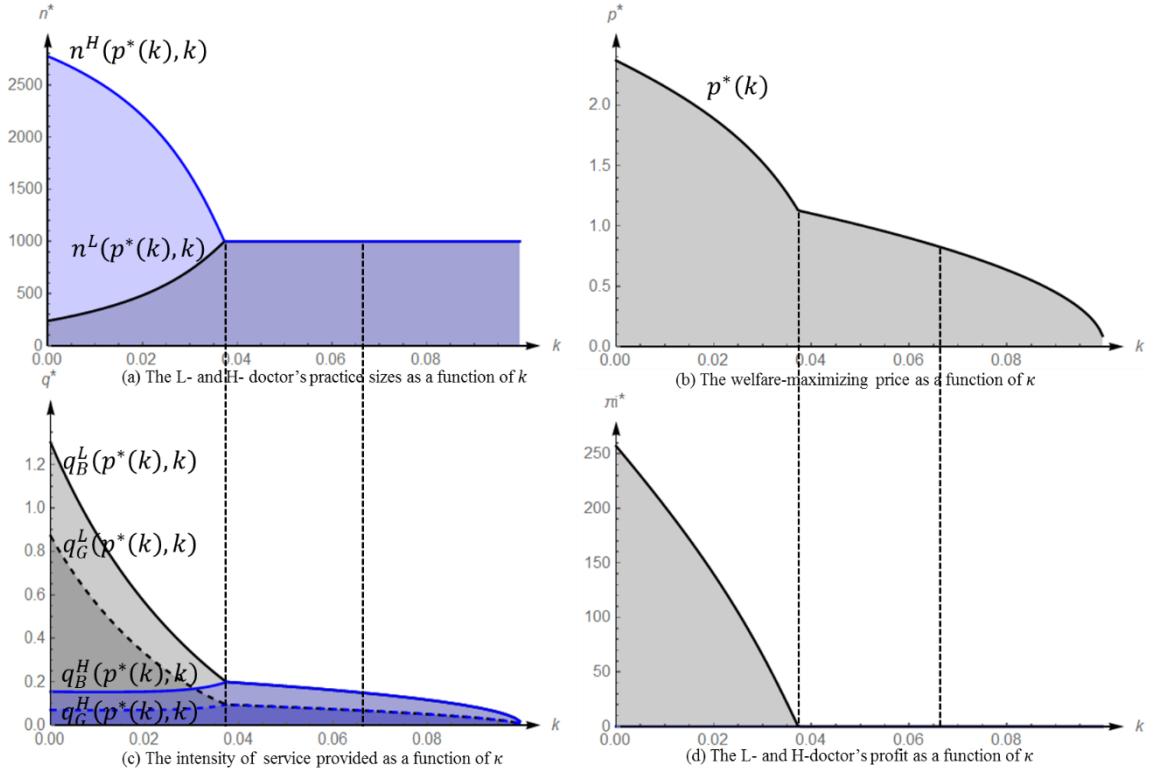


Figure 22. The optimal price and L- and H-doctors' decisions

Figure 21a plots the overall impact of varying k on optimal social well-being, which indicates blending capitation with FFS indeed helps improving welfare, i.e.

$$W(p_N) \leq \max_k W_K(p_K^*(k), k) \quad (113)$$

To illustrate, notice that when $k = 0$, we have $W_K(p_K^*(k), k) = W(p_N)$. When the HI starts raising capitation, both L - and H -type doctors will enrol more patients. Since all patients have already accessed health services (under the welfare maximized FFS), the HI will adjust price downwards (Figure 22b). L -type physicians will increase their practice size but reduce service quality (Figure 22 a, c) as the direct impact of raising capitation outperforms the indirect effect of the price reduction. On the contrary, H -type physicians will reduce their practice size but increase their service quality (Figure 22 a, c) as the indirect effect of price reduction dominates. Overall, both types of doctors reduce their total caseload (Figure 21b) while the less altruistic doctors obtain a lower profit (Figure 22d).

As the introduction of capitation not only reduces both types of doctors' total caseloads but also aligns their incentives, social welfare goes up. This process continues until the capitation rate rises to $k = 0.037$ where both types of doctor select the same roster size

and provide the same intensities of medical service as well as having zero profit. At this capitation rate, any further increases in k would force the HI system to reduce price at a lower rate, as L -type doctors have a zero profit and no incentive to increase their practice size. Since both types of doctor select the same number of patients to be treated and provide the same intensities of medical service, the impact of raising capitation dominates the impact of reducing price. Hence, both types of doctors reduce health service intensities. Finally, when capitation reaches the level $k = 0.066$, all margins align and social welfare reaches the maximum level. Under the selected parameters, I find welfare at the optimal capitation in this numerical exercise is the same as at the optimal revenue restriction.

To summary this subsection, a few remarks are in order. First, my results suggest that the HI scheme can improve social well-being by mixing FFS and capitation. This mechanism substitute price by capitation, which ensures all patients have access to treatment while realigning doctors' different incentives and reducing total caseload. Second, using the optimal capitation yields the same level of social welfare as using the optimal expenditure cap, as both systems exploit physicians' private knowledge and provided a different quality of service to different types of patients. Finally, while the optimal capitation improves the same level of welfare as the optimal revenue cap, it is more tolerant of mistakes made by the regulator (i.e. there are more combinations of price and capitation rate that leads to social well-being at around the maximum level). This is because the substitution between price and capitation allows physicians to have more flexibility in terms of determining the total number of patients to be enrolled and the health service intensities to be provided.

4.6.4 Discussion

The FFS payment has been the predominant physician payment used by many developed and developing countries over the last decades (Porter and Kaplan, 2016). Patients under this system are rarely excluded from treatment and are more likely to receive the intensity of medical service that maximizes their health (Ginsburg, 2015). On the other hand, physicians' income increases as they enrol additional patients and provide more service. FFS therefore successfully aligns incentives between patients and doctors.

However, evidence shows that FFS induces treatment overprovision and drives up healthcare spending (Ginsburg, 2015; Léger, 2008; McGuire, 2000).

The above analysis points out that physician-predicted responses to variations in different systems differ in important ways from traditional wisdom. Specifically, I show physicians increase their practice sizes but do not raise service quality when the price goes up. This result is consistent with studies which have found no evidence of supplier induced demand under a fee-for-service regime in countries with publicly funded healthcare systems (Carlsen and Grytten, 1998; Grytten et al., 2008; Grytten et al., 2001; Grytten and Sørensen, 2001; Madden et al., 2005). While a FFS system can be designed to ensure that all patients have access to healthcare, my numerical finding shows it inevitably incentivizes medical service over-provision and cost escalation. Moreover, the same reimbursement per service to different types of patients leads to inefficiencies as it induces more intrinsically motivated staff to treat too many patients while offering too low levels of quality. Meanwhile, more extrinsically motivated physicians provide too high a level of service quality but treat too few patients.

A logical reaction to this situation is to introduce into FFS system additional restrictions aimed at capping the expenditure or service intensity provided by extrinsically motivated doctors while reducing the price. Doing so helps to incentivise low altruistic providers to enrol more patients and to raise quality provided by more intrinsically motivated doctors. Moreover, the introduction of cost control policies reduces the excessive provision of total caseload by both types of doctors.

In that respect, quantity rationing is found to improve efficiency, nevertheless it does not react to physicians' local knowledge. In contrast to Chapter 3 which suggests that quantity rationing may sometimes worsen social welfare, this chapter demonstrates that restricting service intensity provided per patient can always improve welfare when physicians are able to select their practice size. There are two main departures from the last chapter: (1) it is the less altruistic doctor who provides a higher intensity of medical service and is first restricted by quantity rationing as the value of the restriction imposed becomes lower, and (2) quantity restriction would induce physicians to substitute the number of patients treated and average health service quality rather than substituting between service quality provided between two types of patients. In Chapter 3, as the physician is allocated a fixed number of patients, the introduction of quantity restriction

may induce him to substitute the service provided to less healthy individuals to healthier individuals, resulting in an inefficiency. In this chapter, the use of quantity restriction induces less motivated physicians enrol more patients, leading to an efficiency improvement.

With respect to revenue restriction, this analysis shows that the restriction not only reduces total caseload and realigns different physicians' incentives but also continues to use local doctors' information. However, the expenditure cap in this chapter is per case while in the last chapter it restricts per physician total spending. It can be shown that if this chapter introduces a per physician expenditure cap as in Chapter 3, it would not improve welfare as much as quantity rationing as it does not induce the re-alignment of different physicians' incentives between enrolling more patients and increasing medical service quality.

Finally, I show that the physician reduces the quality of service as the capitation rate increases, although he increases the number of patients treated. This result is consistent with Barham and Milliken (2015), however, it does not require the assumption of the bell-shaped physician altruism. While mixing payment per patient less than fixed cost per patient with price is regarded as the optimal cost control policy in this chapter, I find this mixed system does not yield higher social well-being compared to the revenue cap. This is because both restrictions align physicians' incentives, reduce total caseload, and exploit doctors' local knowledge. The only difference between introducing capitation and a revenue cap is that the former is more tolerant of potential mistakes by health authorities. In the current context, capitation can be considered as a substitute of price which aligns different physicians' incentives rather than a restriction that limits their service provision.

4.7 Summary

In this chapter, I have studied the design of incentive contracting when physicians determine their own practice sizes and the HI scheme can use both financial payments as well as cost-control instruments to align incentives. To formalize this idea, I extend the analytical framework outlined in the last chapter by endogenizing the number of patients allocated and introducing a fixed cost per patient. This setup allows us to analyse how the physician trade-off between enrolling more patients and improving health

service quality without imposing a particular assumption on the physician's preference. Moreover, my model does not assume an exogenously given optimal service provided per patient. Instead, it can be derived as a function of price, capitation rate, as well as other exogenous parameters at the level that maximizes welfare while ensuring all patients have access to healthcare. Finally, this model extends Barham and Milliken's (2015) analysis to the case where the physician needs to treat multiple types of patient while constrained by the non-negative profit condition.

This model generates four main findings: First, raising per patient or per service reimbursement does not lead to improvement in health service quality; Second, the more altruistic the physician is, the higher number of patients he would like to treat but the lower quality he would provide. Third, the FFS system leads to an inefficient provision of health service and medical service over-provision; Finally, cost control policies such as per patient quantity limitation and expenditure control as well as capitation are found to be effective in terms of enhancing efficiency and controlling costs. In particular, blending FFS and capitation outperforms introducing a FFS system a quantity rationing, or revenue cap as it reacts to the local knowledge of doctors and is more tolerant of potential mistakes that could be made by the health insurance scheme. The mix of FFS and per patient reimbursement less than fixed costs can also be regarded as a bonding contract, which sells physician the right to serve patients. This remuneration scheme aligns different types of physician incentives and extracts their informational rents.

4.8 Limitations and Extensions

The generality of the analytical framework in Chapters 3 and 4 offers a simple mechanism that is sufficiently detailed to understand the impact of introducing respective cost control policies in a fee-for-service physician remuneration payment system. However, in addition to the extensions discussed in Chapter 3, there are also other opportunities for further development. In particular, my thesis focuses on investigating physicians' decisions regarding their patient enrolment or the provision of medical service intensity under a FFS system and the subsequent impact of introducing three commonly used cost-control policies. As a result, physicians' decisions under other frequently used single payment systems such as salary, cost-reimbursement and cost sharing have not yet been explored. Hence, future research might adjust my model to

include these systems. In addition, the recently developed cost control policies such as bundled payment, global budgets and payment-for-performance have played an increasingly important role in reducing costs and improving service quality (Offodile et al., 2019, Santos et al., 2019). Accordingly, it will be beneficial to model these policies and evaluate their impact. Future studies can also examine the design of the optimal menu of contracts when cost control policies have been taken into consideration. While my research does not analyse this option, the results I generate can be interpreted as the indirect implementation of the direct revealing mechanism (Gaynor et al., 2018), with a restriction that the payment has to be linear in quantity.

Another potential extension could be to allow patients playing a positive role in determining the service quantity to be provided by their physicians. Studies in this stream of literature have analysed the optimal design of patient insurance and physician payment simultaneously as well as their impact on physicians' health service supply and patients' healthcare demand decisions. For instance, in Ellis and McGuire (1990), the patient and his physician need to reach a Nash-bargaining equilibrium in relation to quantity of service to be supplied. In Ma and McGuire (1997), the patient decides the quantity of service to be supplied while the physician decides the level of effort to be exerted. In both papers, the insurance scheme decides the premiums or patients' copayments to be collected and the remuneration scheme of physicians. Furthermore, as physicians' risk attitude also plays an essential role in determining health service provision (Arrieta et al., 2017; Galizzi et al., 2013), it will be interesting to see how my results would change if patients or physicians are assumed to be risk-averse.

Altogether, theoretical research over the next 5-10 years should extend my setup by including more essential features of the healthcare market, analysing physician behaviour under other most frequently used physician payment mechanisms, examine the impact of imposing recently developed cost-control instruments and investigating the optimal design of insurance for patients and payment for doctors.

Next, it is interesting to see whether my theoretical or numerical predictions hold up to empirical scrutiny. For instance, my thesis shows whether the physician increases his service quality in price depends on whether he is allowed to decide his practice size. Moreover, my findings suggest that price and insurance premiums increase with ageing (represented by the proportion of the population over 65), the proportion of less altruistic

physicians, and science innovations (represented by a technology factor). My results also suggest the need to impose a FFS payment; either a cost control policy such as capitation, an expenditure cap, or quantity rationing which would result in welfare improvement. In particular, using capitation increases social welfare at least to the same level when the other two instruments are used. Econometricians could construct models to test the significance of these hypotheses. To test the impact of using different cost control policies, future studies should find a way to measure social welfare. Subsequently, evidence of physicians' decisions on the supply of health service or patient enrolment before and after introducing the above cost control policies and variations of welfare should be compared with my theoretical findings. Econometric models can also be used to estimate the percentage of respective drivers (i.e. population ageing or technology innovation) contributes to healthcare expenditure growth. Finally, econometric models will allow forecasting or simulating of variations of health service price, insurance premium and the average income of physician over the next decades.

Finally, and as discussed in Chapter 2, laboratory experiments could be one of the best tools to test my theoretical predictions. This is because experiments allow researchers to investigate physician behaviour in a controlled manner and under *ceteris paribus* conditions (Hennig-Schmidt et al., 2011). Moreover, participants in the experiments are randomly assigned to experimental conditions excluding selection bias. Experimental investigations are based on actual decisions associated with monetary rewards that are related to participants' choices. This is a situation real physician face in their daily practice. Experiments also serve as "wind tunnels" before institutional changes such as healthcare reforms are implemented (Hennig-Schmidt et al., 2011).

Future studies can also extend the experiment of Hennig-Schmidt et al. (2011) by modelling cost control policies and analysing their impact on the physician's supply of health service to the individual patient as well as to social welfare. To do that, researchers may divide all subjects into three groups and follow the process of Hennig-Schmidt et al. (2011) to analyse their service supply decisions under FFS only by recording per patient service provided q_i and expenditure incurred r_i while denoting the highest quantity of service offered by \bar{q} and the highest expenditure incurred by \bar{r} . Next, researchers could introduce the first group with a quota $q_i \leq \bar{q}$, the second group with an expenditure cap $r_i \leq \bar{r}$, and the third group with a positive or negative payment per patient in addition to a FFS payment. These would allow researchers to examine how

physicians' supply decisions change after the introduction of respective cost control policies. By measuring patients' health benefits in monetary terms, the experiments can use the welfare function introduced in Chapter 3 to estimate the social well-being generated by each group and conclude the most efficient policy.

4.9 Appendix

B1

Proof. Calculating the first order condition with respect to p and applying the implicit function theorem to system (96), we have:

$$\begin{pmatrix} -q^2 C'' & -nqC'' \\ -qC'' & \alpha f'' - nC'' \end{pmatrix} \begin{pmatrix} \frac{\partial n^*}{\partial p} \\ \frac{\partial q^*}{\partial p} \end{pmatrix} = \begin{pmatrix} -q \\ -1 \end{pmatrix} \quad (114)$$

Inverting the matrix, solving and applying the curvature assumptions yields:

$$\begin{pmatrix} \frac{\partial n^*}{\partial p} \\ \frac{\partial q^*}{\partial p} \end{pmatrix} = \begin{pmatrix} -q^2 C'' & -nqC'' \\ -qC'' & \alpha f'' - nC'' \end{pmatrix}^{-1} \begin{pmatrix} -q \\ -1 \end{pmatrix} \quad (115)$$

Hence,

$$\begin{pmatrix} \frac{\partial n^*}{\partial p} \\ \frac{\partial q^*}{\partial p} \end{pmatrix} = \begin{pmatrix} \frac{1}{qC''} \\ 0 \end{pmatrix} \quad (116)$$

Finally, the definition of Q indicates:

$$\frac{\partial Q^*}{\partial p} = q^* \frac{\partial n^*}{\partial p} + n^* \frac{\partial q^*}{\partial p} = \frac{1}{C''} \quad (117)$$

The respective signs follow from the curvature assumptions $f'' < 0, C'' > 0$. Therefore, we have $\frac{\partial n^*}{\partial p}, \frac{\partial q^*}{\partial p} > 0$ and $\frac{\partial q^*}{\partial p} = 0$, thereby verifying the claim.

The proof can be easily generalized to the case where $z_B = 0$ and $z_G = z$, so in this situation the first order condition with respect to price becomes:

$$\begin{pmatrix} \bar{q}^2 C'' & Q\beta C'' & Q(1-\beta)C'' \\ \bar{q}C'' & n\beta C'' - \alpha f'_B & n(1-\beta)C'' \\ \bar{q}C'' & n\beta C'' & n(1-\beta)C'' - \alpha(1-z)f'_G \end{pmatrix} \begin{pmatrix} \frac{\partial n^*}{\partial p} \\ \frac{\partial q_B^*}{\partial p} \\ \frac{\partial q_G^*}{\partial p} \end{pmatrix} = \begin{pmatrix} \bar{q} \\ 1 \\ 1 \end{pmatrix} \quad (118)$$

Rearranging the above matrix, I have:

$$\begin{pmatrix} \frac{\partial n^*}{\partial p} \\ \frac{\partial q_B^*}{\partial p} \\ \frac{\partial q_G^*}{\partial p} \end{pmatrix} = \begin{pmatrix} \bar{q}^2 C'' & Q\beta C'' & Q(1-\beta)C'' \\ \bar{q}C'' & n\beta C'' - \alpha f_B'' & n(1-\beta)C'' \\ \bar{q}C'' & n\beta C'' & n(1-\beta)C'' - \alpha(1-z)f_G'' \end{pmatrix}^{-1} \begin{pmatrix} \bar{q} \\ 1 \\ 1 \end{pmatrix} \quad (119)$$

Calculating and simplifying, I have:

$$\begin{pmatrix} \frac{\partial n^*}{\partial p} \\ \frac{\partial q_B^*}{\partial p} \\ \frac{\partial q_G^*}{\partial p} \end{pmatrix} = \begin{pmatrix} 1 \\ \bar{q}C'' \\ 0 \\ 0 \end{pmatrix} \quad (120)$$

B2

Proof. Calculating the first order condition with respect to k and using the implicit function theorem, I get:

$$\begin{pmatrix} \frac{\partial n^*}{\partial k} \\ \frac{\partial q^*}{\partial k} \end{pmatrix} = \begin{pmatrix} -q^2 C'' & -nqC'' \\ -qC'' & \alpha f'' - nC'' \end{pmatrix}^{-1} \begin{pmatrix} -1 \\ 0 \end{pmatrix} \quad (121)$$

hence, I have:

$$\begin{pmatrix} \frac{\partial n^*}{\partial k} \\ \frac{\partial q^*}{\partial k} \end{pmatrix} = \begin{pmatrix} nC'' - \alpha f'' \\ -\alpha q^2 f'' C'' \\ 1 \\ \alpha q f'' \end{pmatrix} \quad (122)$$

$$\frac{\partial Q^*}{\partial k} = n^* \frac{\partial q^*}{\partial k} + q^* \frac{\partial n^*}{\partial k} = \frac{1}{qC''} > 0 \quad (123)$$

If there are two types of patients, repeating the foregoing process yields:

$$\begin{pmatrix} \frac{\partial n^*}{\partial k} \\ \frac{\partial q_B^*}{\partial k} \\ \frac{\partial q_G^*}{\partial k} \end{pmatrix} = \begin{pmatrix} \bar{q}^2 C'' & Q\beta C'' & Q(1-\beta)C'' \\ \bar{q}C'' & n\beta C'' - \alpha f_B'' & n(1-\beta)C'' \\ \bar{q}C'' & n\beta C'' & n(1-\beta)C'' - \alpha(1-z)f_G'' \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad (124)$$

hence, we have:

$$\begin{pmatrix} \frac{\partial n^*}{\partial k} \\ \frac{\partial q_B^*}{\partial k} \\ \frac{\partial q_G^*}{\partial k} \end{pmatrix} = \begin{pmatrix} \frac{\alpha^2(1-z)f_B''f_G'' - n\alpha C''[\beta(1-z)f_G'' + (1-\beta)f_B'']}{\alpha^2(1-z)\bar{q}^2 f_B'' f_G'' C''} \\ \frac{\alpha(1-z)f_G''\bar{q}C''}{\alpha^2(1-z)\bar{q}^2 f_B'' f_G'' C''} \\ \frac{\alpha f_B''\bar{q}C''}{\alpha^2(1-z)\bar{q}^2 f_B'' f_G'' C''} \end{pmatrix} \quad (125)$$

The size of $\frac{\partial q_B^*}{\partial k}$ and $\frac{\partial q_G^*}{\partial k}$ depends on the level of f_B'' and $(1-z)f_G''$. Since $f'' < 0$ and $f'(+\infty) = 1$, we know $f''' > 0$. Moreover, as $q_B^* > q_G^* > 0, f'' < 0, 0 > f_B'' \geq (1-z)f_G''$ and therefore we obtain $0 > \frac{\partial q_G^*}{\partial k} > \frac{\partial q_B^*}{\partial k}$, verifying the claim.

B3

From the second equation of (96), we know $p - C' = -\alpha f'$. Substituting this into the first equation of (96) yields:

$$\alpha[f(q) - qf'(q)] + k - F = 0 \quad (126)$$

Applying the implicit function theorem and rearranging, I have:

$$\frac{\partial q^*}{\partial \alpha} = \frac{f - qf'}{\alpha q f''} < 0 \quad (127)$$

Calculating the first order condition with respect to α for the second equation of (96) and substitute the equation (127), I have:

$$\frac{\partial n^*}{\partial \alpha} = \frac{\alpha q f f'' - n C''(f - qf')}{\alpha q^2 f'' C'} > 0 \quad (128)$$

Altogether:

$$\frac{\partial Q^*}{\partial \alpha} = n^* \frac{\partial q^*}{\partial \alpha} + q^* \frac{\partial n^*}{\partial \alpha} = \frac{f(q)}{q C''} > 0 \quad (129)$$

Since $p - C' < 0, k - F < 0, \frac{\partial n^*}{\partial \alpha} > 0$ and $\frac{\partial Q^*}{\partial \alpha} > 0$, we have:

$$\frac{\partial \Pi^*}{\partial \alpha} = (p - C') \frac{\partial Q^*}{\partial \alpha} + (k - F) \frac{\partial n^*}{\partial \alpha} < 0 \quad (130)$$

When there are two types of patients, following the verification procedures from B2 and denoting $\bar{f} = \beta f_B + (1-\beta)[z + (1-z)f_G]$ and $\bar{f}' = \beta f'_B + (1-\beta)(1-z)f'_G$, we have:

$$\begin{pmatrix} \frac{\partial n^*}{\partial \alpha} \\ \frac{\partial q_B^*}{\partial \alpha} \\ \frac{\partial q_G^*}{\partial \alpha} \end{pmatrix} = \begin{pmatrix} \bar{q}^2 C'' & Q \beta C'' & Q(1-\beta)C'' \\ \bar{q} C'' & n \beta C'' - \alpha f'_B & n(1-\beta)C'' \\ \bar{q} C'' & n \beta C'' & n(1-\beta)C'' - \alpha(1-z)f'_G \end{pmatrix}^{-1} \begin{pmatrix} \bar{f} \\ f'_B \\ (1-z)f'_G \end{pmatrix} \quad (131)$$

hence,

$$\begin{pmatrix} \frac{\partial n^*}{\partial \alpha} \\ \frac{\partial q_B^*}{\partial \alpha} \\ \frac{\partial q_G^*}{\partial \alpha} \end{pmatrix} = \begin{pmatrix} \frac{\bar{f}\alpha^2(1-z)f''_B f''_G - n\alpha C''[\beta(1-z)f''_G + (1-\beta)f''_B](\bar{f} - \bar{q}\bar{f}')}{\alpha^2(1-z)\bar{q}^2 f''_B f''_G C''} \\ \frac{\alpha\bar{q}C''(1-z)f''_G(\bar{f} - \bar{q}\bar{f}')}{\alpha^2(1-z)\bar{q}^2 f''_B f''_G C''} \\ \frac{\alpha\bar{q}C''f''_B(\bar{f} - \bar{q}\bar{f}')}{\alpha^2(1-z)\bar{q}^2 f''_B f''_G C''} \end{pmatrix} \quad (132)$$

Similar to Appendix B2, we have $\frac{\partial n^*}{\partial \alpha} > 0$ and $0 > \frac{\partial q_B^*}{\partial \alpha} \geq \frac{\partial q_G^*}{\partial \alpha}$, thereby verifying the claim.

B4

Denoting $\Delta f = f_B - [z + (1-z)f_G]$ and $\Delta q = q_B - q_G$, the first order condition with respect to β and rearranging, I have:

$$\begin{pmatrix} \frac{\partial n^*}{\partial \beta} \\ \frac{\partial q_B^*}{\partial \beta} \\ \frac{\partial q_G^*}{\partial \beta} \end{pmatrix} = \begin{pmatrix} \bar{q}^2 C'' & Q\beta C'' & Q(1-\beta)C'' \\ \bar{q}C'' & n\beta C'' - \alpha f''_B & n(1-\beta)C'' \\ \bar{q}C'' & n\beta C'' & n(1-\beta)C'' - \alpha(1-z)f''_G \end{pmatrix}^{-1} \begin{pmatrix} \alpha\Delta f + \Delta q(p - C' - QC'') \\ -nC''\Delta q \\ -nC''\Delta q \end{pmatrix} \quad (133)$$

hence,

$$\begin{pmatrix} \frac{\partial n^*}{\partial \beta} \\ \frac{\partial q_B^*}{\partial \beta} \\ \frac{\partial q_G^*}{\partial \beta} \end{pmatrix} = \begin{pmatrix} \frac{\alpha\Delta f + \Delta q(p - C' - QC'')}{\bar{q}^2 C''} - \frac{n[\alpha\Delta f + \Delta q(p - C')][\beta(1-z)f''_G + (1-\beta)f''_B]}{\alpha\bar{q}^2(1-z)f''_B f''_G} \\ \frac{\alpha\Delta f + \Delta q(p - C')}{\alpha\bar{q}f''_B} > 0 \\ \frac{\alpha\Delta f + \Delta q(p - C')}{\alpha\bar{q}(1-z)f''_G} > 0 \end{pmatrix} \quad (134)$$

As $0 > f''_B \geq (1-z)f''_G$, we can prove $\frac{\partial q_B^*}{\partial \beta} \geq \frac{\partial q_G^*}{\partial \beta} > 0$. Moreover,

$$\frac{\partial Q^*}{\partial \beta} = n^* \frac{\partial \bar{q}^*}{\partial \beta} + q^* \frac{\partial n^*}{\partial \beta} + n^* \Delta q^* = \frac{\alpha\Delta f + \Delta q(p - C')}{\bar{q}C''} < 0 \quad (135)$$

and finally,

$$\frac{\partial \Pi^*}{\partial \beta} = (p - C') \frac{\partial Q^*}{\partial \beta} + (k - F) \frac{\partial n^*}{\partial \beta} > 0 \quad (136)$$

B5

For $p \notin \tilde{\mathbf{P}}$, the first order condition of (103) with respect to p is:

$$\begin{pmatrix} -qC''(f - qf') & -f''[q(p - C') + k - F] - nC''(f - qf') \\ q(p - C') + k - F & n(p - C') \end{pmatrix} \begin{pmatrix} \frac{\partial n^*}{\partial p} \\ \frac{\partial q^*}{\partial p} \end{pmatrix} = \begin{pmatrix} -(f - qf') \\ -nq \end{pmatrix} \quad (137)$$

Inverting the matrix, I have:

$$\begin{pmatrix} \frac{\partial n^*}{\partial p} \\ \frac{\partial q^*}{\partial p} \end{pmatrix} = - \begin{pmatrix} -qC''(f - qf') & -f''[q(p - C') + k - F] - nC''(f - qf') \\ q(p - C') + k - F & n(p - C') \end{pmatrix}^{-1} \begin{pmatrix} -(f - qf') \\ -nq \end{pmatrix} \quad (138)$$

Finally, applying the implicit function theorem and the curvature assumptions, I will get:

$$\begin{pmatrix} \frac{\partial n^*}{\partial p} \\ \frac{\partial q^*}{\partial p} \end{pmatrix} = \begin{pmatrix} \frac{-n(f - qf')(QC'' + p - C') - Qf''[q(p - C') + k - F]}{f''[q(p - C') + k - F]^2 + nC''(k - F)(f - qf')} > 0 \\ \frac{(f - qf')[q(QC'' + p - C') + k - F]}{f''[q(p - C') + k - F]^2 + nC''(k - F)(f - qf')} < 0 \end{pmatrix} \quad (139)$$

hence,

$$\frac{\partial Q^*}{\partial p} = n^* \frac{\partial q^*}{\partial p} + q^* \frac{\partial n^*}{\partial p} = \frac{n(k - F)(f - qf') - nq^2 f''[q(p - C') + k - F]}{f''[q(p - C') + k - F]^2 + nC''(k - F)(f - qf')} > 0 \quad (140)$$

Moreover,

$$\frac{\partial n^*}{\partial p}(p_0^-) - \frac{\partial n^*}{\partial p}(p_0^+) = \frac{(Q^2 C'' - QC' + C)\{-nC''(f - qf') - [q(p - C') + k - F]f''\}}{QC''\{f''[q(p - C') + k - F]^2 + n(k - F)C''(f - qf')\}} > 0 \quad (141)$$

and finally,

$$\frac{\partial Q^*}{\partial p}(p_0^-) - \frac{\partial Q^*}{\partial p}(p_0^+) = \frac{-f''[q(p - C') + k - F](Q^2 C'' - QC' + C)}{nC''f''[q(p - C') + k - F]^2 + n^2(k - F)(C'')^2(f - qf')} > 0 \quad (142)$$

The above proof can be easily extended to the case when $q = (q_B, q_G)$. In this case, the FOC condition of the physician will be:

$$\begin{cases} (p - C') \left[\beta(f_B - q_B f_B') + (1 - \beta) \left(z + (1 - z)(f_G - q_G f_G') \right) \right] - (k - F)f_B' = 0 \\ f_B' = (1 - z)f_G' \\ pQ - C(Q) + n(k - F) = 0 \end{cases} \quad (143)$$

Denote $\bar{f} = \beta f_B + (1 - z)(z + (1 - z)f_G)$ and $\bar{f}' = \beta f'_B + (1 - \beta)(1 - z)f'_G$.

Calculating the FOC condition with respect to p and rearrange, we have $\begin{pmatrix} \frac{\partial n^*}{\partial p} \\ \frac{\partial q_B^*}{\partial p} \\ \frac{\partial q_G^*}{\partial p} \end{pmatrix} =$

$$\begin{pmatrix} \bar{q}C''(\bar{f} - \bar{q}\bar{f}') & \beta[nC''(\bar{f} - \bar{q}\bar{f}') + f_B''(q_B(p - C') + k - F)] & (1 - \beta)[nC''(\bar{f} - \bar{q}\bar{f}') + (1 - z)f_G''(q_G(p - C') + k - F)] \\ 0 & f_B'' & -(1 - z)f_G'' \\ \bar{q}(p - C') + k - F & n\beta(p - C') & n(1 - \beta)(p - C') \end{pmatrix}^{-1} \begin{pmatrix} \bar{f} - \bar{q}\bar{f}' \\ 0 \\ -Q \end{pmatrix}$$

The above formula is symmetric to the equation (138) and can be solved by following the same process.

B6

Similarly, calculating the first order condition of (103) with respect to k yields:

$$\begin{pmatrix} -qC''(f - qf') & -f''[q(p - C') + k - F] - nC''(f - qf') \\ q(p - C') + k - F & n(p - C') \end{pmatrix} \begin{pmatrix} \frac{\partial n^*}{\partial k} \\ \frac{\partial q^*}{\partial k} \end{pmatrix} = \begin{pmatrix} f' \\ -n \end{pmatrix} \quad (144)$$

Inverting the matrix yields:

$$\begin{pmatrix} \frac{\partial n^*}{\partial k} \\ \frac{\partial q^*}{\partial k} \end{pmatrix} = - \begin{pmatrix} -qC''(f - qf') & -f''[q(p - C') + k - F] - nC''(f - qf') \\ q(p - C') + k - F & n(p - C') \end{pmatrix}^{-1} \begin{pmatrix} f' \\ -n \end{pmatrix} \quad (145)$$

Using the implicit function theorem and curvature assumptions, I obtain:

$$\begin{pmatrix} \frac{\partial n^*}{\partial k} \\ \frac{\partial q^*}{\partial k} \end{pmatrix} = \begin{pmatrix} \frac{n(p - C')(f' - qf'') - nf''(k - F) - n^2C''(f - qf')}{f''[q(p - C') + k - F]^2 + nC''(k - F)(f - qf')} > 0 \\ \frac{QC''(f - qf') - f'[q(p - C') + (k - F)]}{f''[q(p - C') + k - F]^2 + nC''(k - F)(f - qf')} < 0 \end{pmatrix} \quad (146)$$

hence,

$$\frac{\partial Q^*}{\partial k} = n^* \frac{\partial q^*}{\partial p} + q^* \frac{\partial n^*}{\partial p} = \frac{-nq^2(p - C')f'' - n(k - F)(f' + qf'')}{f''[q(p - C') + k - F]^2 + nC''(f - qf')(k - F)} > 0 \quad (147)$$

B7

Consider the production vector for services which solves the system (96). This solution can then be used to define the ensuing total production of services denoted by $\hat{Q}(p, k)$.⁷⁵ The critical price p_0 is implicitly a function of α which is determined by solving:

$$p_0 \hat{Q}(\alpha, p, k) - C(\hat{Q}(\alpha, p, k)) - (k - F)\hat{n}(\alpha, p, k) = 0 \quad (148)$$

Applying the implicit function theorem, I have:

$$\frac{\partial p_0}{\partial \alpha} = \frac{(C' - p_0) \frac{\partial \hat{Q}}{\partial \alpha} - (k - F) \frac{\partial \hat{n}}{\partial \alpha}}{\hat{Q} + (p - C') \frac{\partial \hat{Q}}{\partial p} + (k - F) \frac{\partial \hat{n}}{\partial p}} \quad (149)$$

Following the same steps as in Appendices B1 and B3, it is immediate that $\frac{\partial \hat{Q}}{\partial p} > 0$ and $\frac{\partial \hat{Q}}{\partial \alpha} > 0$. Moreover, the system (96) implies that $C' - p_0 > 0$. Following the same procedures from (100) to (102), it can be proved that $\hat{Q} + (p - C') \frac{\partial \hat{Q}}{\partial p} + (k - F) \frac{\partial \hat{n}}{\partial p} > 0$, thus verifying the claim.

⁷⁵ Slightly abusing notation, observe that $\hat{Q}(p, k) = Q^*(p, k)$ iff $p \in \tilde{P}$.

Chapter 5. Conclusion and Further Research

5.1 Conclusions

This thesis has investigated the optimal provision of a primary care incentive within a health insurance scheme when medical service providers are motivated by patients' health benefits and have private knowledge about their degree of altruism as well as their patients' severity of illness. The main objective has been to provide some new insights which improve our understanding of, respectively, the implementation of cost-containing instruments, the optimal design of incentive contracts, and their impact on PCPs' choices of medical service supply and patient enrolment.

In this chapter, I review the main contributions of each of the three main chapters of this thesis and synthesise the central themes of these research projects. I also discuss the limitations of both the research contained in this thesis and the broader research area. Finally, I briefly summarize several avenues for future research that have been discussed in the previous chapters.

The purpose and contribution of Chapter 2 was to provide an authoritative account of the recently developed literatures of the optimal design of primary care physicians' incentive contracting. To do that, I define and introduce stylised models of physician altruism, which is an essential feature of primary healthcare providers (Arrow, 1963). More importantly, I discuss the optimal design of physician payment for altruistic primary care providers in the presence of moral hazard or adverse selection. Chapter 2 highlights several inconsistencies between the assumptions and predictions of theoretical models, and the results reported in experimental and empirical studies.

I show that the earlier literature focuses mainly on deriving the optimal prospective and cost reimbursement system that induces physicians' "first best" cost-reduction or quality enhancing efforts. These studies captured the moral hazard problem as physicians' non-contractible efforts. I show that studies in this literature do not discuss the alternative approach of modelling moral hazard. Furthermore, the implementation of cost-reimbursement or cost-sharing might be impossible or very costly. As a result, the optimal design of more sophisticated physician payment systems such as fee-for-service or capitation should be examined.

The more recent literature has extended the discussions by further considering physicians' or patients' unknown heterogeneities and showed that the optimal PCP compensation system should offer a menu of contracts. However, this prediction is inconsistent with both theories and real-world applications as using a menu of contracts requires the satisfaction of "single-crossing property" (often difficult to be achieved) and may create a two-tier healthcare system. Moreover, when physicians are offered a menu of contracts and make their choices, the regulator can distinguish each physician's type from the choice he made. As a result, there is no guarantee that the regulator would not exploit information regarding the different types of health service provider in the design of physician payment. Lastly, offering physicians a menu of contracts may incentivise them to cream-skim low-cost patients.

The above difficulties indicate that using a single physician payment could be a better alternative (Allard et al., 2014; Choné and Ma, 2011; Liu and Ma, 2013; Makris and Siciliani, 2013), though not all physicians' incentives could be aligned and some physicians might have informational rent. To align these incentives, studies have investigated introducing a single physician remuneration scheme, a cost control policy such as quantity rationing, or an expenditure cap. However, this literature only includes a few studies with models that can be easily generalized in many ways. First, the analytical frameworks do not consider essential features of the healthcare market; physicians have been assumed to be profit-driven while facing no financial constraints. The models also do not consider information asymmetries between the regulator, patients and physicians. Second, the types of physician and patient involved are assumed to be homogenous. Third, healthcare providers are not constrained by non-negative profit conditions. Finally, these models focus on modelling one particular type of quotas and expenditure cap at a global level.

Section 2.4 discusses empirical studies relating to the aforementioned topics. I have shown that the assumption of physician altruism is well supported by a variety of empirical evidence. Accordingly, altruism is an essential feature in the modelling of physician behaviour. I have also shown that the single physician payment system has been used more frequently than a menu of contracts. Under the commonly used FFS or CAP, I have shown that the theoretical predication that FFS (CAP) drives overproduction (underproduction) and price increase drives service quality do not always consistent with empirical observations. Moreover, I have presented the

respective real-world applications of capitation, quantity rationing as well as revenue restriction. I have found that imposing these policies might help in limiting health spending growth and enhancing resource allocation efficiency, though these effects cannot be always guaranteed.

Research in this thesis further develops the analytical frameworks analysing physicians' decisions on the supply of service and patient enrolment under different types of cost control policies. These frameworks generate predictions of physicians' health service supply decisions which are consistent with different types of empirical observations. Moreover, results from my studies reflect some of the most important impacts of imposing the respective cost control policies in the healthcare market.

Chapter 3 contributes to the literature by providing a stylised model of primary care physician incentive contracting, which focuses on comparing the impact of introducing different cost control policies in the presence of information asymmetry. To do this, I have made a few adjustments to the stylised model discussed in the literature. Most importantly, I assume the physician can choose per patient quantity of service provided only and capture the moral hazard problem by introducing a random element associated with patient benefits after treatment. This prevents the regulator inferring the physician or patient type from the quantity of service provided and therefore the physician's utility. Moreover, I assume the provided medical service intensities can be contracted upon, as physicians are paid on a FFS basis. Finally, I consider the design of payment under a publicly funded health insurance scheme. The objective function of the regulator therefore not only includes patients' benefits but also providers' net financial surpluses. Overall, the analysis based on the model sharpens our understanding of how cost control policies affect physicians' decisions on health service supply and social well-being.

My model demonstrates that the asymmetric information, especially the unknown heterogeneities of physicians—prevents the single FFS system achieving resource allocation efficiency. Moreover, current trends such as population ageing and technological innovations drive up price and health expenditure, exacerbating the inefficiencies. My numerical findings show that quantity rationing, revenue cap and capitation can be introduced to enhance efficiency, as these instruments directly restrict highly motivated physicians' over-production while price can be raised to reduce less motivated physicians' underproduction. In particular, imposing a negative capitation on

FFS improves social well-being the most as it exploits the knowledge of doctors, extracts PCPs' informational rents and aligns heterogeneous PCPs' incentives. These findings are shown to be consistent with the results derived from contract theory and the empirical observations discussed in Chapter 2, although the intuitions are different.

Chapter 4 further develops the previous contribution by allowing physicians to select the size of their practice while introducing a fixed cost associated with enrolling each patient. This generalization, however, requires that the payment system to be designed to ensure the whole population has access to healthcare. The model developed in Chapter 4 allows us to analyse to what extent the physician willing to trade-off the size of his practice and service quality provided per patient, especially when the physician's profit is binding. In consistent with Chapter 3, the model shows that the FFS system fail to align heterogeneous doctors' incentives because of moral hazard and adverse selection; the more altruistic physicians select too many patients but offer too low a level of quality whereas the less altruistic physicians provide excessive level of quality but enrol too few patients. In contrast to Chapter 3 and most studies in the literature, I show the physician does not increase health service intensity/quality provided per patient with a fee-for-service price. This result is consistent with some empirical observations that the physician does not respond to price variations (Madden et al., 2005; Yip, 1998; Zuckerman et al., 1998).

The model also shows that, to ensure the whole population has access to healthcare, the fee-for-service price has to be set sufficiently high, which inevitably leads to excessive provision of health services. Similar to Chapter 3, my numerical findings suggest that quantity rationing, expenditure cap and capitation can still be imposed on a FFS system to enhance efficiency. This is because cost-control instruments force the low altruistic providers to reduce service quality provided per patient but increases the number of patients treated. Meanwhile, by adjusting price downward, high altruistic providers are incentivised to reduce their practice size and to improve medical service quality. In particular, imposing a capitation less than fixed costs or a revenue cap per patient in a FFS system leads to a higher welfare improvement than using quantity rationing. Overall, capitation is considered as the best remuneration mechanism as it is more tolerant of potential mistakes that could be made by the regulator as compared to the revenue cap. The result capitation is superior to other cost control instruments is consistent with the

result I obtained in Chapter 3, reflecting that my previous findings can be further generalized in the current framework with less restrictions.

5.2 Future Studies

To conclude the thesis, I will briefly reflect on some promising directions for future research. The important role that professional ethics play in guiding physicians' behaviour and different types of efficiencies caused by information asymmetries have incentivised economists to extend the classical model to capture these essential features of the healthcare market (Arrow, 1963; Ellis and McGuire, 1986). These models have been applied to the study of physician incentive contracting, thereby improving the theory's predictive capacity.

Chapters 3 and 4 have discussed several possible extensions to my analysis. First, it would be interesting to see whether my current findings can be further generalized. For instance, it should be considered whether introducing more than two types of physicians or patients and multi-dimensional adverse selection or if allowing physicians to choose the types of patients to be treated will change my findings. Second, it is worth investigating how different cost control policies affect PCPs' decisions on the level of cost-reduction, quality enhancing efforts to be provided, as well as their specialised care referral decisions. Third, my thesis focuses mainly on comparing the impact of imposing three specific cost control policies on the FFS system. As a result, physicians' behaviour under other popular payment systems such as salary, menu of contracts and other cost control policies such as global budgets or pay-for-performance have not been discussed. Fourth, it would also be interesting to see how physicians' risk averse preferences would alter their medical service supply and patient enrolment decisions. Finally, patients in reality also play a role in deciding the supply of health service. Accordingly, investigating the optimal design of patients' insurance scheme and how it affects their demand of service along with the optimal design of physician payment mechanism while how it affects physicians' supply of health service better reflects what happens in reality.

This thesis also offers some directions for empirical studies. First, it would be interesting to see whether my findings hold up to empirical scrutiny. This includes testing the significance of physicians' supply and patient enrolment decisions as well as the optimal design of physician payment systems in response to variations of price, the respective

cost control policy, and exogenously given parameters (e.g. the fraction of less healthy individuals and technology factors etc.). Second, econometricians can estimate the percentage that ageing or technology innovation contributes to price, insurance premium or health expenditure growth respectively. Moreover, variations of fee-for-service price, the level of insurance premium and health expenditure over the next few years can also be forecasted. Finally, laboratory experiments can be performed to test physicians' (i.e. as the subjects of experiments) decisions on patients' enrolment and intensity of service provided under different cost control policies as well as welfare generated by using respective cost control policies as discussed in Chapter 4.

Based on theoretical and empirical extensions of this thesis, researchers can also recommend potential policy adjustments in the future. For instance, they can provide qualitative recommendations such as (1) whether cost-containment methods should be introduced, and (2) which specific cost control policies should be used. Researchers can also offer quantitative policy recommendations. Specifically, they can recommend the specific level of quota, budget cap or capitation that the regulator should impose. Furthermore, they can also estimate the potential costs and benefits of using a particular payment system and cost control policy.

References

- Allaby, M. A. (2003) ‘Doctors for the poor in urban Nepal’, *Tropical Doctor*, 33(2), 83-85.
- Allard, M., Jelovac, I. and Léger, P.-T. (2014) ‘Payment mechanism and GP self-selection: Capitation versus fee for service’, *International Journal of Health Care Finance and Economics*, 14(2), 143-160.
- Allard, M., Jelovac, I. and Léger, P. T. (2011) ‘Treatment and referral decisions under different physician payment mechanisms’, *Journal of Health Economics*, 30(5), 880-893.
- Anderson, G., Chalkidou, K. and Herring, B. (2012) ‘High US health-care spending and the importance of provider payment rates’, *Forum for Health Economics and Policy*, 15(3), 1-22.
- Andoh-Adjei, F.-X., Spaan, E., Asante, F. A., Mensah, S. A. and van der Velden, K. (2016) ‘A narrative synthesis of illustrative evidence on effects of capitation payment for primary care: Lessons for Ghana and other low/middle-income countries’, *Ghana Medical Journal*, 50(4), 207-219.
- Arrieta, A., García-Prado, A., González, P. and Pinto-Prades, J. L. (2017) ‘Risk attitudes in medical decisions for others: An experimental approach’, *Journal of Health Economics*, 26(3), 97-113.
- Arrow, K. J. (1963) ‘Uncertainty and the welfare economics of medical care’, *The American Economic Review*, 53(5), 941-973.
- Awad, A. and Fayek, A. R. (2012) ‘Contractor default prediction model for surety bonding’, *Canadian Journal of Civil Engineering*, 39(9), 1027-1042.
- Baker, G. (2002) ‘Distortion and risk in optimal incentive contracts’, *Journal of Human Resources*, 37(4), 728-751.
- Baker, G., Gibbons, R. and Murphy, K. J. (1994) ‘Subjective performance measures in optimal incentive contracts’, *The Quarterly Journal of Economics*, 109(4), 1125-1156.

- Bali, A. S. and Ramesh, M. (2017) ‘Designing effective healthcare: Matching policy tools to problems in China’, *Public Administration and Development*, 37(1), 40-50.
- Barham, V. and Milliken, O. (2015) ‘Payment mechanisms and the composition of physician practices: balancing cost - containment, access, and quality of care’, *Journal of Health Economics*, 24(7), 895-906.
- Barigozzi, F. and Turati, G. (2012) ‘Human health care and selection effects: Understanding labor supply in the market for nursing’, *Health Economics*, 21(4), 477-483.
- Baron, D. P. and Myerson, R. B. (1982) ‘Regulating a monopolist with unknown costs’, *Econometrica: Journal of the Econometric Society*, 50(5), 911-930.
- Barros, P. P. (2003) ‘Cream-skimming, incentives for efficiency and payment system’, *Journal of Health Economics*, 22(3), 419-443.
- Batifoulier, P. and Da Silva, N. (2014) ‘Medical altruism in mainstream health economics: theoretical and political paradoxes’, *Review of Social Economy*, 72(3), 261-279.
- Bauchner, H. and Fontanarosa, P. B. (2018) ‘Health care spending in the United States compared with 10 other high-income countries: What Uwe Reinhardt might have said’, *Journal of the American Medical Association*, 319(10), 990-992.
- BBC (2018) ‘What are opioids and what are the risks?’ [online], available: <https://www.bbc.co.uk/news/health-43462975> [Accessed 21/03 2019].
- Benstetter, F. and Wambach, A. (2006) ‘The treadmill effect in a fixed budget system’, *Journal of Health Economics*, 25(1), 146-169.
- Bickerdyke, I., Dolamore, R., Monday, I. and Preston, R. (2002) *Supplier-induced demand for medical services*, Canberra: Productivity Commission Press.
- Bloom, G., Standing, H. and Lloyd, R. (2008) ‘Markets, information asymmetry and health care: Towards new social contracts’, *Social Science & Medicine*, 66(10), 2076-2087.

- Brown, L. (2016) ‘What does it take to become a doctor?’ [online], available: <http://www.bbc.co.uk/newsbeat/article/37550759/what-does-it-take-to-become-a-doctor> [Accessed 22/03 2019].
- Bories, P., Lamy, S., Simand, C., Bertoli, S., Delpierre, C., Malak, S., Fornecker, L., Moreau, S., Récher, C. and Nebout, A. (2018) ‘Physician uncertainty aversion impacts medical decision making for older patients with acute myeloid leukemia: Results of a national survey’, *Journal of The Ferrata Storti Foundation*, 103(12), 2040-2048.
- Borland, S. (2011) ‘GPs ordered to ration cancer scans: Lives “being put at risk” by bureaucrats’ new cost-saving directive’, *Daily Mail* [online], available: <https://www.dailymail.co.uk/health/article-2034914/GPs-told-ration-cancer-scans-bureaucratic-directive.html>.
- Bowles, S. (2009) *Microeconomics: Behavior, institutions, and evolution*, Princeton: Princeton University Press.
- British Medical Association (2018) ‘The GP practice prescribing budget’, [online], available: <https://www.bma.org.uk/advice/employment/gp-practices/service-provision/prescribing/advice-for-dispensing-gps/the-gp-practice/the-gp-practice-prescribing-budget> [Accessed 07/12 2018].
- Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N. and Wiesen, D. (2016) ‘Using artefactual field and lab experiments to investigate how fee-for-service and capitation affect medical service provision’, *Journal of Economic Behavior & Organization*, 131(Part B), 17-23.
- Brosig-Koch, J., Hennig - Schmidt, H., Kairies - Schwarz, N. and Wiesen, D. (2017) ‘The effects of introducing mixed payment systems for physicians: Experimental evidence’, *Journal of Health Economics*, 26(2), 243-262.
- Buchanan, J. M. (1988) ‘Reviewed work: *The new Palgrave dictionary of economics*’, *Public Choice*, 52(3), 291-293.
- Burani, N. and Palestini, A. (2016) ‘What determines volunteer work? On the effects of adverse selection and intrinsic motivation’, *Economics Letters*, 144, 29-32.

- Busse, R., Schreyögg, J. and Henke, K. D. (2005) ‘Regulation of pharmaceutical markets in Germany: Improving efficiency and controlling expenditures?’, *The International Journal of Health Planning and Management*, 20(4), 329-349.
- Carlsen, F. and Grytten, J. (1998) ‘More physicians: Improved availability or induced demand?’, *Journal of Health Economics*, 7(6), 495-508.
- Carrin, G. (2003) ‘Provider payments and patient charges as policy tools for cost-containment: How successful are they in high-income countries’, *Human Resources for Health*, 1(6), 1-10.
- Chalkley, M. and Malcomson, J. M. (1998) ‘Contracting for health services when patient demand does not reflect quality’, *Journal of Health Economics*, 17(1), 1-19.
- Chalkley, M. and Malcomson, J. M. (1998) ‘Contracting for Health Services with Unmonitored Quality’, *The Economic Journal*, 108(449), 1093-1110.
- Chalkley, M. and Malcomson, J. M. (2002) ‘Cost sharing in health service provision: An empirical assessment of cost savings’, *Journal of Public Economics*, 84(2), 219-249.
- Chen, A. and Goldman, D. (2016) ‘Health care spending: Historical trends and new directions’, *Annual Review of Economics*, 8(1), 291-319.
- Choné, P. and Ma, C.-t. A. (2011) ‘Optimal health care contract under physician agency’, *Annals of Economics and Statistics*, 101, 229-256.
- Chua, K.-P., Brummett, C. M. and Waljee, J. F. (2019) ‘Opioid prescribing limits for acute pain: Potential problems with design and implementation’, *Journal of the American Medical Association*, 321(7), 643-644.
- Clemens, J. and Gottlieb, J. D. (2014) ‘Do physicians’ financial incentives affect medical treatment and patient health?’, *American Economic Review*, 104(4), 1320-49.
- Colombo, F. and Tapay, N. (2004) *Private health insurance in OECD countries: The benefits and costs for individuals and health systems*, OECD Health Working Papers, No. 15, Paris: OECD.
- Cookson, R. (2013) ‘Can the NICE “end-of-life premium” be given a coherent ethical justification?’, *Journal of Health Politics, Policy and Law*, 38(6), 1129-1148.

- Crea, G., Galizzi, M. M., Linnosmaa, I. and Miraldo, M. (2019) ‘Physician altruism and ex-post moral hazard: (No) evidence from Finnish national prescriptions data’, *Journal of Health Economics*, 65, 153-169.
- Croxson, B., Propper, C. and Perkins, A. (2001) ‘Do doctors respond to financial incentives? UK family doctors and the GP fundholder scheme’, *Journal of Public Economics*, 79(2), 375-398.
- Culyer, A. J. (1989) ‘The normative economics of health care finance and provision’, *Oxford Review of Economic Policy*, 5(1), 34-58.
- Cunningham, S. (2011) *Understanding market failures in an economic development context*, Pretoria: Mesopartner.
- Cutler, D. M. and Zeckhauser, R. J. (2000) ‘The anatomy of health insurance’. In A. J. Culyer, and J. P. Newhouse (eds.), *Handbook of health economics*, Dordrecht: Elsevier, 563-643.
- Davidson, S. M., Manheim, L. M., Hohlen, M. M., Werner, S. M., Yudkowsky, B. K. and Fleming, G. V. (1992) ‘Prepayment with office-based physicians in publicly funded programs: Results from the children's Medicaid program’, *Pediatrics*, 89(4), 761-767.
- De Costa, A., Kazmi, T., Lönnroth, K., Uplekar, M. and Diwan, V. (2008) ‘PPM: “Public-private” or “private-public” mix? The case of Ujjain District, India’, *The International Journal of Tuberculosis and Lung Disease*, 12(11), 1333-1335.
- De Fraja, G. (2000) ‘Contracts for health care and asymmetric information’, *Journal of Health Economics*, 19(5), 663-677.
- Delfgaauw, J. and Dur, R. (2007a) ‘Incentives and workers' motivation in the public sector’, *The Economic Journal*, 118(525), 171-191.
- Delfgaauw, J. and Dur, R. (2007b) ‘Signaling and screening of workers' motivation’, *Journal of Economic Behavior & Organization*, 62(4), 605-624.
- Demougin, D. (1989) ‘A renegotiation-proof mechanism for a principal-agent model with moral hazard and adverse selection’, *The Rand Journal of Economics*, 20(2), 256-267.

- Demougin, D. and Fluet, C. (1998) ‘Mechanism sufficient statistic in the risk-neutral agency problem’, *Journal of Institutional and Theoretical Economics*, 154(4), 622-639.
- Demougin, D. and Fluet, C. (2001) ‘Monitoring versus incentives’, *European Economic Review*, 45(9), 1741-1764.
- Desquins, B., Holly, A. and Rochaix, L. (2007) ‘Agent model with a monopoly power: physicians’, *Anglais Working Paper*, 7(28), 1-20.
- Devlin, R. A. and Sarma, S. (2008) ‘Do physician remuneration schemes matter? The case of Canadian family physicians’, *Journal of Health Economics*, 27(5), 1168-1181.
- Dieleman, J. L., Templin, T., Sadat, N., Reidy, P., Chapin, A., Foreman, K., Haakenstad, A., Evans, T., Murray, C. J. and Kurowski, C. (2016) ‘National spending on health by source for 184 countries between 2013 and 2040’, *The Lancet*, 387(10037), 2521-2535.
- Dumont, E., Fortin, B., Jacquemet, N. and Shearer, B. (2008) ‘Physicians’ multitasking and incentives: Empirical evidence from a natural experiment’, *Journal of Health Economics*, 27(6), 1436-1450.
- Durairaj, V. and Evans, D. B. (2010) *Fiscal space for health in resource-poor countries*, Geneva: World Health Organization.
- Eeckhoudt, L., Lebrun, T. and Sainly, J. C. (1985) ‘Risk-aversion and physicians’ medical decision-making’, *Journal of Health Economics*, 4(3), 273-281.
- Egginton, K. (2000) ‘Risk selection and optimal health insurance-provider payment systems’, *The Journal of Risk and Insurance*, 67(2), 173-196.
- Ehrbeck, T., Henke, N. and Kibasi, T. (2010) *The emerging market in health care innovation*, Washington, DC: McKinsey & Co.
- Ellis, R. P. (1998) ‘Creaming, skimping and dumping: Provider competition on the intensive and extensive margins’, *Journal of Health Economics*, 17(5), 537-555.

- Ellis, R. P. and McGuire, T. G. (1986) 'Provider behaviour under prospective reimbursement: Cost sharing and supply', *Journal of Health Economics*, 5(2), 129-151.
- Ellis, R. P. and McGuire, T. G. (1990) 'Optimal payment systems for health services', *Journal of Health Economics*, 9(4), 375-396.
- Emery, J. C. H., Auld, C. J. and Lu, M. (1999) *Paying for physician services in Canada: The institutional, historical and policy contexts*, Institute of Health Economics Working Paper, Calgary, Alberta: University of Calgary.
- Fan, C.-P., Chen, K.-P. and Kan, K. (1998) 'The design of payment systems for physicians under global budget—an experimental study', *Journal of Economic Behavior & Organization*, 34(2), 295-311.
- Farley, P. J. (1986) 'Theories of the price and quantity of physician services: A synthesis and critique', *Journal of Health Economics*, 5(4), 315-333.
- Fischer, K. E., Koch, T., Kostev, K. and Stargardt, T. (2018) 'The impact of physician-level drug budgets on prescribing behavior', *The European Journal of Health Economics*, 19(2), 213-222.
- French, E. and Kelly, E. (2016) 'Medical spending around the developed world', *Fiscal Studies*, 37(3-4), 327-344.
- Galizzi, M. M., Miraldo, M. and Stavropoulou, C. (2013) *Doctor-patient differences in risk preferences, and their links to decision-making: A field experiment*, London: Imperial College London Business School.
- Galizzi, M. M., Tammi, T., Godager, G., Linnosmaa, I. and Wiesen, D. (2015) 'Provider altruism in health economics', *Työpaperi*, 4, 1-28.
- Gaynor, M. and Gertler, P. (1995) 'Moral hazard and risk spreading in partnerships', *The RAND Journal of Economics*, 26(4), 591-613.
- Gaynor, M., Mehta, N. and Richards-Shubik, S. (2018) *Optimal contracting with altruistic agents: A structural model of medicare reimbursements for dialysis drugs*, London, Ontario: University of Western Ontario.

- Ghali, W. (2016) *Physicians as stewards of resources roles, responsibilities, and remuneration*, Calgary, Alberta: University of Calgary.
- Gillon, R. (1994) 'Medical ethics: Four principles plus attention to scope', *British Medical Journal*, 309(6948), 184-187.
- Ginsburg, P. B. (2015) 'Should the U.S. move away from fee-for-service medicine?', *The Wall Street Journal*, 5(2), 1-10.
- Glannon, W. and Ross, L. F. (2002) 'Are doctors altruistic?', *Journal of Medical Ethics*, 28(2), 68-69.
- Glaser, W. A. (1993) 'How expenditure caps and expenditure targets really work', *The Milbank Quarterly*, 71(1), 97-127.
- Godager, G., Hennig-Schmidt, H. and Iversen, T. (2016) 'Does performance disclosure influence physicians' medical decisions? An experimental study', *Journal of Economic Behaviour & Organization*, 131, 36-46.
- Godager, G. and Wiesen, D. (2013) 'Profit or patients' health benefit? Exploring the heterogeneity in physician altruism', *Journal of Health Economics*, 32(6), 1105-1116.
- Gosden, T., Forland, F., Kristiansen, I. S., Sutton, M., Leese, B., Giuffrida, A., Sergison, M. and Pedersen, L. (2000) 'Capitation, salary, fee-for-service and mixed systems of payment: Effects on the behaviour of primary care physicians', *Cochrane Database Syst Rev*, 3(3), 1-25.
- Gosden, T., Forland, F., Kristiansen, I. S., Sutton, M., Leese, B., Giuffrida, A., Sergison, M. and Pedersen, L. (2001) 'Impact of payment method on behaviour of primary care physicians: A systematic review', *Journal of Health Services Research & Policy*, 6(1), 44-55.
- Gosden, T., Pedersen, L. and Torgerson, D. (1999) 'How should we pay doctors? A systematic review of salary payments and their effect on doctor behaviour', *Quarterly Journal of Medicine*, 92(1), 47-55.
- Green, E. P. (2014) 'Payment systems in the healthcare industry: An experimental study of physician incentives', *Journal of Economic Behavior & Organization*, 106(1), 367-378.

- Greenwald, B. C. and Stiglitz, J. E. (1986) ‘Externalities in economies with imperfect information and incomplete markets’, *The Quarterly Journal of Economics*, 101(2), 229-264.
- Grytten, J., Carlsen, F. and Skau, I. (2001) ‘The income effect and supplier induced demand. Evidence from primary physician services in Norway’, *Applied Economics*, 33(11), 1455-1467.
- Grytten, J., Carlsen, F. and Skau, I. (2008) ‘Primary physicians’ response to changes in fees’, *The European Journal of Health Economics*, 9(2), 117-125.
- Grytten, J. and Sørensen, R. (2001) ‘Type of contract and supplier-induced demand for primary physicians in Norway’, *Journal of Health Economics*, 20(3), 379-393.
- Gutiérrez, J. L. G., Puente, C. P., Rodríguez, R. M., López, A. L. and Furlong, L. V. (2006) ‘Nursing motives for helping scale (N-MHS): Reliability and validity’, *The Spanish Journal of Psychology*, 9(1), 103-112.
- Hammond, P. (1987) ‘Altruism’. In Eatwell, J., Milgate, M. and Newman, P. (eds.), *The new Palgrave dictionary of economics*, London: Macmillan Press, pp. 85-87.
- Harris, J. (2018) ‘Altruism: Should it be included as an attribute of medical professionalism?’, *Health Professions Education*, 4(1), 3-8.
- Harsanyi, J. C. (1955) ‘Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility’, *Journal of Political Economy*, 63(4), 309-321.
- Hart, O. and Moore, J. (1988) ‘Incomplete contracts and renegotiation’, *Econometrica*, 56(4), 755-785.
- Hart, O. D. and Holmström, B. (1986) ‘The theory of contracts’, In *Advances in Economic Theory*, Cambridge: Cambridge University Press, pp. 1-132.
- Hawe, E. (2008) *Compendium of Health Statistics 2009*, Abingdon: Radcliffe Publishing.
- Hellerstein, J. K. (1998) ‘The importance of the physician in the generic versus trade-name prescription decision’, *The Rand Journal of Economics*, 29(1), 108-136.
- Hemenway, D., Killen, A., Cashman, S. B., Parks, C. L. and Bicknell, W. J. (1990) ‘Physicians’ responses to financial incentives: Evidence from a for-profit ambulatory care center’, *New England Journal of Medicine*, 322(15), 1059-1063.

- Hennig-Schmidt, H., Selten, R. and Wiesen, D. (2011) ‘How payment systems affect physicians’ provision behaviour: An experimental investigation’, *Journal of Health Economics*, 30(4), 637-646.
- Hennig-Schmidt, H. and Wiesen, D. (2014) ‘Other-regarding behaviour and motivation in health care provision: An experiment with medical and non-medical students’, *Social Science & Medicine*, 108, 156-165.
- Hermalin, B. E. and Katz, M. L. (2009) ‘Information and the hold-up problem’, *The RAND Journal of Economics*, 40(3), 405-423.
- Heyes, A. (2005) ‘The economics of vocation or “why is a badly paid nurse a good nurse”?’’, *Journal of Health Economics*, 24(3), 561-569.
- Hillman, A. L., Pauly, M. V. and Kerstein, J. J. (1989) ‘How do financial incentives affect physicians’ clinical decisions and the financial performance of health maintenance organizations?’, *New England Journal of Medicine*, 321(2), 86-92.
- Holmstrom, B. (1979) ‘Moral hazard and observability’, *Bell Journal of Economics*, 10(1), 74-91.
- Huber, M. and Orosz, E. (2003) ‘Health expenditure trends in OECD countries, 1990-2001’, *Health Care Financing Review*, 25(1), 1-22.
- Iacobucci, G. (2017) ‘GPs call for end to “local rationing” of prescribing’, *The British Medical Journal*, 359, 1.
- Iversen, T. and Lurås, H. (2000) ‘The effect of capitation on GPs’ referral decisions’, *Journal of Health Economics*, 9(3), 199-210.
- Jack, W. (2005) ‘Purchasing health care services from providers with unknown altruism’, *Journal of Health Economics*, 24(1), 73-93.
- Jacobsen, K. J., Eika, K. H., Helland, L., Lind, J. T. and Nyborg, K. (2011) ‘Are nurses more altruistic than real estate brokers?’, *Journal of Economic Psychology*, 32(5), 818-831.
- Jegers, M., Kesteloot, K., De Graeve, D. and Gilles, W. (2002) ‘A typology for provider payment systems in health care’, *Health Policy*, 60(3), 255-273.

- Jelovac, I. and Kembou Nzale, S. (2017) 'Regulation and altruism', *Journal of Public Economic Theory*, 22(1), 49-68.
- Jones, M. R., Viswanath, O., Peck, J., Kaye, A. D., Gill, J. S. and Simopoulos, T. T. (2018) 'A brief history of the opioid epidemic and strategies for pain medicine', *Journal of Pain and Therapy*, 7(1), 13-21.
- Kaarboe, O. and Siciliani, L. (2011) 'Multi-tasking, quality and pay for performance', *Journal of Health Economics*, 20(2), 225-238.
- Kantarevic, J. and Kralj, B. (2016) 'Physician payment contracts in the presence of moral hazard and adverse selection: the theory and its application in Ontario', *Journal of Health Economics*, 25(10), 1326-1340.
- Kantarevic, J., Kralj, B. and Weinkauf, D. (2011) 'Enhanced fee-for-service model and physician productivity: Evidence from family health groups in Ontario', *Journal of Health Economics*, 30(1), 99-111.
- Kaplow, L. (2010) 'Distributive justice and social welfare'. In *The theory of taxation and public economics*, Princeton: Princeton University Press, pp. 347-370.
- Kay, A. (2002) 'The abolition of the GP fundholding scheme: A lesson in evidence-based policy making', *British Journal of General Practice*, 52(475), 141-144.
- Keliddar, I., Mosadeghrad, A. M. and Jafari-Sirizi, M. (2017) 'Rationing in health systems: A critical review', *Medical Journal of the Islamic Republic of Iran*, 31(47), 1-18.
- Kesternich, I., Schumacher, H. and Winter, J. (2015) 'Professional norms and physician behavior: Homo oeconomicus or homo hippocraticus?', *Journal of Public Economics*, 131(1), 1-11.
- Knott, L. (2019) 'Controlled drugs', [online], available: <https://patient.info/doctor/controlled-drugs> [Accessed 05/05 2018].
- Kolstad, J. R. and Lindkvist, I. (2012) 'Pro-social preferences and self-selection into the public health sector: Evidence from an economic experiment', *Journal of Health Policy and Planning*, 28(3), 320-327.

- Krasnik, A., Groenewegen, P. P., Pedersen, P. A., von Scholten, P., Mooney, G., Gottschau, A., Flierman, H. A. and Damsgaard, M. T. (1990) ‘Changing remuneration systems: Effects on activity in general practice’, *British Medical Journal*, 300(6741), 1698-1701.
- Kringos, D. S., Boerma, W. G., Hutchinson, A. and Saltman, R. B. (2015) *Building primary care in a changing Europe: Case studies*, Copenhagen: WHO Regional Office for Europe.
- Kroneman, M. W., Van der Zee, J. and Groot, W. (2009) ‘Income development of general practitioners in eight European countries from 1975 to 2005’, *BMC Health Services Research*, 9(1), 9-26.
- Laberge, M., Wodchis, W. P., Barnsley, J. and Laporte, A. (2016) ‘Efficiency of Ontario primary care physicians across payment models: a stochastic frontier analysis’, *Journal of Health Economics Review*, 6(1), 1-10.
- Laffont, J.-J. and Martimort, D. (2002) ‘The rent extraction-efficiency trade-off’. In *The theory of incentives: The principal agent model*, Princeton: Princeton University Press, pp. 28-81.
- Laffont, J.-J. and Tirole, J. (1986) ‘Using cost observation to regulate firms’, *Journal of Political Economy*, 94(3), 614-641.
- Laffont, J.-J. and Tirole, J. (1993) *A theory of incentives in procurement and regulation*, Cambridge, MA: MIT Press.
- Laffont, J. and Martimort, D. (2002) ‘Moral hazard: The basic trade-offs’. In *The theory of incentives: The principal-agent model*, Princeton: Princeton University Press, pp. 145-186.
- Lagarde, M. and Blaauw, D. (2017) ‘Physicians’ responses to financial and social incentives: A medically framed real effort experiment’, *Journal of Social Science & Medicine*, 179(2), 147-159.
- Lawton, R., Robinson, O., Harrison, R., Mason, S., Conner, M. and Wilson, B. (2019) ‘Are more experienced clinicians better able to tolerate uncertainty and manage risks? A vignette study of doctors in three NHS emergency departments in England’, *British Medical Journal*, 28(5), 382-388.

- Lee, C. (1995) 'Optimal medical treatment under asymmetric information', *Journal of Health Economics*, 14(4), 419-441.
- Léger, P. T. (2008) 'Physician payment mechanisms', In M. Lu and E. Jonsson (eds.), *Financing health care: New ideas for a changing society*, Weinheim: Wiley-VCH Press, pp. 149-176.
- Leonard, D. K., Bloom, G., Hanson, K., O'Farrell, J. and Spicer, N. (2013) 'Institutional solutions to the asymmetric information problem in health and development services for the poor', *World Development*, 48(1), 71-87.
- Lien, H.-M., Ma, C.-T. A. and McGuire, T. G. (2004) 'Provider-client interactions and quantity of health care use', *Journal of Health Economics*, 23(6), 1261-1283.
- Lipari, R. N. and Hughes, A. M. S. (2017) *How people obtain the prescription pain relievers they misuse*, Washington, DC: National Survey on Drug Use and Health.
- Liu, T. and Ma, C.-t. A. (2013) 'Health insurance, treatment plan, and delegation to altruistic physician', *Journal of Economic Behavior & Organization*, 85(1), 79-96.
- Lundin, D. (2000) 'Moral hazard in physician prescription behaviour', *Journal of Health Economics*, 19(5), 639-662.
- Ma, C.-t. A. (1994) 'Health care payment systems: Cost and quality incentives', *Journal of Economics & Management Strategy*, 3(1), 93-112.
- Ma, C.-t. A. (1998) 'Cost and quality incentives in health care: Altruistic providers', *Working Paper*, 84(1), 1-20.
- Ma, C.-T. A. and McGuire, T. G. (1997) 'Optimal health insurance and provider payment', *The American Economic Review*, 87(4), 685-704.
- Madden, D., Nolan, A. and Nolan, B. (2005) 'GP reimbursement and visiting behaviour in Ireland', *Health Economics*, 14(10), 1047-1060.
- Makris, M. (2009) 'Incentives for motivated agents under an administrative constraint', *Journal of Economic Behavior & Organization*, 71(2), 428-440.
- Makris, M. and Siciliani, L. (2013) 'Optimal incentive schemes for altruistic providers', *Journal of Public Economic Theory*, 15(5), 675-699.

- Mannion, R. and Davies, H. T. O. (2008) ‘Payment for performance in health care’, *British Medical Journal*, 336(7639), 306-308.
- Maréchal, F. and Thomas, L. (2018) ‘The optimal contract under adverse selection in a moral-hazard model with a risk-averse agent’, *Journal of Games*, 9(1), 1-22.
- Marshall, L., Charlesworth, A. and Hurst, J. (2014) *The NHS payment system: Evolving policy and emerging evidence*, London: The Nuffield Trust.
- Martinsson, P. and Persson, E. (2019) ‘Physician behaviour and conditional altruism: the effects of payment system and uncertain health benefit’, *Theory and Decision*, 87(1), 365-387.
- Maskin, E. and Moore, J. (1999) ‘Implementation and renegotiation’, *Review of Economic Studies*, 66(1), 39-56.
- Matsaganis, M. and Glennerster, H. (1994) ‘The threat of “cream skimming” in the post-reform NHS’, *Journal of Health Economics*, 13(1), 31-60.
- McGuire, T. G. (2000) ‘Physician agency’. In A. Culyer, and J. P. Newhouse (eds.), *Handbook of health economics*, Dordrecht: Elsevier, pp. 461-536.
- Medical Economics (2015) ‘Understanding physician compensation caps’, [online], available: <https://www.medicaleconomics.com/medical-economics/news/understanding-physician-compensation-caps> [Accessed 20/03 2019].
- MedlinePlus (2017) ‘Choosing a primary care provider’, *Medical Encyclopedia* [online], available: <https://medlineplus.gov/ency/article/001939.htm> [Accessed 23/08 2019].
- Milgrom, P. (1987) ‘Adverse selection without hidden information’, *Working Paper*, 8472(1), 1-19.
- Morris, S., Devlin, N., Parkin, D. and Spencer, A. (2012) ‘Health insurance and healthcare financing’. In *Economic Analysis in Healthcare*, Chichester: Wiley, pp. 133-162.
- Mougeot, M. and Naegelen, F. (2005) ‘Hospital price regulation and expenditure cap policy’, *Journal of Health Economics*, 24(1), 55-72.

- Mougeot, M. and Naegelen, F. (2009) ‘Adverse selection, moral hazard, and outlier payment policy’, *Journal of Risk and Insurance*, 76(1), 177-195.
- Mwachofi, A. and Al-Assaf, A. F. (2011) ‘Health care market deviations from the ideal market’, *Sultan Qaboos University Medical Journal*, 11(3), 328-337.
- National Health Service (2016) ‘About the NHS’, [online], available: <https://www.nhs.uk/using-the-nhs/about-the-nhs/the-nhs/> [Accessed 06/06 2019].
- National Institute on Drug Abuse (2019) ‘Opioid overdose crisis’, [online], available: <https://www.drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis> [Accessed 05/02 2020].
- Nehk, K. (2018) ‘What is fee-for-service in healthcare’, [online], available: <https://prognocis.com/what-is-fee-for-service-in-healthcare/2019>.
- Neudeck, W. (1991) ‘Fee-for-service and quantity rationing in the physician services market’ in G. López-Casasnovas (ed.), *Incentives in health systems*, Berlin: Springer, pp. 99-108.
- Nyman, J. A. (1999) ‘The economics of moral hazard revisited’, *Journal of Health Economics*, 18(6), 811-824.
- OECD (2003) *Society at a glance 2002: OECD social indicators*, Paris: OECD.
- OECD (2013) ‘Aging and long-term care’. In *Health at a glance 2013: OECD indicators*, Paris: OECD, pp. 170-182.
- OECD (2015) *Focus on health spending OECD health statistics 2015*, Paris: OECD.
- OECD (2016) ‘Reforming traditional healthcare provider payments’, In *Better ways to pay for health care*, Paris: OECD, pp. 37-55.
- OECD (2017) ‘Health Expenditure’, In *Health at a glance 2017*, Paris: OECD, pp. 131-144.
- Offodile, A. C., Mehtsun, W., Stimson, C. J. and Aloia, T. (2019) ‘An overview of bundled payments for surgical oncologists: Origins, progress to date, terminology, and future directions’, *Annals of Surgical Oncology*, 26(1), 3-7.

- Papanicolas, I., Woskie, L. R. and Jha, A. K. (2018) ‘Health care spending in the United States and other high-income countries’, *Journal of American Medical Association*, 319(10), 1024-1039.
- Park, M., Braun, T., Carrin, G. and Evans, D. (2007) *Provider payments and cost-containment lessons from OECD countries*, Geneva: World Health Organization.
- Penner, L. A., Fritzsche, B. A., Craiger, J. P. and Freifeld, T. R. (1995) ‘Measuring the prosocial personality’, *Advances in Personality Assessment*, 10(1), 147-163.
- Porter, M. E. and Kaplan, R. S. (2014) ‘How should we pay for health care?’, *Working Paper*, 15(1), 1-26.
- Porter, M. E. and Kaplan, R. S. (2016) ‘How to pay for health care’, *Harvard Business Review*, 94(7), 88-98.
- Poterba, J. M. (1994) ‘A skeptic’s view of global budget caps’, *Journal of Economic Perspectives*, 8(3), 67-74.
- Press Association (2016) ‘Rationing “already widespread in the NHS for a variety of treatments”’ [online], available: <https://www.dailymail.co.uk/wires/pa/article-3789649/Rationing-widespread-NHS-variety-treatments.html> [Accessed 07/02 2019].
- Prospects (2019) ‘General Practice Doctor’ [online], available: <https://www.prospects.ac.uk/job-profiles/general-practice-doctor> [Accessed 07/02 2019].
- Rama, A. (2017) *Payment and delivery in 2016: The prevalence of medical homes, accountable care organizations, and payment methods reported by physicians*, Chicago: American Medical Association.
- Rashidian, A., Omidvari, A. H., Vali, Y., Sturm, H. and Oxman, A. D. (2015) ‘Pharmaceutical policies: Effects of financial incentives for prescribers’, *Cochrane Database of Systematic Reviews*, 4(1), 1-15.
- Rhodes, D. (2018) ‘NHS accused of fuelling rise in opioid addiction’ [online], available: <https://www.bbc.co.uk/news/uk-england-43304375> [Accessed April 11 2019].

- Robineau, D. (2016) ‘Ageing Britain: Two-fifths of NHS budget is spent on over-65s’ [online], available: [Accessed 01/02 2016].
- Rogerson, W. P. (1994) ‘Choice of treatment intensities by a non-profit hospital under prospective pricing’, *Journal of Economics & Management Strategy*, 3(1), 7-51.
- Rosen, B. (1989) ‘Professional reimbursement and professional behaviour: Emerging issues and research challenges’, *Social Science & Medicine*, 29(3), 455-462.
- Rowell, D. and Connelly, L. B. (2012) ‘A history of the term “moral hazard”’, *Journal of Risk and Insurance*, 79(4), 1051-1075.
- Rubin, P. H. (1978) ‘The theory of the firm and the structure of the franchise contract’, *The Journal of Law and Economics*, 21(1), 223-233.
- Rudmik, L., Wranik, D. and Rudisill-Michaelsen, C. (2014) ‘Physician payment methods: A focus on quality and cost control’, *Journal of Otolaryngology-Head & Neck Surgery*, 43(1), 1-5.
- Salter, J. (2017) ‘Express scripts to limit opioids; doctors concerned’, *AP News* (16 August).
- Sandier, S., Paris, V., Polton, D., Thomson, S., Mossialos, E. and Organization, W. H. (2004) *Health care systems in transition: France*, Copenhagen: WHO Regional Office for Europe.
- Santos, R., Barsanti, S. and Seghieri, C. (2019) ‘Pay for performance in primary care: The use of administrative data by health economists. In *Data-driven policy impact evaluation*, Berlin: Springer, pp. 313-332.
- Sawyer, B. (2018) ‘Total health expenditures as percent of GDP, 1970-2017’[online], available: <https://www.healthsystemtracker.org/chart/total-health-expenditures-as-percent-of-gdp-1970-2017/#item-start> [Accessed March 08 2019].
- Scott, A., Sivey, P., Ouakrim, D. A., Willenberg, L., Naccarella, L., Furler, J. and Young, D. (2011) ‘The effect of financial incentives on the quality of health care provided by primary care physicians’, *Cochrane Database of Systematic Reviews*, (9) art. no.: CD008451. DOI: 10.1002/14651858.CD008451.pub2.

- Scully, R. E., Schoenfeld, A. J., Jiang, W., Lipsitz, S., Chaudhary, M. A., Learn, P. A., Koehlmoos, T., Haider, A. H. and Nguyen, L. L. (2018) ‘Defining optimal length of opioid pain medication prescription after common surgical procedures: Optimal length of opioid prescription after common surgical procedures’, *JAMA Surgery*, 153(1), 37-43.
- Serra, D., Serneels, P. and Barr, A. (2011) ‘Intrinsic motivations and the non-profit health sector: Evidence from Ethiopia’ *Personality and Individual Differences*, 51(3), 309-314.
- Shi, L. (2012) ‘The impact of primary care: A focused review’, *Scientifica*, doi: 10.6064/2012/432892.
- Sicsic, J., Le Vaillant, M. and Franc, C. (2012) ‘Intrinsic and extrinsic motivations in primary care: An explanatory study among French general practitioners’, *Health Policy*, 108(2-3), 140-148.
- Siddiqi, A., Hussain, S., Parveen, G., Malik, F., Yasin, F., Akram, T. S., Hameed, A., Riaz, H., Shah, P. A. and Saeed, T. (2011) ‘Relevant influence of promotional tools by pharmaceutical industry on prescribing behaviors of doctors: A cross-sectional survey in Pakistan’, *African Journal of Pharmacy and Pharmacology*, 5(13), 1623-1632.
- Smelser, N. J. and Baltes, P. B. (2001) ‘Health insurance: Economic and risk aspects’. In *International encyclopedia of the social & behavioral sciences*, Amsterdam: Elsevier, pp. 640-645.
- Smith, R., Lagarde, M., Blaauw, D., Goodman, C., English, M., Mullei, K., Pagaiya, N., Tangcharoensathien, V., Erasmus, E. and Hanson, K. (2012) ‘Appealing to altruism: An alternative strategy to address the health workforce crisis in developing countries?’, *Journal of Public Health*, 35(1), 164-170.
- Sood, N., De Vries, H., Gutierrez, I., Lakdawalla, D. N. and Goldman, D. P. (2009) ‘The effect of regulation on pharmaceutical revenues: Experience in nineteen countries’, *Health Affairs*, 28(1), w125-w137.

- Sørensen, R. J. and Grytten, J. (2003) ‘Service production and contract choice in primary physician services’, *Health Policy*, 66(1), 73-93.
- Stabile, M., Thomson, S., Allin, S., Boyle, S., Busse, R., Chevreul, K., Marchildon, G. and Mossialos, E. (2013) ‘Health care cost containment strategies used in four other high-income countries hold lessons for the United States’, *Health Affairs*, 32(4), 643-652.
- Starfield, B. (1991) ‘Primary care and health: A cross-national comparison’, *JAMA*, 266(16), 2268-2271.
- Tan, S. (2018) *Provider payment in health system reforms: Impact, autonomy, capacity*, National University of Singapore, unpublished thesis.
- Vogler, S., Espin, J. and Habl, C. (2009) ‘Pharmaceutical pricing and reimbursement information (PPRI): New PPRI analysis including Spain’, *Pharmaceuticals Policy and Law*, 11(3), 213-234.
- Wichmann, A. B., Adang, E. M., Stalmeier, P. F., Kristanti, S., Van den Block, L., Vernooij-Dassen, M. J., Engels, Y. and PACE (2017) ‘The use of quality-adjusted life years in cost-effectiveness analyses in palliative care: Mapping the debate through an integrative review’, *Palliative Medicine*, 31(4), 306-322.
- Woodward, R. S. and Warren-Boulton, F. (1984) ‘Considering the effects of financial incentives and professional ethics on “appropriate medical care”’, *Journal of Health Economics*, 3(3), 223-237.
- Wu, Y., Chen, Y. and Li, S. (2018) ‘Optimal compensation rule under provider adverse selection and moral hazard’, *Health Economics*, 27(3), 509-524.
- Yip, W. C. (1998) ‘Physician response to Medicare fee reductions: changes in the volume of coronary artery bypass graft (CABG) surgeries in the Medicare and private sectors’, *Journal of Health Economics*, 17(6), 675-699.
- Zuckerman, S., Norton, S. A. and Verrilli, D. (1998) ‘Price controls and Medicare spending: Assessing the volume offset assumption’, *Medical Care Research and Review*, 55(4), 457-478.

Zweigner, J., Meyer, E., Gastmeier, P. and Schwab, F. (2018) ‘Rate of antibiotic prescriptions in German outpatient care: Are the guidelines followed or are they still exceeded?’, *GMS Hygiene und Infection Control*, 13(4), 1-8.